
Supplementary information

**Redesigning algorithms to intervene on
social norm misperceptions during a
national election**

In the format provided by the
authors and unedited

Supplementary information

1.0 Participant procedure details

1.1 Participant instructions

Instructions for Research Participants: Using Bluesky during the study

Welcome to our research study on Bluesky social media use. This study requires you to use Bluesky in certain ways for 8 weeks. We are interested in how different feeds impact your experience of the platform. You may notice your feed seems different than your previous default, or you may not. Please simply use Bluesky as you normally would since your feed will be based partially on your own friendship network and platform behaviors just like on other social media platforms (e.g. X, Instagram, Facebook). At the end of the experiment, we will provide you with full details of the feed and provide you with an option to continue using it, as well as open-source code if you are interested in the technical details.

1. Feed selection
 - At the beginning of the experiment, you will set your default feed to the feed we assign to you. Throughout the experiment, you must maintain this feed as your default feed for 8 weeks of the experiment.
2. Content Interaction:
 - Follow accounts to your liking as you normally do on Bluesky.
 - Engage with content (like, share, comment) as you normally would, but within the guidelines provided by the research team – see next section.
3. Regular Updates and Study Procedure:
 - You will be required to log into Bluesky twice a day. Please use the platform as you normally would.
 - You are required to engage on the BlueSky platform at least 3 times per week (post, like, share, comment). You can engage / post about anything that you want. Posts can be original, or they can be re-posts of other users.

- You will be required to complete a weekly survey asking questions about your experiences on Bluesky. This will be provided by the research team to you every week on Sunday at 9AM Central Time. You will have 24 hours to complete the survey upon receiving it.
- Report any technical issues or concerns to the research team immediately.

If you have any questions, please contact the research team at mindtechnologylab@gmail.com.

1.2 Demographic measures / participant screening materials

With which gender identity do you mostly identify?

☐ Male

☐ Female

☐ Other

☐ Prefer not to say

What is your age?

Is English your first language?

☐ Yes

☐ No, but fluent

☐ No, mostly fluent

☐ No, minimal fluency

Please indicate your education level

- ☐ No school
- ☐ Eighth grade or less
- ☐ More than eighth grade, but less than high school degree
- ☐ High school degree or equivalent
- ☐ Some college
- ☐ 4-year college degree
- ☐ Graduate or professional training

Are you currently...?

- ☐ Employed
- ☐ Self-employed
- ☐ A student
- ☐ Unemployed
- ☐ Other

If you had to choose, which political group do you identify more with?

- ☐ Democrats
- ☐ Republicans

How would you describe your political ideology?

Extremely
liberal
☐

Moderately
liberal
☐

Slightly
liberal
☐

Neither
☐

Slightly
conservative
☐

Moderately
conservative
☐

Extremely
conservative
☐

Generally speaking, how closely do you follow politics?

Not closely at all ☐	☐	☐	Moderately closely ☐	☐	☐	Very closely ☐
---	-----------------------	-----------------------	--	-----------------------	-----------------------	--

To what extent do you agree with the following statement:

"I identify with the **Republican** Party"

1 - Fully Disagree ☐	2 ☐	3 ☐	4 - Neither agree nor disagree ☐	5 ☐	6 ☐	7 - Fully agree ☐
--	----------------------------	----------------------------	--	----------------------------	----------------------------	---

To what extent do you agree with the following statement:

"I identify with the **Democratic** Party"

1 - Fully Disagree ☐	2 ☐	3 ☐	4 - Neither agree nor disagree ☐	5 ☐	6 ☐	7 - Fully agree ☐
--	----------------------------	----------------------------	--	----------------------------	----------------------------	---

How often do you use social media?

☐ Daily or more
☐ 4-6 times a week
☐ 2-3 times a week
☐ Once per week
☐ Between once per week and once per month
☐ Between once per month and once per year
☐ Never

If you use social media daily or more, use the slider to indicate how many times per day you tend to use social media. If you don't use social media daily or more, then leave the slider at 0.

0 10 20 30 40 50 60 70 80 90 100



How often do you use social media specifically to view or learn about political content?

☐ Daily or more

☐ 4-6 times a week

☐ 2-3 times a week

☐ Once per week

☐ Between once per week and once per month

☐ Between once per month and once per year

☐ Never

If you use social media to view or learn about political content daily or more, use the slider to indicate how many times per day you tend to use social media to view or learn about political content. If you don't use social media to view or learn about political content daily or more, then leave the slider at 0.

0 10 20 30 40 50 60 70 80 90 100

Never



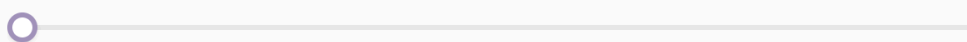
How often do you use social media specifically to post or share political content?

- ☐ Daily or more
- ☐ 4-6 times a week
- ☐ 2-3 times a week
- ☐ Once per week
- ☐ Between once per week and once per month
- ☐ Between once per month and once per year
- ☐ Never

If you use social media to post or share political content daily or more, use the slider to indicate how many times per day you tend to use social media to post or share political content. If you don't use social media to post or share political content daily or more, then leave the slider at 0.

0 10 20 30 40 50 60 70 80 90 100

Never



If you share (e.g. retweet, share, repost) content on social media, why do you typically share it (check all that apply)?

☐ I want people to see the content

☐ I want to make people laugh

☐ I want to change people's opinions

☐ I want to troll people

☐ The content represents my beliefs or attitudes

☐ I want to show people that we share common beliefs or attitudes

☐ Gain social connection

☐ Other

☐ I never share any content

1.3 Measures included in intake survey¹ and/or weekly surveys²

(item order randomized)

Norm Perception outcomes

(1) Perceptions of IME content on Bluesky (0-100).²

What percentage of posts on Bluesky are posting content that describes Democrats or liberals [Republicans or conservatives], or is biased towards Democrats or liberals [Republicans or conservatives] (e.g., praises Democrat or liberals [Republicans or conservatives], or criticizes Republicans or conservatives [Democrats or liberals])?

0	50	100
0 percent		100 percent

What percentage of posts on Bluesky do you think pertain to moral or political topics?

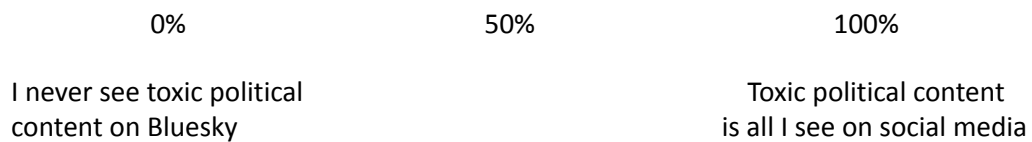
0	50	100
0 percent		100 percent

What percentage of posts on Bluesky do you think are emotional in nature (e.g., content that expresses emotion, or content that evokes emotion)?

0	50	100
0 percent		100 percent

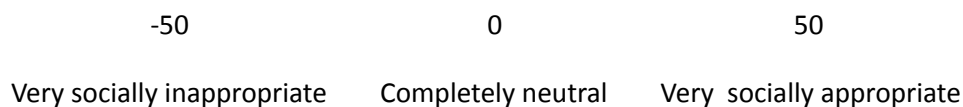
(2) Perception of toxic political discourse on Bluesky (0-100) (descriptive)^{1,2}

What percentage of content on Bluesky is toxic political content that is likely to create conflict among political groups?



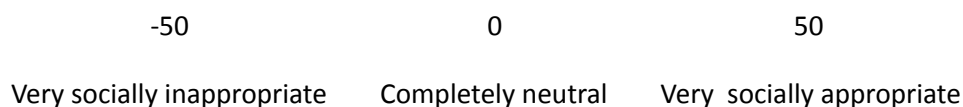
(3a) Perception of toxic political discourse on Bluesky (0-100) (prescriptive - other)^{1,2}

How socially appropriate *do other people on Bluesky feel it is* for you to post toxic political content that is likely to create conflict among political groups?



(3b) Perception of toxic political discourse on Bluesky (0-100) (prescriptive - self)^{1,2}

How socially appropriate *do you feel it is* for other people on Bluesky to post toxic content that is likely to create conflict among political groups?



***Note:** For all measures of toxicity perceptions above, participants are provided with the following definition of toxicity from Google's Perspective API: *A rude, disrespectful or unreasonable comment that is likely to make someone leave a discussion*

(4) Perception of ingroup's dislike of outgroup (perceived partisan animosity) (0-100) (Brady et al., 2023)^{1,2}

Please rate how your political group (Democrats[Republicans]) feels about Republicans [Democrats] on the scale below:

0	50	100
Strongly dislike	Completely neutral	Strongly like

(5) Perception of outgroup's dislike of ingroup (perceived partisan animosity) (0-100) (Brady et al., 2023)^{1,2}

Please rate how (Republicans [Democrats]) feel about Democrats [Republicans] on the scale below:

0	50	100
Strongly dislike	Completely neutral	Strongly like

Exploratory Outcomes

(1) Participant's level of political sectarianism (0-6) (Landry et al., 2024) (9 items averaged to form one composite score)^{1,2}

Othering

I feel distant from Republicans [Democrats]:

0

3

6

I don't feel distant
from Republicans [Democrats]

I feel as if Republicans [Democrats]
and I are on separate planets

I am different from Republicans [Democrats]:

0

3

6

I don't feel distant
from Republicans [Democrats]

I feel as if Republicans [Democrats]
and I are on separate planets

No matter how hard I try, I can't see the world the way Republicans [Democrats] do:

0

3

6

I don't have any difficulty seeing
the world the way Republicans
[Democrats] do

It's impossible for me to see the world
the way Republicans
[Democrats] do

Aversion

My feelings toward Republicans [Democrats] are negative:

0

3

6

My feelings toward Republicans
[Democrats] aren't negative at all

My feelings toward Republicans
[Democrats] are overwhelmingly negative

I hate Republicans [Democrats]:

0

3

6

I don't hate Republicans
[Democrats] at all

I have a fierce hatred for Republicans
[Democrats]

Republicans [Democrats] have lots of negative traits:

0

3

6

Republicans [Democrats]
have no negative traits

Republicans [Democrats] have every negative
trait in the book

Moralization

Republicans [Democrats] are immoral:

0

3

6

There is nothing immoral about
[Democrats]

Republicans [Democrats]
are completely immoral

Republicans [Democrats] are evil:

0

3

6

Republicans [Democrats] are not
evil in any way

Republicans [Democrats] are evil in
every way

Republicans [Democrats] lack integrity:

0

3

6

Republicans [Democrats]
lack integrity

Republicans [Democrats] lack any shred
of integrity

(2) Participant's dislike of political outgroup (0-100) (Iyengar et al., 2019)^{1,2}

Please rate how you feel about Democrats [Republicans] on the feeling thermometer below:

0	50	100
Strongly dislike	Completely neutral	Strongly like

(3) Endorsement of anti-democratic norms (Voelkel et al., 2023)^{1,2}

How would you feel if you learned that the Republican [Democratic] candidate said that Republicans [Democrats] should reduce the number of polling stations in areas that support Democrats [Republicans]?

0	50	100
Definitely vote for the Democratic [Republican] candidate		Definitely vote for the Republican [Democratic] candidate

How would you feel if you learned that the Republican [Democratic] candidate said that Republican [Democratic] governors should ignore unfavourable court rulings by Democratic [Republican]-appointed judges?

0	50	100
Definitely vote for the Democrat [Republican] candidate		Definitely vote for the Republican [Democrat] candidate

How would you feel if you learned that the Republican [Democratic] candidate said that Republican [Democratic] governors should prosecute journalists who accuse Republican [Democratic] politicians of misconduct without revealing sources?

0	50	100
Definitely vote for the Democrat [Republican] candidate		Definitely vote for the Republican [Democrat] candidate

How would you feel if you learned that the Republican [Democratic] candidate said that Republicans [Democrats] should ignore and not accept the results of elections if they lose?

0

50

100

Definitely vote for the Democrat
[Republican] candidate

Definitely vote for the Republican

(4) Measure of offline social network composition¹

In general, what percentage of the people who you know and interact with on a regular basis are Democrats, Republicans, Independents? Please enter the percentage of each group in the space below. The total number must add up to 100:

Democrats_____

Republicans_____

Independents_____

Total_____

(5) Attitude extremity on range of U.S. controversial political issues (Voelkel et al., 2023)^{1,2}

Please indicate how much you oppose or support the following political positions:

- Reducing access to abortion
- Providing a path to citizenship for undocumented immigrants
- Increasing restrictions on gun ownership
- Increasing government regulations to protect the environment
- Raising taxes on the wealthiest Americans
- Expanding Medicaid to cover all currently uninsured Americans

0

50

100

Strongly oppose

Strongly support

(6) Voting intentions for 2024 US Election^{1,2}

Who will you vote for in the 2024 U.S. Presidential election

- Donald Trump
- [Democratic Candidate]
- I will not vote
- I will write in a different candidate (please write below): _____

(7) Willingness to engage in collective action (Cervone et al., 2023)^{1,2}

Consider the moral position you believe the most strongly. Please report how strongly you agree or disagree with the following statements... (1 = strongly disagree; 7 = strongly agree)

- "I would sign a petition in favor of my position."
- "I would attend a rally, a march, or a protest to assert my position."
- "I would write to institutions and newspapers to bring forward my position."
- "I would collaborate in organizing an event to promote my position"

(8) Enjoyment of Bluesky platform experience²

Please report how strongly you agree or disagree with the following statements... (1 = strongly disagree; 7 = strongly agree)

- "I enjoy using BlueSky for my daily social media activities."
- "My overall experience on BlueSky is positive and fulfilling."
- "I look forward to engaging with content on BlueSky."
- "I intend to keep using Bluesky after the study"

(9) Enjoyment of Bluesky content viewed²

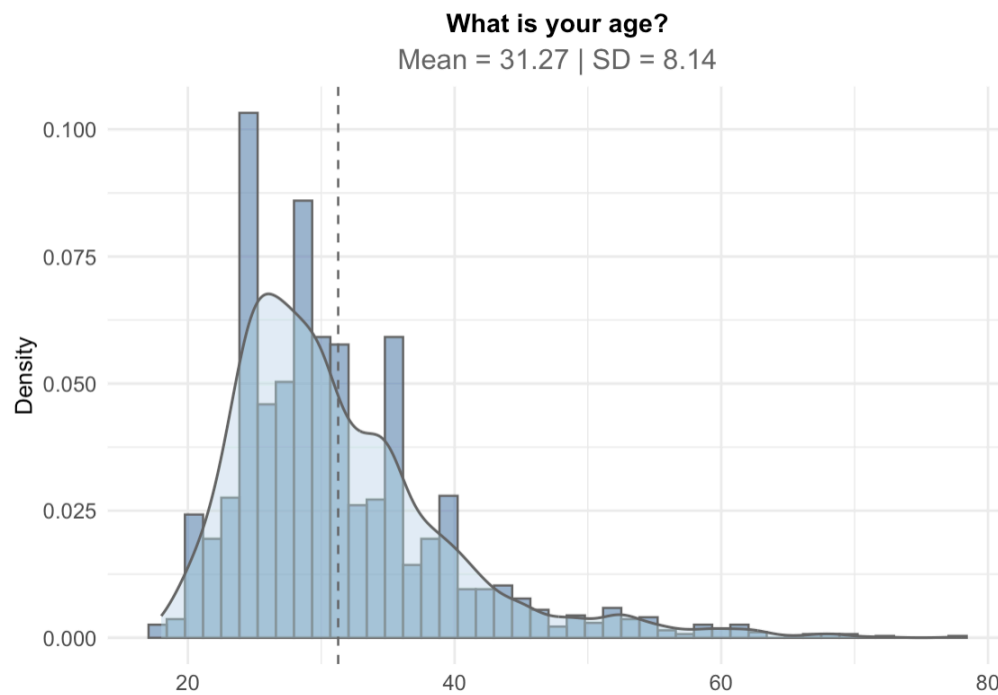
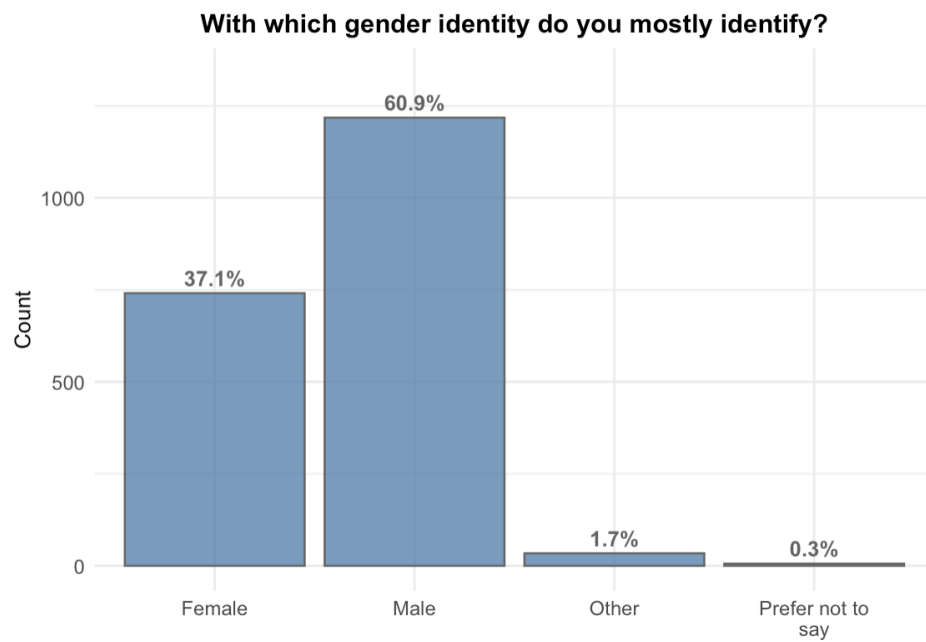
Please report how strongly you agree or disagree with the following statements... (1 = strongly disagree; 7 = strongly agree)

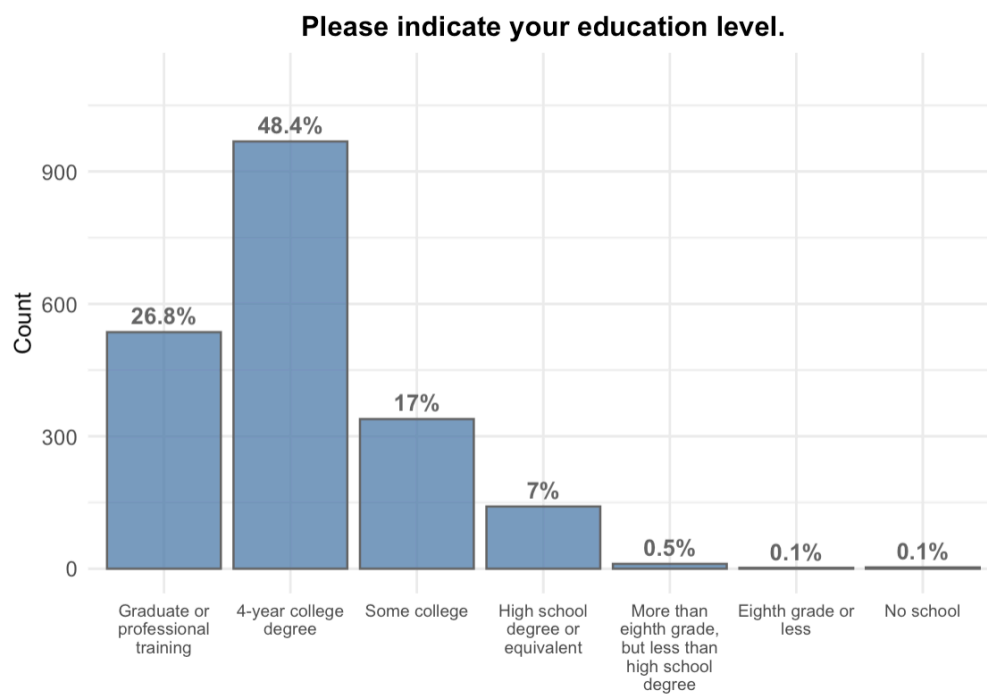
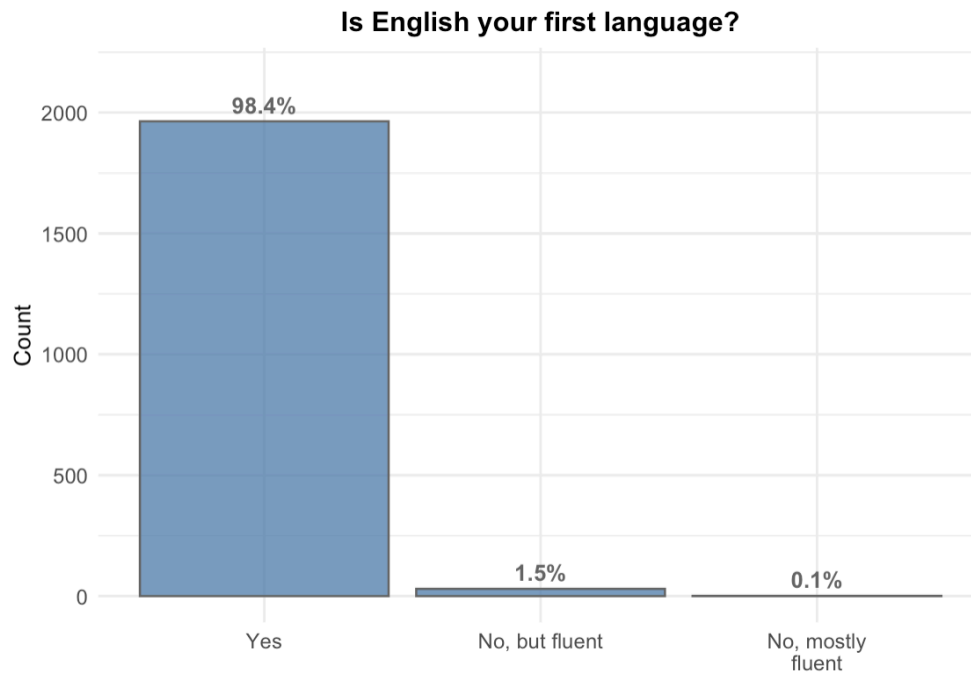
- "The content I come across on BlueSky is often interesting and engaging."
- "I find that the posts and discussions on BlueSky capture my attention and interest."
- "Most of the content I see on BlueSky is relevant to my interests and preferences."

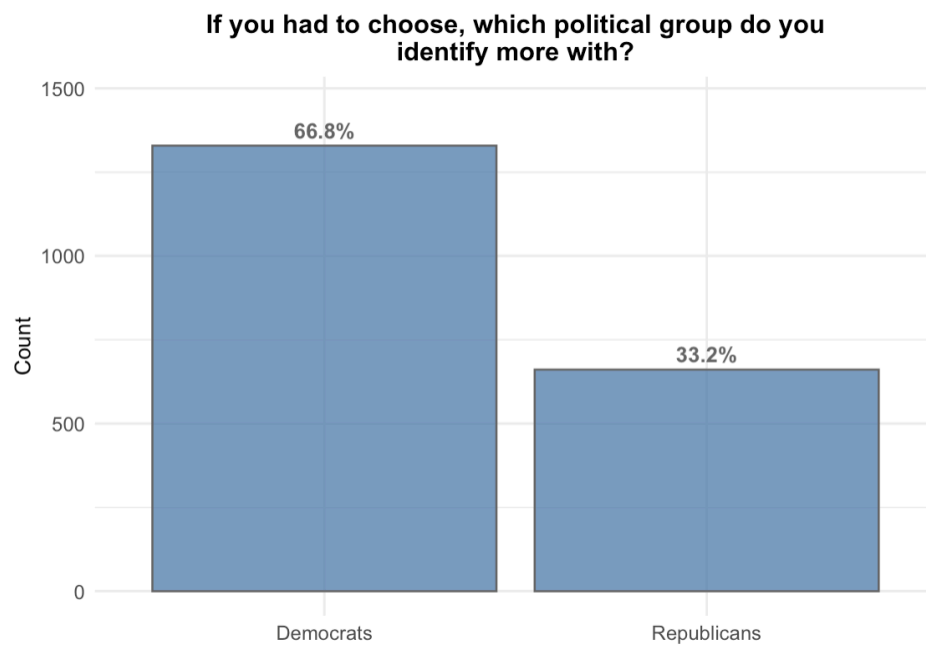
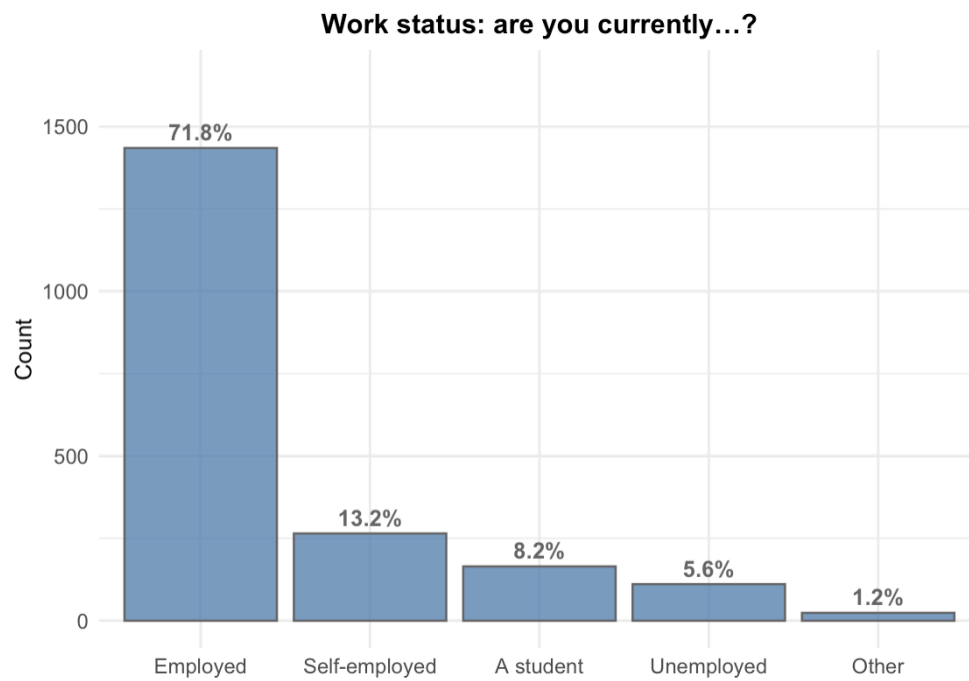
1.4 Steps for identifying non-legitimate sign-ups (Stage 2)

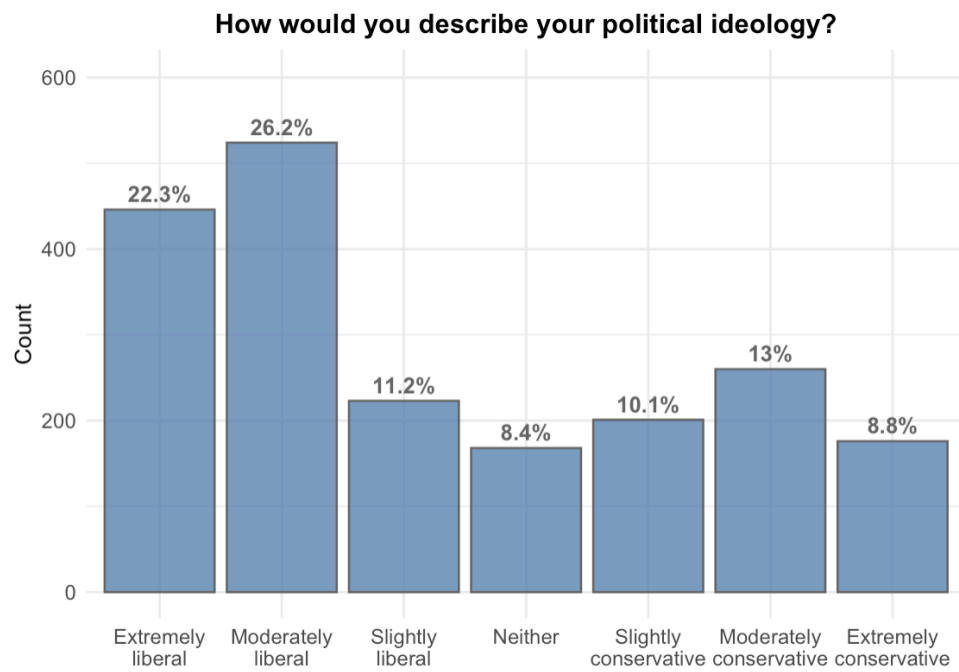
During our initial pre-screening phase, we discovered a large number of non-legitimate sign-ups (i.e., users attempting to sign up for the study with multiple email addresses, or lacking evidence of genuine activity on their accounts). To identify and remove these users, we developed screening criteria that participants had to meet in order to be included in the study. This included removing duplicate IP addresses from the initial intake survey (which eliminated all duplicate sign ups), and including only users who (1) had at least some activity on their account (e.g. posts and a network of followers), (2) a bio or profile picture (or both), and (3) all bio and post text in English.

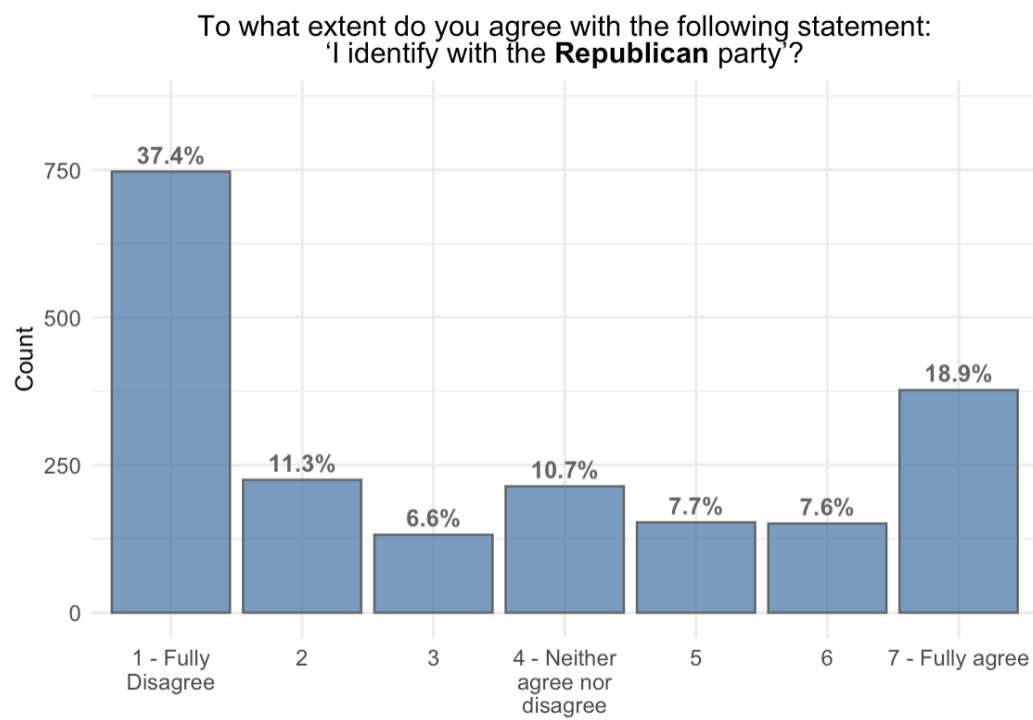
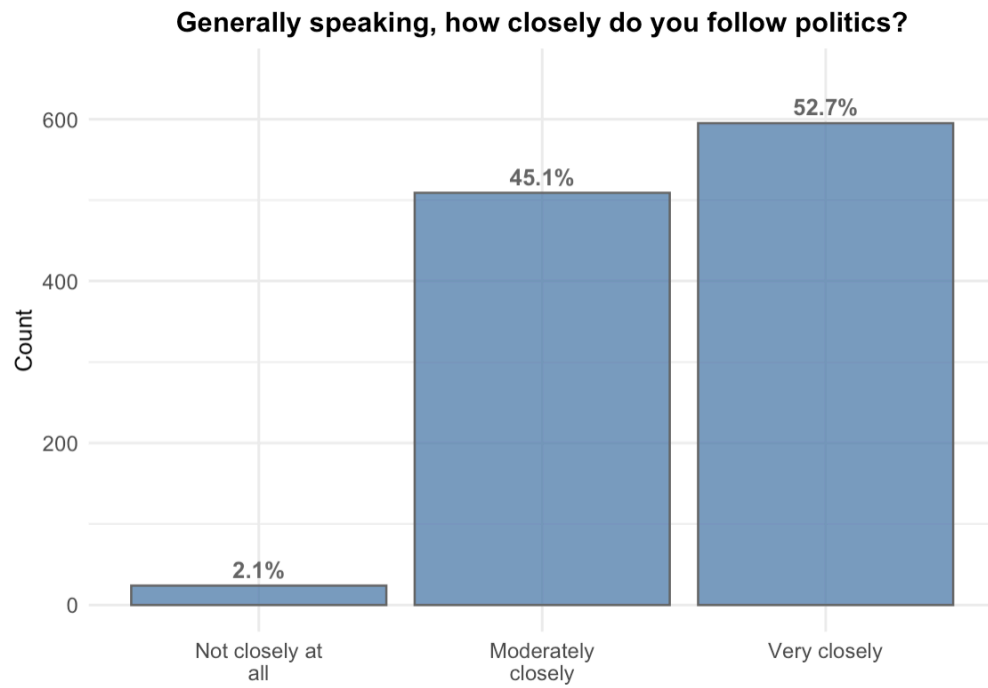
1.5 Descriptive statistics from final participant pool (Stage 2)



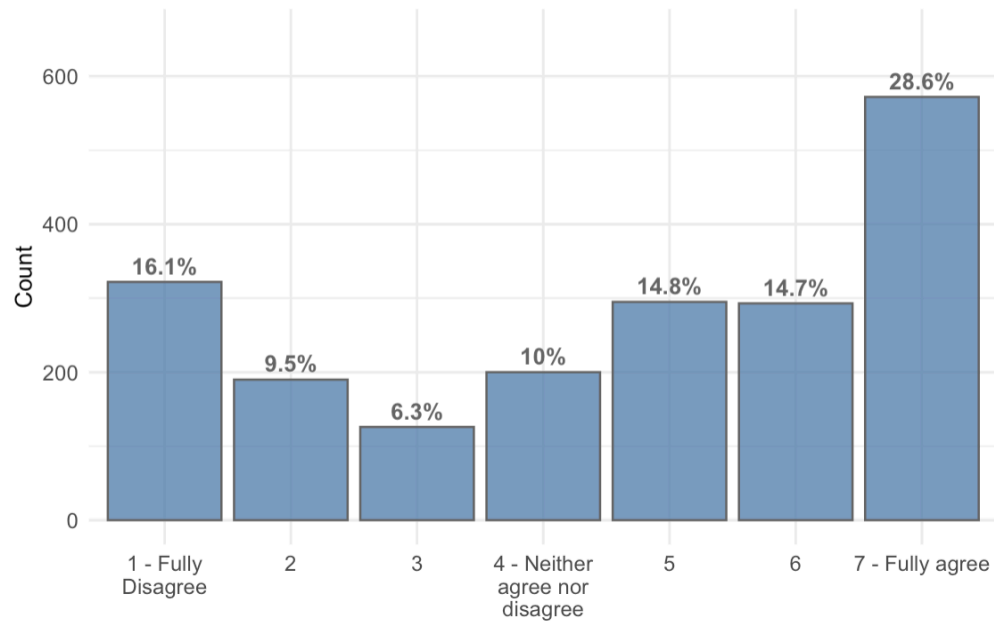




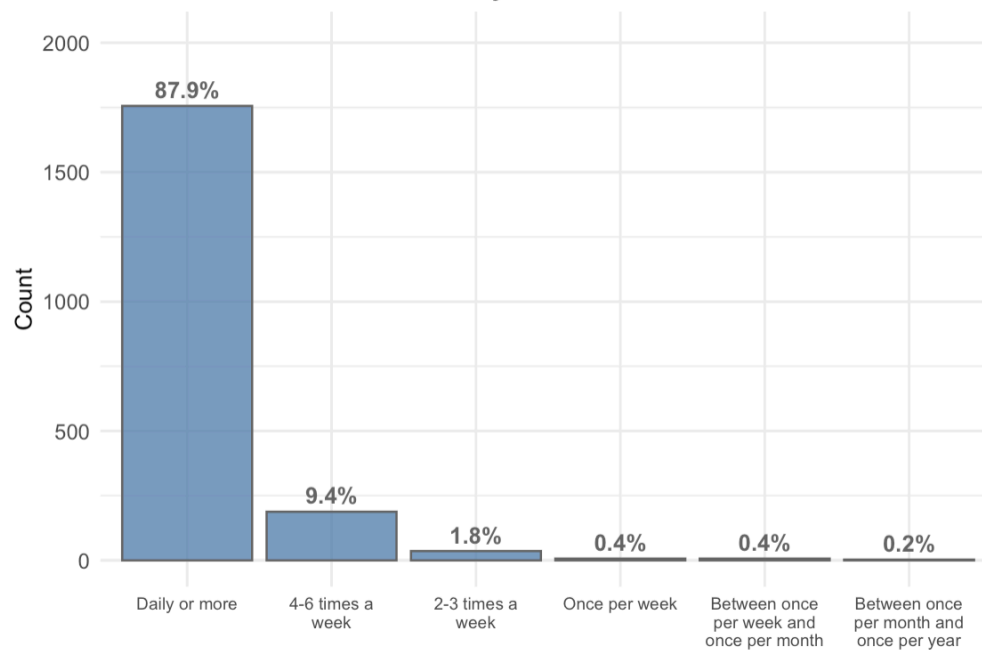




To what extent do you agree with the following statement:
'I identify with the **Democratic** party'?

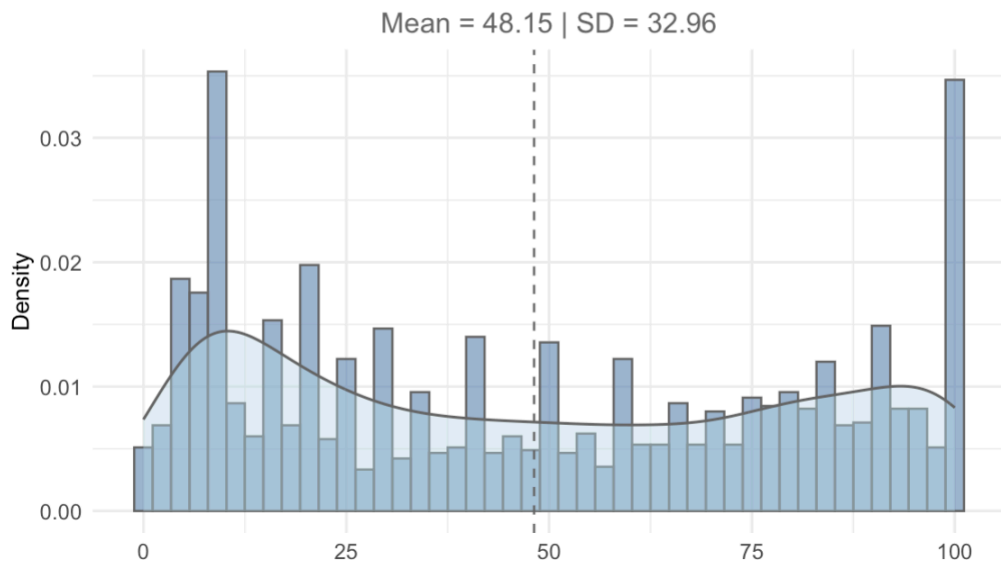


How often do you use social media?

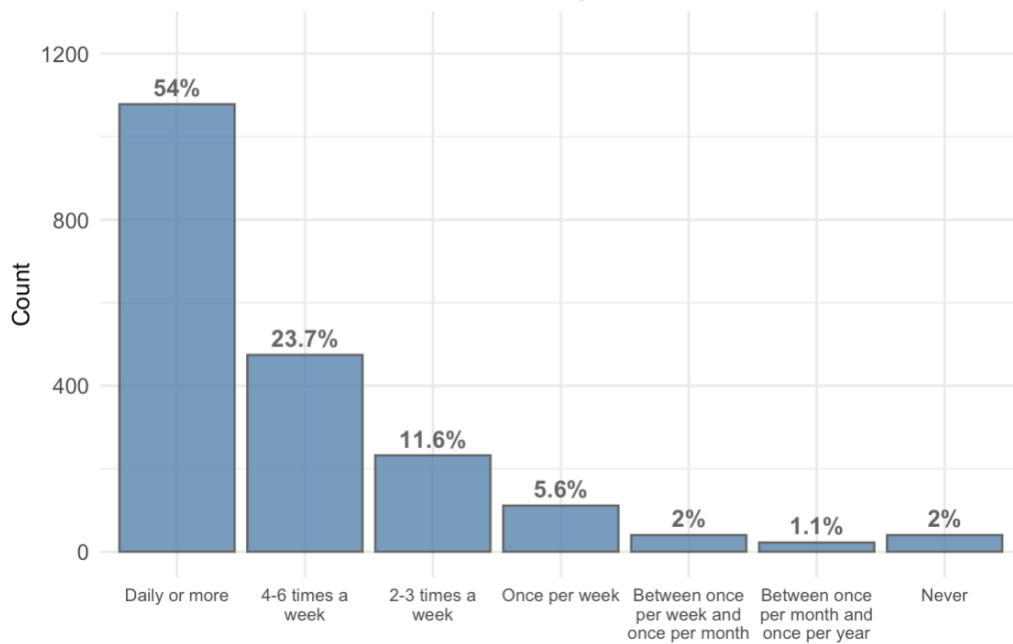


If you use social media **daily or more**, use the slider to indicate how many times per day you tend to use social media.

If you don't use social media daily or more, then leave the slider at 0.

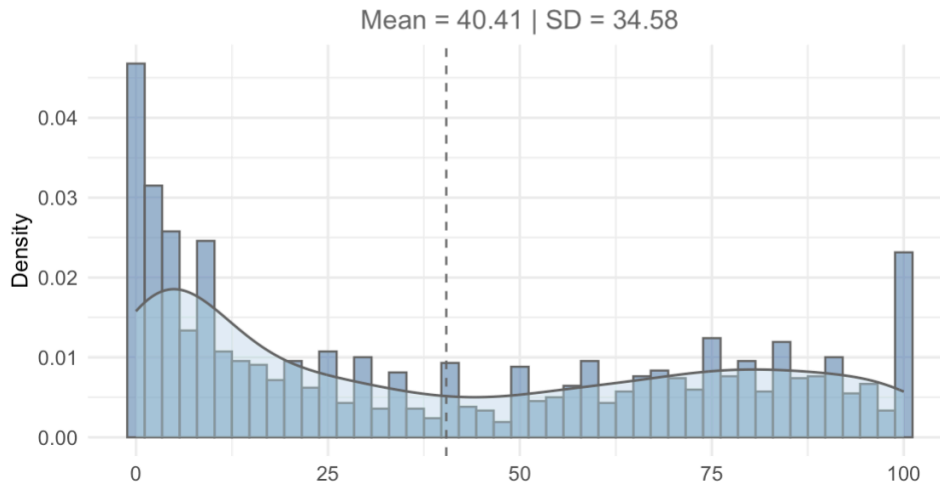


How often do you use social media specifically to **view or learn about political content**?

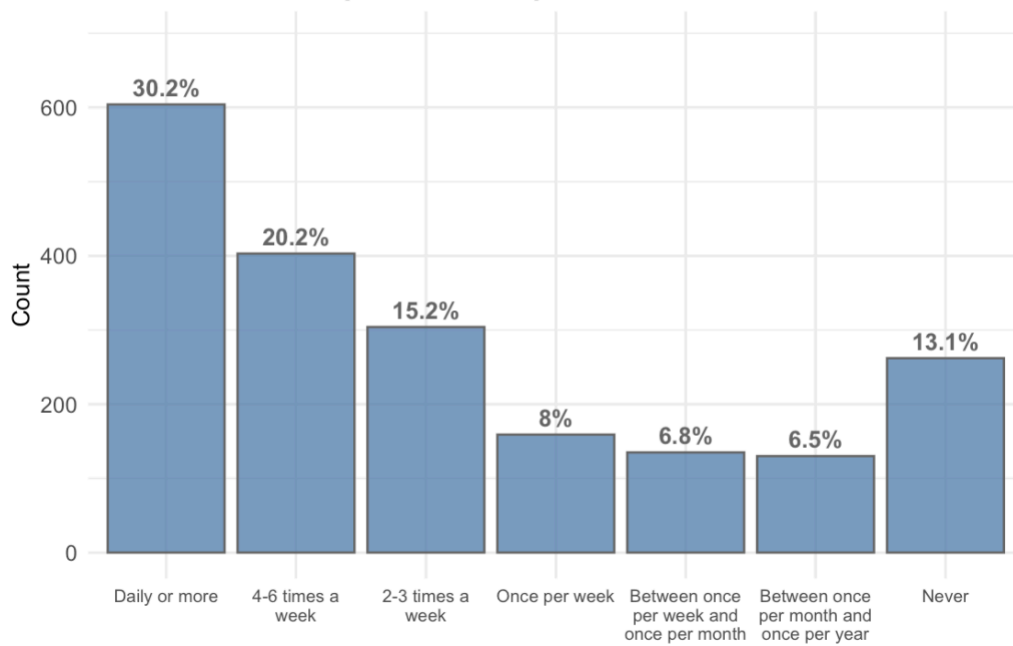


If you use social media to **view or learn about political content daily or more**, use the slider to indicate how many times per day you tend to use social media to view or learn about political content.

If you don't use social media to view or learn about political content daily or more, then leave the slider at 0.

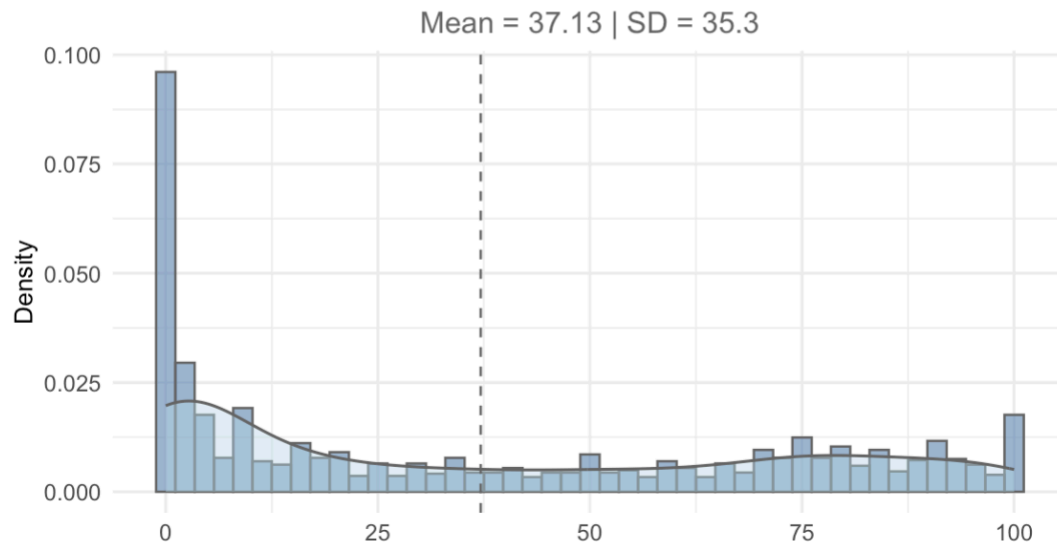


How often do you use social media specifically to **post or share political content**?

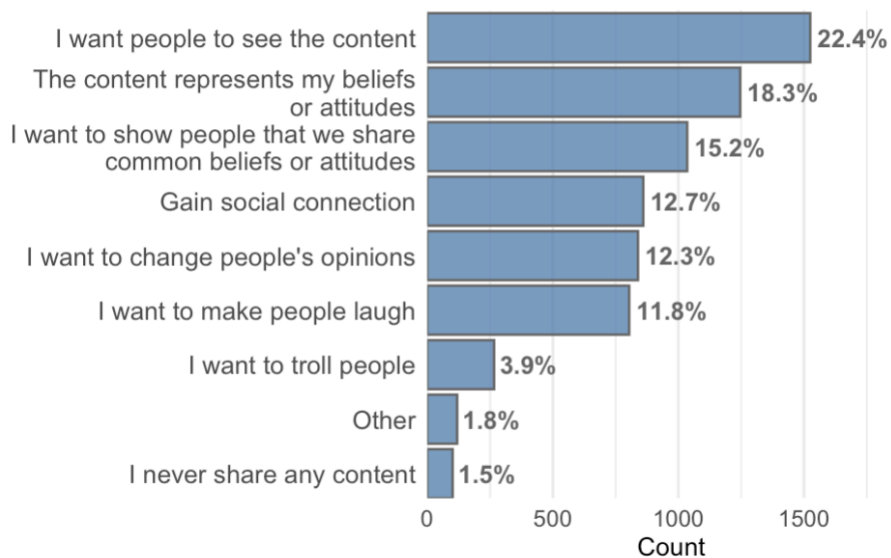


If you use social media to **post or share political content daily or more**, use the slider to indicate how many times per day you tend to use social media to post or share political content.

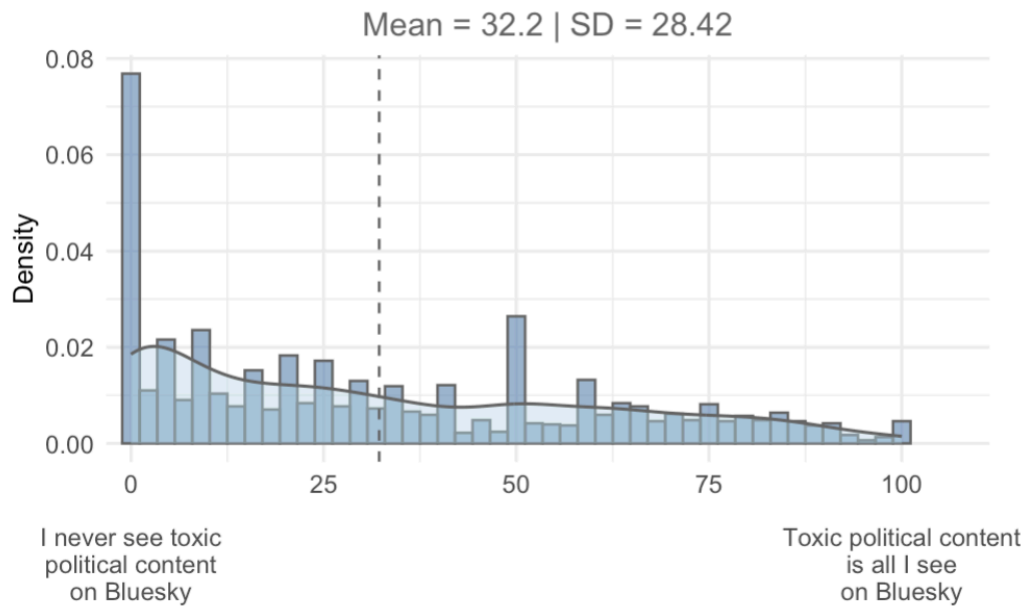
If you don't use social media to post or share political content daily or more, then leave the slider at 0.



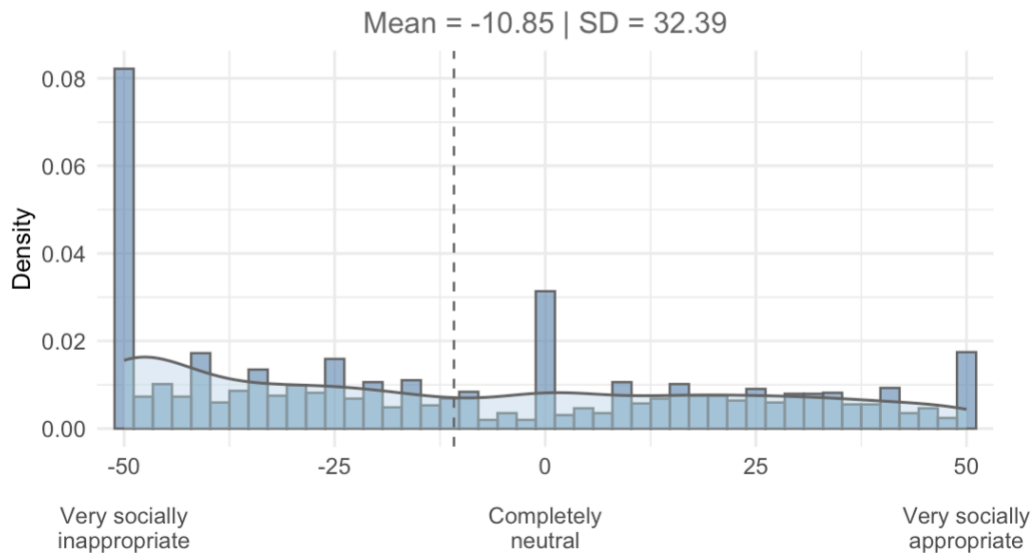
If you share (e.g. retweet, share, repost) content on social media, **why do you typically share it?**
(Check all that apply)



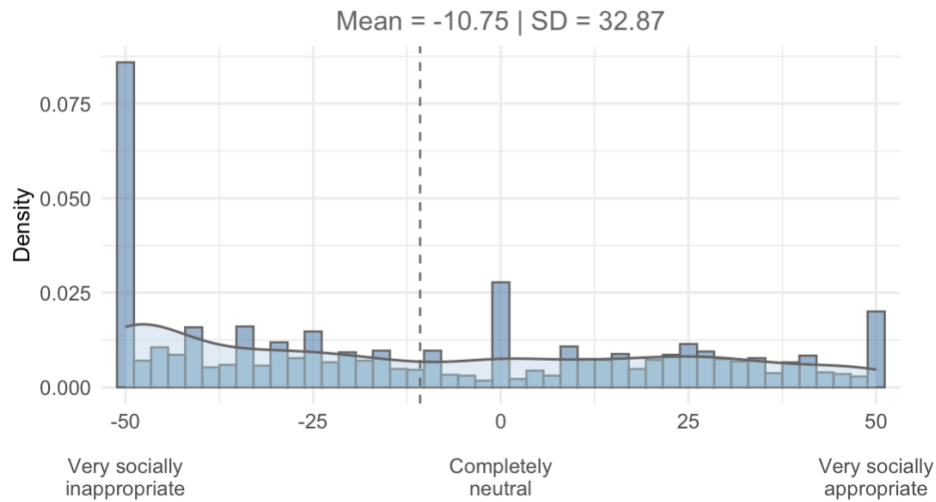
What percentage of posts in your Bluesky social network is **toxic political content** that is likely to create conflict among political groups?



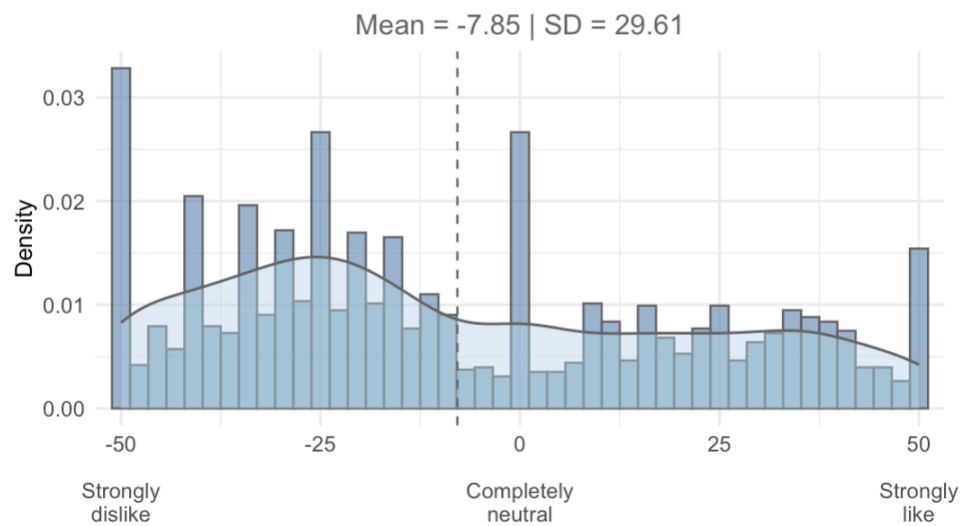
How socially appropriate do **other people on Bluesky** feel it is for **you** to post toxic political content that is likely to create conflict among political groups?



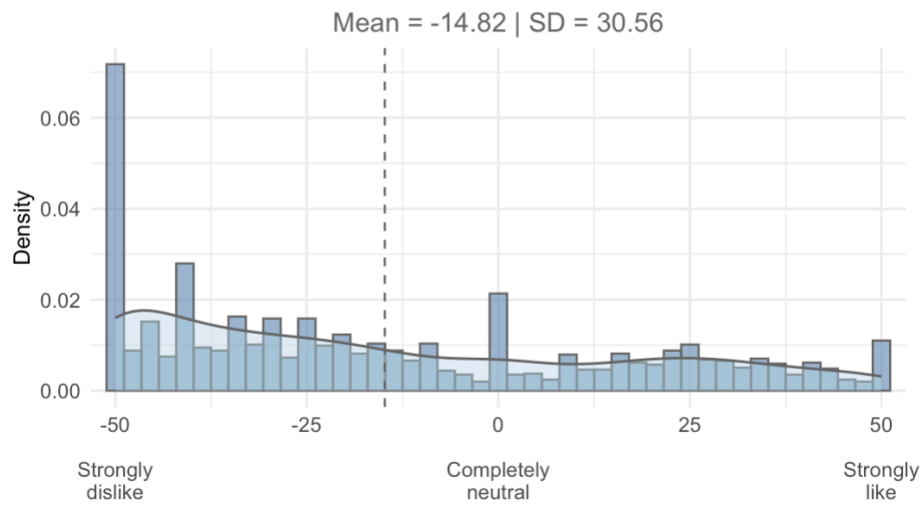
How socially appropriate do **you** feel it is for **other people on Bluesky** to post toxic political content that is likely to create conflict among political groups?



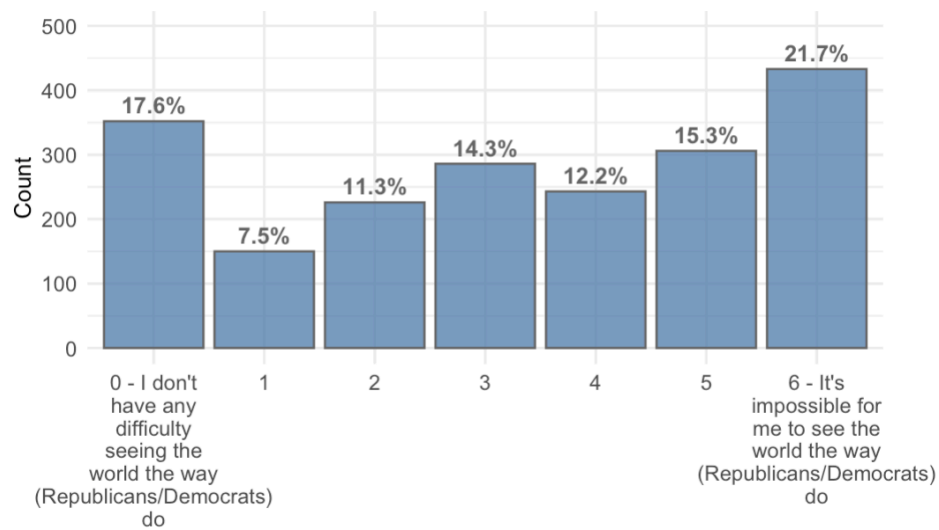
Please rate how you think the **political group you most identify with** (Democrats/Republicans) feels about (Republicans/Democrats) on the scale below:

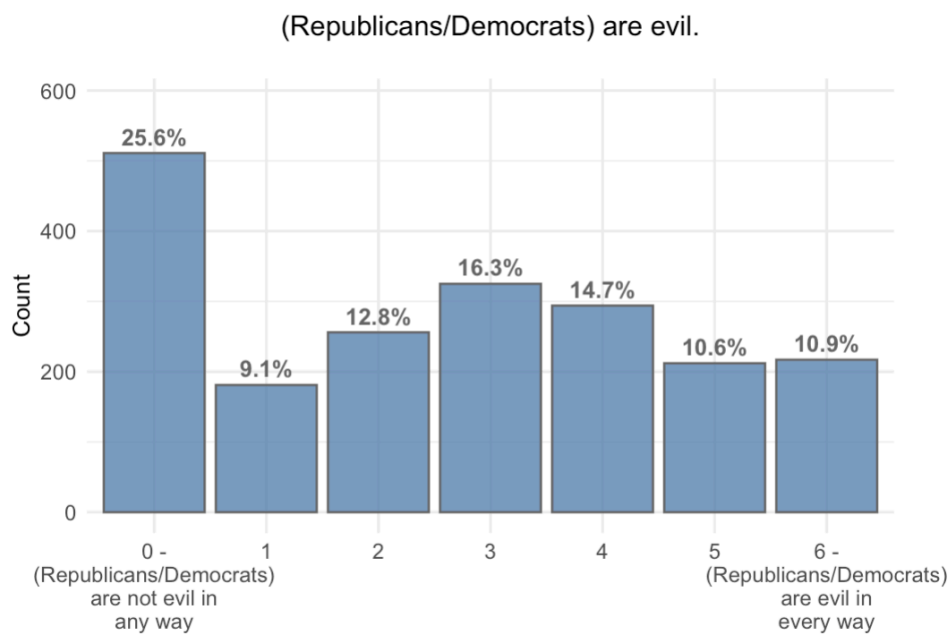
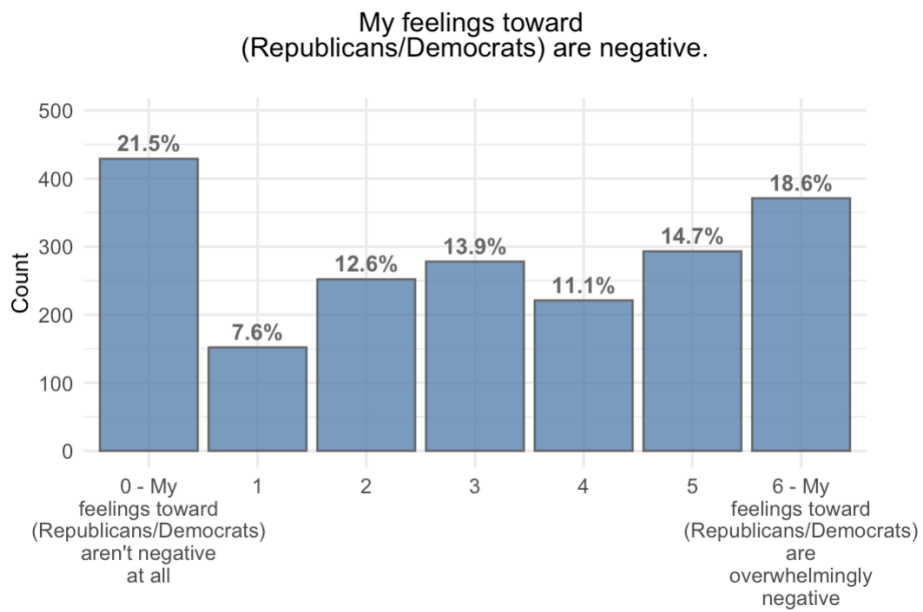


Please rate how you think the **political group you least identify with**
(Democrats/Republicans) feels about (Republicans/Democrats)
on the scale below:

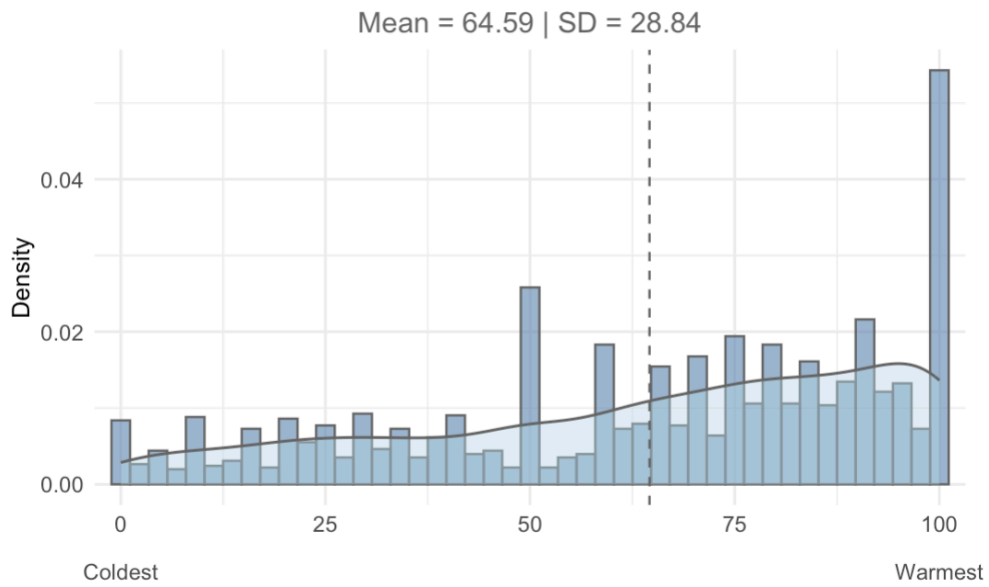


No matter how hard I try, I can't see the world the way
(Democrats/Republicans) do.

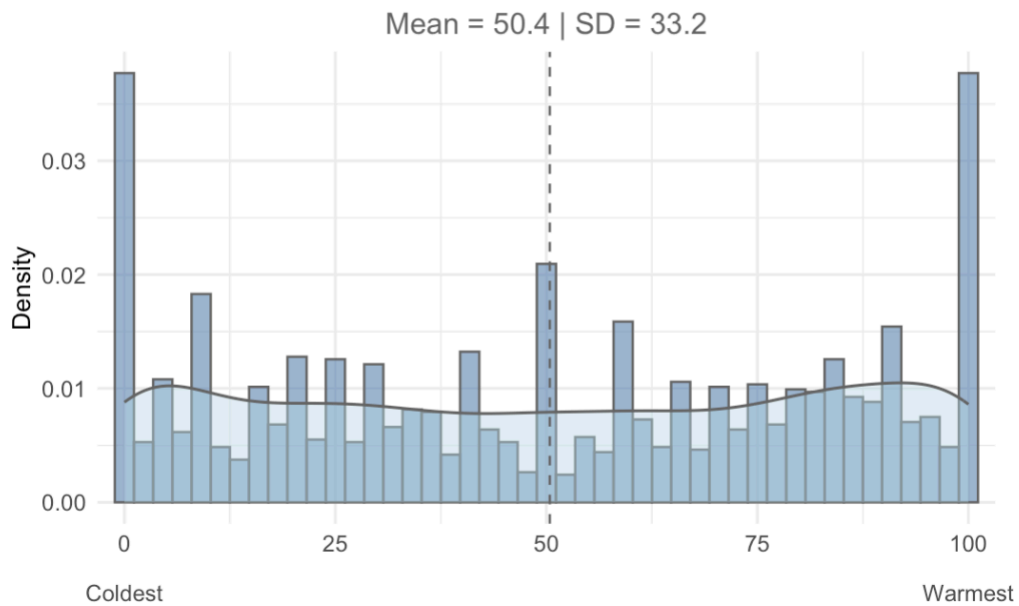




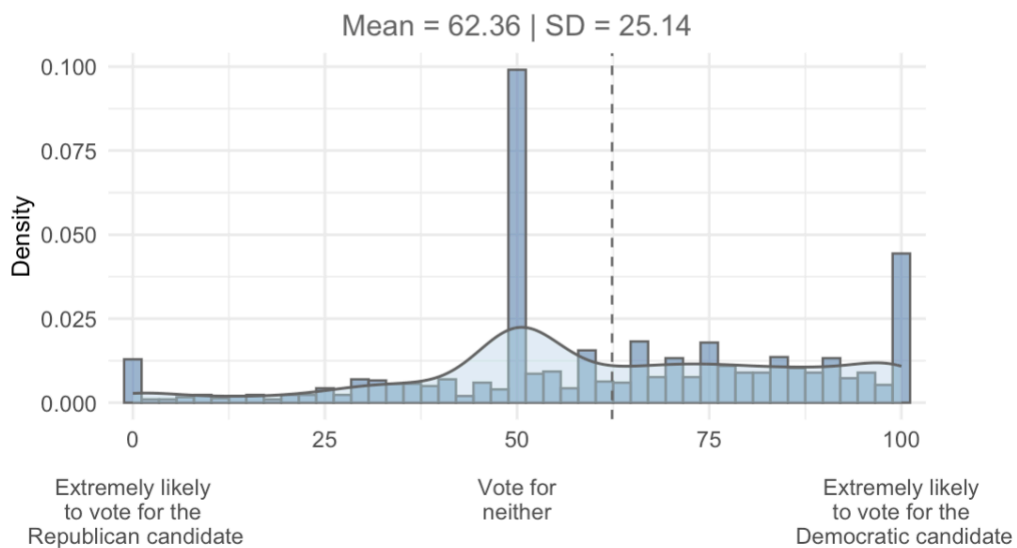
Please rate how you feel about **Democrats** on the feeling thermometer below:

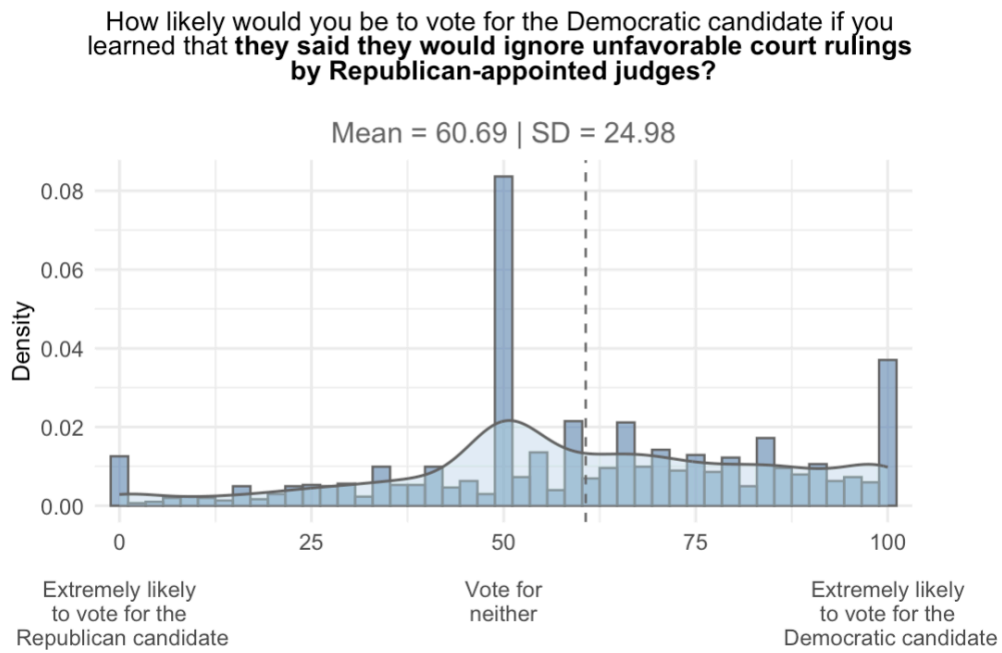


Please rate how you feel about **Republicans** on the feeling thermometer below:

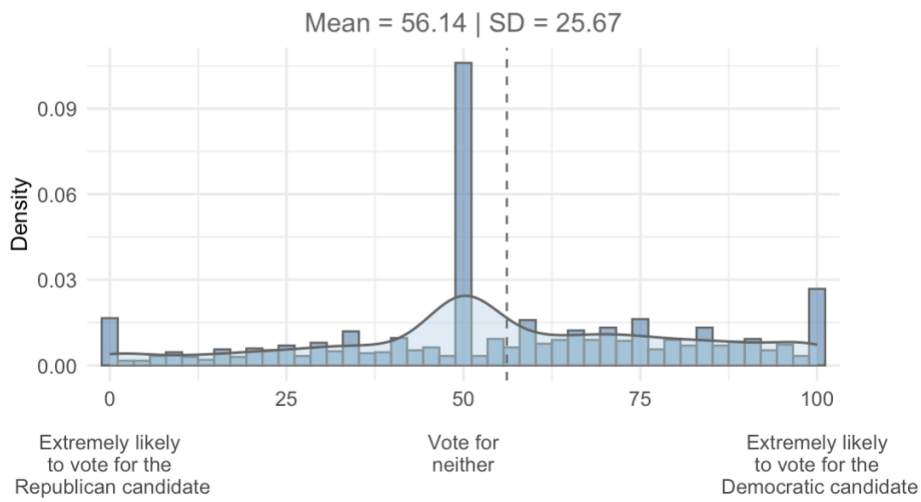


How likely would you be to vote for the Democratic candidate if you learned that **they support a proposal to reduce the number of polling stations in areas that support the Republican party?**

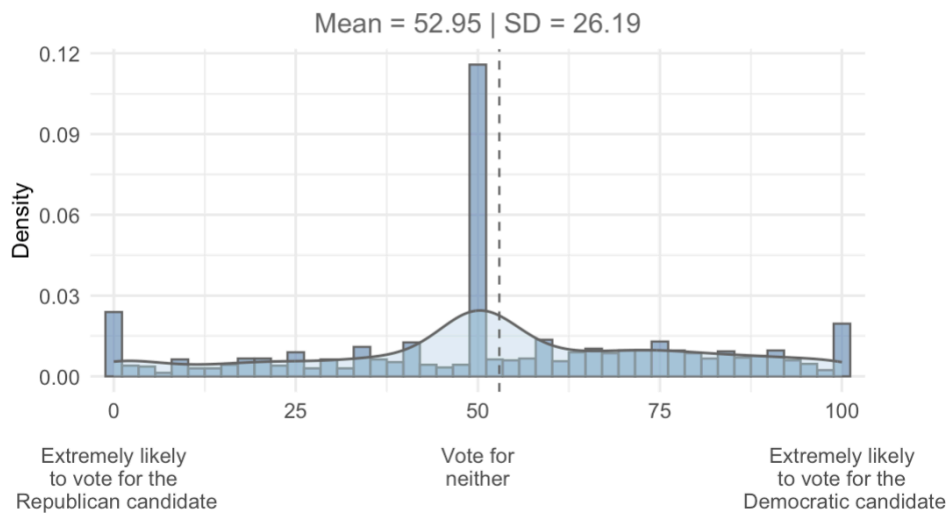




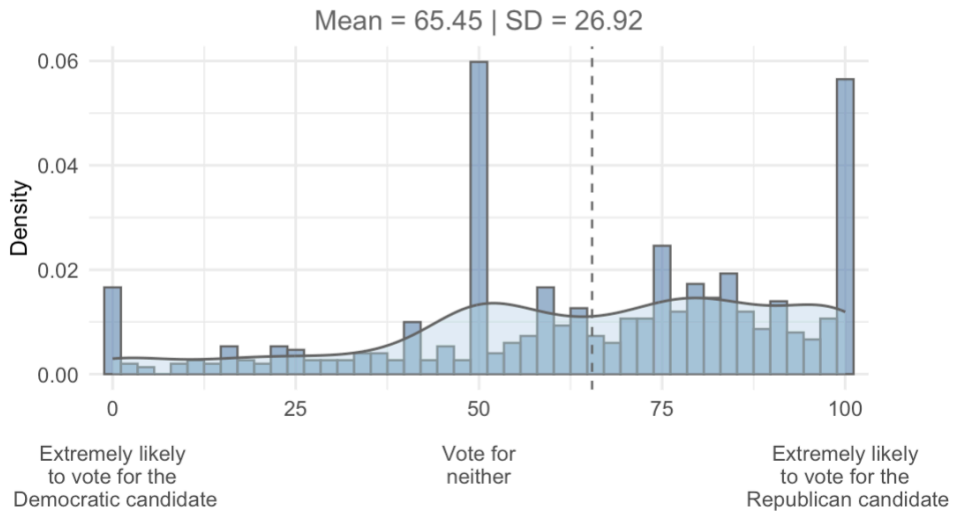
How likely would you be to vote for the Democratic candidate if you learned that **they said they would prosecute journalists who accuse them of misconduct if the journalists won't reveal their sources?**



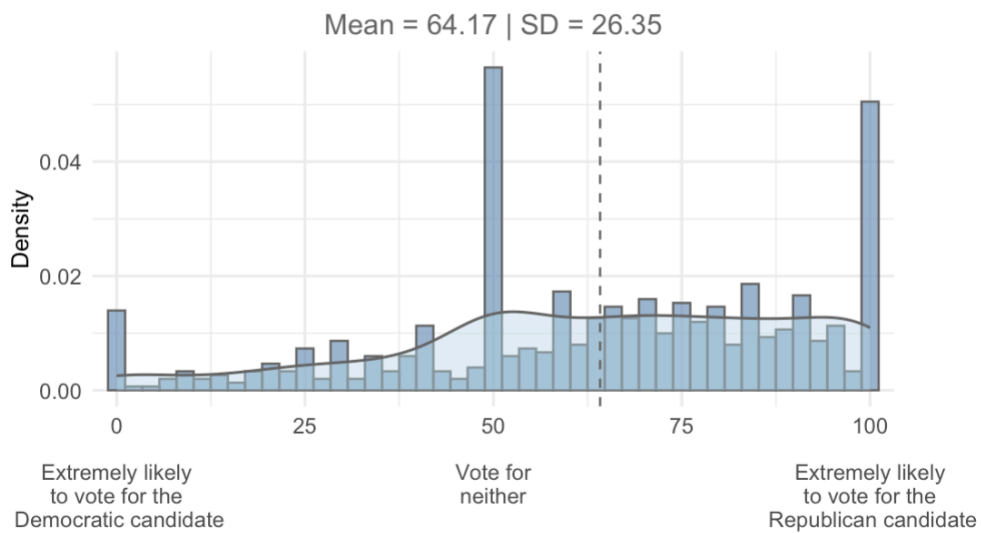
How likely would you be to vote for the Democratic candidate if you learned that they said that **Democrats should not accept election results if they do not win?**



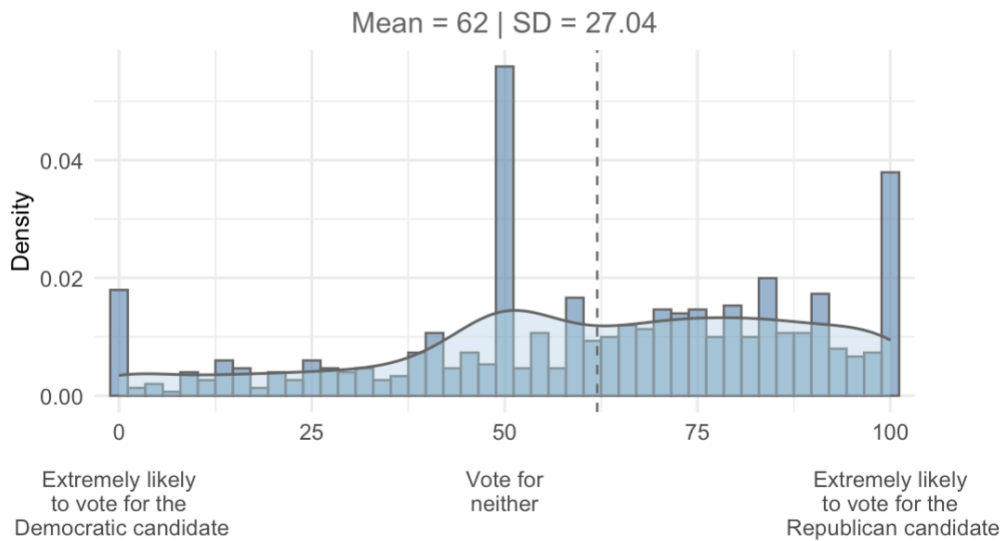
How likely would you be to vote for the Republican candidate if you learned that **they support a proposal to reduce the number of polling stations in areas that support the Democratic party?**



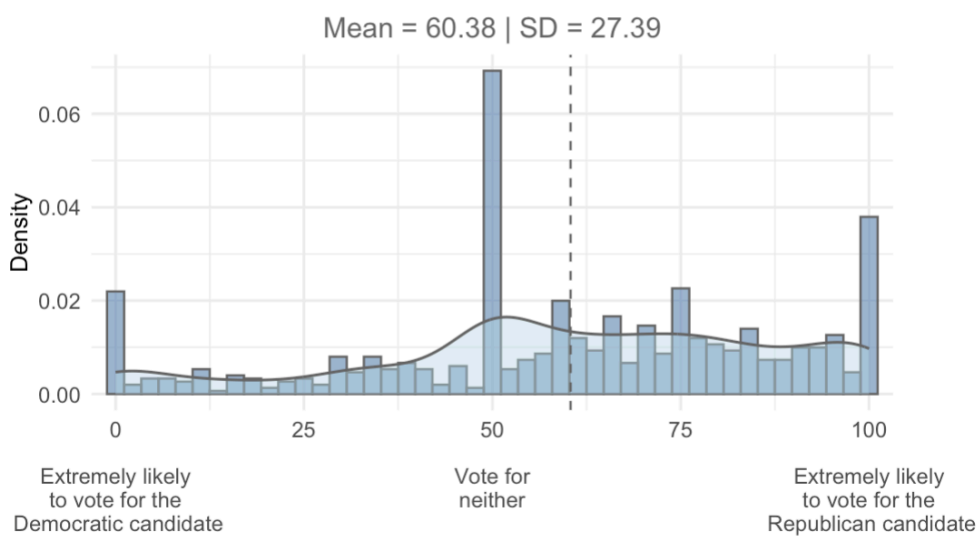
How likely would you be to vote for the Republican candidate if you learned that **they said they would ignore unfavorable court rulings by Democratic-appointed judges?**

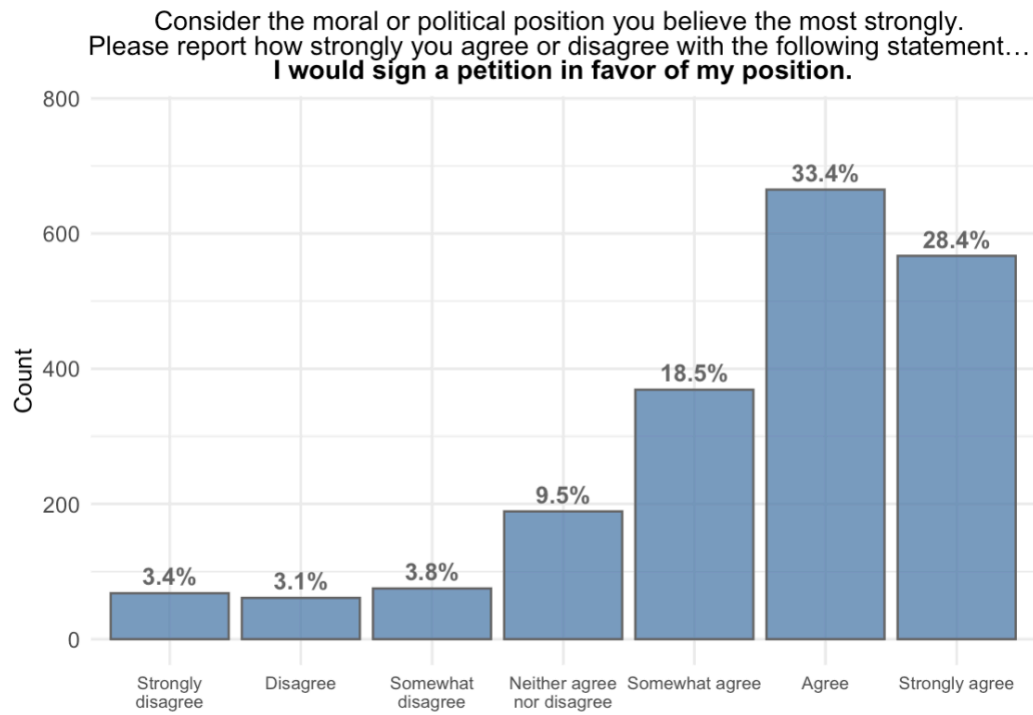


How likely would you be to vote for the Republican candidate if you learned that **they said they would prosecute journalists who accuse them of misconduct if the journalists won't reveal their sources?**

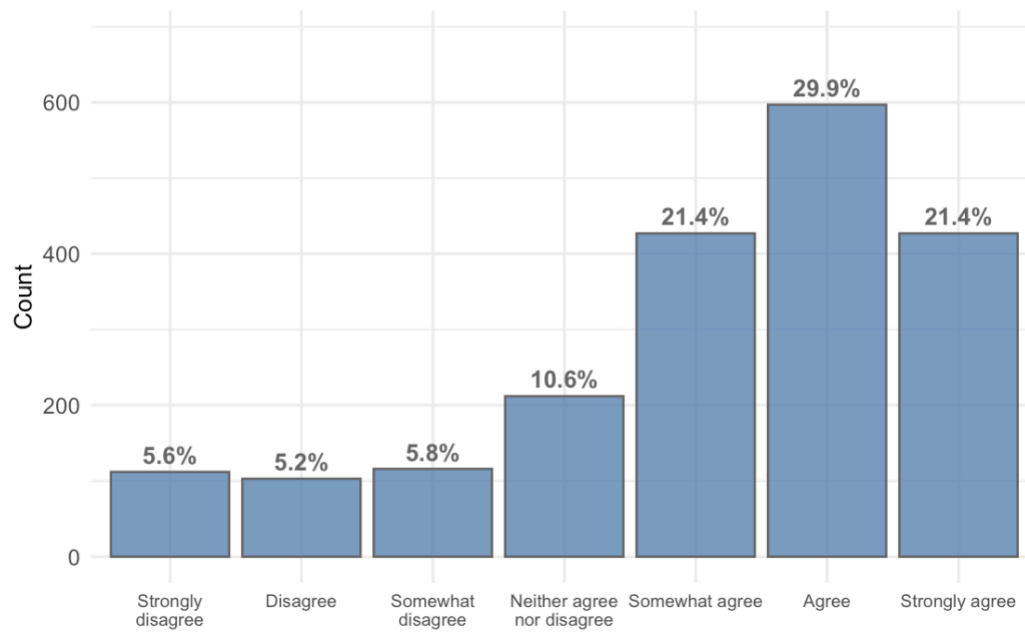


How likely would you be to vote for the Republican candidate if you learned that they said that **Republicans should not accept election results if they do not win?**

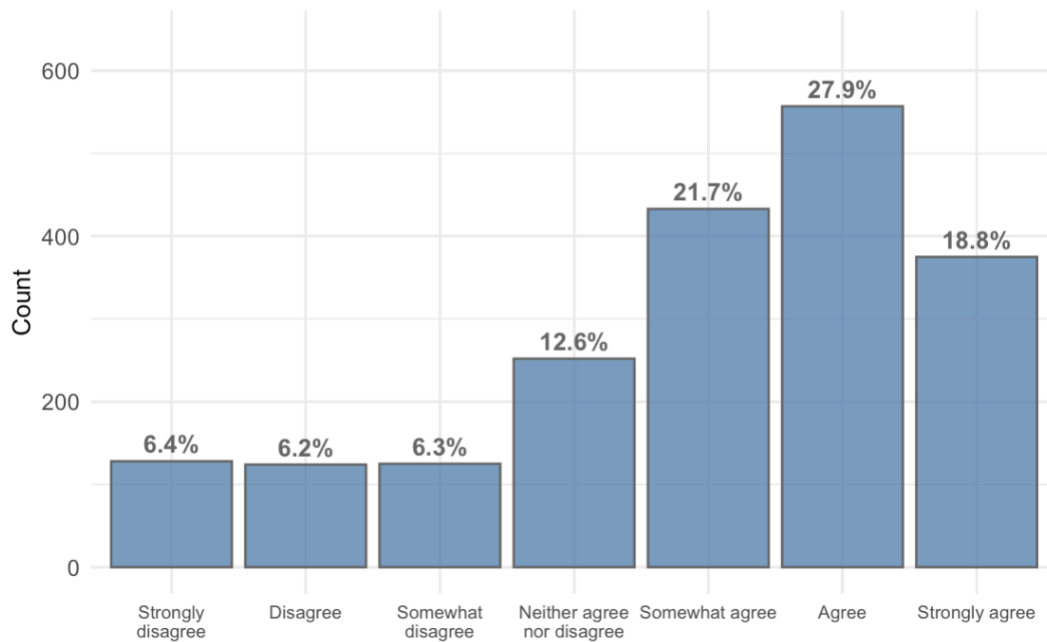




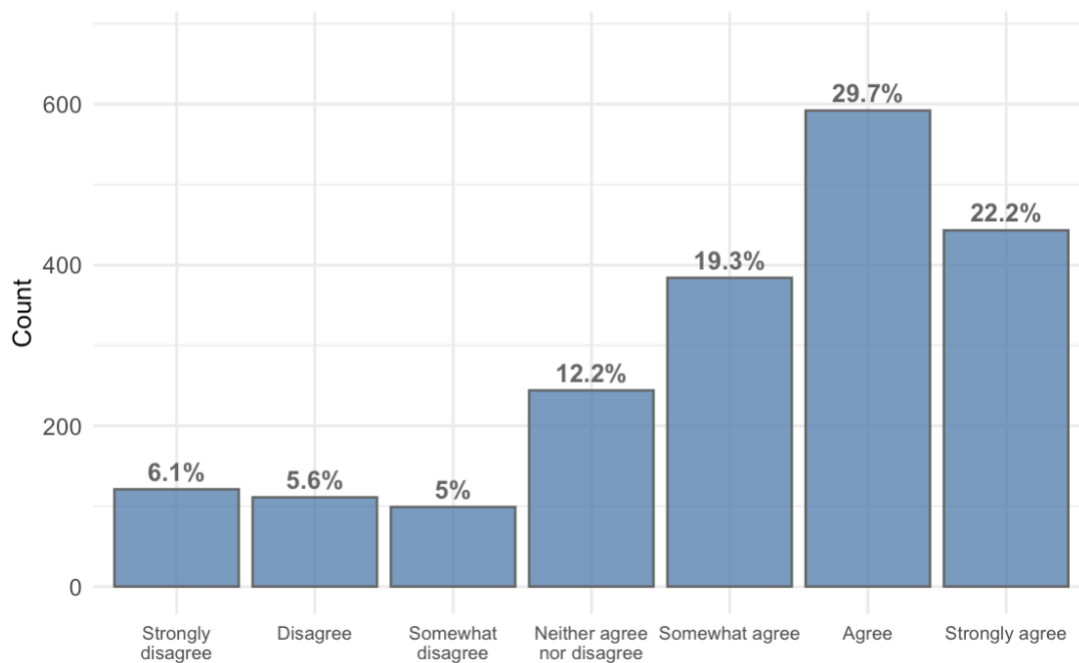
Consider the moral or political position you believe the most strongly.
Please report how strongly you agree or disagree with the following statement...
I would attend a rally, a march, or a protest to assert my position.



Consider the moral or political position you believe the most strongly.
Please report how strongly you agree or disagree with the following statement...
I would write to institutions and newspapers to bring forward my position.



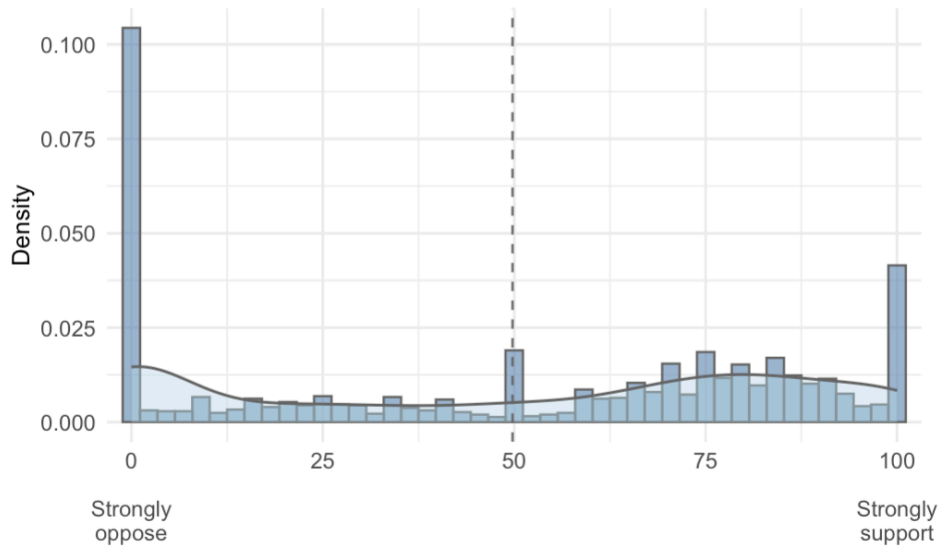
Consider the moral or political position you believe the most strongly.
Please report how strongly you agree or disagree with the following statement...
I would collaborate in organizing an event to promote my position.



Please indicate how much you oppose or support the following political position:

Reducing access to abortion

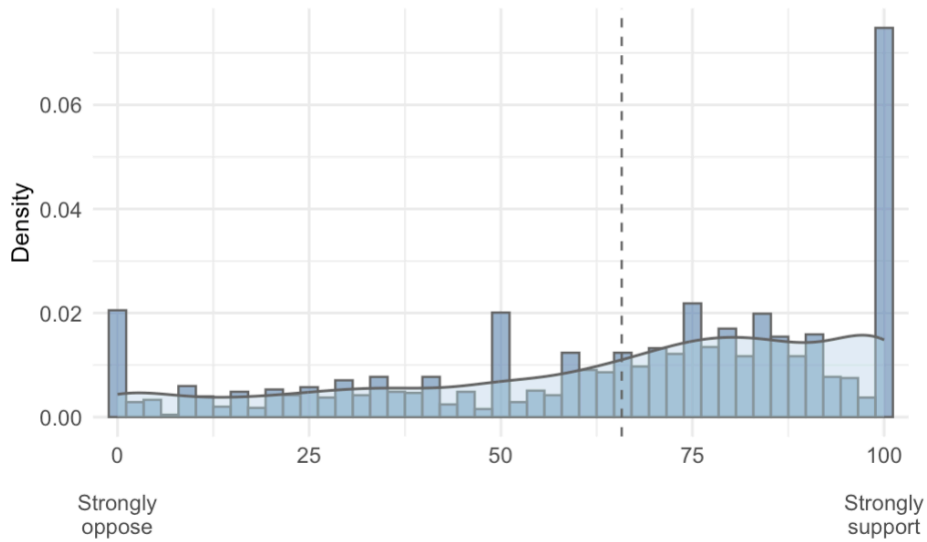
Mean = 49.78 | SD = 36.91



Please indicate how much you oppose or support the following political position:

Providing a path to citizenship for undocumented immigrants

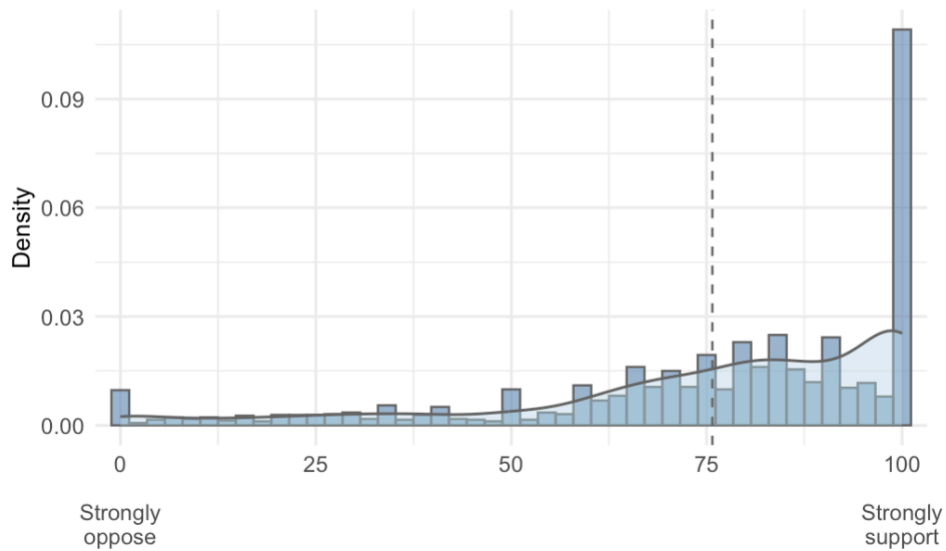
Mean = 65.77 | SD = 29.72



Please indicate how much you oppose or support the following political position:

Increasing restrictions on gun ownership

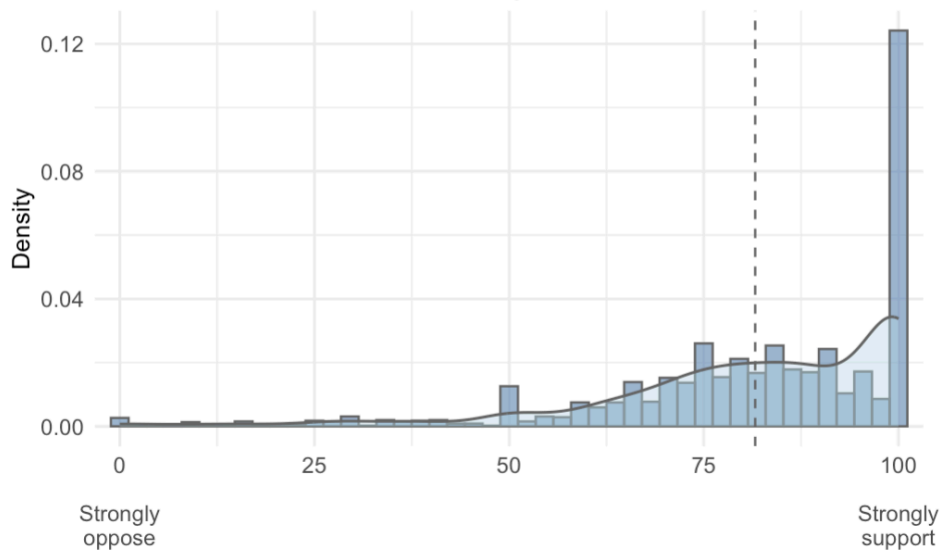
Mean = 75.73 | SD = 25.56



Please indicate how much you oppose or support the following political position:

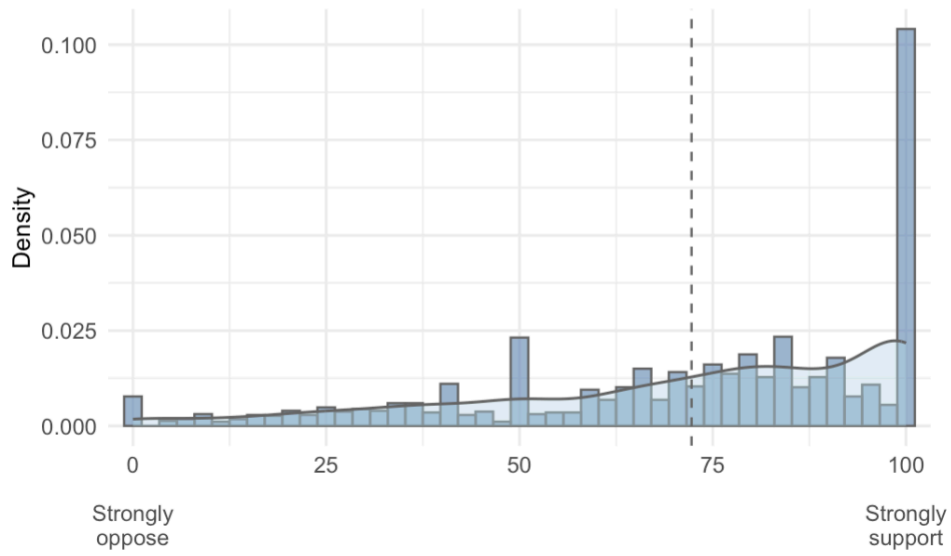
Increasing government regulations to protect the environment

Mean = 81.6 | SD = 19.55



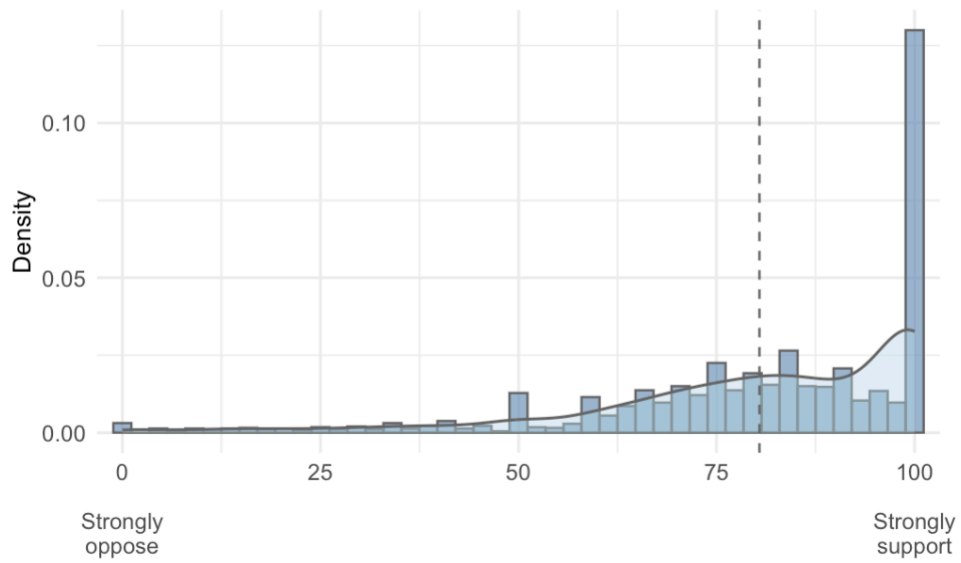
Please indicate how much you oppose or support the following political position:
Raising taxes on the wealthiest Americans

Mean = 72.26 | SD = 26.53

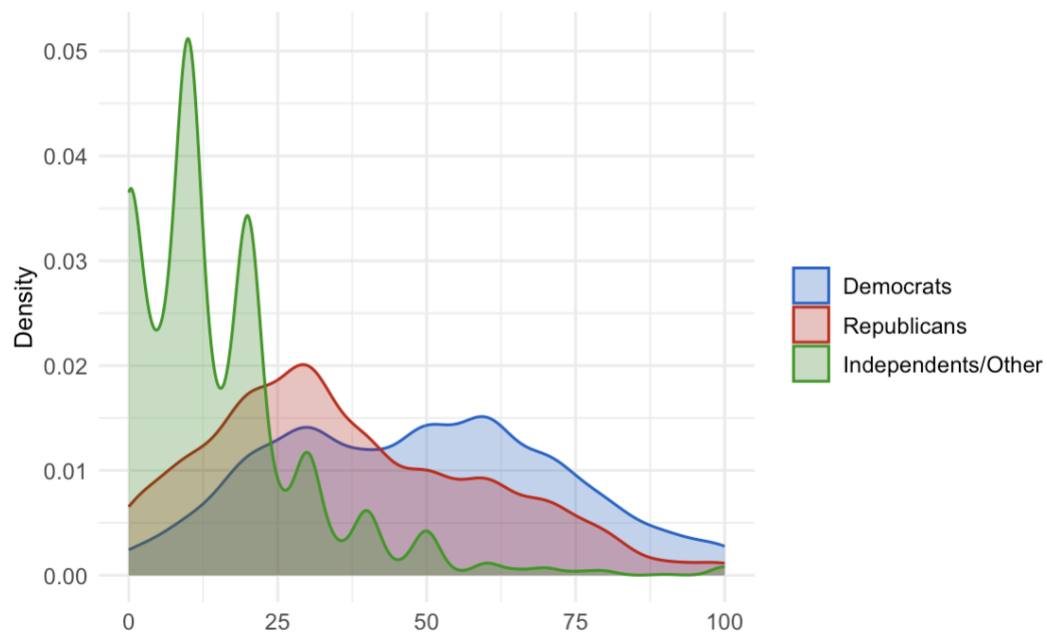


Please indicate how much you oppose or support the following political position:
Expanding Medicaid to cover all currently uninsured Americans

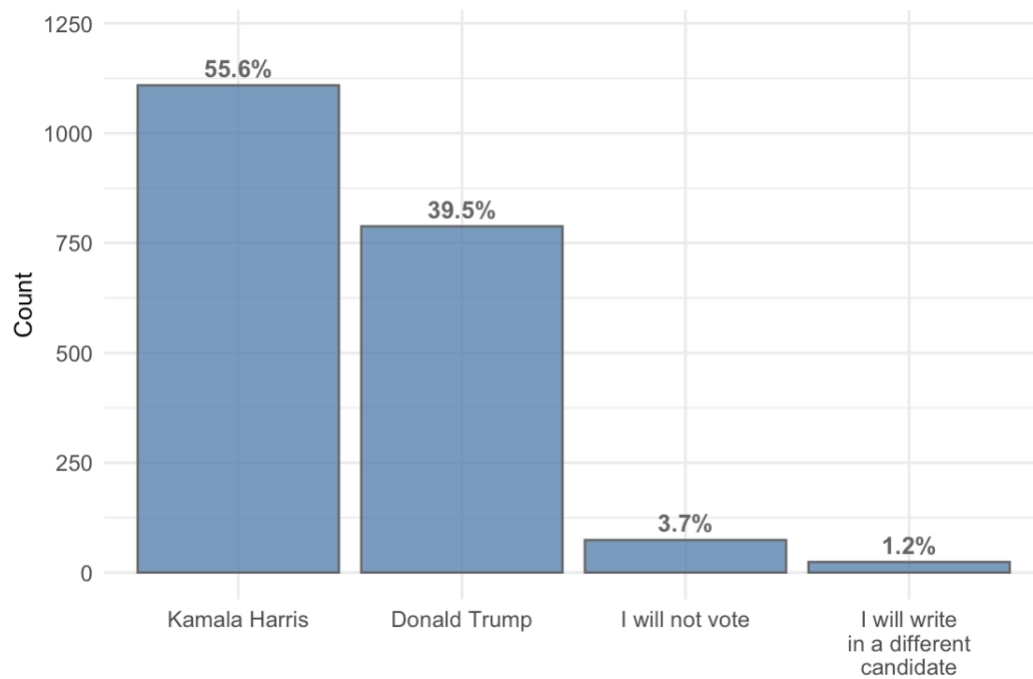
Mean = 80.41 | SD = 21.29



In general, **what % of the people who you know and interact with on a regular basis offline** are Democrats, Republicans, and Independents?



Who will you vote for in the 2024 U.S. Presidential Election?



2.0 Bluesky Platform details

2.1 Bluesky Social Details

Bluesky Social (henceforth “Bluesky”) is an open-sourced social media app. According to their blog and terms of service, Bluesky prioritizes decentralization away from a single ownership entity by giving users ownership of their own data, and developers complete control of the information ecosystems that they help build. Key for this research project, Bluesky has a feature to implement custom feed-ranking pages that leave the question of content moderation and curation in the hands of specific users, developers, and communities.

Bluesky is rapidly expanding and has over 5 million registered users as of February 2024. Bluesky is becoming an important platform for news and politics, as several prominent outlets and news organizations, such as the New York Times, NPR, USA Today, Washington Post, and Bloomberg recently joined. Bluesky also hosts prominent political elites, such as Alexandria Ocasio-Cortez, Arnold Schwarzenegger, and several U.S. Congressmen. See Fig. S1 for an example of Bluesky’s user interface.

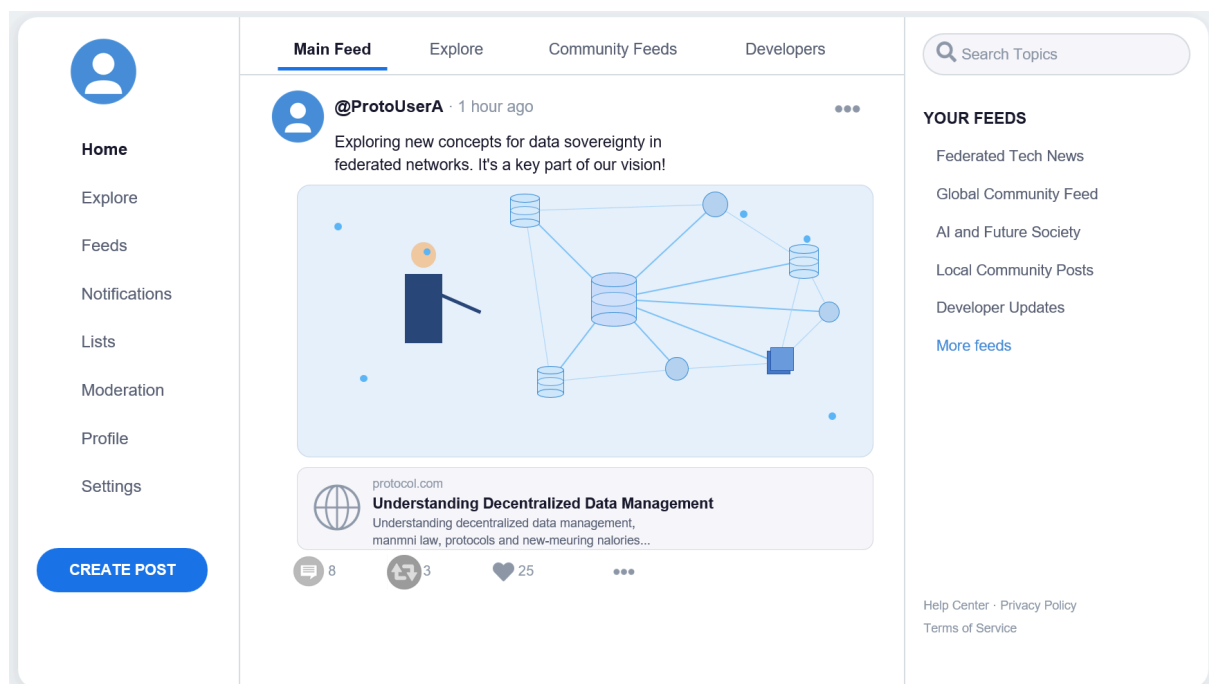


Fig. S1: Conceptual rendering of Bluesky’s homepage (web version). A conceptual rendering (not actual screenshot) of how Bluesky implements a user interface similar to Twitter/X, but has additional functionalities such as “Feeds”, “Lists”, and “Moderation”. Bluesky also defaults users to the “Following” feed, a reverse-chronological feed of posts from those that the user follows, but gives them the option to view a variety of custom feeds, such as ‘Discover’, and ‘Developers’. These characteristics align with Bluesky’s policy of giving the user full ownership over their experience.

2.2 Why Bluesky?

We used Bluesky for our research project for two main reasons. First, Bluesky is a fully open-sourced social media platform. All of the application code is available on Github and details of every public user and public post are available and queryable using Bluesky's firehose.

Second, Bluesky allows custom feeds: In an effort to promote algorithmic choice as a foundational principle, Bluesky allows developers to create custom feeds and for users to subscribe to the custom feeds. By default, Bluesky pushes users to a reverse-chronological follower feed but allows users to enhance their experience by subscribing to any of many thousands of custom-hosted feeds, such as the most popular posts in the past day or week and topical feeds such as a developer feed, political science feed, or feed featuring up-and-coming artists. For the purposes of our study, users were defaulted into their randomly assigned feed-ranking condition, and were told to only view their feed as part of the participant instructions (we tracked their feed viewing activity)..

We had full visibility into the Bluesky codebase and API and were able to design our own custom feed-ranking algorithms. This allowed us to design algorithms with full transparency and full control over the specific features that determine how that algorithm affected the experience / content of the participants in our study.

3.0 Details of feed ranking algorithm development

3.1 Feed-ranking algorithm pipeline overview

We built an end-to-end recommendation system pipeline to support this project. The steps of the pipeline were broken into four stages: (1) Data collection, (2) Feature Generation, (3) Candidate Generation, and (4) Feed Ranking. We summarize each stage below, and reference sections of the SI containing full details for each corresponding stage.

Data Collection (Section 3.2)

1. Collect raw data from the Bluesky firehose (i.e., public messages that are being posted in real-time)
2. Filter raw data (e.g., spam, bots, hate speech, NSFW)

Feature Generation (Section 3.3)

3. Measure specific features of each post (e.g., whether it is a political post (and then its ideological lean), whether it is toxic, whether it displays constructive dialogue)

Candidate Generation (Sections 3.4)

4. Estimate what posts a user is most likely to engage with.

Feed Ranking (Sections 3.5)

5. Rank order posts given the ranking criteria depending on experimental condition
6. Display rank-order posts (feed) to participants

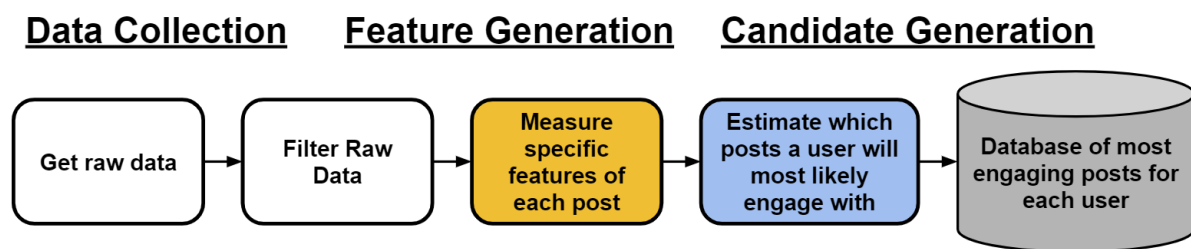


Fig. S2: Steps for collecting raw data from Bluesky firehose and estimating which posts a user is most likely to engage with.

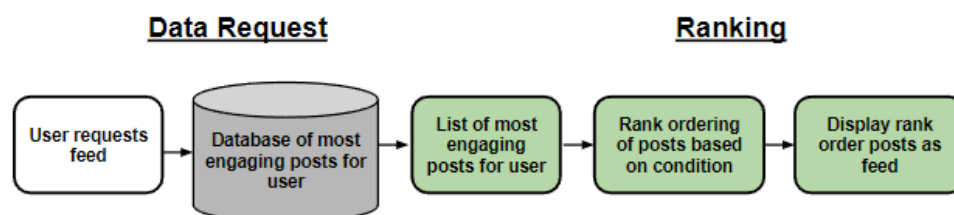


Fig. S3: Overview of steps for recommending a post to a user. When a user logs into Bluesky, we create their feed by querying our database to find the posts that a user is most likely to engage with, rank-ordering the posts given the criteria of their experimental condition, and then displaying the rank-order posts as a feed.

3.2 Data collection

3.2.1 Raw data collection

We collected ~20M posts during the course of the 8-week study. We collected these posts by subscribing to the Bluesky firehose, which publishes every update to every post in real time (including new posts, new comments, new likes on a post, etc.).

We collected user platform data through the Bluesky firehose. We collected likes, comments, shares, and new posts through the firehose. We collected information about each user's social network. We built a social graph of users and their followers and used the information about the social graph as part of our recommendation algorithms. These data were stored in a de-identified manner: the premade usernames are randomly assigned, and before downloading the data, we ran a personally identifiable information (PII) classifier on all text data to ensure no user posted PII in their public posts. All data was stored on Kellogg's secure Linux Cluster as well as on Kellogg's managed Amazon Web Services (AWS) service, as per IRB approval.

3.2.2 Filtering raw data

We filtered out posts that were spam, were posted by bots, contained hate speech, contained NSFW content, and were not English text. We used conventional off-the-shelf machine learning models to first filter for only posts written in English. Then, we filtered out posts that contained hate speech, spam, or NSFW content by blocking content from particular profiles or groups of profiles. We maintained denylists of these profiles by tracking profiles that consistently triggered our filtering classifiers, making use of content moderation lists curated by Bluesky, and by filtering out certain hashtags and communities (for example, Bluesky is home to niche communities such as FurryList that publish NSFW content, and such communities provide ways for filtering out their content for those not interested in such content).

3.3 Feature Generation

3.3.1 Metrics of on-platform behavior

We enriched the raw post data sourced from Bluesky with the following measurements of on-platform behavior which went into our model that predicts which posts individual users are most likely to engage with:

- Aggregate user-level engagement metrics: for each user, we aggregated their engagement metrics across a sliding window of three (3) days
 -
 - Number of posts that a user has engaged with
 - Number of posts that a user has engaged with that are written by authors who are in their network
 - Number of sociopolitical posts that a user has engaged with
 - Number of likes that a user has given
 - Number of comments that a user has written
 - Number of posts that a user has written
- Network metrics: for each user, we tracked details about their social network. We tracked the following:
 - Follower count
 - Number of new posts written by those in the user's network

3.3.2 Machine learning-powered feature generation

In addition to tracking on-platform behavior, we used machine learning to enrich our data with additional features. We used a combination of open-source classifiers, industry-standard models, and open-source large-language models (LLMs) for creating these features.

1. **Classification using Google Jigsaw's Perspective API**: Google Jigsaw's Perspective API is an industry standard for AI-enabled content moderation at scale ¹. The API exposes classifiers that fine-tune Google's PaLM 2 LLM² with online conversations annotated for the presence of toxicity, outrage, and other conversation attributes. For our use case, two classifiers were relevant:
 - a. Detecting toxicity: The Perspective API defines toxicity as "a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion", and returns the probability that a rater would rate a certain text as having toxicity.³ In Table S1, we list messages pulled from Bluesky between May 22nd, 2024, and May 29th, 2024, as part of our pilot.

Classified as Toxic	Not Classified as Toxic
hey george bush!!! suck my unit you bitch!!!	Fabulous Swedish poster for the 1929 German silent movie version of the Sherlock Holmes' adventure "Der Hund von Baskerville".
fuck you, you're not sorry at all	Weekend, let's goooooo!
HOLY FUCK SOMEONE FIXED BLENDERS SHIT UI AND MADE IT USABLE BY HUMANS WITHOUT INVESTING A SHIT TON OF TIME JUST KNOWING WHERE THE SINGLE OPTION YOU NEED IT FOR IS OR LEARN 300 GODSDAMNED SHORTCUTS RESKEETED TO THE MAX HOLY SHIT	In a first, the statewide median sales price for a previously owned single-family house in California has surpassed \$900,000
ohhh ohoho oh fuck this shit all the way off	A remarkable thing happened in March this year. For the first time, the fossil fuel rich state of Texas generated more electricity from solar power than from coal.
He would still be a bad person if he wasn't stupid, but God, he's so fucking stupid	"One year into the [Basic Income for Artists] pilot, research suggests that the BIA payment is having a consistent, positive impact on almost all indicators; positively impacting practice development, sectoral retention, well-being, and lessening deprivation."

Table S1: An example of a "toxic" (left) and "non-toxic" (right) post pulled from Bluesky's firehose and classified by the Perspective API.

- b. Detecting constructiveness: The Perspective API defines constructiveness as "makes specific or well-reasoned points to provide a fuller understanding of the topic without disrespect or provocation", and returns the probability that a rater would rate a certain text as having constructiveness. In Table S2, we list messages pulled from Bluesky between May 22nd, 2024, and May 29th, 2024, as part of our pilot.

Classified as constructive	Not classified as constructive
I often think about how Haber–Bosch, which is used to make nitrogen fertilizer (ammonia) from atmospheric nitrogen, uses 1% of the world's energy; a crazy amount for one process. But Shell's scenario for the year	He's a petty asshole?

2100 has direct air capture (DAC) using 13%(!) of the world's energy.

Air conditioning is an equity intervention for climate resilience, full stop. Every school should be safe, have healthy air & solar on the roof, & have air conditioning. In the next inflation reduction act, we need large investments in clean & green schools. Btw- they're in every political district

If you're job-hunting in CA & see a posting that doesn't disclose the job's pay band, as required by CA pay transparency law (for any employer with 15+ employees): You can file a complaint with the state Labor Commissioner's Office. You can even report anonymously!

Current studies show 18th cent construction trades were 30% women which is over twice as many women (by percentage) in construction today! The 19th cent is when we see a massive loss of women from trades they'd worked for centuries. Gendered notions of work are nonsense and often very recent.

Okay, this one is legit terrifying. DO NOT COOK CHICKEN TO 102° F. Chicken must be cooked to 165° F by conventional cooking.

There are so many reasons to support a universal basic income but the one I think about the most is how much domestic violence we might be able to prevent if people felt financially able to leave. That isn't the only reason but it seems like it is almost always one of the reasons.

memorial day weekend reminder that the river water in the PNW is absolutely not warm enough to swim in and won't be for months and you can very easily go into cold water shock which causes your muscles to lock up and you'll drown even if you're a strong swimmer.

you're telling me this was peak american culture? this?!?!?

ANGRY FACT-FREE POLITICAL OPINION!

There's your fucking winning strategy. Now please will you just...do that.

what all this gibberish of yours is about, I don't know, and I don't much care

Dems: we alone can save democracy
Also Dems: hearings? About treason adjacent SCOTUS justices? Surely ye jest

How many fucking articles can journalists write about these Nazis?

Table S2: Examples of "constructive" (left) and "non-constructive" (bottom) posts pulled from

Bluesky's firehose and classified by the Perspective API.

- c. Toxicity and constructiveness as mutually exclusive: Our intervention aimed to promote posts that were constructive and penalize posts that had toxicity. Our initial pilot testing showed that, except for very specifically contrived posts, toxicity and constructiveness will not co-occur; conditioning on a post being toxic, it will have low constructiveness (and vice versa), see Section 5.0 for more details on the pilot procedures. However, in order to remove rare cases of overlap (e.g., someone expressing hate speech in a way that uses "constructive features" for example, "All X should burn in hell because they are nasty, but I respect others who think differently than me"), we simply removed toxicity first, and then classify constructive dialogue second. Thus, examples like above would be removed from the pool of posts to uprank for constructiveness.

Post example	Toxicity binary label (and probability)	Constructiveness binary label (and probability)
the thing about Bari Weiss is she's a lying sack of shit	1 (0.823)	0 (0.146)
"Even affluent people would enjoy a better quality of life if they lived in a country with a more equal distribution of wealth, similar to a Scandinavian nation." We really need to reduce inequality. Universal basic income plus a mix of taxes is how. Everyone would benefit.	0 (0.015)	1 (0.743)

Table S3: Ratings of posts by both toxicity and constructiveness. When a post has a high probability of being toxic, it has a low probability of being constructive (and vice versa).

2. **Few-shot inference using GPT4o-mini**: OpenAI's GPT4o-mini was used to classify the following:
 - a. Sociopolitical classifier: We plan to predict whether a given post is sociopolitical in nature. We use the same definition as in recent work (Guess et al., 2024), and define sociopolitical as "relating to politics (government, elections, politicians, activism, etc.) or social issues (major issues that affect a large group of people, such as the economy, inequality, racism, education, immigration, human rights, the environment, etc.)". In Table S4, we list messages pulled from Bluesky's firehose based on their classification as either sociopolitical or not sociopolitical.

Classified as Sociopolitical	Not classified as Sociopolitical
The fact that Bill Maher has Republicans defending him should tell you everything you need to know about Bill Maher.	First few steps of making dinner pretty much every night: 1. Dice onion. 2. Put in pan to sauté until browned and

	translucent. 3. Get distracted by a cat and burn the onions.
TikTok might be violating our privacy, which is why we need to make sure it's owned by the people who think you should be tracked across state lines if you want an abortion.	Shel Silverstein doesn't get nearly enough credit for predicting ChatGPT.
the thing about the vances, the hawleys, the cottons and the desantises is that they all have loser energy, meaning they all have politics that seem to emerge from deep seated feelings of social inadequacy and resentment toward people with the status they crave	Star Wars reflects George Lucas's passion for hot rods, sci-fi pulp serials, samurai movies and WWII dogfights. Ghostbusters reflects Dan Aykroyd's passion for New Age paranormalism, New York architecture and HP Lovecraft. These sequels, mostly, reflect fandom's passion for fandom.
If JD Vance becomes vice president the blame is squarely on the shoulders of every liberal who bought Hillbilly Elegy and didn't listen to actual people from Appalachia telling you what a fraud he was.	Pretty depressing day at work today with lots of staff updating their resumes as we grapple with NASA's decision to shut down Chandra, the world's only ever high resolution X-ray space telescope, still returning fabulous science discoveries. Still hoping this can be reversed
"Just because they're appointing Kavanaugh doesn't mean Roe is dead, don't get so hysterical." "They're only coming for abortion, they're never going to touch birth control"	I saw a critique of The Boy and the Heron that said, the movie failed the audience because it felt like a movie that Miyazaki did just for himself. And I wish there was more room in our media landscape for artists to just tell the stories they want to tell and that's for them.

Table S4: An example of "sociopolitical" (left) and "non-sociopolitical" (right) posts pulled from Bluesky firehose.

- b. Political Ideology classifier: We classified the political ideology of the posts in the user's feed. For each user in the social network, we collected a subset of their posts classified as "sociopolitical" by our sociopolitical classifier and then passed them to the LLM for classification. We used four labels for our ideology classifier: (1) left-leaning, (2) moderate, (3) right-leaning, and (4) unclear. We had ideology information (left-leaning vs. right-leaning) for each of our study participants as part of the pre-study questionnaire. We could then measure the proportion of posts viewed by participants that is political ingroup vs. outgroup content.

Other classifications using open-source models: We have other classification use cases that we used open-sourced models for:

- c. Text embedding generation using embedding models: We used embedding models to represent the text within the posts. We used embedding models including OpenAI's GPT4⁴ and Meta's Llama embeddings⁵.
- d. Filtering non-English posts using machine learning: We used Facebook's fastText library for detecting which posts were written in English.

3.3.3 In-house classification tools for ingroup, moral and emotional content

Overview

We aimed to develop comprehensive dictionaries of intergroup, moral, and emotion words. There have been previous efforts to develop these kinds of dictionaries, such as the extended Moral Foundations Dictionary and the Moral Foundations Dictionary 2.0⁶. However, these dictionaries have been model-based (e.g., only focused on the dimensions of moral variance from moral foundations theory). We sought to develop dictionaries that were model-free, but also contained a large number of words that ordinary people would agree to communicate intergroup, moral, and emotional content.

To do this, we first fine-tuned large-language model classifiers to recognize intergroup words, morality words, and emotion words. We then curated a large corpus of language from social media (using Reddit's Pushshift dataset) and the news (using News API), and classified every word in these corpora using our models. This process yielded a large dictionary comprised of 72,287 words, with every word receiving an intergroup probability, an emotion probability, a morality probability.

Training Stimulus Curation

We curated our training wordlists using a combination of GPT stimulus generation and manual selection.

Intergroup Words. We prompted GPT4 to generate a list of social identities, such as gender, age, ethnicity, and socioeconomic status. After GPT4 identified these dimensions, we asked it to list words for each social dimension (an example of a gender identity word would be "woman"). We then manually selected words that are relevant to the social dimension and excluded ambiguous or repetitive words. The final list included 326 intergroup words.

Moral Words. We prompted GPT4 to generate words related to morality and their different part of speech. We then manually selected relevant words and excluded duplicated words and polysemies. The final list included 211 moral words.

Emotion Words. We relied on an existing list of emotion words—the word list from Shaver⁷—as our training data, which has been used in previous computational analyses. This list comprised 353 words.

Data Pre-Processing

Before model fine-tuning, we used a five-step procedure to curate our training data. First, we transformed all words into lowercase. Second, we removed compound terms that contained hyphens or multiple words. Third, we added a further 350 words that we deemed irrelevant to the emotion, moral, and intergroup categories. Fourth, we attached available definitions to words to provide word usage context using Wiktionary. We found that 81.9% of intergroup words, 85.3% of moral words, 97.5% of emotion words, and 95.4% of "other" words had available definitions. Fifth, we split all words into a training set (56.25%), an evaluation set (18.75%), and a testing set (25%). We

oversampled the intergroup and moral words to match the average of emotion and other words in the training set.

Model Performance

We fine-tuned 10 BERT-family models and identified distilbert-base-uncased and roberta-base as the two best models. Table 1 lists the F1 scores for each semantic dimension, as distinguished from words outside the dimension. For example, emotion words could either be correctly classified as an emotion word, or incorrectly classified as a moral word, intergroup word, or a word belonging to none of the categories. We set the classification probability at 0.50.

Table S5

F1 score on the Testing Set with a 0.50 Probability Classification Threshold

Model	Emotion	Intergroup	Moral
distilbert-base-uncased	0.83	0.82	0.83
roberta-base	0.82	0.81	0.81

Model Validation in a New Sample of Words

We next conducted an out-of-sample validation by applying the model to a further 1,200 words which were not included in our original training materials, and comparing our model classifications of these words with ratings made by 1,200 human subjects recruited from Prolific ($M_{\text{age}} = 40.88$; 562 men, 632 women, 13 non-binary). We sourced the 1,200 words from the corpus described in the next section. Each participant rated 10 words with the instructions “Is the word ____ a word about social groups/morality/emotion”, and response options of “yes,” “no,” and “I don’t know” (converted to NA). This meant that each word was rated 10 times.

We found that our model classifications were most accurate when we averaged the probabilities across the two best-performing models (distilbert-base-uncased & roberta-base). These model-averaged probabilities correlated at $r = 0.63$ with average ratings of intergroup, $r = 0.59$ with average ratings of morality, and $r = 0.58$ with average ratings of emotion.

We also compared the performance of our combined model probabilities with the performance of the average human rater. To do this, we combined the item-total correlation of each human subject in our study (the correlation between their ratings and the other people who rated the same words), and also computed the item-total correlation of the combined model probabilities (the correlation between the model probabilities and the human subjects).

We found that the average item-total correlation among human subjects ($r = 0.66$) was highly similar and slightly lower than the model item-total correlation ($r = 0.67$), suggesting that our models performed on par with an average human subject on Prolific.

Model Application to a Corpus of Language on News and Social Media

Corpus Curation. We collected all words used in 468 political subreddits. Following Rajadesingan and colleagues⁸, we define subreddits with more than 25% political comments as political subreddits. On Reddit, there are many user-created words. Therefore, we limited the scope to words that appeared at least 2000 times in those subreddits, which captured 99.5% of the words used. After deduplication, this left a sample of 71,879 words.

We next expanded our corpus by adding news articles from various news sources between January 1, 2022, and January 31, 2023, which we sourced using the News API tool. We added a total of 2,292 articles, using ratings by AllSides [see⁹] to ensure a politically balanced sample. We extracted the full vocabulary of words in these news articles after removing punctuation, numbers, single letters, and misspelled words from full article content. This process identified a further 408 words, leaving a full sample of 72,287 words.

Model Application. After compiling our full vocabulary, we submitted it to our models. Table 2 summarizes the number of words which were rated as having a 0.50, 0.75, and a 0.90 probability of communicating intergroup, moral, and emotional information. These probabilities have been averaged across the distilbert-base-uncased & roberta-base models.

Table S6

Number of Words in Each Dictionary According to Three Probability Thresholds

Probability Threshold	Emotion Words	Intergroup Words	Moral Words
0.5	3,961	19,453	4,022
0.75	956	7,535	1,307
0.9	91	669	60

Summary

There are two approaches to using these comprehensive dictionaries. The first approach involves identifying the proportion of intergroup, moral, or emotion words in some posts based on the number of words that exceed some threshold divided by the total number of words. We recommend a value of 0.75, since this appears to balance comprehensiveness (including a large number of words) with accuracy (minimizing the number of words that are falsely classified as IME). The second approach involves averaging the probabilities across all words in a post. For example, a

post where words have an average emotion probability of 0.18 is conveying more emotion than a post where the words have an average emotion probability of 0.11.

3.3.4 Defining “super posters”

We define a super poster as someone who posts more than 5 times in a 24-hour window, which is determined by calculating the top 1% according to daily post tracking here: <https://bsky.jazco.dev/stats>. These users were penalized such that the probability that they appear in the top rank posts will decrease if they post beyond 5 posts. Each day, a users' post count was reset so the penalization decayed. The goal with this approach was to reduce the number of people who are attempting to dominate a conversation in a given time period, rather than over time. Some users posted about the same topic everyday, but they were not penalized in our model unless they post over 5 or more times in a specific 24-hour window.

3.4 Candidate generation

For candidate generation, we pulled posts that had been the most engaged with Bluesky in the past 24-hour period as well as the past 7 days. From this base pool, we used an approximate nearest neighbors (ANN) algorithm to surface candidate posts that the user was most likely to engage with given the posts that they engaged with in the past.

3.5 Feed ranking and scoring

For our model we will use a simple linear weighted sum (detailed in Section 4.2 below).

3.6 Feed postprocessing

After candidate generation and feed ranking, we will do feed postprocessing for the diversified extremity condition. In this step, we will perform calibration as well as do quality assurance checks. In this step, we will:

- Split posts 50:50 in-network:out-of-network. We plan to follow the procedure from Twitter's own recommendation algorithm¹⁰, which does a similar 50:50 split of in-network and out-of-network posts for a given user.
- Remove any NSFW or spam content that our upstream filters may have missed.

4.0 Algorithm details

4.1 Suggesting posts using approximate nearest neighbors (ANN)

To generate candidates during inference, we used approximate nearest neighbors (ANN). ANN allowed us to efficiently approximately determine the nearest neighbors of a given vector by partitioning the vector space, thereby exponentially reducing the search space. ANN is commonly used in recommendation systems in industry, such as YouTube, Twitter, Amazon, and Alibaba, among others. We used an open-source implementation of ANN from Facebook’s Faiss¹¹ and Spotify’s Annoy¹² algorithms. We plan to take the embeddings of posts that have been most liked as well as the embeddings of posts that the user has previously liked, perform ANN, and learn which posts out of those that are most liked are most similar to content that they’ve previously liked.

4.2 Ranking and scoring algorithms

For our model, we used a simple linear weighted sum.

$$Score = \alpha P(toxic) + \beta P(constructive) + \gamma(is\ superposter)$$

We scored posts based on maximizing the probability of engagement, but penalized civic posts that were toxic or written by super posters, and upranked civic posts that were constructive. For non-civic posts, we maximized the probability of engagement. For two posts that have the same score, we selectively upranked the post that was more recent.

5.0 Pre-pipeline pilot (conducted during Stage 1 report)

5.1 Data

The goal of our pilot was to provide a proof-of-concept that our proposed criteria for ranking in the diversified extremity condition in fact creates feeds with content distributions similar to the base rates for all Bluesky posts, as compared to an engagement-based algorithm. To do so, we created “simulated feeds” where we pulled data from the firehose and created two different top 250 rank-ordered feeds: the engagement-based feed, and the representative-diversification feed. These simulated feeds represent an estimate of the final output of our pipeline once it is fully implemented to run in real-time for the study as described in sections 3.0 and 4.0.

We collected posts from the Bluesky firehose from May 22nd, 2024, to May 29th, 2024. In that same time period, we also collected posts from the “Most Liked” feed on Bluesky, which aggregates the most liked posts every 24 hours. After filtering out non-English, spam, and NSFW content, we had a baseline pool of 160,046 posts from the firehose and 5,117 posts from the most liked feeds. For each post, we classified their toxicity and constructiveness using the Perspective API, as previously detailed.

We calculated the base rate of toxicity and constructiveness in the baseline pool of all Bluesky posts based on the sample of 160,046 posts that we obtained from the firehose. We then created two feeds, an engagement feed of the top 250 most liked posts, and a diversified extremity feed of the top 250 posts after being re-ranked based on our scoring algorithm.

We used the following scoring algorithm:

1. Take the natural log of the likes on a post (so that posts with more likes don’t disproportionately get biased in the scoring model).
2. Calculate a normalized weighted sum of the toxicity and constructiveness probabilities returned by the Perspective API.

$$\alpha P(\text{toxic}) + \beta P(\text{constructive}) / P(\text{toxic}) + P(\text{constructive})$$

3. Return the normalized weighted sum multiplied by the natural log of the likes on a post.

$$[\alpha P(\text{toxic}) + \beta P(\text{constructive}) / P(\text{toxic}) + P(\text{constructive})] * \log(\text{likes})$$

We used $\alpha = 0.965$ and $\beta = 1.02$ for our weights.

Our scoring algorithm weighs the number of likes on a post (a measure of how engaging a post is) by how toxic and constructive that post is.

First, we report several metrics for the two feeds. Data and code are available here: <http://github.com/METResearchGroup/bluesky-research>. We use a threshold of $p=0.7$ for classification.

Table	Metric (proportions of top 250 posts in the feed)
-------	---

	Toxicity	Change in toxicity % compared to baseline	Constructiveness	Change in constructiveness % compared to baseline
Baseline rate on Bluesky	2.894%	0%	1.434%	0%
Engagement-based	4.310%	+48.929%	0.862%	-39.888%
Diversified Extremity	2.8%	-3.248%	2.0%	+39.470%

Table S7. Metrics for the top 250 posts in the feeds for each of the three conditions.

We observe two results here: (1) an engagement-based feed has a higher rate of toxic posts and a lower rate of constructive posts as compared to Bluesky's baseline rate, and (2) our diversified extremity algorithm produces content with toxicity and constructiveness rates more closely aligned with the natural base rates on Bluesky.

Below, we also display the top messages in order of their ranking (i.e., order by which would be displayed to a user) for each of the conditions.

Engagement-based	Diversified Extremity
Google, 1998: We're going to organize the world's information and make it universally accessible and useful. Imagine the whole world's knowledge in your pocket! Google 2024: Drink your own piss. Now eat glue.	I'd like to dispute the basis of this kind of meme: it isn't that there are no longer 30 year olds who look like this, they just aren't permitted to be on tv or in movies anymore
Oh, so literally NO ONE is supposed to do war crimes?!?	CEO of Sony Picture group, an executive finally saying publicly almost all the stuff that needs to be done to make movies profitable again
Low level government employee: fuck, they're gonna figure out I accepted this free chicken wing Alito: today I put strobe lights on my confederate flag	*JUSTICE DEPARTMENT TO SEEK BREAKUP OF LIVE NATION-TICKETMASTER
One of the most important things you can do as an ally and as fighter of fascism is stop laughing at evangelical Christians and realize they mean every single word they say. They're not funny, they're absolutely terrifying. Take them extremely seriously and act accordingly.	One of the most important things you can do as an ally and as fighter of fascism is stop laughing at evangelical Christians and realize they mean every single word they say. They're not funny, they're absolutely terrifying. Take them extremely seriously and act accordingly.
Americans will use anything but the metric system	I hope people understand that what we are seeing is the systematic destruction of the civil war amendments by the Supreme Court, which are what made America and actual democracy

	and upon which all minority rights in the United States rely
I hope people understand that what we are seeing is the systematic destruction of the civil war amendments by the Supreme Court, which are what made America and actual democracy and upon which all minority rights in the United States rely	Sorry to say, archive.org is under a ddos attack. The data is not affected, but most services are unavailable. We are working on it & will post updates in comments.
The idea that Alito isn't willing to tell a woman what to do is a bit of a howler	Who called a JAWS rip off set in France 'Under Paris', and not 'Shark de Triomphe'?
Students who were pro-Apartheid in the 1980s probably also had trouble making friends in college, who gives a shit	if already wasn't apparent from his behavior on the court, Alito is straightforwardly an insurrectionist and should not be there.
whenever I think about how a movie's financial success has been reduced to essentially it's first weekend in theaters, I am reminded there used to be a thriving economy based on months-long schedules, rentals, VHS and DVD sales, regional theaters, syndication on TBS, and international release dates	Low level government employee: fuck, they're gonna figure out I accepted this free chicken wing Alito: today I put strobe lights on my confederate flag
People will point to something and say "see voting doesn't do anything" and it's literally just the consequences of Republicans winning elections	whenever I think about how a movie's financial success has been reduced to essentially it's first weekend in theaters, I am reminded there used to be a thriving economy based on months-long schedules, rentals, VHS and DVD sales, regional theaters, syndication on TBS, and international release dates
Bad News: I've been laid off from mmfa, along with a dozen colleagues. There's a reason far-right billionaires attack Media Matters with armies of lawyers: They know how effective our work is, and it terrifies them (him).	Ireland, along with [at least] two other European Union member-states, will officially recognise the State of Palestine tomorrow morning.
Seeing a lot of "this story is depressing not inspiring!" And yeah. Here's an actually inspiring story you may not know about: school lunch (and breakfast!) in Massachusetts is free for all students starting in 23-24 thanks to Democrats using money from the voter-passed millionaire's tax to fund it	People will point to something and say "see voting doesn't do anything" and it's literally just the consequences of Republicans winning elections

Table S8: Example feeds for the engagement versus diversified extremity conditions. Posts in the engagement condition that are toxic are penalized and appear lower, if at all, in the diversified extremity feed. In their place we see posts being upranked and placed higher in the feed if they are

constructive and non-toxic.

6.0 RQ1 supplementary information

6.1 Examples of IME content classifications by category

Below, we show examples of content classifications for each category in our IME framework.

Classified as Intergroup	Not Classified as Intergroup
Trump voters lacking cognitive sophistication will believe their cult leader rather than a successful businessman or economists.	his is my new cat, he's a bit grumpy but I love him! #Alart #CatsOfBluesky #GrumpyCat
"While barred from most government healthcare programs, undocumented residents pay more in healthcare taxes than the entire cost of their own care."	My newest creation is a little ghost bag. She uses her pouch to hold her baby, tiny books, or autumn leaves. Her insides are swimming with various souls that she's swallowed over the years. The souls quite enjoy hanging out there. I designed and sewed everything here.
Ironically enough, the vast majority of my American family were in the USA for centuries speaking an entirely other language that they had to be forced out of using in the 20th century and now they're some of the hardest-core Republican voters in the nation	After missing it entirely the first time around, I found a cheap DVD boxset of the entire run of the Hannibal TV show and OH MY GOD What an absolutely exquisite thing it was. Brain's awlirl.
Classified as Moral	Not Classified as Moral
Why would he, a con man fraudster, go out of his way to have that specific screwdriver in his toolbox, when his goal was political power in the US. And why did it work so effectively for so long. And what does that mean. Just don't think about it though. Shhh no thinky that's trouble.	The sequel to Sucker Punch's Ghost of Tsushima, Ghost of Yotei, being set in 1603 Hokkaido has some interesting implications.
The media is complicit, through malice or incompetence, in this trend where Trump's freakish bigoted ignorance and volatility is treated as an accepted baseline. label_emotion	Overall I'm still super happy the app's growing! I just have to remember to switch back into twitter mode of operations... dying
The racism is right out in the open	Same here all week... in fact much of

September... and I moved from the UK to get away from all that wet stuff!	
Classified as Positive Emotion	Classified as Negative Emotion
His voice!!!! Unforgettable <3 <3 <3	Kids these days don't even know how to come in clutch, they only know how to come in automatic. *This is a joke about driving, please comment and then carry on, thank you I'm so alone.
Well we'll either be delighted or totally horrified lol	sorry i am very frustrated about this.
There will never come a day when this doesn't make me laugh.	like, i am not gonna spend my whole night on this, but the purpose of this is to spread fear, but fear is not the only response, ANGER IS AN ENERGY

Table S9. Examples of posts for each label within our IME framework. For each classification, we use $p=0.5$ as our threshold.

6.1 Descriptive statistics and correlations

Content Type	Reverse Chronological		Diversified Extremity		Engagement-Based		Baseline Firehose	
	M (SD)	Median	M (SD)	Median	M (SD)	Median	M (SD)	Median
Intergroup Content	0.005 (0.004)	0.004	0.006 (0.008)	0.004	0.006 (0.007)	0.004	0.006 (0.001)	0.006
Moral Content	0.010 (0.006)	0.009	0.009 (0.007)	0.009	0.011 (0.009)	0.010	0.010 (0.002)	0.009
Negative Valence Content	0.237 (0.037)	0.235	0.249 (0.077)	0.237	0.264 (0.082)	0.252	0.252 (0.032)	0.244
Positive Valence Content	0.482 (0.033)	0.484	0.444 (0.066)	0.448	0.447 (0.063)	0.453	0.471 (0.027)	0.477
Toxic Content	0.062 (0.015)	0.061	0.065 (0.029)	0.061	0.072 (0.033)	0.069	0.063 (0.009)	0.063
Sociopolitical Content	0.220 (0.087)	0.207	0.276 (0.119)	0.266	0.305 (0.129)	0.298	0.242 (0.070)	0.228
Moral Outrage Content	0.126 (0.041)	0.122	0.180 (0.087)	0.160	0.201 (0.090)	0.188	0.137 (0.047)	0.126
Constructive Content	0.085 (0.022)	0.087	0.076 (0.032)	0.073	0.083 (0.034)	0.082	0.090 (0.011)	0.092

Note. Proportions range from 0 to 1. M = Mean; SD = Standard Deviation.

Table S10. Means, standard deviations and medians for content proportions by algorithmic condition and baseline firehose.

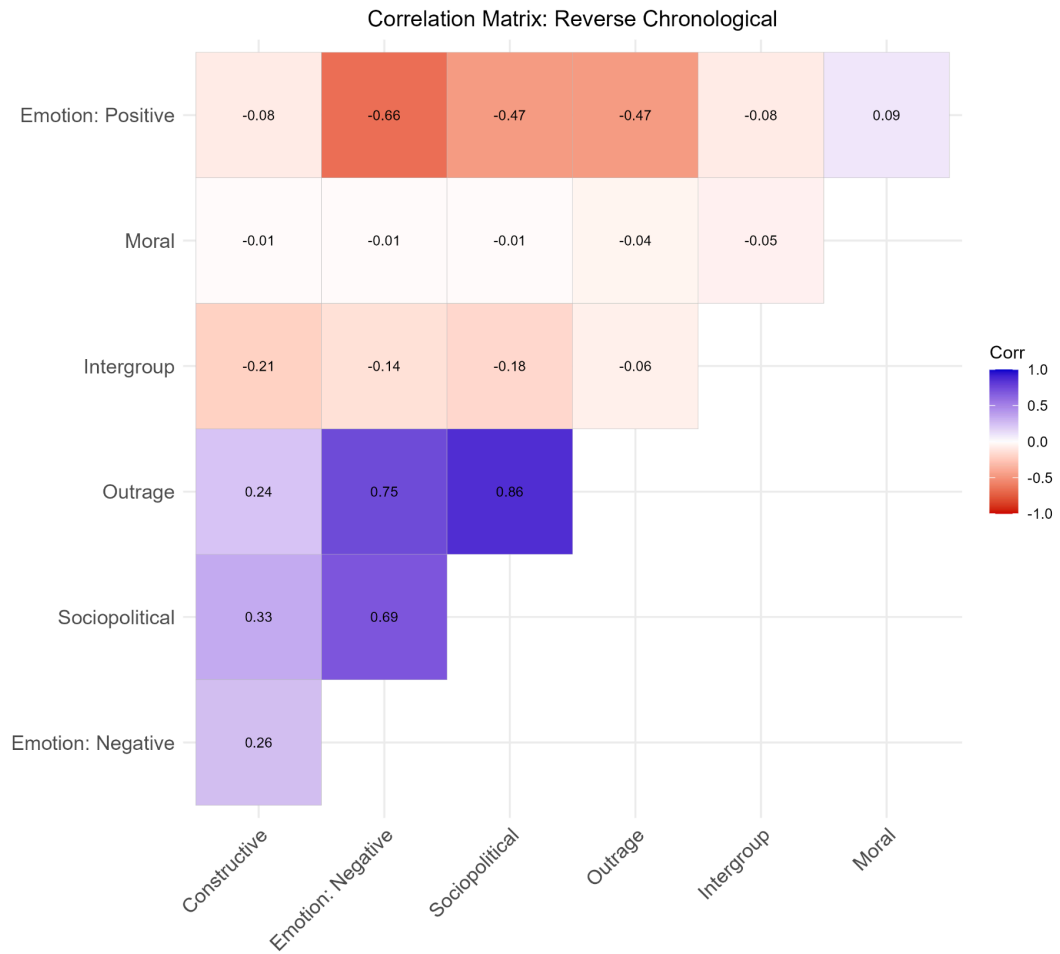


Fig. S11. Correlation matrix for all content types measured in feeds, reverse-chronological condition.

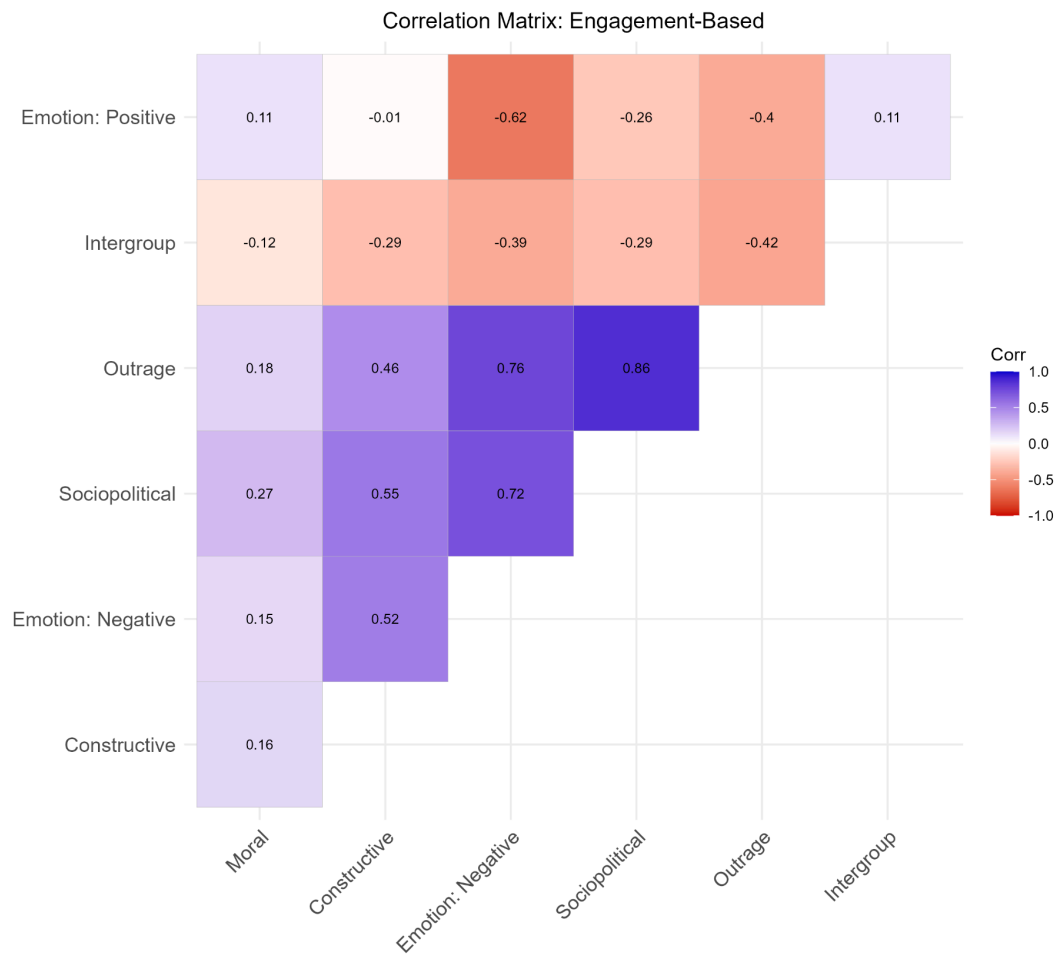


Fig. S12. Correlation matrix for all content types measured in feeds, engagement-based algorithm condition.

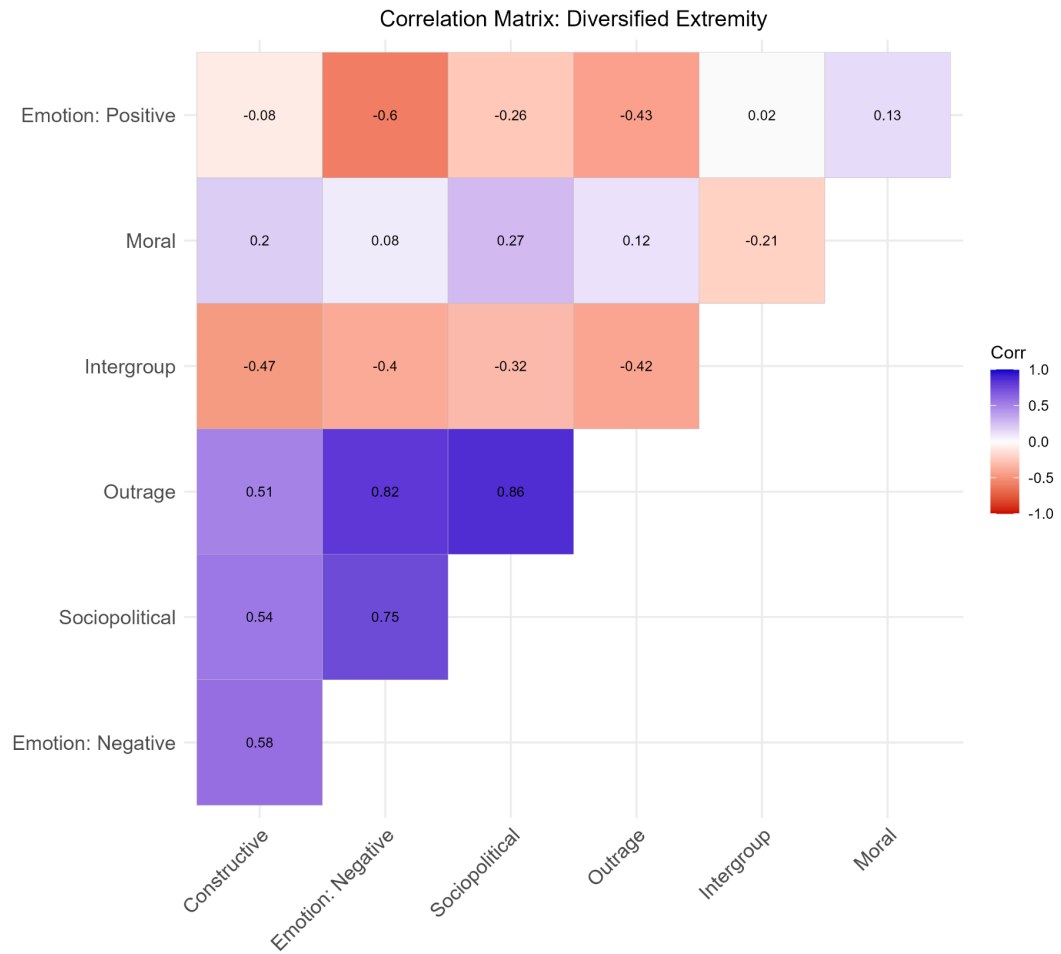


Fig. S13. Correlation matrix for all content types measured in feeds, diversified extremity condition.

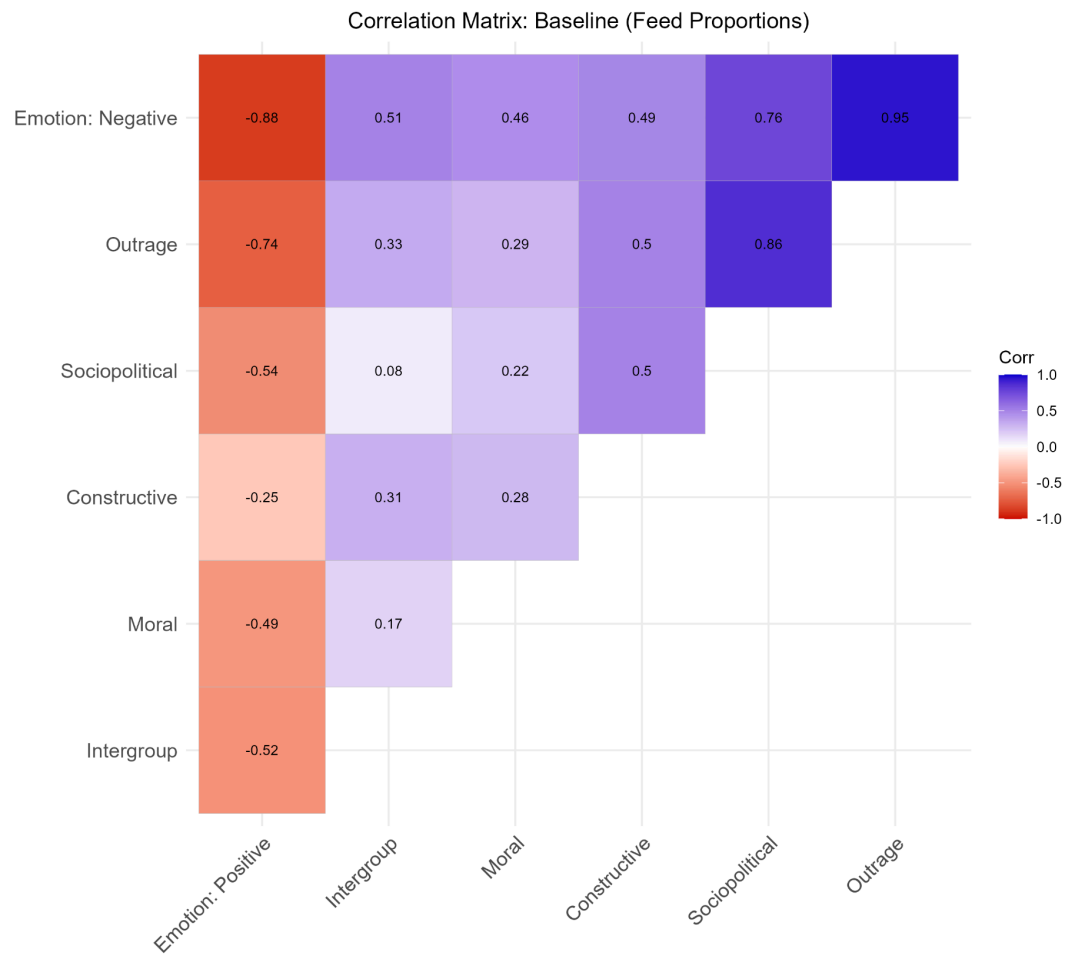


Fig. S14. Correlation matrix for all content types measured in feeds, baseline estimate from Bluesky firehose.

6.2 Estimating baseline proportions of content on Bluesky

To estimate the baseline proportions of content on Bluesky, we sampled approximately 100,000 posts daily published from the Bluesky firehose throughout the course of our study duration. We filtered out posts that were non-English or marked as having spam/NSFW content. We used this pool of ~6M posts in order to estimate the baseline proportions of content on Bluesky for the 2-month duration of our study, averaging the proportions of content per day.

We used this estimate of baseline content to assess how representative each condition was of the true base rate for each category. Overall, the reverse-chronological condition closely reflected the true base rate, the engagement-based algorithm substantially overrepresented IME content, and the diversified extremity condition did so to a lesser degree. Figs S15-S22 show daily proportions of content over the study period for each condition.

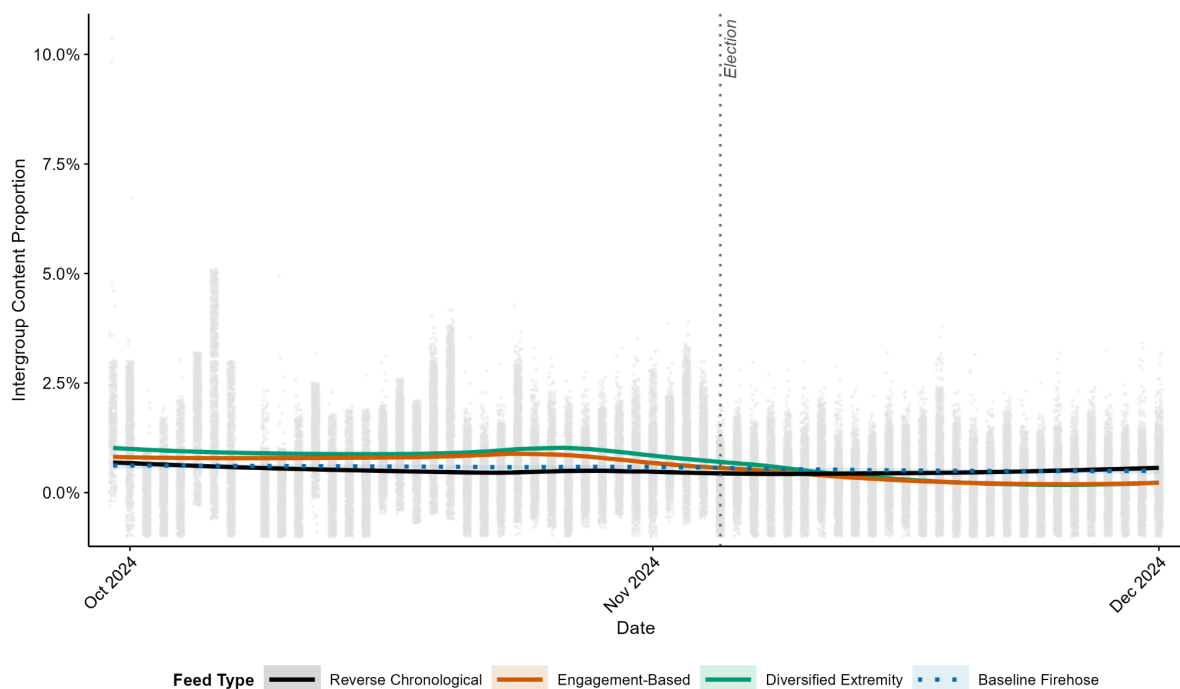


Fig. S15. Daily proportions of intergroup content by condition, compared to the baseline firehose (estimate of true base rate of daily content in English-speaking Bluesky).

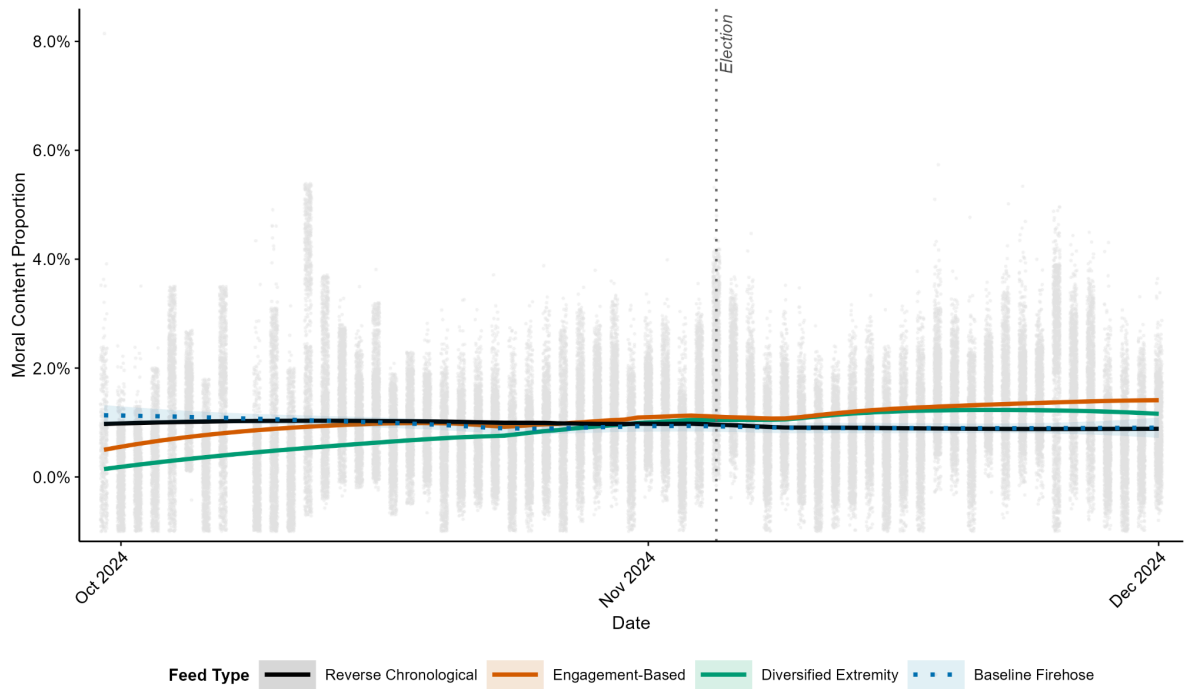


Fig. S16. Daily proportions of moral content by condition, compared to the baseline firehose (estimate of true base rate of daily content in English-speaking Bluesky).

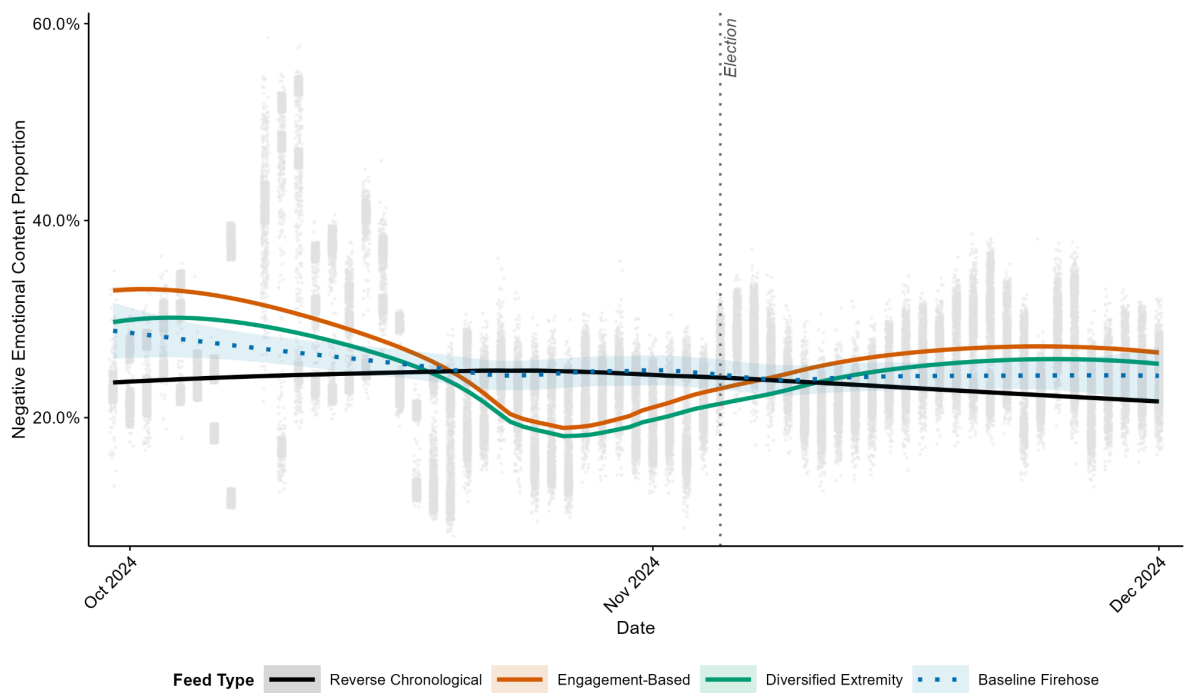


Fig. S17. Daily proportions of negative emotional content by condition, compared to the baseline firehose (estimate of true base rate of daily content in English-speaking Bluesky).

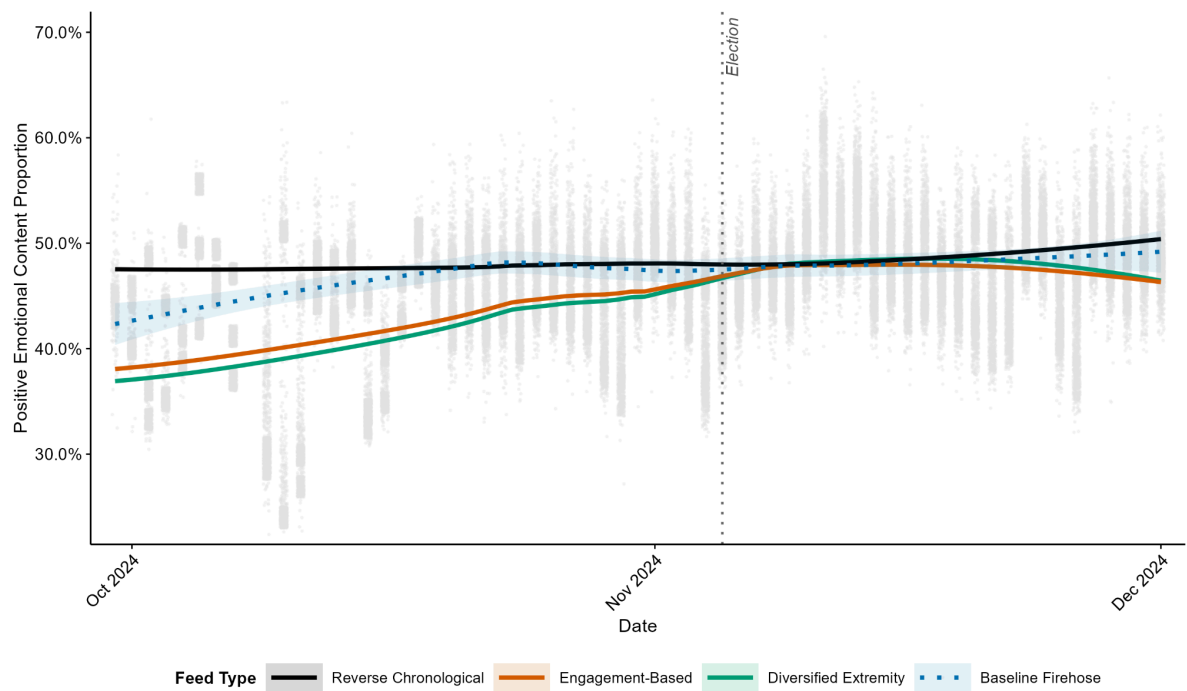


Fig. S18. Daily proportions of positive emotional content by condition, compared to the baseline firehose (estimate of true base rate of daily content in English-speaking Bluesky).

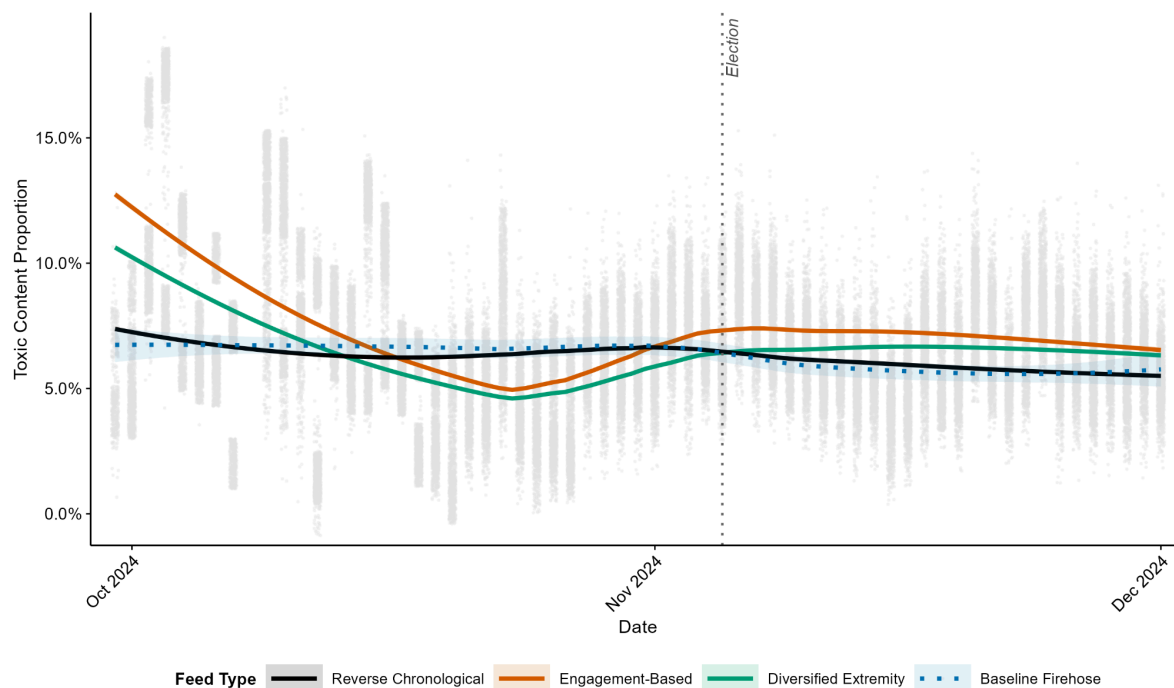


Fig. S19. Daily proportions of toxic content by condition, compared to the baseline firehose (estimate of true base rate of daily content in English-speaking Bluesky).

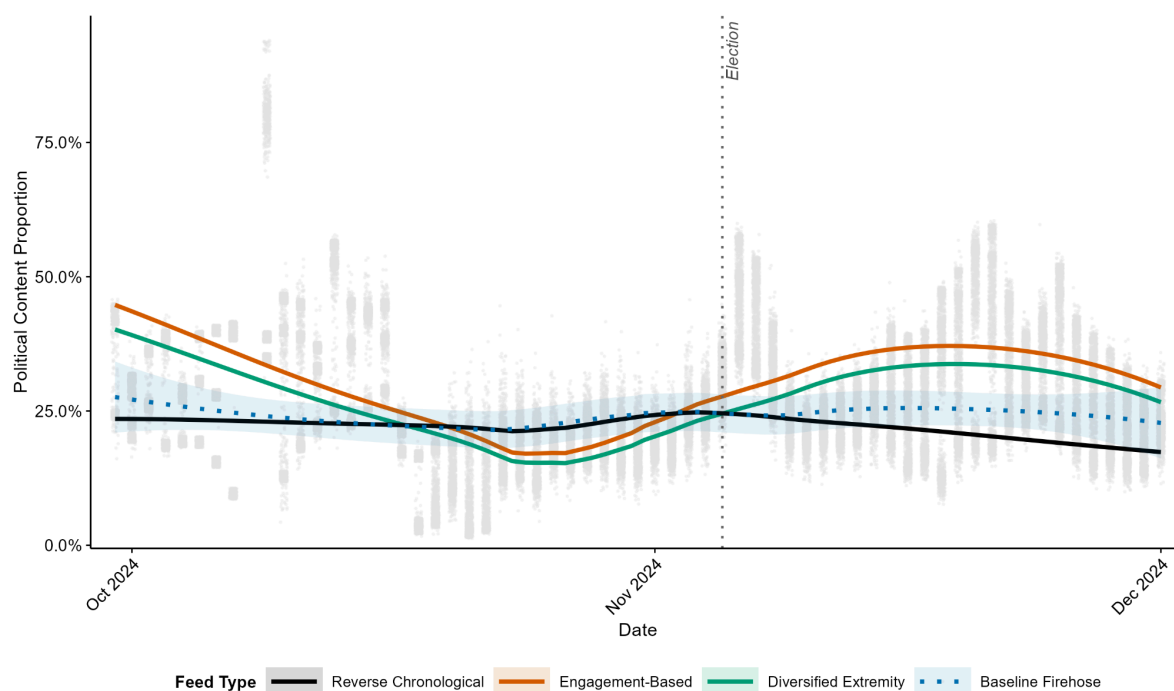


Fig. S20. Daily proportions of political content by condition, compared to the baseline firehose (estimate of true base rate of daily content in English-speaking Bluesky).

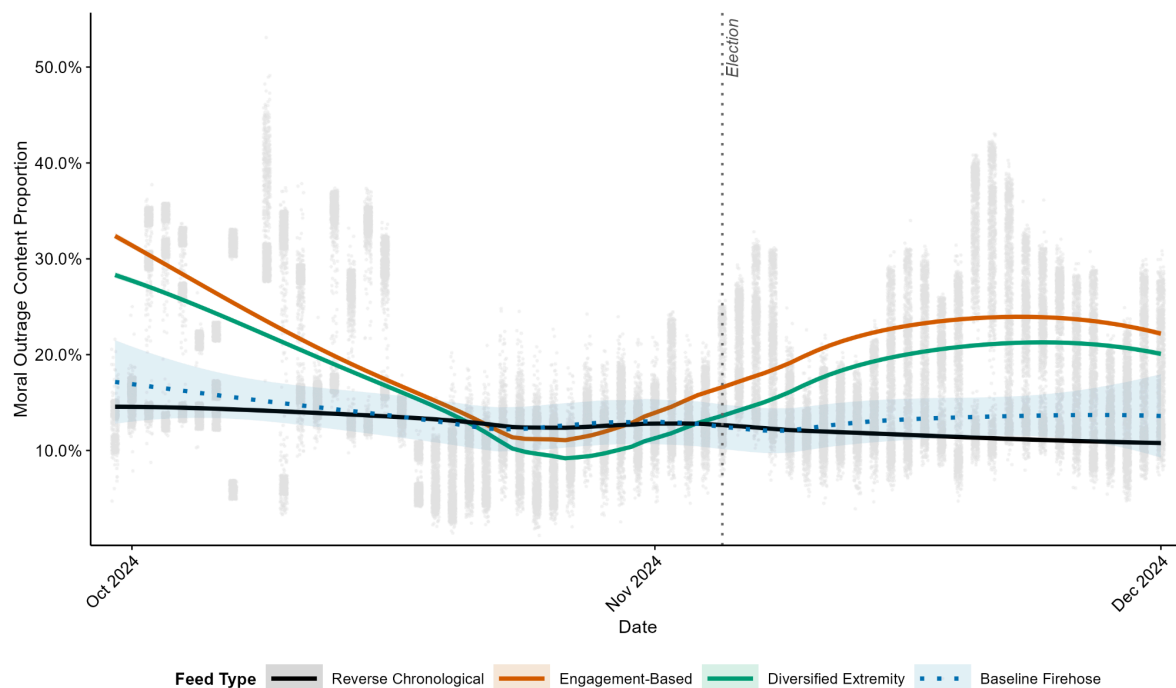


Fig. S21. Daily proportions of moral outrage content by condition, compared to the baseline firehose (estimate of true base rate of daily content in English-speaking Bluesky).

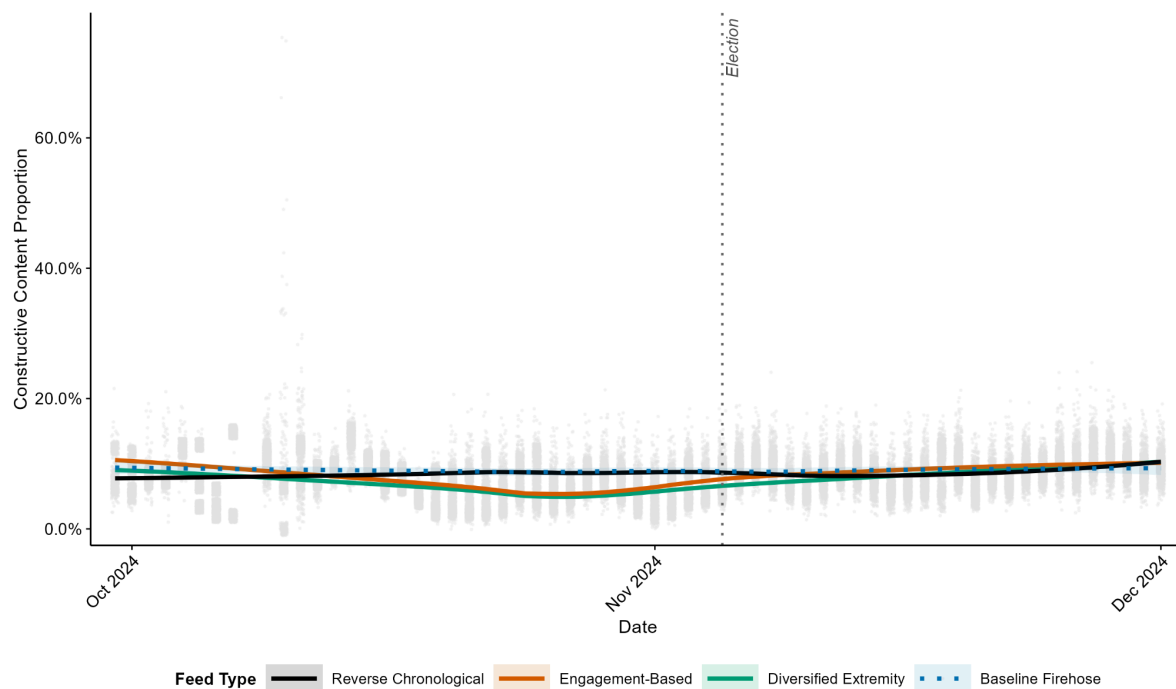


Fig. S22. Daily proportions of constructive content by condition, compared to the baseline firehose (estimate of true base rate of daily content in English-speaking Bluesky).

6.3 Model fit analyses

To assess whether the 2024 U.S. presidential election should be modeled as an exogenous shock in our estimation of condition effects on feed content, we compared model fit using AIC and BIC, and conducted a likelihood ratio test between models with and without a binary pre-/post-election indicator entered as a fixed effect. Table S6 displays results.

Model Comparison: Base vs. Pre/Post Election Models							
Lower AIC/BIC values indicate better model fit							
Content Type	AIC		BIC		Likelihood Ratio Test		Preferred Model [†]
	Base Model	Pre/Post Model	Base Model	Pre/Post Model	χ^2	p-value	
Toxic	−390,285.62	−390,714.33	−390,238.64	−390,639.16	477.59	< .001	**Pre/Post**
Intergroup	−642,070.13	−655,516.96	−642,023.15	−655,441.79	13,505.06	< .001	**Pre/Post**
Moral	−618,843.45	−622,251.41	−618,796.47	−622,176.24	3,464.74	< .001	**Pre/Post**
Negative Emotion	−223,947.66	−224,331.81	−223,900.68	−224,256.64	427.41	< .001	**Pre/Post**
Positive Emotion	−261,771.75	−282,534.71	−261,724.77	−282,459.54	20,808.23	< .001	**Pre/Post**
Outrage	−205,689.65	−209,663.77	−205,642.67	−209,588.60	4,016.89	< .001	**Pre/Post**
Constructive	−372,103.51	−380,239.64	−372,056.57	−380,164.54	8,184.77	< .001	**Pre/Post**
Political	−135,105.42	−142,473.59	−135,058.44	−142,398.42	7,408.62	< .001	**Pre/Post**

[†] Bold indicates statistically preferred model based on AIC / BIC

Table S11. Model fit comparison between the base model (without a pre-/post-election indicator) and the pre-/post-election model. Fit metrics consistently favored the pre-/post-election specification.

6.4 Fully tabulated model results (pre/post models)

Tabulated results for RQ1 models estimating the effect of condition on various types of IME content in user feeds.

RQ1: Intergroup content — Pre/Post model			
Entries are β (SE). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	0.003 (0.000)	0.004 (0.000)	-0.001 (0.000)
Post-election	-0.002 (0.000)	-0.002 (0.000)	0.000 (0.000)

Table S12: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate higher content proportions in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ1: Moral content — Pre/Post model			
Entries are β (SE). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	0.000 (0.000)	-0.003 (0.000)	0.003 (0.000)
Post-election	0.003 (0.000)	0.002 (0.000)	0.001 (0.000)

Table S13: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate higher content proportions in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ1: Negative emotion content — Pre/Post model			
Entries are β (SE). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	0.019 (0.001)	0.003 (0.001)	0.016 (0.001)
Post-election	0.037 (0.001)	0.024 (0.001)	0.013 (0.001)

Table S14: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate higher content proportions in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ1: Positive emotion content — Pre/Post model			
Entries are β (SE). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	-0.054 (0.001)	-0.062 (0.001)	0.008 (0.001)
Post-election	-0.012 (0.001)	-0.007 (0.001)	-0.005 (0.001)

Table S15: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate higher content proportions in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ1: Toxic content — Pre/Post model			
Entries are β (SE). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	0.008 (0.000)	-0.001 (0.000)	0.009 (0.000)
Post-election	0.011 (0.000)	0.006 (0.000)	0.005 (0.000)

Table S16: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate higher content proportions in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ1: Sociopolitical content — Pre/Post model			
Entries are β (SE). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	0.043 (0.001)	0.014 (0.001)	0.029 (0.001)
Post-election	0.139 (0.001)	0.109 (0.001)	0.030 (0.001)

Table S17: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate higher content proportions in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ1: Moral outrage content — Pre/Post model			
Entries are β (SE). Bold indicates Holm-Bonferroni-adjusted $p < .05$ (exploratory outcome).			
	EB – RC	DE – RC	EB – DE
Pre-election	0.050 (0.001)	0.032 (0.001)	0.018 (0.001)
Post-election	0.109 (0.001)	0.083 (0.001)	0.026 (0.001)

Table S18: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate higher content proportions in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ1: Constructive content — Pre/Post model			
Entries are β (SE). Bold indicates Holm-Bonferroni-adjusted $p < .05$ (exploratory outcome).			
	EB – RC	DE – RC	EB – DE
Pre-election	-0.010 (0.001)	-0.017 (0.001)	0.008 (0.001)
Post-election	0.007 (0.001)	0.001 (0.001)	0.007 (0.001)

Table S19: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate higher content proportions in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

6.5 Fully tabulated model results (base models)

RQ1: Intergroup content — Base model		
Entries are β (SE). Bold indicates $p < .05$.		
EB – RC	DE – RC	EB – DE
0.001 (0.000)	0.001 (0.000)	-0.001 (0.000)

Table S20: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate higher content proportions in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ1: Moral content — Base model

Entries are β (SE). Bold indicates $p < .05$.

EB – RC	DE – RC	EB – DE
0.001 (0.000)	-0.001 (0.000)	0.002 (0.000)

Table S21: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate higher content proportions in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ1: Negative emotion content — Base model		
Entries are β (SE). Bold indicates $p < .05$.		
EB – RC	DE – RC	EB – DE
0.027 (0.001)	0.012 (0.001)	0.015 (0.001)

Table S22: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate higher content proportions in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ1: Positive emotion content — Base model		
Entries are β (SE). Bold indicates $p < .05$.		
EB – RC	DE – RC	EB – DE
-0.036 (0.001)	-0.038 (0.001)	0.003 (0.001)

Table S23: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate higher content proportions in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ1: Toxic content — Base model		
Entries are β (SE). Bold indicates $p < .05$.		
EB – RC	DE – RC	EB – DE
0.009 (0.000)	0.002 (0.000)	0.007 (0.000)

Table S24: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate higher content proportions in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ1: Sociopolitical content — Base model

Entries are β (SE). Bold indicates $p < .05$.

EB – RC	DE – RC	EB – DE
0.084 (0.001)	0.055 (0.001)	0.029 (0.001)

Table S25: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate higher content proportions in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ1: Moral outrage content — Base model

Entries are β (SE). Bold indicates Holm-Bonferroni-adjusted $p < .05$ (exploratory outcome).

EB – RC	DE – RC	EB – DE
0.075 (0.001)	0.054 (0.001)	0.021 (0.001)

Table S26: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate higher content proportions in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ1: Constructive content — Base model

Entries are β (SE). Bold indicates Holm-Bonferroni-adjusted $p < .05$ (exploratory outcome).

EB – RC	DE – RC	EB – DE
-0.002 (0.001)	-0.009 (0.001)	0.007 (0.001)

Table S27: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate higher content proportions in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

6.6 Topic modeling and content analyses

We performed a deeper investigation into the themes and content patterns present in the content displayed to users in our study. We constrained our analysis to posts that we previously classified as ‘sociopolitical’ since we were interested in the IME content amplified by the engagement-based algorithm and reduced in the diversified extremity algorithm. This analysis allows us to move from broad classifications of political tone to concrete topics, terms, and symbols present in user social media feeds before and after the election and across the study conditions.

6.6.1 Topic Modeling

We fit topic models on the posts shown to users in their feeds. We selected posts that were classified as sociopolitical, as fitting on all posts led to noisy topics being generated and we were focused on IME content. Past work^{13,14} has demonstrated the efficacy of using BERTopic for topic modeling of social media content.

Methods

We used the ‘all-MiniLM-L6-v2’ SentenceTransformer¹⁵ as our embedding model, UMAP¹⁶ as our dimensionality reduction algorithm, and HDBSCAN¹⁷ for our clustering algorithm, in line with approaches from past work¹⁸. For computational efficiency, we sampled 500 daily posts shown in feeds, evenly sampled from each condition, and used this to train our BERTopic model (total n=27,288). We then used the trained BERTopic model to classify our remaining corpus of n=496,699 sociopolitical posts shown to users in feeds. We then fitted a BERTopic model¹⁹ to our dataset. We report that our model uncovered 33 topics and had a c_v score of 0.54. Code and hyperparameters for our model can be found [here](#).

Results

We first evaluated the topics generated by our topic model across our entire dataset of sociopolitical posts used in feeds. Table S28 displays our results.

Topic 1: Political Opinions and Perspectives (n=125,253)	“people”, “trump”, “harris”, “us”, “gaza”, “russia”, “ukraine”, “war”, “nato”
Topic 2: Literature and Reading (n=3,990)	“books”, “library”, “novels”, “writing”
Topic 3: Tech Installation Guides (n=3,570)	“starter”, “pack”, “bluesky”, “install”, “bunch”
Topic 4: Severe Weather Alerts (n=2,978)	“hurricane”, “weather”, “flood”, “advisory”, “helene”, “milton”, “flooding”
Topic 5: Education and Teaching (n=2,310)	“education”, “teaching”, “schools”, “students”, “faculty”, “university”

Table S28: The top keywords in each of the top 5 topics uncovered by our BERTopic model, sorted by counts descending. We also prompted ChatGPT to generate short topic descriptions based on the top 10 keywords for each topic. We filter out posts that were not cleanly assigned to a topic (n=337,641). Topic 1, “Political Opinions and Perspectives”, dominates all other topics, accounting for 78.7% of posts assigned to a topic. Topic 3 is likely a mislabeling on ChatGPT’s part, as the topic pertains to posts discussing how to add new users using Bluesky’s “Starter Packs” feature (which

was popularized during the one point in the study and resulted in substantial related activity).

Our first topic, “Political Opinions and Perspectives”, disproportionately represented the posts in our sample. Other topics were uncovered that represented topical discourse on Bluesky at the time (such as “Literature and Reading”), but are less relevant to our current analysis. We also see time-sensitive topics related to news cycles or Bluesky platform-specific events (e.g, “Hurricane Milton”, “starter packs”) that are only found in brief periods during the 2-month study. Therefore, we focused our downstream analysis on posts within the “Political Opinions and Perspectives” topic.

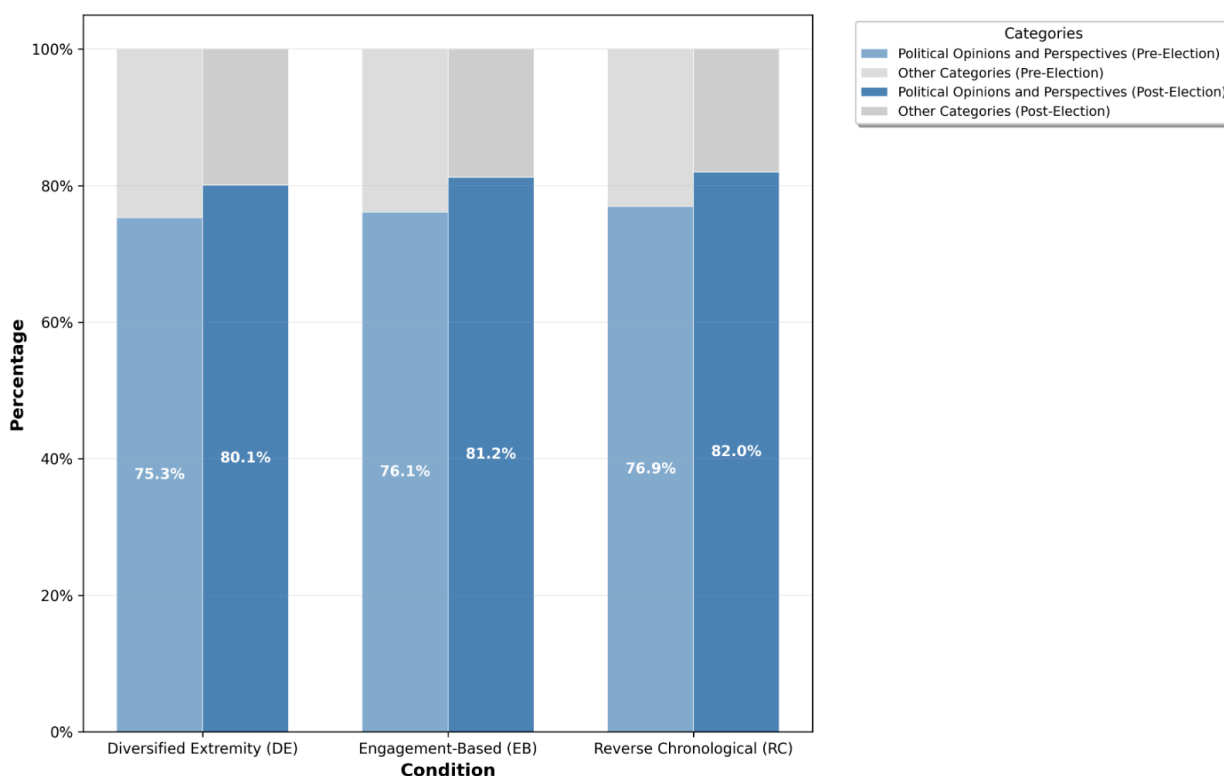


Fig. S23. Percentage of feed content shown to users falling under the “Political Opinions and Perspectives”, across conditions and pre/post-election. Across conditions, we observed an increase in posts classified as “Political Opinions and Perspectives” post-election within our sampled subset of sociopolitical posts. This pattern displays that political discourse on Bluesky increased, platform-wide, after the 2024 U.S. election.

6.6.2 Named Entity Recognition (NER)

We also fit a named entity recognition (NER) model on all sociopolitical posts shown to users in feeds. Doing so gives us insight on the entities that were mentioned in the posts shown to study users.

We used a transformer-based English NER model (``spaCy en_core_web_trf``), which identifies spans labeled as PERSON, ORG, GPE (countries/states/cities), EVENT, and NORP (nationalities/political/religious groups). We retain only these types to focus on actors, institutions, places, and election-linked events. We ensured that any PII related to non-public individuals and

accounts was removed prior to analysis. We generated canonical representations of common entities for consolidated representations downstream. We counted an entity once if it is present in a post, avoiding double-counting entities that appear multiple times within a post.

For our analysis, we report the total number of times that each entity occurs.

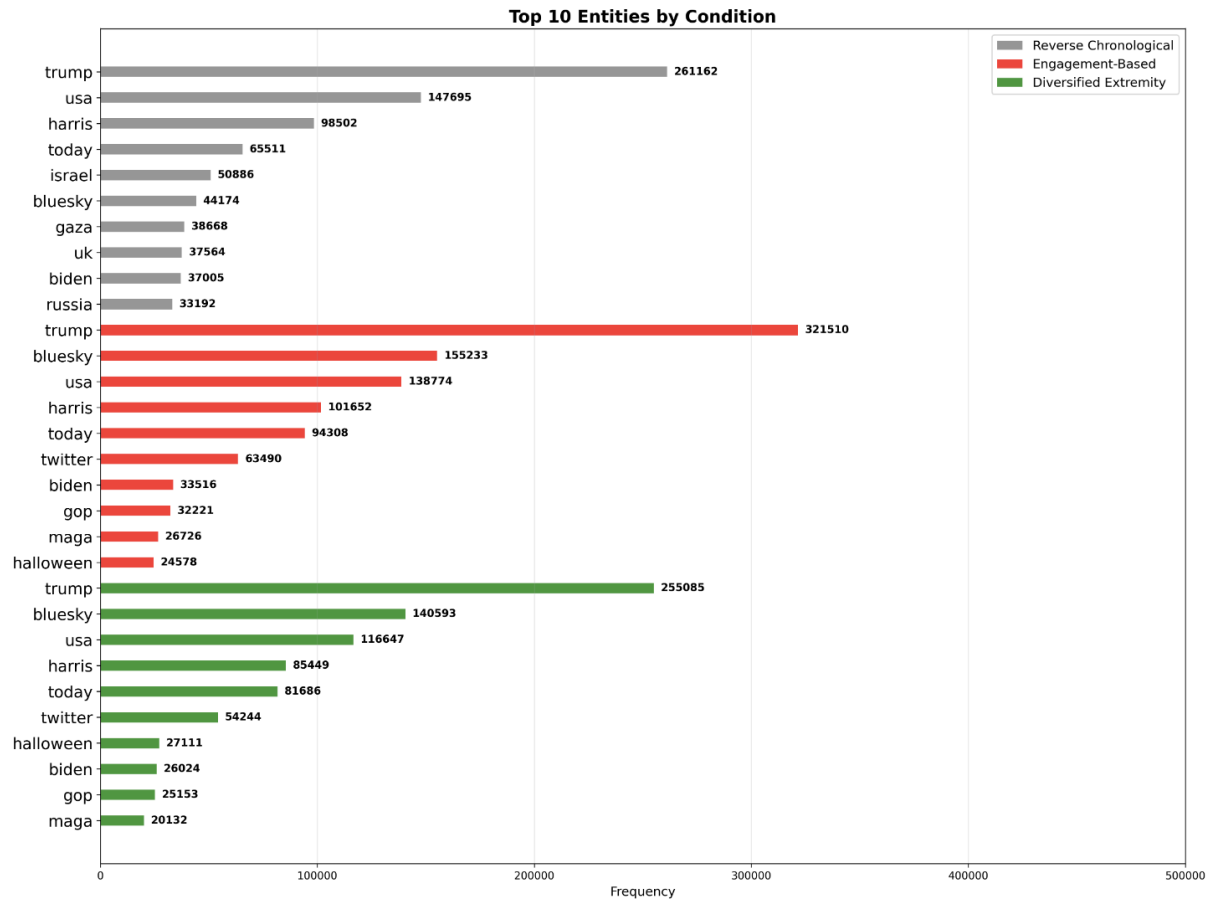


Fig. S24. Distribution of most common entities that appeared in posts shown to users, split by condition. We see that across conditions, we see similar political entities mentioned. We observed a high overlap of entities found across all 3 feeds. When looking at the entities shared specifically in both the EB and DE feeds, these entities generally had a higher frequency in the EB feeds. One possible explanation for this is our algorithm's penalization of toxic and superposter-related content, which may have reduced the presence of some posts related to political entities. We see a slightly lower frequency of political entities in RC feeds as compared to EB and DE feeds, both across all entities and also for entities found commonly across both RC feeds and EB/DE feeds.

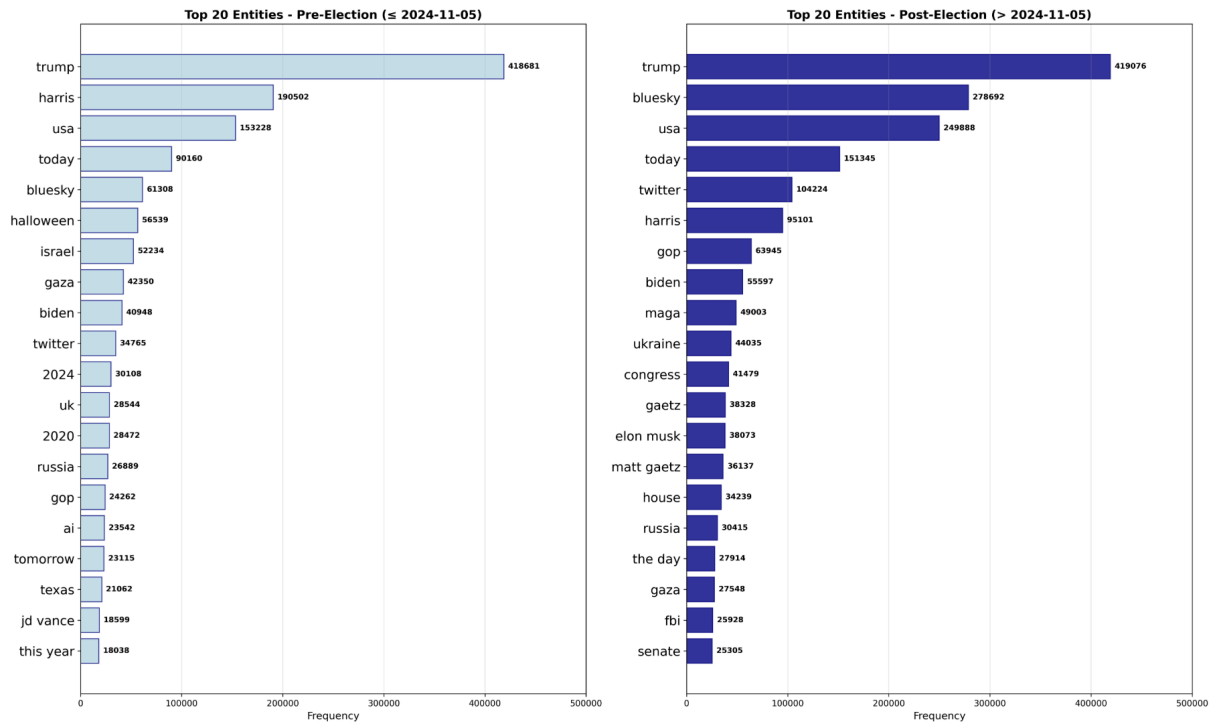


Fig. S25. Distribution of most common entities that appeared in posts shown to users, separated by pre-election versus post-election feeds. We observed that the most common entities detected are related closely to trends in news cycle coverage. We see a high frequency of mentions related to political elites (e.g., “trump”, “harris”, “biden”) across both pre-election and post-election feeds. We also see a general increased frequency of political entities post-election, both across all entities and for each entity individually showing the increased emphasis on political discourse on the platform.

Our exploratory analysis of entity distribution shows that entities found in posts shown in feeds. We see that entities found in posts closely track trends in news cycle coverage, across all 3 study conditions. We also observed that political entities were mentioned in higher frequency after the 2024 US election as compared to before.

6.6.3 Hashtags

We also analyzed patterns in the hashtags on all sociopolitical posts shown to users in feeds. Like above, we preprocessed the posts and performed our analysis on canonicalized representations of hashtags (e.g., consolidating “#trump” and “#Trump” into a single hashtag unit). We collected the most mentioned hashtags per day and then stratified our analysis by looking at hashtag trends across conditions and pre-election versus post-election.

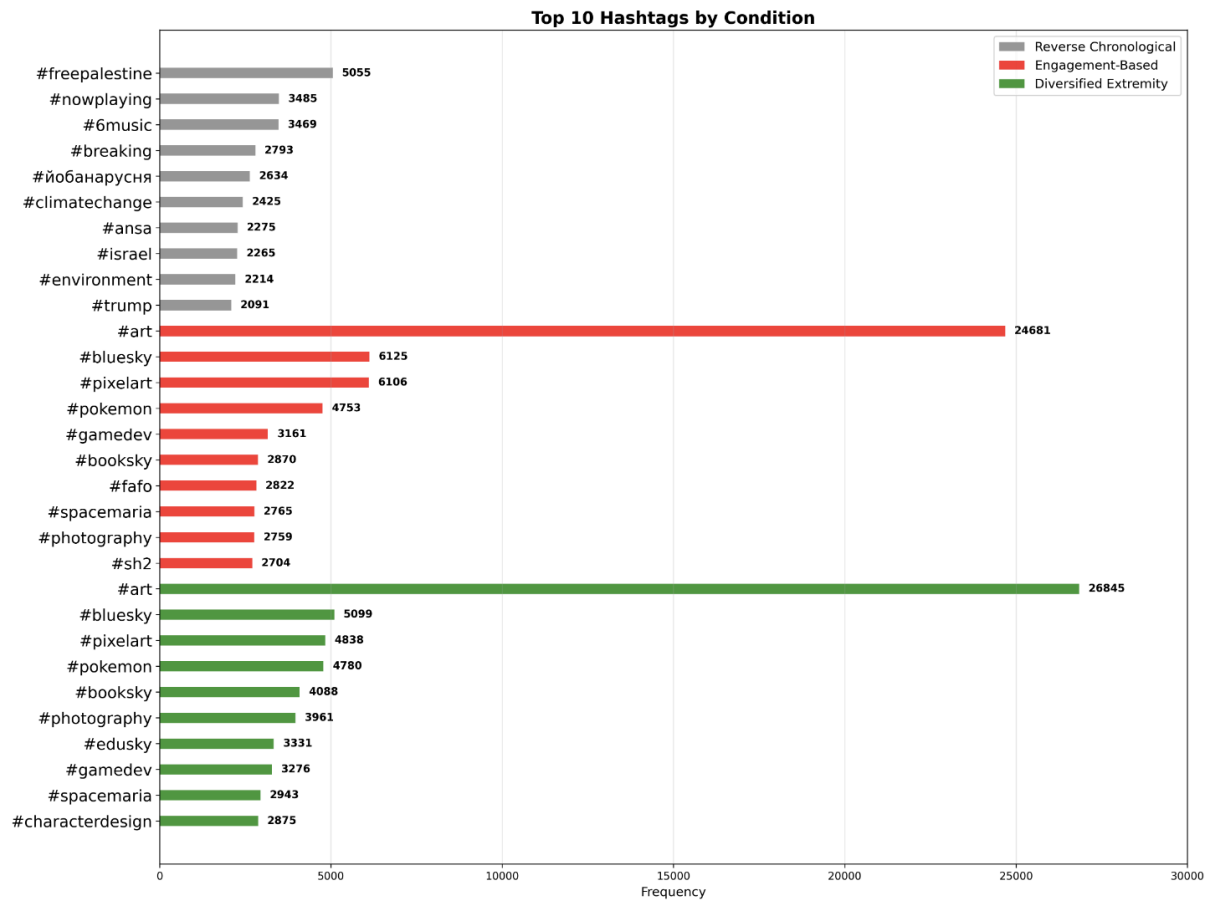


Fig. S26. Distribution of most common hashtags that appeared in posts shown to users, split by condition. We see a significant difference in the hashtags present in content displayed to users in our Reverse Chronological (RC) condition as compared to our Engagement-Based (EB) and Diversified Extremity (DE). Our EB and DE feeds were based on posts that were most engaging throughout the Bluesky platform, while our RC condition was based on posts sampled from the general Bluesky firehose. We see that throughout the course of the study, the hashtags most common in content most widely engaged with by the wider Bluesky network were primarily related to specific early-adopter networks on Bluesky (such as the art and literature communities) that were both prominent and highly active on Bluesky. However, we see that the most common hashtags from content widely sampled from Bluesky users in general, irrespective of engagement, better represented a snapshot of political sentiments and headlines from various news cycles. This hashtag distribution more closely aligns with expectations around political discourse on social media.

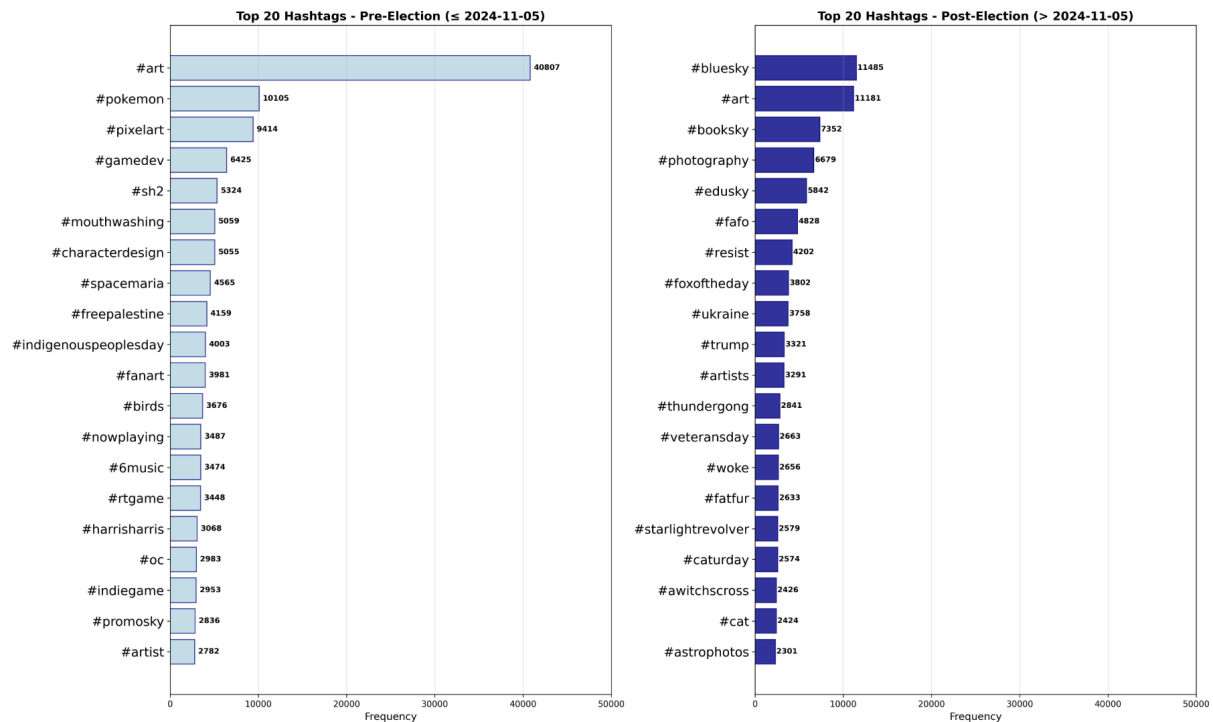


Fig. S27. Distribution of most common hashtags that appeared in posts shown to users, separated by pre-election versus post-election feeds. We see a significant change in the types of hashtags present in posts after the 2024 US election as compared to before the election. We see an increased number of political-related hashtags as well as the rise of new hashtags related to partisan sentiments (e.g., “#resist” and “#fafo”). However, we also see, both pre-election and post-election, the high proliferation of non-political hashtags related to community-building on the Bluesky platform (e.g., “#bluesky”, “#art”, “#booksky”), which is related to community norms on the platform at the time, a trend that was already present pre-election but was exacerbated post-election due to both norm creation and community-building on Bluesky as well as a broader exodus of ex-Twitter users onto the platform.

Our exploratory analysis of hashtag distribution across study conditions and pre-election vs. post-election displays another dimension of discourse on the Bluesky platform during the course of our study. From our EB and DE feeds, we see a strong network effect of early-adopter communities on Bluesky (e.g., the art and literature communities) driving the proliferation of certain hashtags by the most widely engaged with content on Bluesky. However, we see in our RC feeds a more representative distribution of hashtags in content posted by the wider Bluesky community, and in these hashtags we observed a higher proliferation of political and partisan discourse. This hashtag distribution gives a clearer snapshot of what average users were posting on the platform, even if this content wasn’t necessarily as algorithmically amplified through existing network effects as content posted by early-adopter communities. Observing hashtag distribution from pre-election to post-election, we see an increase, across all 3 study conditions, in politically-based hashtags included in posted content.

6.6 Plots of raw proportions of content over the study period

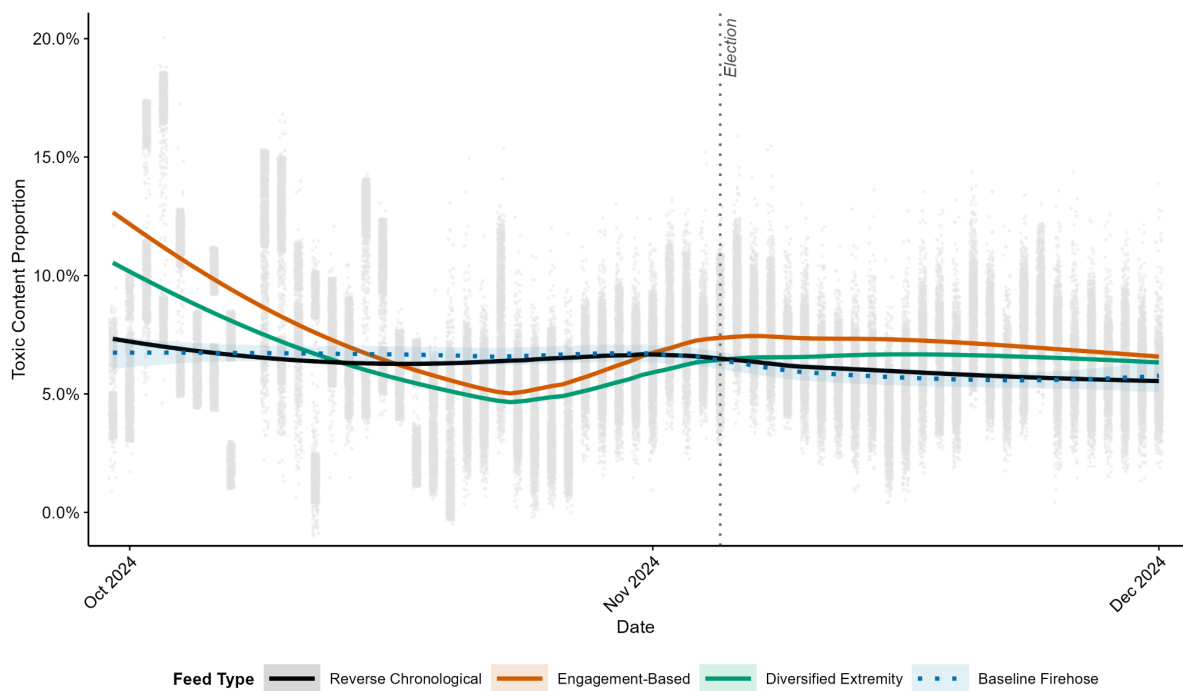


Fig. S28. Mean proportions of toxic content in feeds by condition, compared to a large sample from English-speaking Bluesky acquired daily from the firehose. Error bands represent 95% confidence intervals.

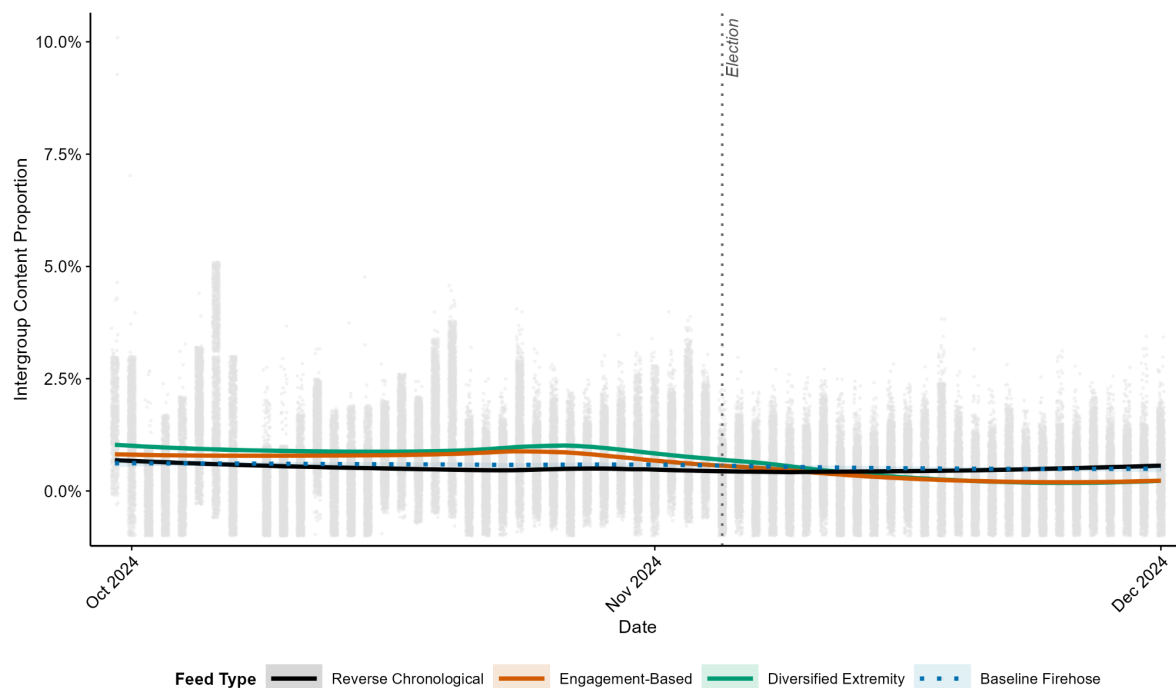


Fig. S29. Mean proportions of intergroup content in feeds by condition, compared to a large sample from English-speaking Bluesky acquired daily from the firehose. Error bands represent 95% confidence intervals.

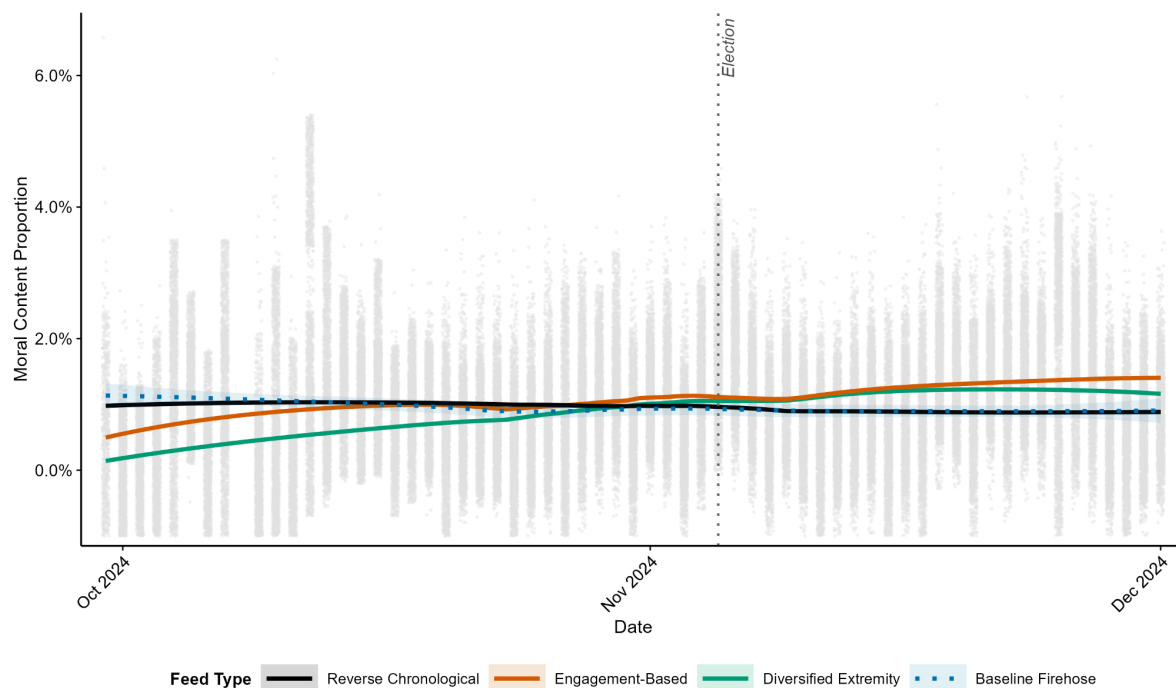


Fig. S30. Mean proportions of moral content in feeds by condition, compared to a large sample from English-speaking Bluesky acquired daily from the firehose. Error bands represent 95% confidence intervals.

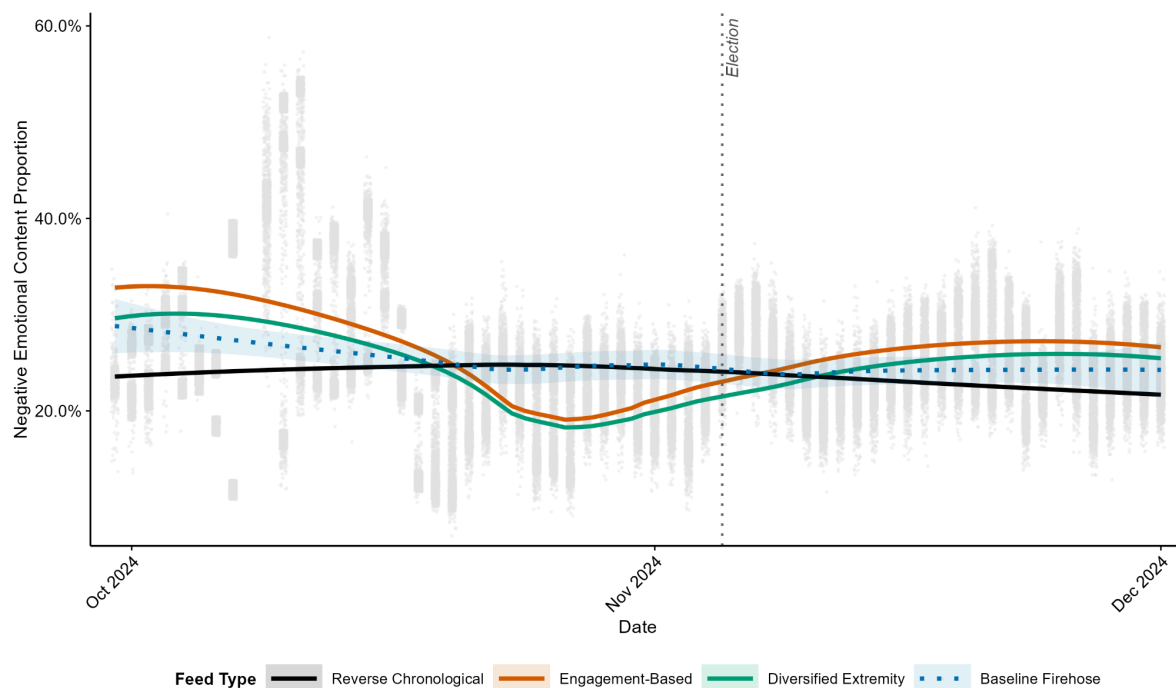


Fig. S31. Mean proportions of negative emotional content in feeds by condition, compared to a large sample from English-speaking Bluesky acquired daily from the firehose. Error bands represent 95% confidence intervals.

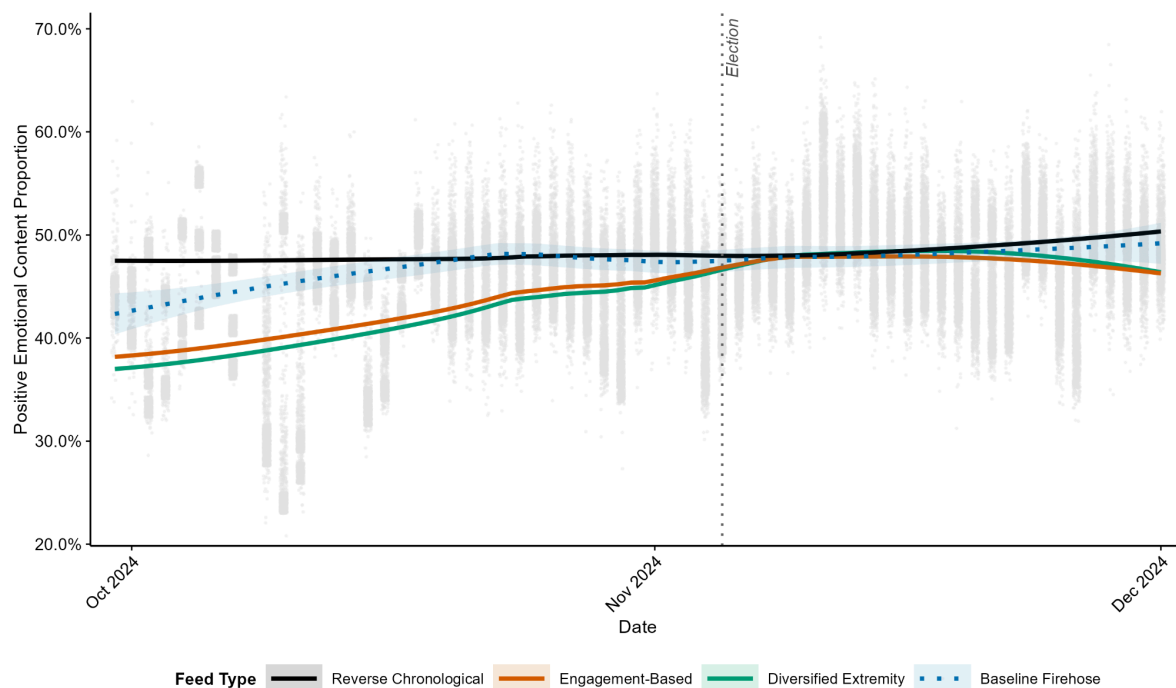


Fig. S32. Mean proportions of positive emotional content in feeds by condition, compared to a large sample from English-speaking Bluesky acquired daily from the firehose. Error bands represent 95% confidence intervals.

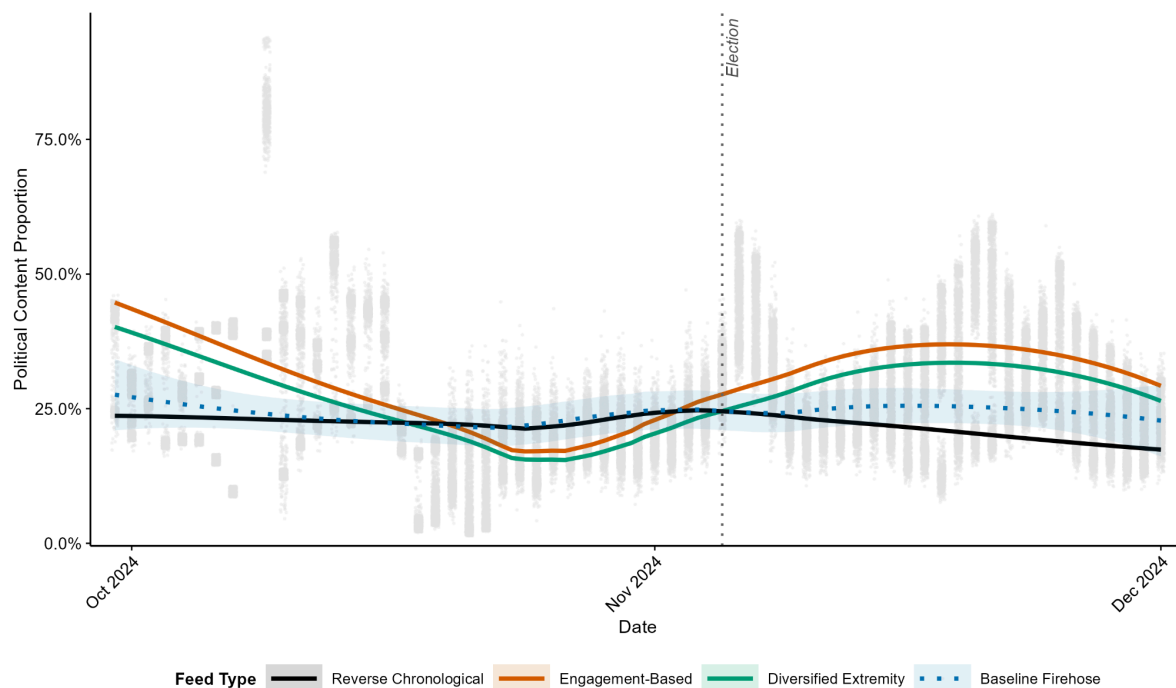


Fig. S33. Mean proportions of political content in feeds by condition, compared to a large sample from English-speaking Bluesky acquired daily from the firehose. Error bands represent 95% confidence intervals.

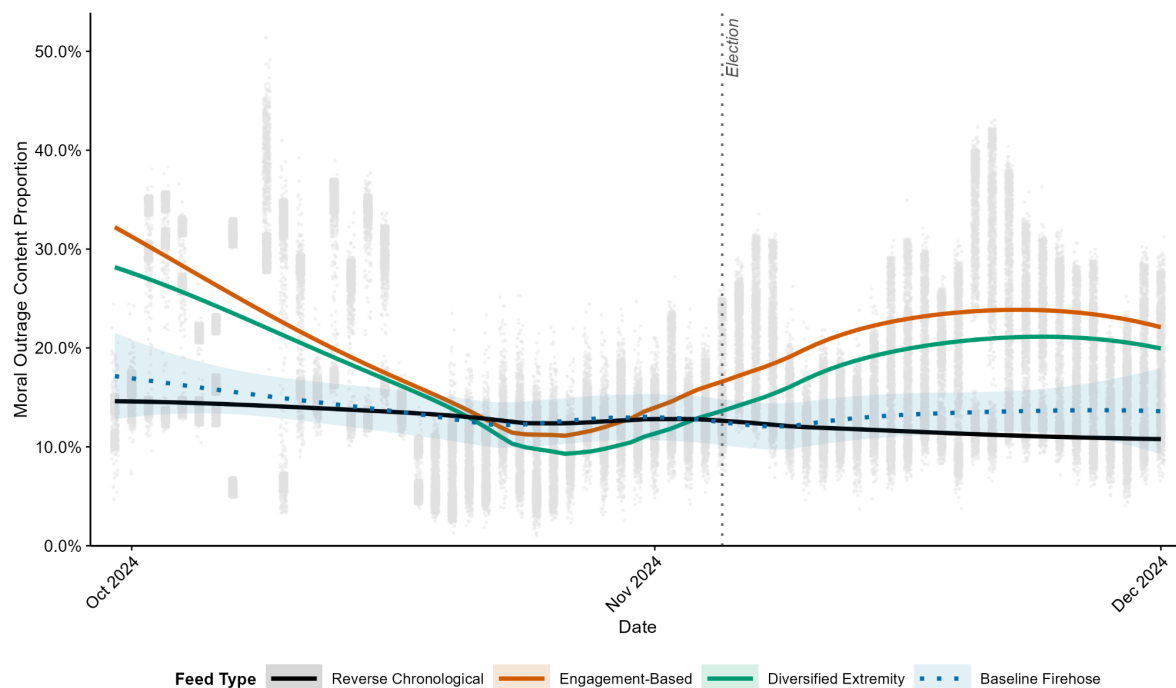


Fig. S34. Mean proportions of moral outrage content in feeds by condition, compared to a large sample from English-speaking Bluesky acquired daily from the firehose. Error bands represent 95% confidence intervals.

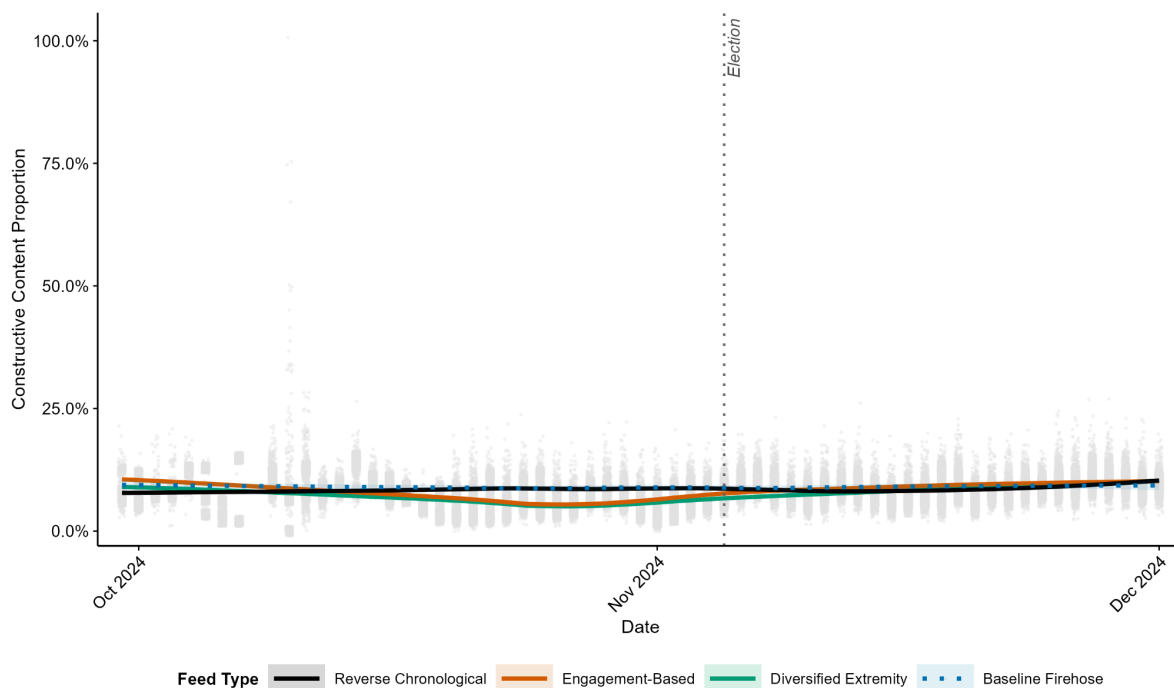


Fig. S35. Mean proportions of constructive content in feeds by condition, compared to a large sample from English-speaking Bluesky acquired daily from the firehose. Error bands represent 95% confidence intervals.

6.7 Examining polynomial time models

Given the observed raw data trends shown in section 6.6, we explored models that accounted for time and how condition effects may have changed over time. We did so in a data-driven manner, exploring various polynomial orders to determine which models fit the data better as assessed by AIC/BIC criteria. We explored 4 models: a base model with condition but not time, a linear model with condition, time and their interaction, a quadratic polynomial model, and a cubic polynomial model. Table S29 displays AIC/BIC fit statistics. The Cubic model consistently produced the greatest model fit, with the largest change occurring from linear to quadratic. Thus, below we report predicted values and tabulated results for all cubic models, see Fig S36 and Tables S29-37.

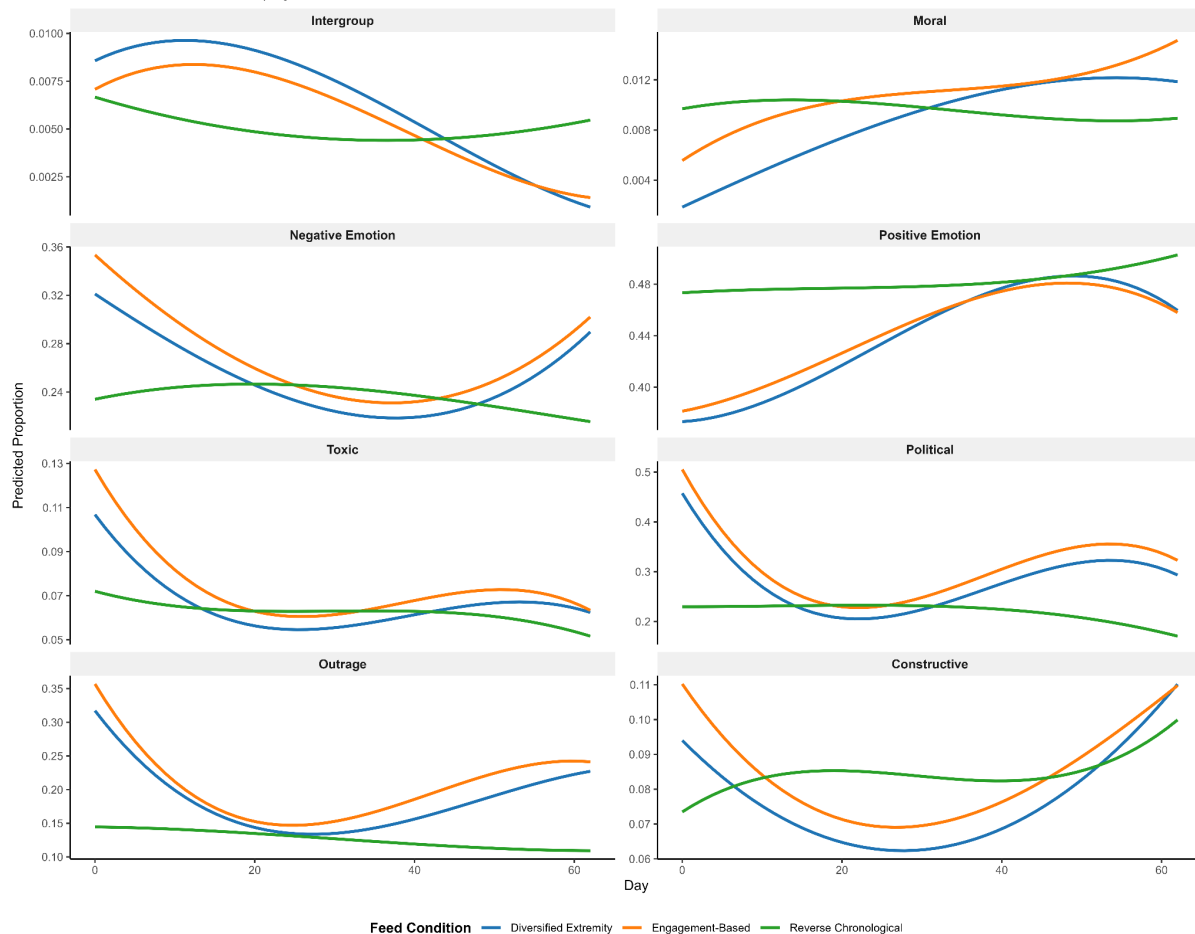
Detailed Polynomial Model Selection Summary

Dependent Variable	Best Models		AIC Delta Values				BIC Delta Values			
	Best Model (AIC)	Best Model (BIC)	Base	Linear	Quadratic	Cubic	Base	Linear	Quadratic	Cubic
Intergroup	Cubic	Cubic	1593.7505	1541.1562	318.4679	0	1535.5047	1482.9104	289.3450	0
Moral	Cubic	Cubic	920.4764	867.8821	249.4800	0	862.2307	809.6364	220.3571	0
Negative Emotion	Cubic	Cubic	13822.6503	13770.0560	160.5338	0	13764.4046	13711.8103	131.4109	0
Positive Emotion	Cubic	Cubic	7086.4585	7033.8642	1922.9031	0	7028.2128	6975.6185	1893.7802	0
Toxic	Cubic	Cubic	15075.5011	15022.9068	5573.6609	0	15017.2554	14964.6611	5544.5380	0
Political	Cubic	Cubic	22872.6879	22820.0936	9847.7984	0	22814.4422	22761.8479	9818.6755	0
Outrage	Cubic	Cubic	31706.3369	31653.7426	5349.7607	0	31648.0912	31595.4969	5320.6378	0
Constructive	Cubic	Cubic	15283.9008	15231.3065	697.9025	0	15225.7072	15173.1129	668.8057	0

Table S29. Exploration of various polynomial fit statistics.

Cubic Time Trends Across All Content Types

Predicted values from cubic polynomial models



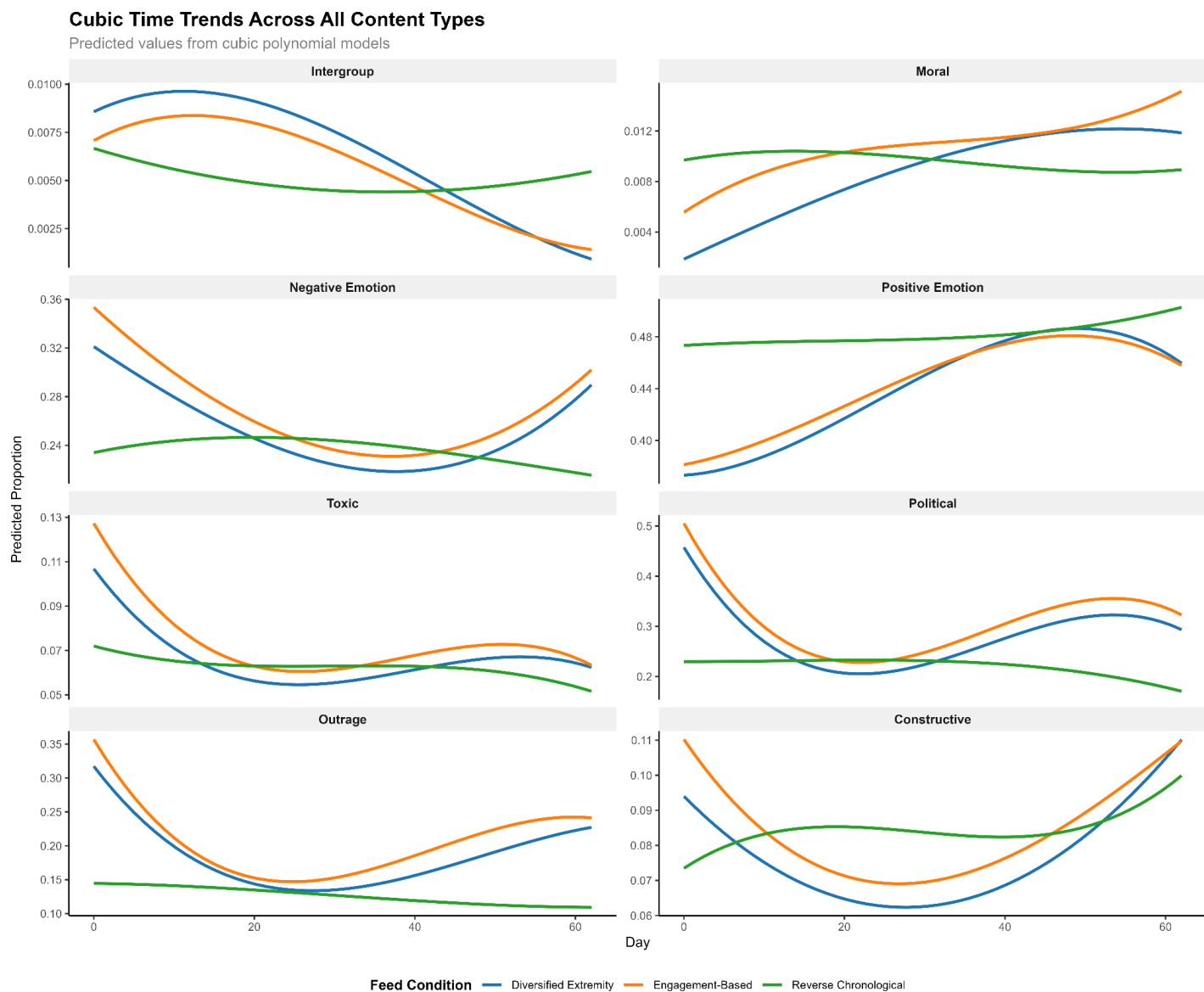


Fig. S36. Predicted proportion of content in feeds by condition, based on cubic model fits.

Model Summary: Intergroup Content - Cubic Model

	Term	Coefficient	95% CI	t	p-value
(Intercept)	(Intercept)	0.006	[0.006, 0.006]	207.79	p < .001
conditionEngagement-Based	conditionEngagement-Based	-0.001	[-0.001, -0.001]	-18.08	p < .001
conditionReverse Chronological	conditionReverse Chronological	-0.001	[-0.001, -0.001]	-32.46	p < .001
poly(day, 3)1	poly(day, 3)1	-0.985	[-1.006, -0.963]	-88.10	p < .001
poly(day, 3)2	poly(day, 3)2	-0.285	[-0.307, -0.263]	-25.31	p < .001
poly(day, 3)3	poly(day, 3)3	0.136	[0.114, 0.158]	12.16	p < .001
conditionEngagement-Based:poly(day, 3)1	conditionEngagement-Based:poly(day, 3)1	0.180	[0.149, 0.211]	11.46	p < .001
conditionReverse Chronological:poly(day, 3)1	conditionReverse Chronological:poly(day, 3)1	0.866	[0.835, 0.897]	55.31	p < .001
conditionEngagement-Based:poly(day, 3)2	conditionEngagement-Based:poly(day, 3)2	0.049	[0.018, 0.080]	3.08	p = 0.002
conditionReverse Chronological:poly(day, 3)2	conditionReverse Chronological:poly(day, 3)2	0.461	[0.431, 0.492]	29.29	p < .001
conditionEngagement-Based:poly(day, 3)3	conditionEngagement-Based:poly(day, 3)3	0.019	[-0.012, 0.050]	1.22	p = 0.222
conditionReverse Chronological:poly(day, 3)3	conditionReverse Chronological:poly(day, 3)3	-0.139	[-0.169, -0.108]	-8.84	p < .001

Table S30. Tabulated model results from condition X time with cubic polynomial term, intergroup content.

Model Summary: Moral Content - Cubic Model

	Term	Coefficient	95% CI	t	p-value
(Intercept)	(Intercept)	0.009	[0.009, 0.009]	246.29	p < .001
conditionEngagement-Based	conditionEngagement-Based	0.002	[0.002, 0.002]	39.89	p < .001
conditionReverse Chronological	conditionReverse Chronological	0.001	[0.001, 0.001]	14.70	p < .001
poly(day, 3)1	poly(day, 3)1	1.097	[1.072, 1.123]	84.07	p < .001
poly(day, 3)2	poly(day, 3)2	-0.321	[-0.346, -0.295]	-24.41	p < .001
poly(day, 3)3	poly(day, 3)3	-0.049	[-0.075, -0.024]	-3.76	p < .001
conditionEngagement-Based:poly(day, 3)1	conditionEngagement-Based:poly(day, 3)1	-0.373	[-0.408, -0.337]	-20.33	p < .001
conditionReverse Chronological:poly(day, 3)1	conditionReverse Chronological:poly(day, 3)1	-1.293	[-1.329, -1.257]	-70.71	p < .001
conditionEngagement-Based:poly(day, 3)2	conditionEngagement-Based:poly(day, 3)2	0.242	[0.206, 0.279]	13.15	p < .001
conditionReverse Chronological:poly(day, 3)2	conditionReverse Chronological:poly(day, 3)2	0.273	[0.237, 0.309]	14.84	p < .001
conditionEngagement-Based:poly(day, 3)3	conditionEngagement-Based:poly(day, 3)3	0.240	[0.204, 0.276]	13.06	p < .001
conditionReverse Chronological:poly(day, 3)3	conditionReverse Chronological:poly(day, 3)3	0.134	[0.098, 0.169]	7.29	p < .001

Table S31. Tabulated model results from condition X time with cubic polynomial term, moral content.

Model Summary: Negative Emotion - Cubic Model

	Term	Coefficient	95% CI	t	p-value
(Intercept)	(Intercept)	0.251	[0.250, 0.251]	752.95	p < .001
conditionEngagement-Based	conditionEngagement-Based	0.016	[0.015, 0.016]	33.29	p < .001
conditionReverse Chronological	conditionReverse Chronological	-0.014	[-0.015, -0.013]	-29.34	p < .001
poly(day, 3)1	poly(day, 3)1	-4.909	[-5.136, -4.682]	-42.31	p < .001
poly(day, 3)2	poly(day, 3)2	9.047	[8.818, 9.276]	77.47	p < .001
poly(day, 3)3	poly(day, 3)3	1.285	[1.058, 1.513]	11.06	p < .001
conditionEngagement-Based:poly(day, 3)1	conditionEngagement-Based:poly(day, 3)1	-1.250	[-1.569, -0.931]	-7.67	p < .001
conditionReverse Chronological:poly(day, 3)1	conditionReverse Chronological:poly(day, 3)1	2.404	[2.086, 2.723]	14.79	p < .001
conditionEngagement-Based:poly(day, 3)2	conditionEngagement-Based:poly(day, 3)2	1.120	[0.799, 1.441]	6.83	p < .001
conditionReverse Chronological:poly(day, 3)2	conditionReverse Chronological:poly(day, 3)2	-11.091	[-11.411, -10.770]	-67.81	p < .001
conditionEngagement-Based:poly(day, 3)3	conditionEngagement-Based:poly(day, 3)3	-0.568	[-0.888, -0.249]	-3.48	p < .001
conditionReverse Chronological:poly(day, 3)3	conditionReverse Chronological:poly(day, 3)3	-0.881	[-1.200, -0.562]	-5.41	p < .001

Table S32. Tabulated model results from condition X time with cubic polynomial term, negative emotion

Model Summary: Positive Emotion - Cubic Model

	Term	Coefficient	95% CI	t	p-value
(Intercept)	(Intercept)	0.442	[0.441, 0.443]	804.53	p < .001
conditionEngagement-Based	conditionEngagement-Based	0.003	[0.001, 0.004]	3.81	p < .001
conditionReverse Chronological	conditionReverse Chronological	0.040	[0.039, 0.042]	52.54	p < .001
poly(day, 3)1	poly(day, 3)1	13.033	[12.868, 13.198]	154.58	p < .001
poly(day, 3)2	poly(day, 3)2	-4.043	[-4.210, -3.877]	-47.67	p < .001
poly(day, 3)3	poly(day, 3)3	-2.995	[-3.160, -2.829]	-35.53	p < .001
conditionEngagement-Based:poly(day, 3)1	conditionEngagement-Based:poly(day, 3)1	-2.201	[-2.433, -1.969]	-18.59	p < .001
conditionReverse Chronological:poly(day, 3)1	conditionReverse Chronological:poly(day, 3)1	-10.599	[-10.831, -10.367]	-89.54	p < .001
conditionEngagement-Based:poly(day, 3)2	conditionEngagement-Based:poly(day, 3)2	0.021	[-0.212, 0.255]	0.18	p = 0.857
conditionReverse Chronological:poly(day, 3)2	conditionReverse Chronological:poly(day, 3)2	5.086	[4.853, 5.319]	42.81	p < .001
conditionEngagement-Based:poly(day, 3)3	conditionEngagement-Based:poly(day, 3)3	0.857	[0.625, 1.089]	7.24	p < .001
conditionReverse Chronological:poly(day, 3)3	conditionReverse Chronological:poly(day, 3)3	3.442	[3.211, 3.674]	29.15	p < .001

Table S33. Tabulated model results from condition X time with cubic polynomial term, positive emotion

Model Summary: Political Content - Cubic Model

	Term	Coefficient	95% CI	t	p-value
(Intercept)	(Intercept)	0.279	[0.277, 0.280]	403.39	p < .001
conditionEngagement-Based	conditionEngagement-Based	0.029	[0.028, 0.031]	30.34	p < .001
conditionReverse Chronological	conditionReverse Chronological	-0.058	[-0.060, -0.056]	-60.42	p < .001
poly(day, 3)1	poly(day, 3)1	0.745	[0.380, 1.111]	4.00	p < .001
poly(day, 3)2	poly(day, 3)2	16.158	[15.790, 16.525]	86.14	p < .001
poly(day, 3)3	poly(day, 3)3	-12.613	[-12.979, -12.248]	-67.63	p < .001
conditionEngagement-Based:poly(day, 3)1	conditionEngagement-Based:poly(day, 3)1	-0.152	[-0.665, 0.361]	-0.58	p = 0.561
conditionReverse Chronological:poly(day, 3)1	conditionReverse Chronological:poly(day, 3)1	-5.607	[-6.120, -5.095]	-21.45	p < .001
conditionEngagement-Based:poly(day, 3)2	conditionEngagement-Based:poly(day, 3)2	1.550	[1.034, 2.066]	5.89	p < .001
conditionReverse Chronological:poly(day, 3)2	conditionReverse Chronological:poly(day, 3)2	-19.543	[-20.058, -19.029]	-74.39	p < .001
conditionEngagement-Based:poly(day, 3)3	conditionEngagement-Based:poly(day, 3)3	-1.240	[-1.753, -0.727]	-4.74	p < .001
conditionReverse Chronological:poly(day, 3)3	conditionReverse Chronological:poly(day, 3)3	11.677	[11.165, 12.189]	44.70	p < .001

Table S34. Tabulated model results from condition X time with cubic polynomial term, political content

Model Summary: Toxic Content - Cubic Model

	Term	Coefficient	95% CI	t	p-value
(Intercept)	(Intercept)	0.066	[0.065, 0.066]	330.35	p < .001
conditionEngagement-Based	conditionEngagement-Based	0.008	[0.007, 0.008]	27.03	p < .001
conditionReverse Chronological	conditionReverse Chronological	-0.003	[-0.004, -0.003]	-10.98	p < .001
poly(day, 3)1	poly(day, 3)1	-1.752	[-1.840, -1.663]	-38.82	p < .001
poly(day, 3)2	poly(day, 3)2	3.152	[3.063, 3.241]	69.43	p < .001
poly(day, 3)3	poly(day, 3)3	-2.018	[-2.106, -1.929]	-44.71	p < .001
conditionEngagement-Based:poly(day, 3)1	conditionEngagement-Based:poly(day, 3)1	-1.142	[-1.266, -1.018]	-18.03	p < .001
conditionReverse Chronological:poly(day, 3)1	conditionReverse Chronological:poly(day, 3)1	0.512	[0.388, 0.636]	8.09	p < .001
conditionEngagement-Based:poly(day, 3)2	conditionEngagement-Based:poly(day, 3)2	0.526	[0.401, 0.651]	8.25	p < .001
conditionReverse Chronological:poly(day, 3)2	conditionReverse Chronological:poly(day, 3)2	-3.281	[-3.405, -3.156]	-51.59	p < .001
conditionEngagement-Based:poly(day, 3)3	conditionEngagement-Based:poly(day, 3)3	-0.623	[-0.748, -0.499]	-9.84	p < .001
conditionReverse Chronological:poly(day, 3)3	conditionReverse Chronological:poly(day, 3)3	1.392	[1.268, 1.515]	22.01	p < .001

Table S35. Tabulated model results from condition X time with cubic polynomial term, toxic content

Model Summary: Moral Outrage - Cubic Model

	Term	Coefficient	95% CI	t	p-value
(Intercept)	(Intercept)	0.182	[0.181, 0.183]	331.44	p < .001
conditionEngagement-Based	conditionEngagement-Based	0.022	[0.020, 0.023]	28.22	p < .001
conditionReverse Chronological	conditionReverse Chronological	-0.056	[-0.058, -0.055]	-73.08	p < .001
poly(day, 3)1	poly(day, 3)1	-2.687	[-2.923, -2.452]	-22.37	p < .001
poly(day, 3)2	poly(day, 3)2	14.929	[14.692, 15.166]	123.51	p < .001
poly(day, 3)3	poly(day, 3)3	-4.629	[-4.864, -4.393]	-38.52	p < .001
conditionEngagement-Based:poly(day, 3)1	conditionEngagement-Based:poly(day, 3)1	1.324	[0.993, 1.655]	7.85	p < .001
conditionReverse Chronological:poly(day, 3)1	conditionReverse Chronological:poly(day, 3)1	-1.536	[-1.866, -1.205]	-9.11	p < .001
conditionEngagement-Based:poly(day, 3)2	conditionEngagement-Based:poly(day, 3)2	0.869	[0.536, 1.201]	5.12	p < .001
conditionReverse Chronological:poly(day, 3)2	conditionReverse Chronological:poly(day, 3)2	-14.853	[-15.184, -14.521]	-87.73	p < .001
conditionEngagement-Based:poly(day, 3)3	conditionEngagement-Based:poly(day, 3)3	-2.851	[-3.182, -2.521]	-16.90	p < .001
conditionReverse Chronological:poly(day, 3)3	conditionReverse Chronological:poly(day, 3)3	5.047	[4.718, 5.377]	29.99	p < .001

Table S36. Tabulated model results from condition X time with cubic polynomial term, moral outrage content.

Model Summary: Constructive Content - Cubic Model

	Term	Coefficient	95% CI	t	p-value
(Intercept)	(Intercept)	0.076	[0.076, 0.077]	227.96	p < .001
conditionEngagement-Based	conditionEngagement-Based	0.007	[0.006, 0.008]	15.55	p < .001
conditionReverse Chronological	conditionReverse Chronological	0.008	[0.007, 0.009]	17.87	p < .001
poly(day, 3)1	poly(day, 3)1	1.774	[1.676, 1.871]	35.65	p < .001
poly(day, 3)2	poly(day, 3)2	4.299	[4.202, 4.396]	86.65	p < .001
poly(day, 3)3	poly(day, 3)3	-0.032	[-0.129, 0.065]	-0.65	p = 0.516
conditionEngagement-Based:poly(day, 3)1	conditionEngagement-Based:poly(day, 3)1	-0.830	[-0.967, -0.693]	-11.88	p < .001
conditionReverse Chronological:poly(day, 3)1	conditionReverse Chronological:poly(day, 3)1	-0.575	[-0.711, -0.439]	-8.27	p < .001
conditionEngagement-Based:poly(day, 3)2	conditionEngagement-Based:poly(day, 3)2	0.098	[-0.039, 0.234]	1.40	p = 0.160
conditionReverse Chronological:poly(day, 3)2	conditionReverse Chronological:poly(day, 3)2	-3.953	[-4.089, -3.817]	-56.94	p < .001
conditionEngagement-Based:poly(day, 3)3	conditionEngagement-Based:poly(day, 3)3	-0.638	[-0.774, -0.502]	-9.21	p < .001
conditionReverse Chronological:poly(day, 3)3	conditionReverse Chronological:poly(day, 3)3	1.142	[1.006, 1.277]	16.54	p < .001

Table S37. Tabulated model results from condition X time with cubic polynomial term, constructive content

6.8 Estimating the impact of Bluesky's new user growth on content

The duration of our study coincided with a dramatic increase and interest in the Bluesky platform. Throughout the study, various exogenous events (e.g., an exodus of users off Twitter, the 2024 US election, etc.) led to massive spikes in user growth on the Bluesky platform.

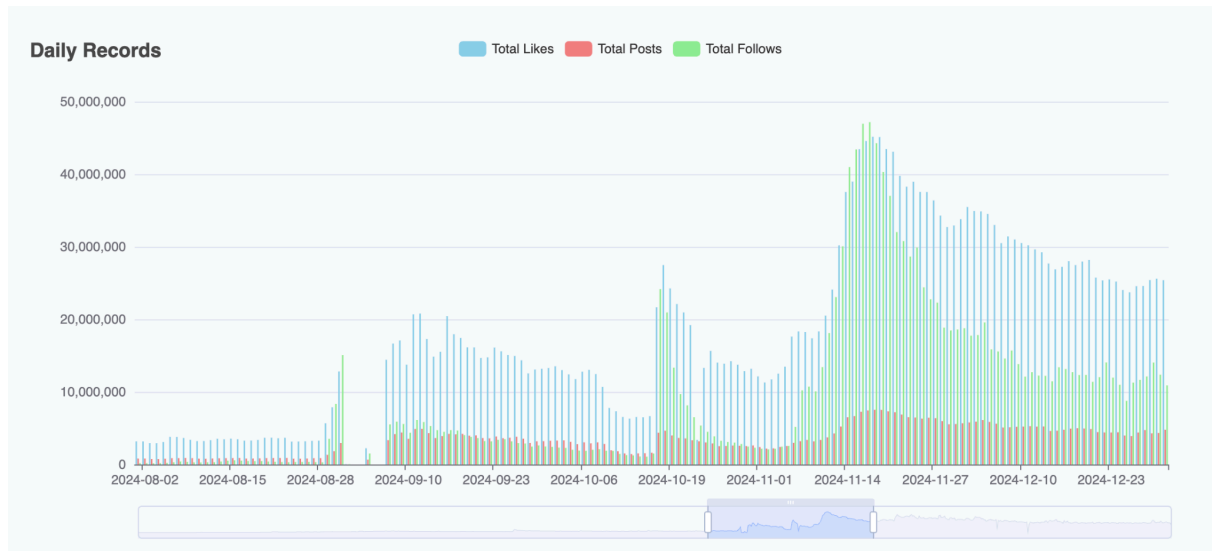


Fig. S37. A histogram of daily total likes, posts, and follows on the Bluesky platform, sourced from Bluesky's own metrics dashboard, from the period August 1st, 2024 (1 month prior to the study) to December 31st, 2024 (1 month after the study). This histogram shows a >10x increase in total posts, total network follows, and total likes on posts during this span, denoting the exponential growth in activity, interest, and engagement on the platform that overlapped with the span of our study²⁰.

We examined the possible impact of Bluesky's new user growth on the measures for our study. Specifically, we reviewed whether the content posted by users varied as a function of when they joined the platform. As Bluesky moved along the adoption curve away from niche early adopters to becoming more of a mainstream application, we hypothesized that the distribution of new users could possibly correlate with a difference in the type of content posted.

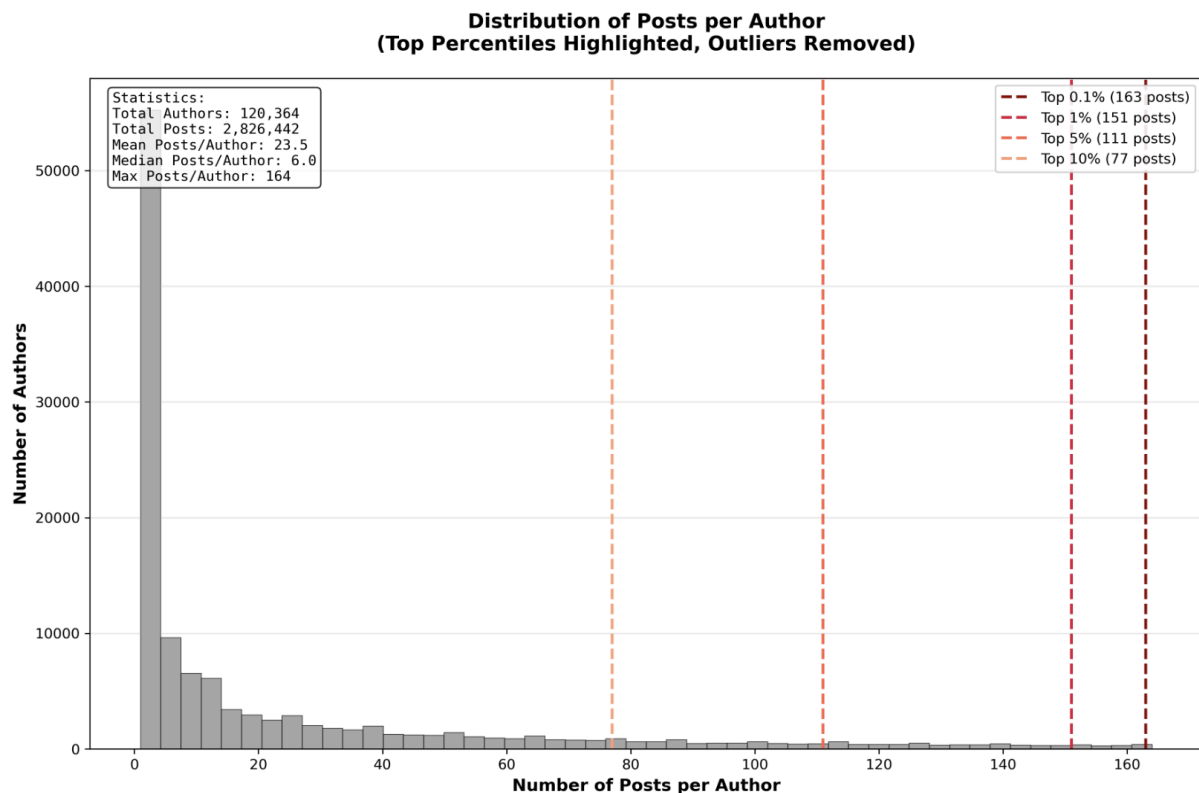


Fig. S38. A histogram of the number of posts written per author, across all posts in our database. We remove outliers ($1.5 \times \text{IQR}$) and plot the distribution of posts per user as well as associated percentiles. We see a heavy right-skew in the distribution, where a set of power users dominate the distribution of total posts written per author. This tracks with existing research describing the influence of “superposters” on social media discourse.

We sampled 3,000 Bluesky users (not users in our study) who posted ≥ 20 times throughout the date range of our study, and correlated the average toxicity and moral outrage measures of their content with when they joined the platform. After removing users with deactivated profiles, we had a sample size of $n=2,862$ users. We then split users between two groups, those who joined prior to September 1st, 2024 ($n=2,356$), and those who joined afterwards ($n=506$). We chose this date because it corresponds to the first massive spike in new users on the Bluesky platform²¹, a result of Brazil banning Twitter, with the increased traffic being large enough to cause outages on Bluesky itself. Although the new user base was primarily from Brazil, it is the first time that mainstream media began to actively cover Bluesky²², leading to increased awareness and interest from general social media users that preceded the subsequent user growth resulting from US-based exogenous events (e.g., exodus from Twitter, the 2024 US election, etc).

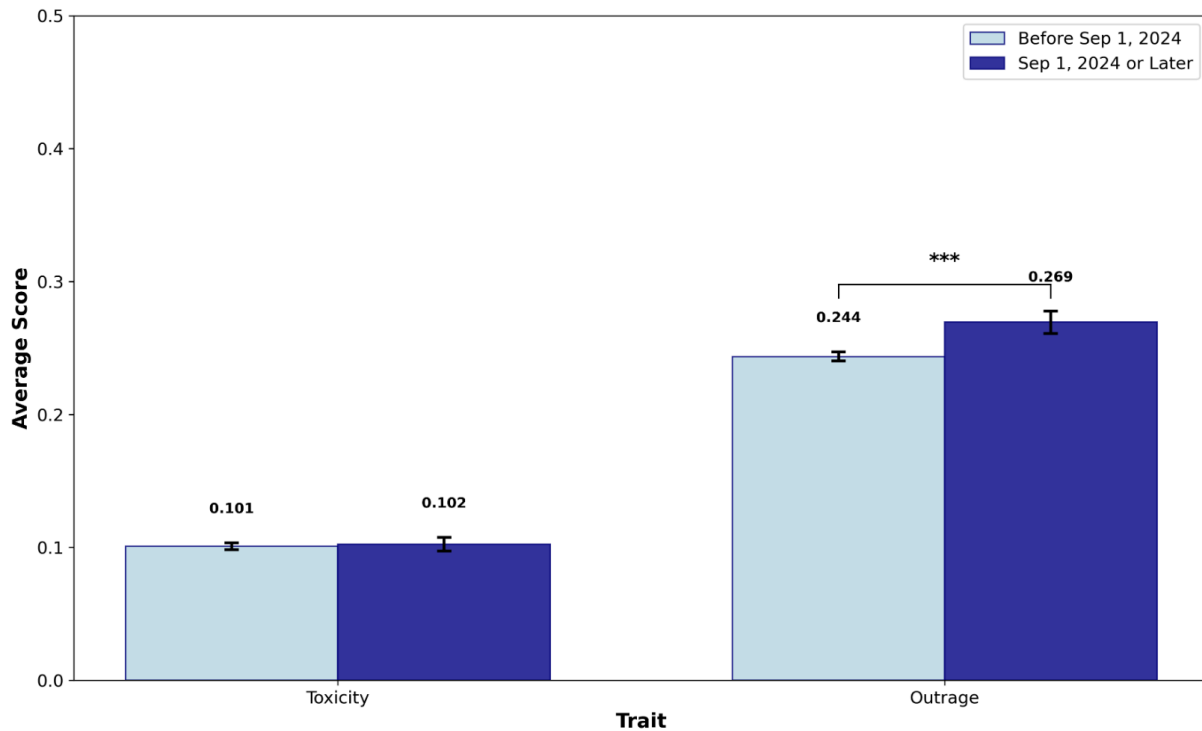


Fig. S39. A graph of the average toxicity and moral outrage classification probabilities of content posted by users who joined before September 1st, 2024, and those who joined after, with associated error bars. We see a statistically significant increase in the average moral outrage of content posted by users who joined after September 1st, 2024, as compared to content by users posted before. Error bars represent 95% confidence intervals. Asterisks indicate significant group differences from independent-samples t-tests ($***p < .001$).

We examined whether average toxicity and outrage scores differed before and after September 1, 2024. For toxicity, average scores were nearly identical across the two time periods (Before: $M = 0.101$, $SD = 0.062$, $n = 2,356$; After: $M = 0.102$, $SD = 0.060$, $n = 506$). An independent-samples t-test indicated no significant difference between groups, $t(2860) = 0.51$, $p = .61$, Cohen's $d = 0.02$, 95% CI $[-0.004, 0.007]$. The nonparametric Mann–Whitney U test converged on the same conclusion, $U = 611,863$, $p = .35$.

For outrage, scores were modestly higher after September 1, 2024 (Before: $M = 0.244$, $SD = 0.086$, $n = 2,356$; After: $M = 0.269$, $SD = 0.095$, $n = 506$). Welch's t-test (applied due to unequal variances, Levene's $p < .001$) revealed a statistically significant difference, $t(\sim 717) = 5.61$, $p < .001$, Cohen's $d = 0.29$, 95% CI $[0.017, 0.035]$. The Mann–Whitney U test confirmed this result, $U = 690,390$, $p < .001$.

Taken together, these analyses suggest that Bluesky's rapid influx of new users during the study period coincided with a modest but reliable and statistically significant increase in moral outrage expressed in posts, whereas levels of toxicity remained stable. This finding highlights that, in addition to algorithmic amplification effects reported above, changes in platform composition during exogenous shocks may also shape the tone of discourse online.

6.9 Comparison of in- vs. out-network toxic content

A key advantage of building our own engagement-based algorithm is that we had the ability to disentangle key components used in the algorithms of major social media platforms. Specifically, we engineered our engagement-based ranking so that two families of signals could be isolated and tested separately: (i) community (out-of-network) engagement, which represented platform-wide popularity/trending signals derived from posts outside a user's follow network; and (ii) personalized (in-network) engagement, which represented signals based on a user's prior interactions within their follow network. This decomposition, common in modern recommender systems (popularity vs. personalization), allowed us to estimate which component primarily amplifies toxic political content.

We discovered that the community (out-of-network) component disproportionately drove amplification of toxic political content. At its peak during the study window, the share of posts across conditions delivered in the feeds that were toxic reached 15.8% for the community-engagement stream, compared with 10.0% for the personalized (in-network) stream. See Fig. S40.

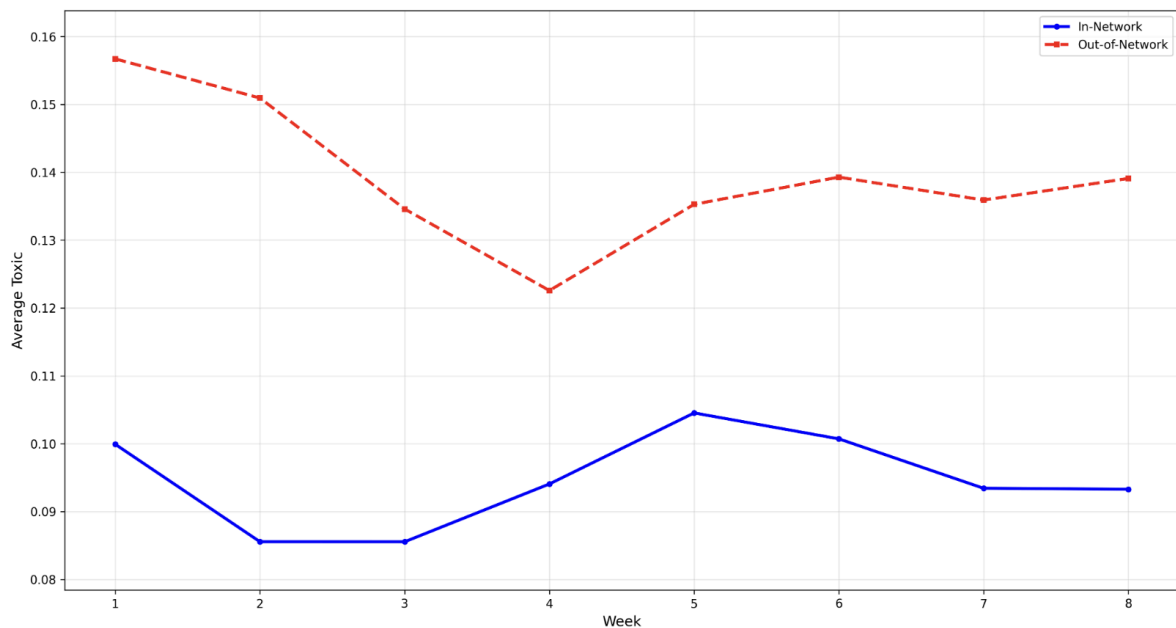


Fig. S40 A graph of the average toxicity of feeds, comparing whether the posts were derived from in-network or out-of-network sources. We see that the average toxicity of out-of-network posts was noticeably greater, across all weeks, than posts derived from in-network sources.

As shown in **Fig. S40**, out-of-network content was consistently more toxic than in-network content throughout the study period. While both series fluctuated week to week, the gap between

in-network and out-of-network posts remained stable, with out-of-network content carrying substantially higher average toxicity scores.

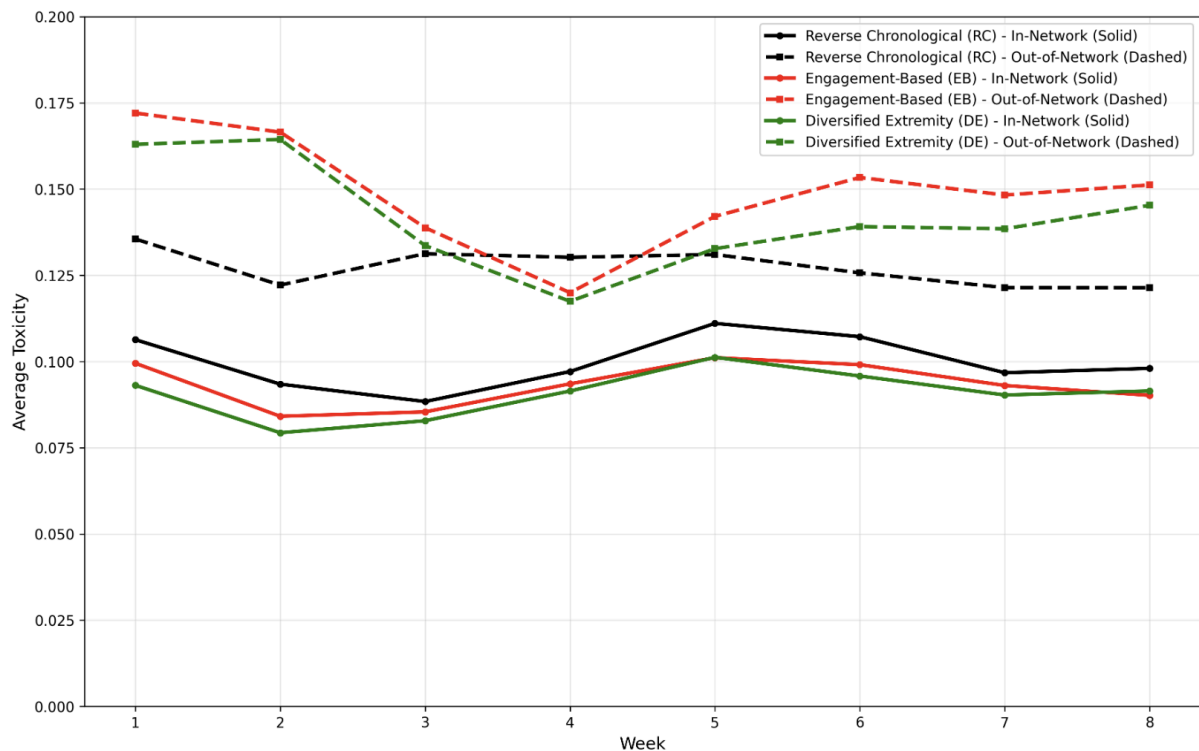


Fig. S41. Visualizing the average toxicity of in-network versus out-of-network content in feeds, across time and by condition. Across conditions, we see that the toxicity of in-network content is consistently lower than out-of-network content. As seen in our previous results, we show that the diversified extremity (DE) feeds, across both in-network and out-of-network content shown in feeds, were less toxic than engagement-based (EB) feeds.

Taken together, our analysis here sheds light on the trends in our previous results. We see a difference in toxicity measures between in-network and out-of-network content. More generally, we have further insight into the difference in content between those presented algorithmically to the user (out-of-network) versus content derived from a user's own curated network (in-network). Feed-ranking algorithms can mitigate (as with diversification) or exacerbate (as with engagement-based ranking) exposure to such content.

6.10 Further analyses of constructive dialogue classification

In the main analyses, we found that diversified extremity (DE) feeds served slightly less content classified as *constructive dialogue* than engagement-based (EB) feeds, despite the fact that constructiveness was one of the explicit ranking criteria in the DE algorithm. At face value, this result appears counterintuitive. Here, we report additional exploratory analyses that clarify why EB can appear higher than DE on this measure in practice.

Overlap between constructive dialogue and moral outrage classifications. First, we examined whether posts classified as “constructive” were distinct from other IME dimensions amplified by EB. Using Google’s Perspective API classification thresholds (probability > .50), we found that approximately 10% of posts classified as constructive dialogue were also classified as moral outrage. Inspection of these posts revealed a consistent pattern: they typically presented a reasoned political argument or explanation but also included a brief sentence expressing moral condemnation or indignation (see examples in Table S38). These posts can therefore plausibly register as both constructive dialogue and moral outrage according to the classifier.

Because EB strongly amplifies outrage content in practice (see main results), it can indirectly amplify this overlapping subset of posts, thereby inflating the apparent prevalence of constructive content in EB feeds relative to DE feeds. To test this interpretation, we reran the constructiveness models after excluding posts that were simultaneously classified as constructive dialogue and moral outrage. As shown in Fig. S42, removing these overlap posts attenuated the EB–DE difference in constructive content, and EB no longer showed higher levels of constructiveness than the reverse-chronological condition. This pattern suggests that part of the EB advantage on the constructiveness metric is driven by outrage amplification rather than by the amplification of purely constructive dialogue.

Effects of superposter downranking on measured constructiveness. Second, we examined whether the DE algorithm’s downranking of superposters contributed to the observed pattern. A core design feature of DE is to reduce the influence of highly prolific accounts, given prior evidence that superposters disproportionately contribute to toxic political content. However, this filtering step also has the side effect of downranking some highly active accounts that post frequent, “reasoned” political commentary.

In exploratory analyses, we found that superposters who appeared in the EB and DE candidate pools posted higher levels of content classified as constructive dialogue, on average, than non-superposters, despite also posting more toxic content overall. As shown in Table S39, this pattern holds across the EB and DE conditions. Consequently, downranking superposters in DE can mechanically reduce the overall proportion of posts classified as constructive, even though DE explicitly upranks constructiveness conditional on posts remaining in the candidate pool.

Summary. Taken together, these analyses help explain why EB can appear to serve slightly more constructive content than DE in practice. EB’s strong amplification of outrage indirectly boosts a subset of posts that overlap between constructive dialogue and moral outrage, while DE’s superposter downranking removes some highly prolific accounts that contribute more classifier-labeled constructiveness. These findings illustrate how non-orthogonal content labels and upstream filtering decisions can interact in unintuitive ways in multi-objective feed-ranking systems, and they underscore the importance of examining classifier overlap and user-level contribution patterns when interpreting algorithmic effects on complex content dimensions.

	Messages Classified as containing both constructive dialogue and outrage expression
1.	Hearing public figures speak honestly about reproductive health highlights how much real suffering gets erased in policy debates. Decisions about abortion should reflect lived experience and medical reality, not nostalgia or ideology. It's hard not to feel angry when politicians casually push people back into pain they've never had to endure.
2.	A serious investigation is needed into how Florida officials are using emergency management resources in relation to migrant relocation. If public agencies are being repurposed to score political points, that deserves scrutiny. Treating human beings like logistical problems is morally disturbing.
3.	Before signing petitions that promise quick benefits, people should consider how those signatures might later be used or misused. Even if your vote stays the same, the data you hand over can empower actors you strongly oppose. It's infuriating how often manipulation hides behind convenience.
4.	Environmental reporting requires consultation with experts who understand fragile ecosystems, especially when development threatens long-term damage. Ignoring that responsibility isn't neutrality—it's negligence. The casual disregard for environmental harm here is honestly outrageous.
5.	Balance and restraint are important, but they stop being virtues when one actor repeatedly violates democratic norms. Calling that "controversial" rather than illegitimate blurs a critical line. At some point, pretending this is normal becomes deeply offensive.

Table S38. Examples of messages classified as both constructive dialogue and moral outrage (probability greater than .5) by Google's Perspective API. In order to preserve user privacy, messages were paraphrased from the original text but maintain the same idea and expression style.

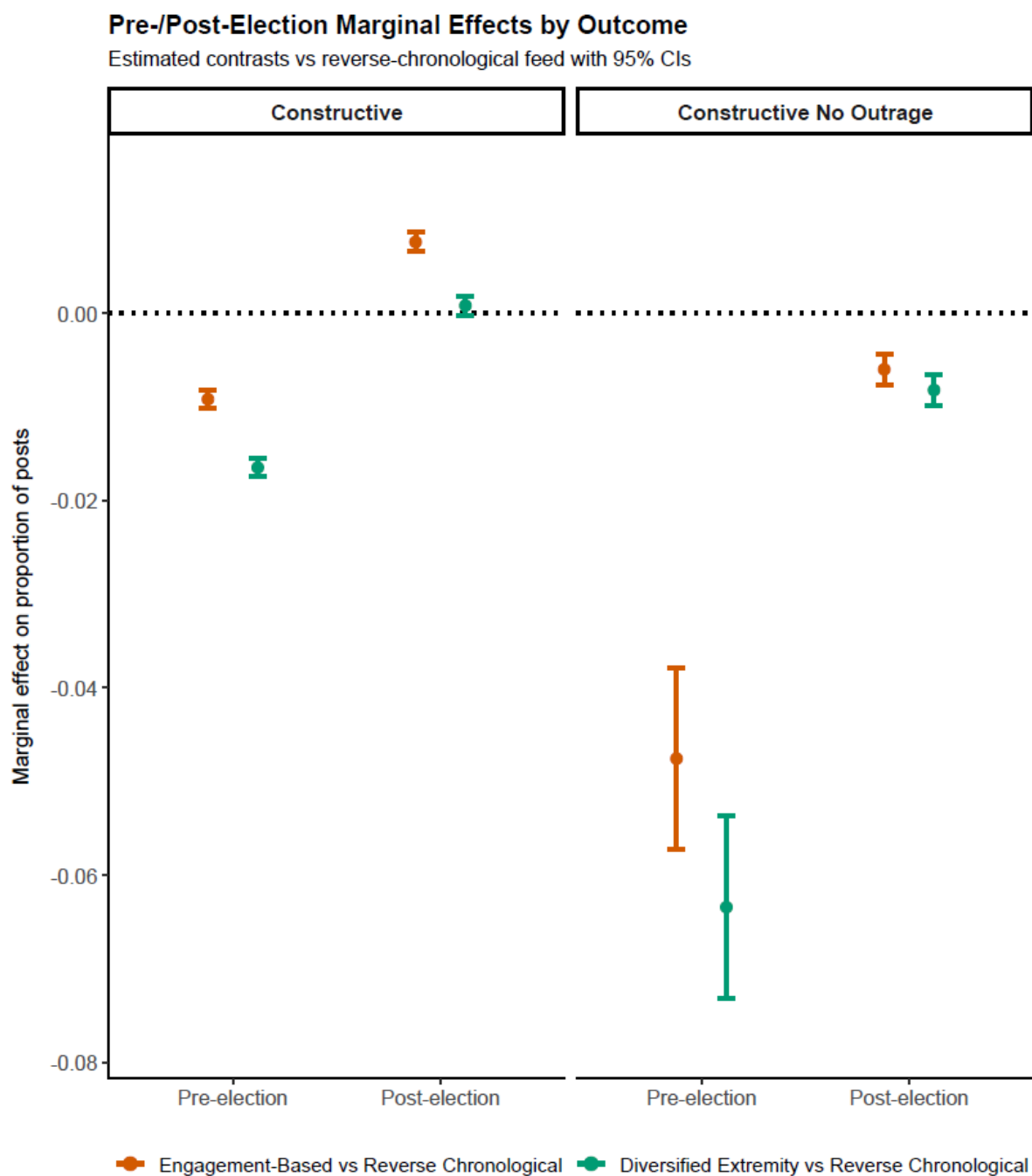


Fig. S42. Comparing marginal effects of condition on constructive content displayed in feeds, comparing all constructive content to constructive content with moral outrage removed.

Superposter status	Diversified extremity	Engagement-based	Reverse chronological
Not a superposter	0.1509	0.1593	0.2278
Superposter	0.1935	0.1950	0.1921

Table S39. Average probability of constructive content classifications by condition and superposter status. SI Appendix Section 3.3.4 details how we quantified superposters in our study.

6.11 Re-classification of intergroup content

To evaluate the performance of our in-house classifiers for intergroup, moral, and emotional (IME) content at the scale of the full experiment, we conducted a post-study quality control analysis. Using the original pre-registered definitions of IME content developed during classifier construction (see SI Section 3.3.3), two members of the research team independently manually coded random samples of posts drawn from the study corpus.

For each content category (intergroup, moral, emotional), we selected three independent samples of 50 posts (150 posts per category). Within each sample, we approximated class balance by selecting 50% of posts classified as positive instances and 50% classified as negative instances according to the original automated classifier. Positive and negative instances were defined based on the classifier's binary output.

Inter-rater reliability between the two human coders was high across categories ($ICC(2,2) = .88$), indicating strong agreement in applying the predefined coding scheme. We then compared classifier outputs to the human-coded labels. Following prior validation standards (e.g., Brady et al., 2021), we adopted a 70% accuracy benchmark as a minimum threshold for acceptable performance warranting no further investigation. Mean accuracy relative to human coding was as follows:

- Intergroup content: 62%
- Moral content: 82%
- Emotional content: 76%

Both the moral and emotional classifiers exceeded the 70% benchmark. However, the intergroup classifier fell below this threshold, indicating insufficient performance at scale and necessitating follow-up analysis.

To address this issue, we implemented a revised intergroup classifier using a few-shot prompt-tuned large language model (LLM). The prompt included a definition of intergroup content consistent with our preregistered construct, along with examples of positive and negative cases drawn from the human-coded sample. When evaluated against the same human-coded test set, the revised classifier achieved 80% accuracy, exceeding the predefined benchmark and substantially improving upon the original model.

We then reclassified all posts in the study corpus using the revised intergroup classifier and recomputed daily mean proportions of intergroup content in participants' feeds. These updated results are reported in the main text (RQ1 section). Fig. S43 below displays the corresponding marginal effects. Substantively, the revised intergroup results more closely mirrored those observed for the aggregate IME classifier reported in the main text (RQ1), suggesting that the pattern of results using the pre-registered classifier may have underestimated how much intergroup content was amplified in the study period.

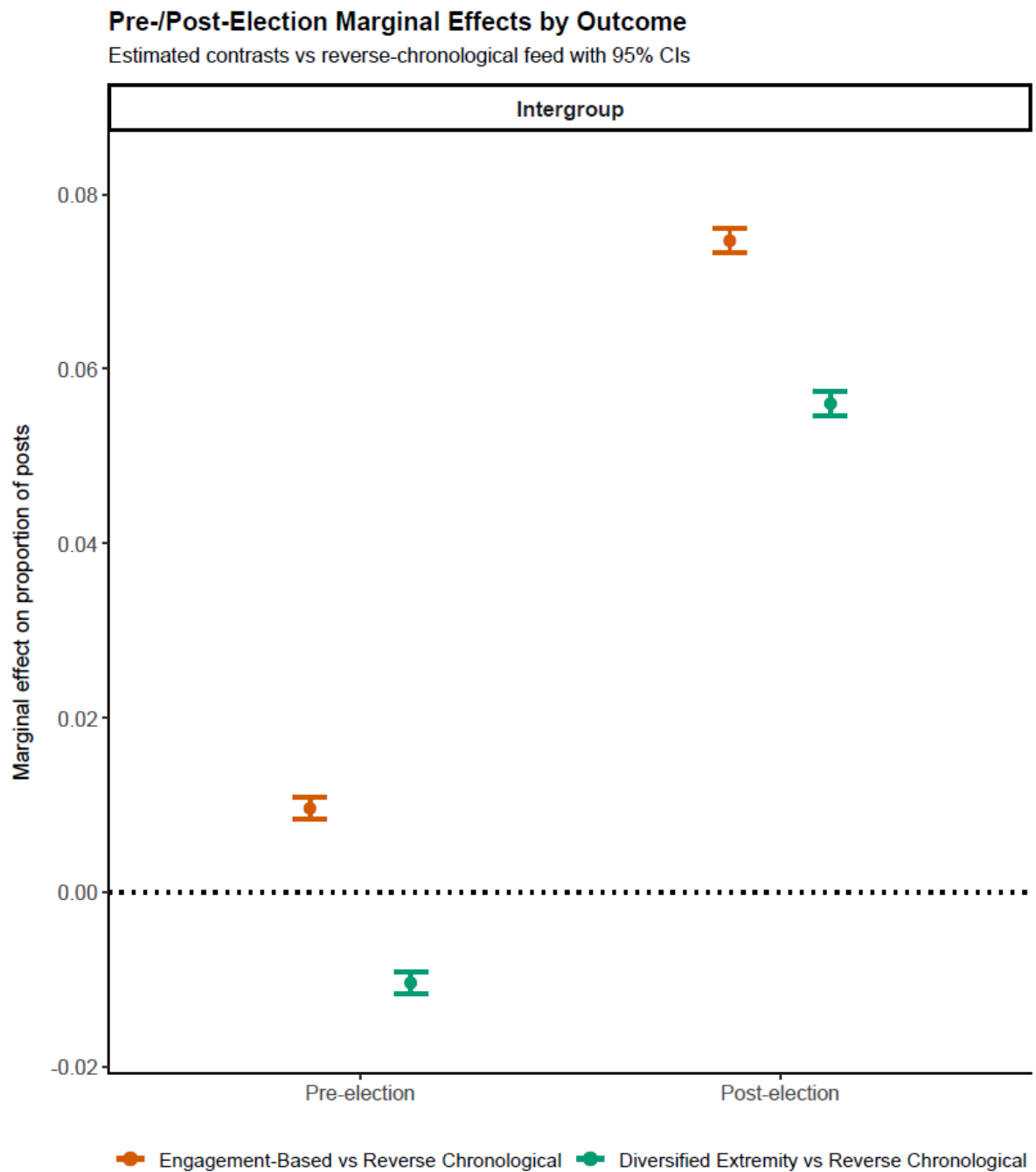


Fig. S43. Effects of engagement-based (EB) and diversified extremity (DE) algorithm feeds compared to a reverse-chronological feed on intergroup feed content before and after the 2024 U.S. Presidential election. The plot displays marginal effects of EB vs. RC and DE vs. RC on the proportion of intergroup content served in participants' feeds. Values greater than 0 indicate higher greater proportions of content compared to the reverse-chronological feed.

7.0 RQ2 supplementary information

7.1 Attrition analyses

Differential attrition occurs when participants drop out of a study at different rates across experimental conditions, which can bias results and threaten the validity of between-group comparisons. To assess whether attrition varied systematically across our three experimental conditions, we conducted a formal test for differential attrition.

We recruited 2,000 participants to the study assuming a 20% attrition rate, with participants defined as having dropped out if they did not complete 63% of the study. After pre-screening for eligibility criteria (U.S. citizens, Democrat or Republican identification, and active Bluesky users logging in at least once per week), we excluded participants from data analysis who failed to complete 63% of the weekly surveys (i.e., missed more than 3 of the 8 weeks) or did not engage with the platform for at least 5 of the 8 weeks of the experiment.

We defined study completion as appearing in both the weekly login tracking data and meeting the minimum engagement thresholds described above. Users who logged into the platform but did not meet these completion criteria were classified as having dropped out. We calculated attrition rates for each of the three experimental conditions (engagement-based, diversified extremity, and reverse-chronological) and tested whether these rates differed significantly using Pearson's chi-square test of independence. This test examines whether there is a significant association between experimental condition and study completion status, with a significant result indicating differential attrition that could bias between-group comparisons.

Of the 2000 users who enrolled in the study and completed surveys, 1834 users actually ever logged into our feeds. Of the 1834 users who logged into the feeds, 1,475 (80.4%) completed the study, resulting in an overall attrition rate of 19.6% relative to ever attempting to login to the study feeds. Attrition rates varied descriptively across conditions: 19.7% in the engagement-based condition, 17.2% in the diversified extremity condition, and 21.8% in the reverse-chronological condition (see Table S40).

Condition	Initial N	Retained N	Dropped N	Attrition Rate
<i>Engagement-Based (EB)</i>	614	493	121	19.7%
<i>Diversified Extremity (DE)</i>	600	497	103	17.2%
<i>Reverse-Chronological (RC)</i>	620	485	135	21.8%
Total	1,834	1,475	359	19.6%

Table S40. Attrition rates by condition. The table displays attrition rates for users who actually logged into the condition feeds (166 of the full 2000 sample did not attempt to log into feeds).

The chi-square test revealed no statistically significant association between experimental condition and study completion ($\chi^2 = 4.122$, $df = 2$, $p = .127$). This indicates that attrition rates did not differ significantly across the three experimental conditions. The absence of differential attrition suggests that participant dropout was not systematically related to experimental condition assignment. This supports the validity of between-group comparisons in our main analyses and justifies our use of condition-specific mean imputation for handling missing data on dependent variables, as specified in our pre-registered analysis plan.

7.2 Model fit analyses

Table S39 displays a summary of model fit statistics for each model reported in the main text for RQ2. The pre/post model with an election indicator was preferred for 6/8 models.

Variable	AIC_Base	AIC_PrePost	BIC_Base	BIC_PrePost	LLR	LLR_p	Preferred_Model
Descriptive Accuracy – Political	103012.4	103002.6	103049.3	103061.6	13.404	0.004	Pre/Post
Descriptive Accuracy – Republican	104129.1	104125.2	104166.0	104184.2	7.289	0.063	Pre/Post
Descriptive Accuracy – Democratic	104126.8	104046.6	104163.6	104105.6	83.565	0.000	Pre/Post
Descriptive Accuracy – Toxic	105482.2	105473.8	105519.0	105532.8	11.482	0.009	Pre/Post
Descriptive Accuracy – Emotion	101636.4	101605.6	101673.3	101664.6	34.781	0.000	Pre/Post
Prescriptive Norm Accuracy	105450.0	105452.1	105486.9	105511.1	1.159	0.763	Base
Ingroup Dislike	102734.2	102736.3	102771.1	102795.3	2.094	0.553	Base
Outgroup Dislike	102874.0	102871.1	102910.8	102930.2	6.878	0.076	Pre/Post

Table S41. Model fit statistics comparing base vs. pre/post model for RQ2 analyses appearing in main text.

7.3 Fully tabulated results (pre/post models)

RQ2: Descriptive Accuracy: Political Content — Pre/Post model			
Entries are β (SE). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	-3.234 (0.979)	-3.322 (0.977)	0.088 (0.973)
Post-election	-1.054 (1.119)	-1.856 (1.117)	0.802 (1.114)

Table S42: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate overestimation relative to true baseline content prevalence; negative values indicate underestimation. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Descriptive Accuracy: Right-Leaning Content — Pre/Post model			
Entries are β (SE). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	-1.685 (1.092)	-1.255 (1.090)	-0.431 (1.085)
Post-election	0.691 (1.229)	0.020 (1.227)	0.671 (1.223)

Table S43: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate overestimation relative to true baseline content prevalence; negative values indicate underestimation. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Descriptive Accuracy: Left-Leaning Content — Pre/Post model			
Entries are β (SE). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	-1.378 (1.112)	-2.645 (1.109)	1.267 (1.105)
Post-election	0.135 (1.245)	-0.769 (1.242)	0.904 (1.239)

Table S44: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate overestimation relative to true baseline content prevalence; negative values indicate underestimation. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Descriptive Accuracy: Toxic Content — Pre/Post model			
Entries are β (SE). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	-3.750 (1.333)	-2.233 (1.331)	-1.517 (1.325)
Post-election	-1.277 (1.456)	-0.514 (1.453)	-0.763 (1.448)

Table S45: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate overestimation relative to true baseline content prevalence; negative values indicate underestimation. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Descriptive Accuracy: Emotional Content — Pre/Post model			
Entries are β (SE). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	-1.074 (0.939)	-0.623 (0.937)	-0.451 (0.933)
Post-election	-1.694 (1.069)	-3.187 (1.067)	1.493 (1.064)

Table S46: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate overestimation relative to true baseline content prevalence; negative values indicate underestimation. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Prescriptive Norm Accuracy — Pre/Post model			
Entries are β (SE). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	-5.200 (1.562)	-1.778 (1.559)	-3.422 (1.552)
Post-election	-4.750 (1.663)	-1.618 (1.660)	-3.132 (1.654)

Table S47: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate overestimation of permissiveness of posting toxic content; negative values indicate underestimation of permissiveness of posting toxic content. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Ingroup Dislike — Pre/Post model			
Entries are β (SE). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	-2.918 (1.662)	-2.595 (1.659)	-0.323 (1.652)
Post-election	-3.467 (1.734)	-2.159 (1.731)	-1.308 (1.725)

Table S48: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Negative values indicate greater perceived dislike in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Outgroup Dislike — Pre/Post model			
Entries are β (SE). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	-2.825 (1.659)	-1.376 (1.656)	-1.449 (1.649)
Post-election	-4.539 (1.732)	-3.003 (1.729)	-1.535 (1.723)

Table S49: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Negative values indicate greater perceived dislike in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Enjoy Bluesky — Pre/Post model			
Entries are β (SE). Bold indicates Holm–Bonferroni-adjusted $p < .05$ (exploratory outcome).			
	EB – RC	DE – RC	EB – DE
Pre-election	-0.102 (0.051)	-0.133 (0.051)	0.031 (0.051)
Post-election	-0.587 (0.057)	-0.033 (0.057)	-0.554 (0.057)

Table S50: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate higher enjoyment ratings in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Enjoy Content — Pre/Post model			
Entries are β (SE). Bold indicates Holm–Bonferroni-adjusted $p < .05$ (exploratory outcome).			
	EB – RC	DE – RC	EB – DE
Pre-election	-0.053 (0.049)	-0.086 (0.049)	0.034 (0.048)
Post-election	-0.061 (0.055)	-0.589 (0.055)	0.528 (0.055)

Table S51: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate higher enjoyment ratings in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

7.4 Fully tabulated model results (base models)

RQ2: Descriptive Accuracy: Political Content — Base model		
Entries are β (SE). Bold indicates $p < .05$.		
EB – RC	DE – RC	EB – DE
-2.586 (0.947)	-2.883 (0.945)	0.297 (0.941)

Table S52: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate overestimation relative to true baseline content prevalence; negative values indicate underestimation. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Descriptive Accuracy: Right-Leaning Content — Base model

Entries are β (SE). Bold indicates $p < .05$.

EB – RC	DE – RC	EB – DE
-0.983 (1.061)	-0.876 (1.058)	-0.107 (1.054)

Table S53: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate overestimation relative to true baseline content prevalence; negative values indicate underestimation. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Descriptive Accuracy: Left-Leaning Content — Base model

Entries are β (SE). Bold indicates $p < .05$.

EB – RC	DE – RC	EB – DE
-0.915 (1.080)	-2.075 (1.078)	1.159 (1.074)

Table S54: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate overestimation relative to true baseline content prevalence; negative values indicate underestimation. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Descriptive Accuracy: Toxic Content — Base model		
Entries are β (SE). Bold indicates $p < .05$.		
EB – RC	DE – RC	EB – DE
-3.021 (1.306)	-1.726 (1.304)	-1.295 (1.298)

Table S55: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate overestimation relative to true baseline content prevalence; negative values indicate underestimation. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Descriptive Accuracy: Emotional Content — Base model

Entries are β (SE). Bold indicates $p < .05$.

EB – RC	DE – RC	EB – DE
-1.267 (0.909)	-1.387 (0.907)	0.120 (0.903)

Table S56: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate overestimation relative to true baseline content prevalence; negative values indicate underestimation. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Prescriptive Norm Accuracy — Base model

Entries are β (SE). Bold indicates $p < .05$.

EB – RC	DE – RC	EB – DE
-5.066 (1.540)	-1.728 (1.537)	-3.337 (1.530)

Table S57: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate overestimation of permissiveness of posting toxic content; negative values indicate underestimation of permissiveness of posting toxic content. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Ingroup Dislike — Base model		
Entries are β (SE). Bold indicates $p < .05$.		
EB – RC	DE – RC	EB – DE
-3.078 (1.646)	-2.467 (1.643)	-0.612 (1.636)

Table S58: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Negative values indicate greater perceived dislike in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Outgroup Dislike — Base model		
Entries are β (SE). Bold indicates $p < .05$.		
EB – RC	DE – RC	EB – DE
-3.331 (1.643)	-1.857 (1.640)	-1.474 (1.633)

Table S59: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Negative values indicate greater perceived dislike in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Enjoy Bluesky — Base model

Entries are β (SE). Bold indicates Holm–Bonferroni-adjusted $p < .05$ (exploratory outcome).

EB – RC	DE – RC	EB – DE
-0.247 (0.049)	-0.106 (0.049)	-0.140 (0.049)

Table S60: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate higher enjoyment ratings in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Enjoy Content — Base model

Entries are β (SE). Bold indicates Holm–Bonferroni-adjusted $p < .05$ (exploratory outcome).

EB – RC	DE – RC	EB – DE
-0.058 (0.047)	-0.236 (0.047)	0.179 (0.047)

Table S61: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate higher enjoyment ratings in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

7.5 Effects of condition on raw perception variables (ignoring accuracy) - pre/post models

RQ2, H1

We examined whether engagement-based algorithm feeds directly altered participants' perceptions of content prevalence and appropriateness, irrespective of accuracy. We report marginal effects for engagement-based vs. reverse-chronological feeds before and after the 2024 U.S. Presidential election; fully tabulated results are reported below, see also Fig. S42.

Descriptive norm perception: political content. Engagement-based feeds led participants to perceive significantly less political content in their feeds prior to the election compared to reverse-chronological feeds (3.22 points lower, $p < .001$). This difference was no longer significant after the election (0.98 points lower, $p = .380$).

Descriptive norm perception: democratic-leaning content. Engagement-based feeds did not significantly affect perceived prevalence of Democratic-leaning content before the election (1.39 points lower, $p = .210$) or after the election (0.16 points higher, $p = .900$).

Descriptive norm perception: republican-leaning content. Engagement-based feeds likewise did not significantly alter perceived Republican-leaning content either before the election (1.68 points lower, $p = .124$) or after the election (0.70 points higher, $p = .569$).

Descriptive norm perception: emotional content. Perceived emotional content did not significantly differ between engagement-based and reverse-chronological feeds before the election (1.08 points lower, $p = .250$) or after the election (1.70 points lower, $p = .112$).

Descriptive norm perception: toxic content. Engagement-based feeds significantly reduced perceived toxic content before the election (3.75 points lower, $p = .005$). This effect did not persist after the election (1.28 points lower, $p = .381$).

Perception of toxicity appropriateness (others). Engagement-based feeds significantly reduced perceived norms of content appropriateness for others. Before the election, perceived appropriateness was 5.19 points lower relative to reverse-chronological feeds ($p < .001$), and this reduction persisted after the election (4.78 points lower, $p = .004$).

Own judgment of toxicity appropriateness (self). Engagement-based feeds also reduced perceived appropriateness for oneself both before the election (4.15 points lower, $p = .009$) and after the election (4.05 points lower, $p = .016$).

Summary (RQ2, H1). Engagement-based feeds reduced perceived prevalence of political and toxic content before the election, but not after the election. At the same time, engagement-based feeds consistently reduced perceived norms of content appropriateness for both others and oneself before and after the election. Together, these findings indicate that engagement-based algorithms altered some of participants' perceptions of content prevalence and social norms, but in an unexpected direction (see discussion in main text for some interpretations of this trend).

RQ2, H2

We next examined whether diversified extremity algorithm feeds decreased perceived IME content prevalence and increased perceived content appropriateness relative to engagement-based feeds. We report marginal effects for diversified extremity vs. engagement-based feeds before and after the 2024 U.S. Presidential election; fully tabulated results are reported below, see also Fig. S42.

Descriptive norm perception: political content. Diversified extremity feeds did not

significantly alter perceived prevalence of political content relative to engagement-based feeds before the election (0.09 points lower, $p = .929$) or after the election (0.80 points lower, $p = .472$).

Descriptive norm perception: democratic-leaning content. Diversified extremity feeds did not significantly affect perceived prevalence of Democratic-leaning content either before the election (1.27 points lower, $p = .251$) or after the election (0.90 points lower, $p = .466$).

Descriptive norm perception: republican-leaning content. Diversified extremity feeds likewise did not significantly differ from engagement-based feeds in perceived Republican-leaning content before the election (0.43 points higher, $p = .691$) or after the election (0.67 points lower, $p = .584$).

Descriptive norm perception: emotional content. Perceived emotional content did not significantly differ between diversified extremity and engagement-based feeds before the election (0.45 points higher, $p = .629$) or after the election (1.49 points lower, $p = .160$).

Descriptive norm perception: toxic content. Diversified extremity feeds did not significantly differ from engagement-based feeds in perceived toxic content either before the election (1.52 points higher, $p = .252$) or after the election (0.76 points higher, $p = .598$).

Perception of toxicity appropriateness (others). Diversified extremity feeds significantly increased perceived norms of content appropriateness for others relative to engagement-based feeds before the election (3.42 points higher, $p = .027$). This effect trended but was not significant after the election (3.13 points higher, $p = .059$).

Own judgment of toxicity appropriateness (self). Diversified extremity feeds also increased perceived appropriateness for oneself before the election (3.93 points higher, $p = .013$) and after the election (3.66 points higher, $p = .029$).

Summary (RQ2, H2). Diversified extremity feeds did not significantly reduce perceived prevalence of political, toxic, partisan, or emotional content relative to engagement-based feeds either before or after the election. However, diversified extremity feeds consistently increased perceived norms of content appropriateness for both others and oneself, providing partial support for RQ2, H2. These findings suggest that diversified extremity algorithms primarily influenced perceived social norms rather than perceived content prevalence (see Discussion).

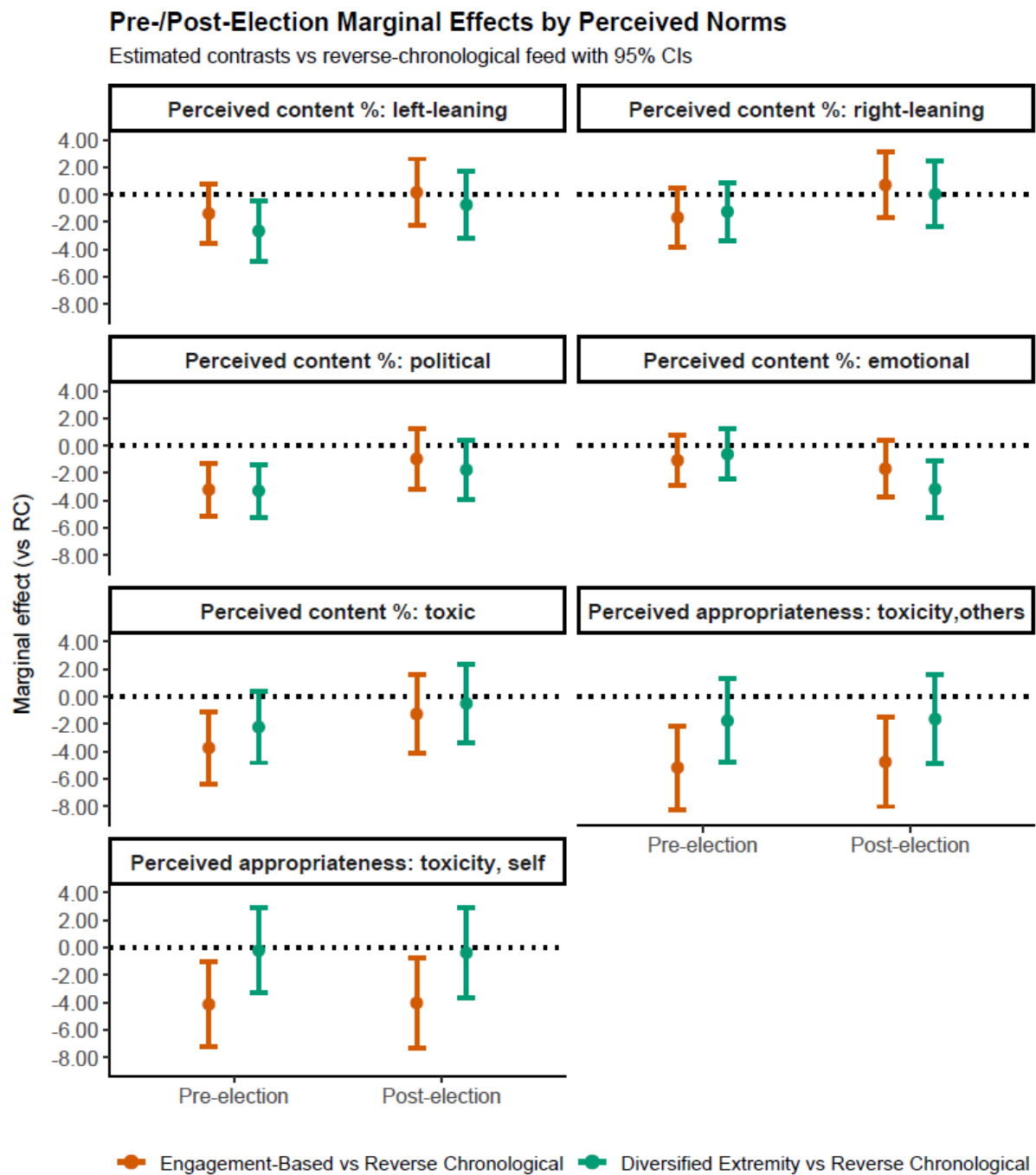


Fig. S44. Effects of engagement-based (EB) and diversified extremity (DE) algorithm feeds compared to a reverse-chronological feed on perceived content prevalence and appropriateness before and after the 2024 U.S. Presidential election. The plot displays marginal effects of EB vs. RC and DE vs. RC on participants' perceptions of the proportion of political content overall, right-leaning political content, left-leaning political content, emotional content, and toxic content in their feeds, as well as on perceived norms of content appropriateness for others and for oneself. Values greater

than 0 indicate higher perceived prevalence or greater perceived appropriateness compared to the reverse-chronological feed.

RQ2: Perceived Left-Leaning Content — Pre/Post model			
Entries are β (SE). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	-1.394 (1.112)	-2.662 (1.109)	1.268 (1.105)
Post-election	0.157 (1.243)	-0.746 (1.241)	0.903 (1.237)

Table S61. Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate higher perceived prevalence of left-leaning content in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Perceived Right-Leaning Content — Pre/Post model			
Entries are β (SE). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	-1.680 (1.092)	-1.249 (1.090)	-0.431 (1.085)
Post-election	0.701 (1.229)	0.031 (1.227)	0.670 (1.223)

Table S63: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate higher perceived prevalence of right-leaning content in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Perceived Political Content — Pre/Post model			
Entries are β (SE). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	-3.219 (0.977)	-3.306 (0.976)	0.086 (0.971)
Post-election	-0.976 (1.112)	-1.773 (1.110)	0.797 (1.107)

Table S64: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate higher perceived prevalence of political content in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Perceived Emotional Content — Pre/Post model			
Entries are β (SE). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	-1.081 (0.939)	-0.630 (0.937)	-0.450 (0.933)
Post-election	-1.696 (1.068)	-3.189 (1.066)	1.494 (1.064)

Table S65: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate higher perceived prevalence of emotional content in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Perceived Toxic Content — Pre/Post model			
Entries are β (SE). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	-3.749 (1.333)	-2.233 (1.331)	-1.517 (1.325)
Post-election	-1.277 (1.456)	-0.513 (1.453)	-0.763 (1.448)

Table S66: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate higher perceived prevalence of toxic content in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Others Appropriate — Pre/Post model			
Entries are β (SE). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	-5.190 (1.562)	-1.767 (1.559)	-3.423 (1.553)
Post-election	-4.777 (1.664)	-1.647 (1.661)	-3.130 (1.655)

Table S67: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate higher perceived appropriateness of posting toxic political content (as judged for others) in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Self Appropriate — Pre/Post model			
Entries are β (SE). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	-4.149 (1.585)	-0.218 (1.581)	-3.930 (1.575)
Post-election	-4.049 (1.684)	-0.385 (1.681)	-3.664 (1.675)

Table S68: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate higher perceived appropriateness of posting toxic political content (as judged for oneself) in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

7.6 Effects of condition on raw perception variables (ignoring accuracy) - base models

Engagement-based ranking primarily shifted how participants perceived the appropriateness of posting toxic political content rather than altering their perceptions of the partisan or emotional composition of posts. Before the election, EB participants did not differ significantly from RC in their perceptions of the share of Democratic posts ($-1.39, p = 0.210$) or Republican posts ($-1.68, p = 0.124$), and no differences were evident after the election. Similarly, EB did not alter perceptions of emotional posts either before ($-1.08, p = 0.250$) or after ($-1.70, p = 0.112$) the election. By contrast, EB participants perceived a smaller share of political posts compared to RC pre-election ($-3.22, p < .001$), though this difference diminished post-election. EB participants also perceived fewer toxic posts prior to the election ($-3.75, p = 0.005$), but this effect attenuated afterward.

EB participants judged others toxic posts as less appropriate than RC participants both before ($-5.19, p < .001$) and after ($-4.78, p = 0.004$) the election, and they also thought others viewed own toxic posts as less appropriate under EB compared to RC pre-election ($-4.15, p = 0.009$) and post-election ($-4.05, p = 0.016$). Relative to EB, DE participants rated others' toxic political posts (pre: $+3.42, p = 0.027$; post: $+3.13, p = 0.059$) and thought others viewed their own toxic political posts (pre: $+3.93, p = 0.013$; post: $+3.66, p = 0.029$) as more appropriate.

RQ2: Perceived Left-Leaning Content — Base model

Entries are β (SE). Bold indicates $p < .05$.

EB – RC	DE – RC	EB – DE
-0.928 (1.081)	-2.088 (1.079)	1.160 (1.075)

Table S69. Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate higher perceived prevalence of left-leaning content in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Perceived Right-Leaning Content — Base model

Entries are β (SE). Bold indicates $p < .05$.

EB – RC	DE – RC	EB – DE
-0.980 (1.060)	-0.872 (1.058)	-0.108 (1.054)

Table S70: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate higher perceived prevalence of right-leaning content in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Perceived Political Content — Base model

Entries are β (SE). Bold indicates $p < .05$.

EB – RC	DE – RC	EB – DE
-2.559 (0.946)	-2.854 (0.944)	0.295 (0.940)

Table S71: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate higher perceived prevalence of political content in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Perceived Emotional Content — Base model

Entries are β (SE). Bold indicates $p < .05$.

EB – RC	DE – RC	EB – DE
-1.275 (0.909)	-1.395 (0.907)	0.120 (0.903)

Table S72: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate higher perceived prevalence of emotional content in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Perceived Toxic Content — Base model

Entries are β (SE). Bold indicates $p < .05$.

EB – RC	DE – RC	EB – DE
-3.016 (1.306)	-1.721 (1.303)	-1.296 (1.298)

Table S73: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate higher perceived prevalence of toxic content in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Others Appropriate — Base model

Entries are β (SE). Bold indicates $p < .05$.

EB – RC	DE – RC	EB – DE
-5.066 (1.540)	-1.728 (1.537)	-3.337 (1.530)

Table S74: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate higher perceived appropriateness of posting toxic political content (as judged for others) in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Self Appropriate — Base model		
Entries are β (SE). Bold indicates $p < .05$.		
EB – RC	DE – RC	EB – DE
-4.118 (1.563)	-0.265 (1.559)	-3.852 (1.553)

Table S75: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate higher perceived appropriateness of posting toxic political content (as judged for oneself) in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

7.7 Other model specifications

We noticed variation in how often users logged in throughout the study, suggesting that users had different exposure to feed content which could impact norm perceptions. Thus, we ran models controlling for the login amounts per user, but results did not change.

Descriptive accuracy - political posts				
Term	Coefficient	SE	95% CI	p-value
RC baseline (pre-election)	36.64	0.69	[35.29, 38.00]	<.001
EB vs RC (pre-election)	-3.27	0.97	[-5.17, -1.36]	<.001
DE vs RC (pre-election)	-3.35	0.97	[-5.25, -1.45]	<.001
RC change (post-election)	-2.02	0.60	[-3.20, -0.84]	<.001
Login count	-0.04	0.01	[-0.06, -0.02]	<.001
EB x Post-election	2.24	0.85	[0.58, 3.90]	0.008
DE x Post-election	1.50	0.85	[-0.16, 3.16]	0.077

Table S76. Effect of condition X pre/post election period on descriptive norm accuracy, controlling for login counts, political posts.

Descriptive accuracy - Republican posts				
Term	Coefficient	SE	95% CI	p-value
RC baseline (pre-election)	43.73	0.77	[42.22, 45.25]	<.001
EB vs RC (pre-election)	-1.71	1.09	[-3.84, 0.41]	0.005
DE vs RC (pre-election)	-2.25	1.32	[-4.85, 0.34]	0.089
RC change (post-election)	-0.62	0.65	[-1.89, 0.65]	0.342
Login count	-0.03	0.01	[-0.05, -0.01]	0.005
EB x Post-election	2.53	0.91	[0.72, 4.31]	0.006
DE x Post-election	1.74	0.91	[-0.05, 3.53]	0.056

Table S77. Effect of condition X pre/post election period on descriptive norm accuracy, controlling for login counts, Republican posts.

Descriptive accuracy - Democratic posts				
Term	Coefficient	SE	95% CI	p-value
RC baseline (pre-election)	9.50	0.78	[7.95, 11.03]	<.001
EB vs RC (pre-election)	-1.41	1.10	[-3.58, 0.75]	0.200
DE vs RC (pre-election)	-2.67	1.10	[-4.83, -0.52]	0.015
RC change (post-election)	-4.21	0.62	[-5.42, -2.99]	<.001
Login count	-0.05	0.01	[-0.07, -0.03]	<.001
EB x Post-election	1.58	0.88	[-0.14, 3.30]	0.071
DE x Post-election	1.91	0.87	[0.20, 3.62]	0.029

Table S78. Effect of condition X pre/post election period on descriptive norm accuracy, controlling for login counts, Democratic posts.

Descriptive accuracy - Toxic posts				
Term	Coefficient	SE	95% CI	p-value
RC baseline (pre-election)	32.77	0.94	[30.93, 34.62]	<.001
EB vs RC (pre-election)	-3.77	1.33	[-6.37, -1.17]	0.005
DE vs RC (pre-election)	-2.25	1.32	[-4.85, 0.34]	0.089
RC change (post-election)	-0.62	0.65	[-1.89, 0.65]	0.342
Login count	-0.03	0.01	[-0.05, -0.01]	0.005
EB x Post-election	2.52	0.91	[0.72, 4.31]	0.006
DE x Post-election	1.74	0.91	[-0.05, 3.53]	0.056

Table S79. Effect of condition X pre/post election period on descriptive norm accuracy, controlling for login counts, Toxic posts.

Descriptive accuracy - Emotion posts				
Term	Coefficient	SE	95% CI	p-value
RC baseline (pre-election)	-11.64	0.66	[-12.94, -10.34]	<.001
EB vs RC (pre-election)	-1.10	0.94	[-2.93, 0.74]	0.241
DE vs RC (pre-election)	-0.64	0.93	[-2.47, 1.19]	0.492
RC change (post-election)	2.70	0.57	[1.59, 3.81]	<.001
Login count	-0.03	0.01	[-0.05, -0.01]	<.001
EB x Post-election	-0.58	0.80	[-2.14, 0.99]	0.469
DE x Post-election	-2.54	0.80	[-4.10, -0.98]	0.001

Table S80. Effect of condition X pre/post election period on descriptive norm accuracy, controlling for login counts, Emotion posts.

Prescriptive norm accuracy				
Term	Coefficient	SE	95% CI	p-value
RC baseline (pre-election)	2.05	1.11	[-0.11, 4.22]	0.063
EB vs RC (pre-election)	-5.23	1.56	[-8.28, -2.18]	<.001
DE vs RC (pre-election)	-1.80	1.55	[-4.84, 1.24]	0.247
RC change (post-election)	-0.47	0.63	[-1.71, 0.76]	0.453
Login count	-0.04	0.01	[-0.06, -0.01]	<.001
EB x Post-election	0.50	0.89	[-1.25, 2.25]	0.575
DE x Post-election	0.18	0.89	[-1.56, 1.93]	0.836

Table S81. Effect of condition X pre/post election period on prescriptive norm accuracy of posting toxic political content, controlling for login counts.

We also ran models estimating the impact of condition on prescriptive norm accuracy that adjusted for prescriptive norm perceptions users reported during the intake phase of our study.

Prescriptive norm accuracy - pre/post adjusted for baseline prescriptive norm judgments (other)				
Term	Coefficient	SE	95% CI	p-value
RC baseline (pre-election)	8.08	0.84	[6.43, 9.73]	<.001
EB vs RC (pre-election)	-4.21	1.16	[-6.49, -1.93]	<.001
DE vs RC (pre-election)	-0.92	1.16	[-3.19, 1.36]	0.430
RC change (post-election)	-0.50	0.63	[-1.74, 0.74]	0.427
Login count	0.51	0.01	[0.48, 0.54]	<.001
EB x Post-election	0.37	0.89	[-1.39, 2.12]	0.682
DE x Post-election	0.11	0.89	[-1.64, 1.86]	0.900

Table S82. Effect of condition X pre/post election period on prescriptive norm accuracy of posting toxic political content, controlling for baseline prescriptive norm judgments (other).

Prescriptive norm accuracy - pre/post adjusted for baseline prescriptive norm judgments (self)				
Term	Coefficient	SE	95% CI	p-value
RC baseline (pre-election)	7.79	0.85	[6.12, 9.47]	<.001
EB vs RC (pre-election)	-3.58	1.18	[-5.89, -1.27]	0.002
DE vs RC (pre-election)	-1.02	1.18	[-3.32, 1.29]	0.388
RC change (post-election)	-0.51	0.63	[-1.75, 0.73]	0.418
Login count	0.49	0.01	[0.46, 0.52]	<.001
EB x Post-election	0.38	0.89	[-1.36, 2.13]	0.668
DE x Post-election	0.10	0.89	[-1.64, 1.85]	0.907

Table S83. Effect of condition X pre/post election period on prescriptive norm accuracy of posting toxic political content, controlling for baseline prescriptive norm judgments (self).

7.7.1 Heterogeneity by Degree of IME Removal

It is possible that null or weak effects on social norm perceptions could reflect heterogeneity in the strength of the algorithmic intervention, rather than the absence of any meaningful impact. In particular, if diversified extremity (DE) feeds varied in how much IME content they actually removed relative to engagement-based (EB) feeds, then any downstream effects on norm perceptions might be concentrated among participants whose feeds experienced larger changes in IME exposure.

To address this possibility, we conducted a heterogeneity analysis focusing on descriptive norm accuracy for IME content, using an aggregate IME accuracy measure that averaged political, partisan, toxic, and emotional accuracy scores. We operationalized intervention strength as the amount of IME content removed from participants' feeds relative to the engagement-based algorithm, computed at the participant-week level. Positive values indicate that a participant's feed contained less IME content than the contemporaneous engagement-based feed baseline.

Importantly, because this "IME removed" measure is defined relative to engagement-based feeds, it is conceptually meaningful only for contrasts involving the diversified extremity condition. Accordingly, we restricted this heterogeneity analysis to the DE vs. EB comparison, avoiding counterfactual contrasts for conditions that do not plausibly vary on the moderator.

Analytic approach. We estimated a linear mixed-effects model predicting aggregated IME descriptive norm accuracy from algorithm condition (DE vs. EB), pre- vs. post-election period, the centered IME-removed measure, and their interactions, with random intercepts for participants. To facilitate interpretation, we report marginal DE-vs-EB contrasts evaluated at low (-1 SD) and high ($+1$ SD) levels of IME removal, separately for the pre- and post-election periods.

Results. Across both time periods and levels of IME removal, diversified extremity feeds did not significantly differ from engagement-based feeds in descriptive norm accuracy for IME content. Before the election, at low levels of IME removal, the DE-vs-EB contrast was not significant, $b = 1.32$, $SE = 1.02$, 95% CI $[-0.68, 3.32]$, $p = .196$; the same was true at high levels of IME removal, $b = 0.70$, $SE = 1.01$, 95% CI $[-1.28, 2.68]$, $p = .487$. After the election, no significant differences emerged at either low IME removal, $b = 0.41$, $SE = 1.08$, 95% CI $[-1.70, 2.53]$, $p = .701$, or high IME removal, $b = -0.20$, $SE = 1.10$, 95% CI $[-2.36, 1.95]$, $p = .853$.

Interpretation. These findings indicate that variation in the amount of IME content removed by the diversified extremity algorithm did not meaningfully moderate its effect on descriptive norm accuracy, relative to engagement-based feeds. Within the range of IME removal achieved in this study—corresponding to roughly a 2.5 percentage-point difference in IME exposure between low and high removal—participants did not exhibit more accurate perceptions of IME prevalence when assigned to the diversified extremity feed. Thus, the largely null effects observed for descriptive norm accuracy in RQ2 do not appear to be driven by modest heterogeneity in intervention strength across participants. At the same time, these results do not rule out the possibility that substantially larger or more sustained reductions in IME exposure could produce detectable changes in social norm perceptions.

7.7.2 Controlling for Baseline IME Exposure and Login Frequency

For the RQ2 descriptive norm accuracy results, between-condition differences could partly reflect differences in the *theoretical opportunity to be exposed to IME content* (i.e., week-specific baseline IME prevalence on Bluesky) and/or differences in *actual exposure intensity* driven by how often participants logged in. To address this concern, we re-estimated the descriptive norm accuracy model for the aggregated IME outcome while controlling for (a) baseline IME exposure and (b) login

frequency. This provides a robustness check that condition effects on descriptive norm accuracy are not attributable to participants being exposed to systematically more (or less) IME content due to platform-level shifts or differential usage.

Operationalization of controls. Baseline IME exposure was computed as a weekly aggregate (0–100 scale) from the Bluesky firehose baseline, averaging political, partisan (left/right), toxic, and emotional baseline prevalence for each wave-by-week cell. Login frequency was measured as the number of logins per participant-week and entered as a centered covariate. We then estimated a mixed-effects model predicting aggregated descriptive norm accuracy for IME content from algorithm condition (RC, EB, DE), pre/post-election period, and their interaction, with the baseline and login controls and participant random intercepts. For interpretability, we report model-implied contrasts with both controls held at their means.

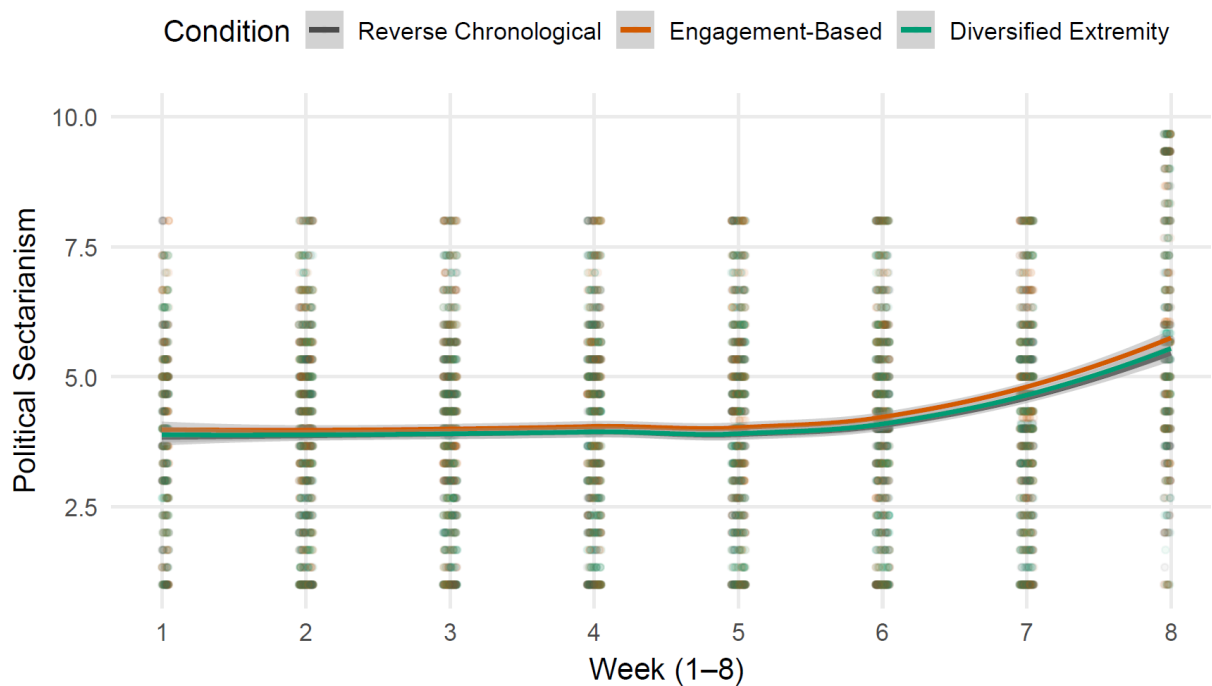
Results. Controlling for baseline IME exposure and login frequency, engagement-based feeds (EB) yielded lower descriptive norm accuracy than reverse-chronological feeds (RC) before the election, $b = -2.25$, $SE = 0.92$, 95% CI $[-4.04, -0.45]$, $p = .014$. After the election, the EB vs. RC difference was not significant, $b = -0.61$, $SE = 1.00$, 95% CI $[-2.57, 1.35]$, $p = .541$.

Diversified extremity (DE) did not significantly differ from engagement-based feeds (EB) in descriptive norm accuracy in either period. Before the election, the DE vs. EB contrast was not significant, $b = 0.22$, $SE = 0.91$, 95% CI $[-1.57, 2.00]$, $p = .812$; after the election, the DE vs. EB contrast was also not significant, $b = -0.64$, $SE = 0.99$, 95% CI $[-2.59, 1.30]$, $p = .517$.

Interpretation. The pattern of results closely mirrors the primary RQ2 findings: engagement-based feeds reduced descriptive norm accuracy for the aggregated IME outcome before the election but not after, and diversified extremity feeds did not significantly improve accuracy relative to engagement-based feeds. Importantly, these conclusions hold even after accounting for both the platform-level baseline prevalence of IME content and participants' login intensity, suggesting that the observed condition differences are not explained by differential opportunity for exposure or differential usage.

7.8 Exploratory models (political sectarianism)

Political Sectarianism Over Weeks by Condition



RQ2: Sectarianism — Pre/Post model

Entries are β (SE). Bold indicates Holm–Bonferroni-adjusted $p < .05$ (exploratory outcome).

	EB – RC	DE – RC	EB – DE
Pre-election	0.122 (0.103)	0.026 (0.103)	0.096 (0.103)
Post-election	0.250 (0.111)	0.071 (0.111)	0.179 (0.111)

Table S84: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate higher perceived political sectarianism in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Sectarianism — Base model

Entries are β (SE). Bold indicates Holm–Bonferroni-adjusted $p < .05$ (exploratory outcome).

EB – RC	DE – RC	EB – DE
0.156 (0.102)	0.035 (0.102)	0.121 (0.101)

Table S85: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate higher perceived political sectarianism in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

7.9 Exploratory models (attitude extremity)

Across all six attitude domains and the aggregate measure of extremity, differences between conditions were consistently small and non-significant. For abortion, EB participants were slightly less extreme than RC both pre-election (-0.09 points, 95% CI $[-1.76, 1.57]$, $p = 0.911$) and post-election (-0.68 , $[-2.47, 1.10]$, $p = 0.453$), while DE tended to be marginally more extreme than EB (pre: $+0.69$, $[-0.97, 2.34]$, $p = 0.416$; post: $+0.92$, $[-0.85, 2.70]$, $p = 0.309$). For immigration, EB leaned somewhat more extreme than RC (pre: $+1.02$, $[-0.60, 2.64]$, $p = 0.218$; post: $+0.96$, $[-0.76, 2.68]$, $p = 0.275$), while DE was essentially indistinguishable from EB (pre: $+0.31$, $[-1.30, 1.93]$, $p = 0.702$; post: $+0.10$, $[-1.62, 1.81]$, $p = 0.912$). For gun control, EB again appeared somewhat higher than RC (pre: $+1.09$, $[-0.49, 2.68]$, $p = 0.176$; post: $+0.83$, $[-0.85, 2.50]$, $p = 0.333$), while DE vs. EB contrasts hovered near zero (pre: $+0.11$, $[-1.46, 1.69]$, $p = 0.888$; post: $+0.15$, $[-1.51, 1.82]$, $p = 0.857$). For environmental attitudes, EB was slightly more extreme than RC (pre: $+0.52$, $[-0.97, 2.02]$, $p = 0.493$; post: $+0.54$, $[-1.04, 2.13]$, $p = 0.500$), and DE did not differ from EB (pre: -0.07 , $[-1.55, 1.42]$, $p = 0.929$; post: -0.26 , $[-1.84, 1.31]$, $p = 0.743$). For progressive taxation, EB and RC did not differ (pre: -0.17 , $[-1.91, 1.57]$, $p = 0.847$; post: -0.10 , $[-1.92, 1.72]$, $p = 0.915$), while DE was modestly higher than EB, especially post-election ($+0.99$, $[-0.82, 2.79]$, $p = 0.285$). For universal healthcare, EB was somewhat higher than RC (pre: $+0.43$, $[-1.14, 1.99]$, $p = 0.593$; post: $+0.66$, $[-0.99, 2.32]$, $p = 0.432$), while DE tended to be lower than EB (pre: -0.16 , $[-1.71, 1.40]$, $p = 0.841$; post: -0.79 , $[-2.44, 0.86]$, $p = 0.346$). Finally, analyses of the aggregate extremity measure showed EB modestly higher than RC (pre: $+0.46$, $[-0.84, 1.77]$, $p = 0.485$; post: $+0.37$, $[-0.98, 1.72]$, $p = 0.589$), while DE vs. EB contrasts were negligible (pre: $+0.21$, $[-1.09, 1.50]$, $p = 0.755$; post: $+0.18$, $[-1.16, 1.53]$, $p = 0.788$). Taken together, estimated effects were consistently small (generally within ± 1 point on the 0–100 extremity scale), confidence intervals all included zero, and no contrasts reached significance, indicating that attitude extremity remained stable across conditions and over time.

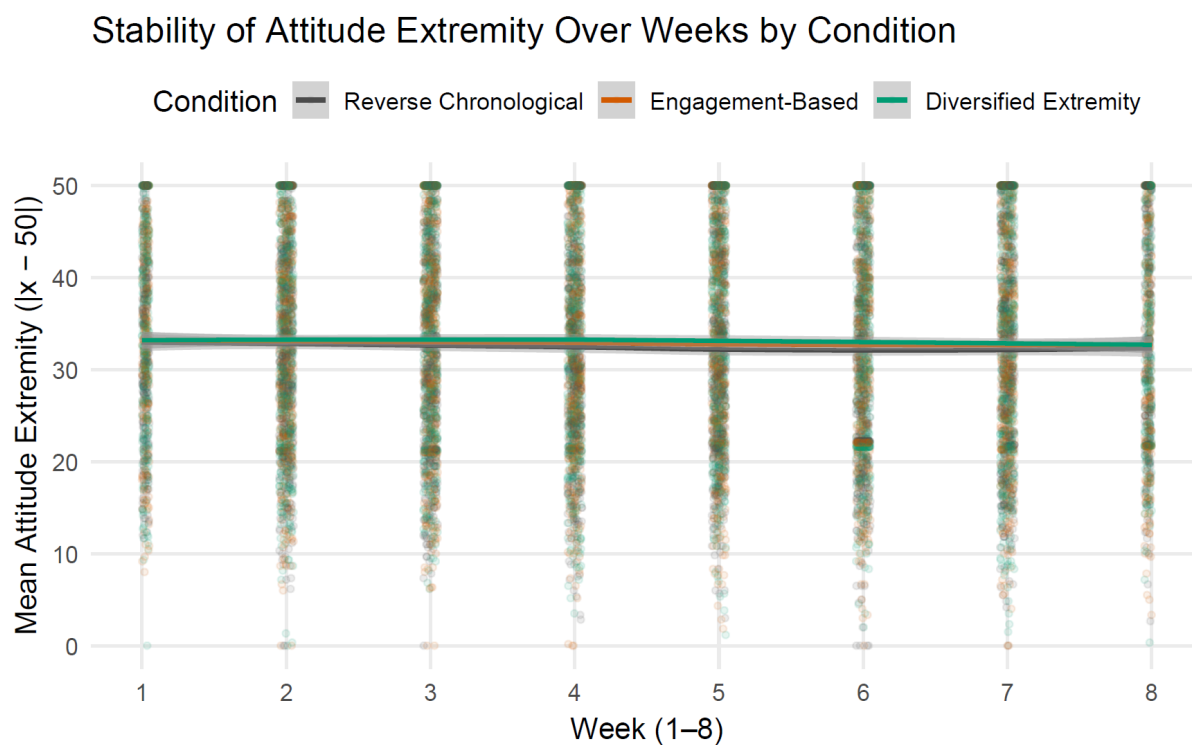


Fig. S44. Mean attitude extremity over the 8-week study period by condition. Greater values indicate greater attitude extremity.

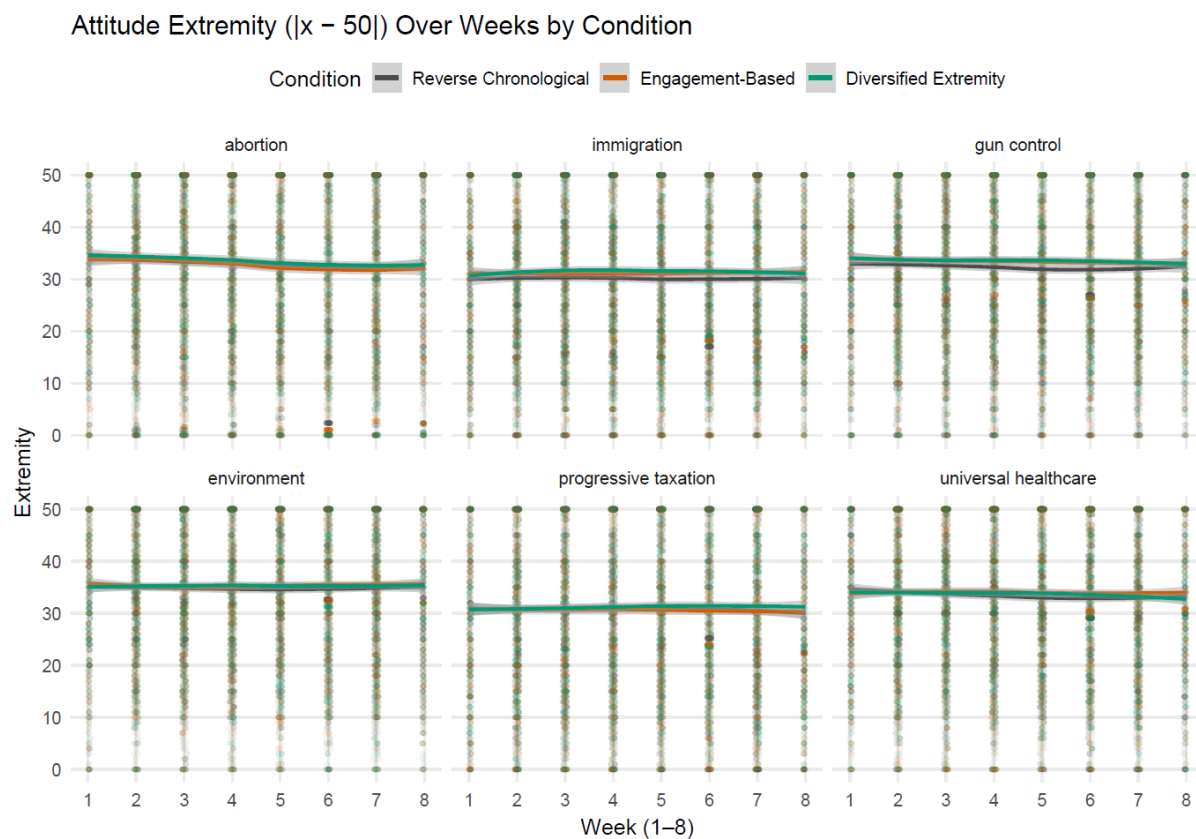


Fig. S45. Attitude extremity over the 8-week study period by condition and each issue. Greater values indicate greater attitude extremity.

RQ2: Attitude Extremity ($|x - 50|$) — Pre/Post model

Entries are β (SE). Bold indicates Holm–Bonferroni-adjusted $p < .05$ (exploratory outcome).

	EB – RC	DE – RC	EB – DE
Pre-election	0.465 (0.666)	0.671 (0.665)	-0.207 (0.662)
Post-election	0.372 (0.689)	0.557 (0.687)	-0.184 (0.685)

Table S86: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Positive values indicate higher perceived attitude extremity in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ2: Attitude Extremity ($|x - 50|$) — Base model

Entries are β (SE). Bold indicates Holm–Bonferroni-adjusted $p < .05$ (exploratory outcome).

EB – RC	DE – RC	EB – DE
0.440 (0.662)	0.640 (0.660)	-0.200 (0.658)

Table S87: Estimates represent condition effects from mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Positive values indicate higher perceived attitude extremity in the first condition relative to the second condition in each contrast. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

8.0 RQ3 supplementary information

8.1 Descriptive statistics and correlations

8.2 Conditional models

RQ3, H1

Our first hypothesis (RQ3, H1) was that engagement-based algorithm feeds would increase engagement with intergroup, moral, emotional and toxic content compared to reverse-chronological feeds. We report marginal effects of engagement-based vs. reverse-chronological feeds below. Fully tabulated results are displayed in SI, Section 8.4.

Intergroup content. Consistent with RQ3, H1, engagement-based feeds significantly increased engagement with intergroup content compared to reverse-chronological feeds (0.51 percentage points higher, $p = .012$).

Moral content. Contrary to H1, engagement-based feeds did not significantly increase engagement with moral content compared to reverse-chronological feeds (0.29 points lower, $p = .133$).

Negative valence content. Inconsistent with H1, engagement-based feeds actually reduced engagement with negative-valence content compared to reverse-chronological feeds (4.39 points lower, $p < .001$).

Positive valence content. Engagement-based feeds did not significantly alter engagement with positive-valence content relative to reverse-chronological feeds (0.80 points lower, $p = .516$). Diversified extremity feeds also did not differ from reverse-chronological feeds (1.33 points higher, $p = .290$).

Sociopolitical content. No significant differences emerged between engagement-based and reverse-chronological feeds in engagement with sociopolitical content (1.64 points lower, $p = .392$).

Toxic content. Inconsistent with H1, engagement-based feeds significantly reduced engagement with toxic content compared to reverse-chronological feeds (1.56 points lower, $p = .009$).

Summary (RQ3, H1). Engagement-based feeds increased engagement with intergroup content compared to reverse-chronological feeds, but reduced engagement with negative-valence and toxic content, and did not significantly alter engagement with moral, positive-valence, or sociopolitical content. Thus, RQ3, H1 received mixed support.

RQ3, H2

Our second hypothesis (RQ3, H2) predicted that diversified extremity algorithm feeds would reduce engagement with intergroup, moral, and toxic content compared to engagement-based feeds. Fully tabulated results are displayed in SI, Section 8.4.

Intergroup content. Diversified extremity feeds did not significantly reduce engagement with intergroup content relative to engagement-based feeds (0.23 points lower, $p = .283$).

Moral content. Diversified extremity feeds yielded somewhat higher engagement with moral content than engagement-based feeds (0.31 points higher), but this difference was not significant ($p = .127$).

Negative emotional content. Diversified extremity feeds did not significantly differ from

engagement-based feeds in engagement with negative-valence content (0.45 points lower, $p = .718$).

Positive emotional content. Diversified extremity feeds showed marginally higher engagement with positive-valence content compared to engagement-based feeds (2.13 points higher), but this effect only approached significance ($p = .093$).

Sociopolitical content. No differences emerged between diversified extremity and engagement-based feeds in engagement with sociopolitical content (0.08 points higher, $p = .968$).

Toxic content. Diversified extremity feeds did not significantly reduce engagement with toxic content compared to engagement-based feeds (0.67 points lower, $p = .270$).

Summary (RQ3, H2). Contrary to RQ3, H2, diversified extremity feeds did not reduce engagement with intergroup, moral, or toxic content relative to engagement-based feeds.

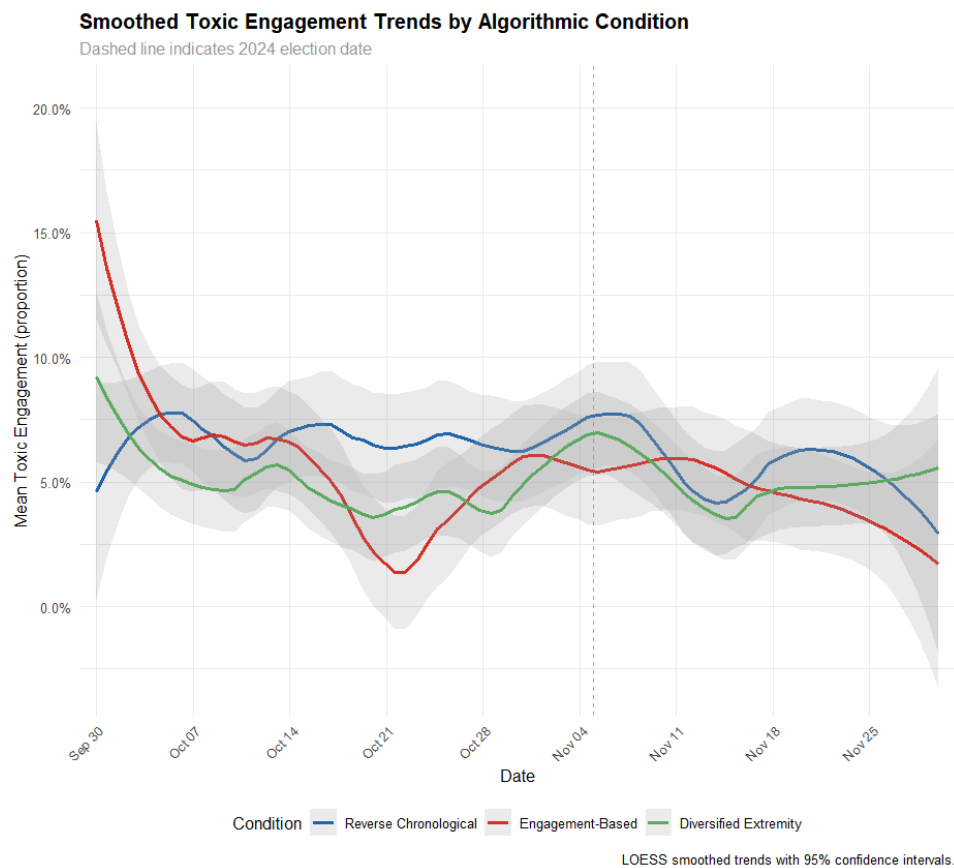


Fig. S46. Smooth toxic engagement trends by condition. Greater values indicate a greater percentage of toxic content with which a participant engaged.

8.3 Model fit comparisons for RQ3

Model fit: Base vs Pre/Post (winners per outcome)

outcome	AIC_base	AIC_prepost	AIC winner	BIC_base	BIC_prepost	BIC winner
intergroup	3871.9	3912.3	base	3909.6	3978.3	base
moral	4480.7	4499.4	base	4518.4	4565.5	base
sociopolitical	14892.2	15020.7	base	14929.9	15086.7	base
toxic	10222.3	10323.2	base	10260.0	10389.2	base
valence_negative	14609.3	14715.3	base	14647.0	14781.4	base
valence_positive	17312.1	17462.8	base	17349.8	17528.9	base

Table S88. Model fit comparisons for each outcome for research question 3.

8.4 Fully tabulated results (base models)

RQ3: Intergroup engagement (any) — Base model		
Entries are OR (SE for log-odds contrast). Bold indicates $p < .05$.		
EB – RC	DE – RC	EB – DE
1.012 (0.465)	1.143 (0.451)	0.886 (0.457)

Table S89: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Odds ratios greater than 1 indicate higher odds of engagement in the first condition relative to the second condition in each contrast. Engagement is measured as any occurrence of like/share/comment/post interactions. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ3: Moral engagement (any) — Base model

Entries are OR (SE for log-odds contrast). Bold indicates $p < .05$.

EB – RC	DE – RC	EB – DE
1.023 (0.477)	0.857 (0.486)	1.194 (0.487)

Table S90: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Odds ratios greater than 1 indicate higher odds of engagement in the first condition relative to the second condition in each contrast. Engagement is measured as any occurrence of like/share/comment/post interactions. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ3: Negative emotion engagement (any) — Base model		
Entries are OR (SE for log-odds contrast). Bold indicates $p < .05$.		
EB – RC	DE – RC	EB – DE
0.948 (0.419)	0.890 (0.408)	1.065 (0.423)

Table S91: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Odds ratios greater than 1 indicate higher odds of engagement in the first condition relative to the second condition in each contrast. Engagement is measured as any occurrence of like/share/comment/post interactions. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ3: Positive emotion engagement (any) — Base model		
Entries are OR (SE for log-odds contrast). Bold indicates $p < .05$.		
EB – RC	DE – RC	EB – DE
0.998 (0.375)	0.928 (0.386)	1.076 (0.405)

Table S92: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Odds ratios greater than 1 indicate higher odds of engagement in the first condition relative to the second condition in each contrast. Engagement is measured as any occurrence of like/share/comment/post interactions. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ3: Sociopolitical engagement (any) — Base model

Entries are OR (SE for log-odds contrast). Bold indicates $p < .05$.

EB – RC	DE – RC	EB – DE
0.956 (0.418)	0.891 (0.478)	1.073 (0.450)

Table S93: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Odds ratios greater than 1 indicate higher odds of engagement in the first condition relative to the second condition in each contrast. Engagement is measured as any occurrence of like/share/comment/post interactions. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ3: Toxic engagement (any) — Base model		
Entries are OR (SE for log-odds contrast). Bold indicates $p < .05$.		
EB – RC	DE – RC	EB – DE
0.902 (0.000)	0.812 (0.000)	1.110 (0.000)

Table S94: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period. Odds ratios greater than 1 indicate higher odds of engagement in the first condition relative to the second condition in each contrast. Engagement is measured as any occurrence of like/share/comment/post interactions. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

8.5 Fully tabulated results (pre/post models)

RQ3: Intergroup engagement (any) — Pre/Post model			
Entries are OR (SE for log-odds contrast). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	1.158 (0.481)	1.316 (0.466)	0.880 (0.464)
Post-election	0.694 (0.526)	0.768 (0.502)	0.904 (0.516)

Table S95: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Odds ratios greater than 1 indicate higher odds of engagement in the first condition relative to the second condition in each contrast. Engagement is measured as any occurrence of like/share/comment/post interactions. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ3: Moral engagement (any) — Pre/Post model			
Entries are OR (SE for log-odds contrast). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	0.863 (0.476)	0.721 (0.491)	1.196 (0.498)
Post-election	1.255 (0.483)	1.053 (0.498)	1.192 (0.498)

Table S96: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Odds ratios greater than 1 indicate higher odds of engagement in the first condition relative to the second condition in each contrast. Engagement is measured as any occurrence of like/share/comment/post interactions. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ3: Negative emotion engagement (any) — Pre/Post model

Entries are OR (SE for log-odds contrast). Bold indicates $p < .05$.

	EB – RC	DE – RC	EB – DE
Pre-election	0.877 (0.398)	0.836 (0.409)	1.049 (0.425)
Post-election	1.096 (0.404)	1.001 (0.414)	1.095 (0.429)

Table S97: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Odds ratios greater than 1 indicate higher odds of engagement in the first condition relative to the second condition in each contrast. Engagement is measured as any occurrence of like/share/comment/post interactions. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ3: Positive emotion engagement (any) — Pre/Post model			
Entries are OR (SE for log-odds contrast). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	1.037 (0.407)	0.932 (0.416)	1.112 (0.417)
Post-election	0.933 (0.411)	0.921 (0.419)	1.013 (0.420)

Table S98: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Odds ratios greater than 1 indicate higher odds of engagement in the first condition relative to the second condition in each contrast. Engagement is measured as any occurrence of like/share/comment/post interactions. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ3: Sociopolitical engagement (any) — Pre/Post model			
Entries are OR (SE for log-odds contrast). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	0.897 (0.410)	0.832 (0.424)	1.079 (0.433)
Post-election	1.065 (0.415)	1.001 (0.428)	1.064 (0.437)

Table S99: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Odds ratios greater than 1 indicate higher odds of engagement in the first condition relative to the second condition in each contrast. Engagement is measured as any occurrence of like/share/comment/post interactions. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ3: Toxic engagement (any) — Pre/Post model			
Entries are OR (SE for log-odds contrast). Bold indicates $p < .05$.			
	EB – RC	DE – RC	EB – DE
Pre-election	0.861 (0.447)	0.772 (0.445)	1.116 (0.471)
Post-election	1.002 (0.453)	0.910 (0.452)	1.101 (0.476)

Table S100: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group for the time period before the 2024 U.S. Presidential election, and the time period after. Odds ratios greater than 1 indicate higher odds of engagement in the first condition relative to the second condition in each contrast. Engagement is measured as any occurrence of like/share/comment/post interactions. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

8.6 Fully tabulated results (exposure models)

RQ3: Intergroup engagement (any) — Exposure model		
Entries are OR (SE for log-odds contrast). Bold indicates $p < .05$.		
EB – RC	DE – RC	EB – DE
0.760 (0.000)	0.629 (0.000)	1.207 (0.001)

Table S101: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period, controlling for daily exposure to each content type. Odds ratios greater than 1 indicate higher odds of engagement in the first condition relative to the second condition in each contrast. Engagement is measured as any occurrence of like/share/comment/post interactions. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ3: Moral engagement (any) — Exposure model

Entries are OR (SE for log-odds contrast). Bold indicates $p < .05$.

EB – RC	DE – RC	EB – DE
0.914 (0.486)	0.858 (0.502)	1.066 (0.499)

Table S102: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period, controlling for daily exposure to each content type. Odds ratios greater than 1 indicate higher odds of engagement in the first condition relative to the second condition in each contrast. Engagement is measured as any occurrence of like/share/comment/post interactions. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ3: Negative emotion engagement (any) — Exposure model

Entries are OR (SE for log-odds contrast). Bold indicates $p < .05$.

EB – RC	DE – RC	EB – DE
0.892 (0.434)	0.871 (0.436)	1.024 (0.438)

Table S103: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period, controlling for daily exposure to each content type. Odds ratios greater than 1 indicate higher odds of engagement in the first condition relative to the second condition in each contrast. Engagement is measured as any occurrence of like/share/comment/post interactions. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ3: Positive emotion engagement (any) — Exposure model

Entries are OR (SE for log-odds contrast). Bold indicates $p < .05$.

EB – RC	DE – RC	EB – DE
0.937 (0.421)	0.885 (0.423)	1.058 (0.422)

Table S104: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period, controlling for daily exposure to each content type. Odds ratios greater than 1 indicate higher odds of engagement in the first condition relative to the second condition in each contrast. Engagement is measured as any occurrence of like/share/comment/post interactions. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ3: Sociopolitical engagement (any) — Exposure model		
Entries are OR (SE for log-odds contrast). Bold indicates $p < .05$.		
EB – RC	DE – RC	EB – DE
0.801 (0.440)	0.789 (0.444)	1.015 (0.445)

Table S105: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period, controlling for daily exposure to each content type. Odds ratios greater than 1 indicate higher odds of engagement in the first condition relative to the second condition in each contrast. Engagement is measured as any occurrence of like/share/comment/post interactions. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

RQ3: Toxic engagement (any) — Exposure model

Entries are OR (SE for log-odds contrast). Bold indicates $p < .05$.

EB – RC	DE – RC	EB – DE
0.782 (0.469)	0.767 (0.475)	1.019 (0.480)

Table S106: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period, controlling for daily exposure to each content type. Odds ratios greater than 1 indicate higher odds of engagement in the first condition relative to the second condition in each contrast. Engagement is measured as any occurrence of like/share/comment/post interactions. RC = Reverse Chronological, EB = Engagement-Based, DE = Diversified Extremity.

8.7 Precision analysis of null engagement effects

To contextualize the null effects reported for RQ3, we quantified the precision of the estimated algorithm-condition effects by identifying what magnitudes of engagement change are incompatible with the data. For each outcome and contrast, we began with the mixed-effects logistic regression models reported in SI Section 8.4, which estimate condition effects on the log-odds scale. Using model-based marginal means, we first computed the baseline probability of engagement under the relevant comparison condition (e.g., reverse-chronological for EB–RC contrasts, engagement-based for DE–EB contrasts).

We then translated the 95% confidence intervals for each condition contrast from the log-odds scale to the probability scale. Specifically, for a given contrast with an estimated log-odds coefficient (β) and a 95% confidence interval with lower and upper bounds (β_L , β_U), we computed the implied engagement probabilities by adding each quantity to the logit of the baseline probability and then applying the inverse-logit transformation. The resulting values define a confidence interval for the change in engagement probability (Δp). From this interval, we derived the largest proportional increase and decrease in engagement likelihood, relative to the baseline probability, that are inconsistent with the data at the 95% confidence level. From this interval, we derived the largest proportional increase and decrease in engagement likelihood (relative to the baseline probability) that are inconsistent with the data at the 95% confidence level.

Because engagement probabilities were extremely small, we report these bounds as percent changes relative to baseline rates, rather than as absolute probability differences. These quantities correspond to the “ruled-out” effect sizes reported in Table S107. For toxic engagement, standard mixed-effects models exhibited convergence and identifiability issues due to extreme sparsity; therefore, we estimated confidence intervals using handle-level bootstrap resampling and applied the same probability-scale transformation to the resulting bootstrap intervals.

Together, this approach provides an interpretable measure of precision that complements null hypothesis tests by clarifying which magnitudes of engagement change are incompatible with the observed data.

Outcome	Contrast	Baseline	Ruled-out decrease (% of baseline)	Ruled-out increase (% of baseline)
Intergroup engagement (any)	EB – RC	RC	59%	152%
Intergroup engagement (any)	DE – EB	EB	54%	176%
Moral engagement (any)	EB – RC	RC	60%	161%
Moral engagement (any)	DE – EB	EB	68%	117%
Negative emotion engagement (any)	EB – RC	RC	58%	116%
Negative emotion engagement (any)	DE – EB	EB	59%	115%
Positive emotion engagement (any)	EB – RC	RC	52%	108%
Positive emotion engagement (any)	DE – EB	EB	58%	106%
Sociopolitical engagement (any)	EB – RC	RC	58%	117%
Sociopolitical engagement (any)	DE – EB	EB	61%	125%
Toxic engagement (any)	EB – RC	RC	42%	18%
Toxic engagement (any)	DE – EB	EB	35%	12%

Table S107. Precision bounds for null effects on engagement outcomes (RQ3). Values show the largest proportional decrease and the largest proportional increase in engagement likelihood (each expressed as a percentage of the baseline engagement rate in the relevant comparison condition) that are inconsistent with the data at the 95% confidence level. Baseline indicates the condition used as the reference rate for each contrast (RC for EB – RC; EB for DE – EB). For toxic engagement, confidence intervals were obtained via bootstrap due to convergence issues in the base mixed-effects models.

8.8 Condition effects of on general engagement

Weekly login activity did not differ meaningfully by algorithm condition in either time period. Before the election, engagement-based feeds did not significantly change weekly logins relative to reverse-chronological feeds, $b = -0.76$, $SE = 1.29$, $p = .555$, and diversified extremity feeds did not differ from engagement-based feeds, $b = 0.17$, $SE = 1.28$, $p = .897$. After the election, engagement-based feeds again did not significantly differ from reverse-chronological feeds, $b = 0.63$, $SE = 1.38$, $p = .649$, and diversified extremity feeds did not differ from engagement-based feeds, $b = -0.54$, $SE = 1.37$, $p = .696$. Overall, these results suggest the interventions altered feed composition (RQ1) without producing detectable changes in general platform usage intensity, at least defined as weekly login count.

9.0 Additional attrition analyses and exclusion robustness tests

9.1 Analyzing results across various exclusion thresholds

9.1.1 RQ1 results

Robustness across inclusion thresholds: Intergroup content						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1824)	0.00408	0.00270	0.00138	-0.00218	-0.00217	-0.00001
≥ 2 weeks (n=1766)	0.00408	0.00271	0.00137	-0.00218	-0.00217	-0.00001
≥ 3 weeks (n=1728)	0.00408	0.00271	0.00137	-0.00219	-0.00218	-0.00001
≥ 4 weeks (n=1636)	0.00407	0.00271	0.00137	-0.00218	-0.00218	0.00000
≥ 5 weeks (n=1475)	0.00406	0.00269	0.00137	-0.00217	-0.00214	-0.00003

Table S108. Effect of treatment condition on content using progressively relaxed exclusion criteria (intergroup content). The table displays the coefficients for each treatment (diversified extremity algorithm vs reverse chronological, DE - RC; engagement-based algorithm vs. reverse chronological, EB - RC) at different levels of exclusion criteria. Valid weeks were defined as the number of weeks participants logged in and filled out a survey. Bold values represent significant values at the $p < .05$.

Robustness across inclusion thresholds: Moral content						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1824)	-0.00290	0.00002	-0.00292	0.00251	0.00322	-0.00071
≥ 2 weeks (n=1766)	-0.00290	0.00002	-0.00292	0.00250	0.00321	-0.00071
≥ 3 weeks (n=1728)	-0.00291	0.00001	-0.00292	0.00249	0.00320	-0.00072
≥ 4 weeks (n=1636)	-0.00292	0.00001	-0.00293	0.00251	0.00324	-0.00073
≥ 5 weeks (n=1475)	-0.00292	-0.00003	-0.00289	0.00248	0.00322	-0.00073

Table S109. Effect of treatment condition on content using progressively relaxed exclusion criteria (moral content). The table displays the coefficients for each treatment (diversified extremity algorithm vs reverse chronological, DE - RC; engagement-based algorithm vs. reverse chronological, EB - RC) at different levels of exclusion criteria. Valid weeks were defined as the number of weeks participants logged in and filled out a survey. Bold values represent significant values at the $p < .05$.

Robustness across inclusion thresholds: Negative emotion content						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1824)	0.00241	0.01907	-0.01667	0.02425	0.03765	-0.01340
≥ 2 weeks (n=1766)	0.00249	0.01899	-0.01651	0.02405	0.03755	-0.01350
≥ 3 weeks (n=1728)	0.00250	0.01900	-0.01649	0.02397	0.03743	-0.01347
≥ 4 weeks (n=1636)	0.00248	0.01905	-0.01656	0.02406	0.03758	-0.01352
≥ 5 weeks (n=1475)	0.00260	0.01900	-0.01641	0.02384	0.03703	-0.01319

Table S110. Effect of treatment condition on content using progressively relaxed exclusion criteria (negative emotional content). The table displays the coefficients for each treatment (diversified extremity algorithm vs reverse chronological, DE - RC; engagement-based algorithm vs. reverse chronological, EB - RC) at different levels of exclusion criteria. Valid weeks were defined as the number of weeks participants logged in and filled out a survey. Bold values represent significant values at the $p < .05$.

Robustness across inclusion thresholds: Positive emotion content						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1824)	-0.06281	-0.05449	-0.00832	-0.00781	-0.01222	0.00441
≥ 2 weeks (n=1766)	-0.06273	-0.05453	-0.00819	-0.00753	-0.01212	0.00459
≥ 3 weeks (n=1728)	-0.06273	-0.05452	-0.00820	-0.00740	-0.01197	0.00457
≥ 4 weeks (n=1636)	-0.06265	-0.05452	-0.00813	-0.00714	-0.01198	0.00484
≥ 5 weeks (n=1475)	-0.06228	-0.05393	-0.00835	-0.00696	-0.01161	0.00465

Table S111. Effect of treatment condition on content using progressively relaxed exclusion criteria (positive emotional content). The table displays the coefficients for each treatment (diversified extremity algorithm vs reverse chronological, DE - RC; engagement-based algorithm vs. reverse chronological, EB - RC) at different levels of exclusion criteria. Valid weeks were defined as the number of weeks participants logged in and filled out a survey. Bold values represent significant values at the $p < .05$.

Robustness across inclusion thresholds: Toxic content						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1824)	-0.00065	0.00824	-0.00889	0.00658	0.01127	-0.00469
≥ 2 weeks (n=1766)	-0.00063	0.00824	-0.00888	0.00650	0.01125	-0.00476
≥ 3 weeks (n=1728)	-0.00061	0.00827	-0.00888	0.00647	0.01126	-0.00479
≥ 4 weeks (n=1636)	-0.00056	0.00829	-0.00885	0.00651	0.01127	-0.00476
≥ 5 weeks (n=1475)	-0.00053	0.00822	-0.00876	0.00641	0.01092	-0.00451

Table S112. Effect of treatment condition on content using progressively relaxed exclusion criteria (toxic content). The table displays the coefficients for each treatment (diversified extremity algorithm vs reverse chronological, DE - RC; engagement-based algorithm vs. reverse chronological, EB - RC) at different levels of exclusion criteria. Valid weeks were defined as the number of weeks participants logged in and filled out a survey. Bold values represent significant values at the $p < .05$.

Robustness across inclusion thresholds: Political content						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1824)	0.01361	0.04248	-0.02886	0.11032	0.14040	-0.03008
≥ 2 weeks (n=1766)	0.01388	0.04258	-0.02870	0.10986	0.14067	-0.03081
≥ 3 weeks (n=1728)	0.01398	0.04260	-0.02863	0.10951	0.14031	-0.03081
≥ 4 weeks (n=1636)	0.01399	0.04279	-0.02880	0.10994	0.14055	-0.03061
≥ 5 weeks (n=1475)	0.01432	0.04313	-0.02880	0.10913	0.13894	-0.02981

Table S113. Effect of treatment condition on content using progressively relaxed exclusion criteria (political content). The table displays the coefficients for each treatment (diversified extremity algorithm vs reverse chronological, DE - RC; engagement-based algorithm vs. reverse chronological, EB - RC) at different levels of exclusion criteria. Valid weeks were defined as the number of weeks participants logged in and filled out a survey. Bold values represent significant values at the $p < .05$.

Robustness across inclusion thresholds: Outrage content						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1824)	0.03187	0.05015	-0.01828	0.08413	0.11009	-0.02595
≥ 2 weeks (n=1766)	0.03199	0.05021	-0.01822	0.08380	0.11019	-0.02639
≥ 3 weeks (n=1728)	0.03205	0.05026	-0.01821	0.08353	0.10997	-0.02644
≥ 4 weeks (n=1636)	0.03203	0.05031	-0.01828	0.08376	0.11007	-0.02630
≥ 5 weeks (n=1475)	0.03209	0.05021	-0.01811	0.08305	0.10865	-0.02559

Table S114. Effect of treatment condition on content using progressively relaxed exclusion criteria (moral outrage content). The table displays the coefficients for each treatment (diversified extremity algorithm vs reverse chronological, DE - RC; engagement-based algorithm vs. reverse chronological, EB - RC) at different levels of exclusion criteria. Valid weeks were defined as the number of weeks participants logged in and filled out a survey. Bold values represent significant values at the $p < .05$.

Robustness across inclusion thresholds: Constructive content						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1824)	-0.01725	-0.00995	-0.00730	0.00048	0.00714	-0.00666
≥ 2 weeks (n=1766)	-0.01715	-0.00984	-0.00731	0.00050	0.00724	-0.00673
≥ 3 weeks (n=1728)	-0.01714	-0.00982	-0.00732	0.00047	0.00720	-0.00673
≥ 4 weeks (n=1636)	-0.01722	-0.00981	-0.00742	0.00062	0.00734	-0.00672
≥ 5 weeks (n=1475)	-0.01710	-0.00957	-0.00753	0.00055	0.00729	-0.00674

Table S115. Effect of treatment condition on content using progressively relaxed exclusion criteria (constructive content). The table displays the coefficients for each treatment (diversified extremity algorithm vs reverse chronological, DE - RC; engagement-based algorithm vs. reverse chronological, EB - RC) at different levels of exclusion criteria. Valid weeks were defined as the number of weeks participants logged in and filled out a survey. Bold values represent significant values at the $p < .05$.

9.1.2 RQ2 results

Robustness across inclusion thresholds: descriptive_accuracy_political						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1815)	-3.411	-2.696	0.715	-1.606	-0.803	0.803
≥ 2 weeks (n=1760)	-3.388	-2.694	0.694	-1.570	-0.786	0.785
≥ 3 weeks (n=1722)	-3.383	-2.741	0.642	-1.559	-0.786	0.772
≥ 4 weeks (n=1630)	-3.396	-2.812	0.584	-1.669	-0.839	0.830
≥ 5 weeks (n=1470)	-3.212	-3.124	0.088	-1.909	-1.107	0.802

Table S116. Effect of treatment condition on content using progressively relaxed exclusion criteria (political content). The table displays the coefficients for each treatment (diversified extremity algorithm vs reverse chronological, DE - RC; engagement-based algorithm vs. reverse chronological, EB - RC) at different levels of exclusion criteria. Valid weeks were defined as the number of weeks participants logged in and filled out a survey. Bold values represent significant values at the $p < .05$.

Robustness across inclusion thresholds: descriptive_accuracy_republican						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1815)	-1.275	-1.563	-0.288	-0.131	0.069	0.200
≥ 2 weeks (n=1760)	-1.353	-1.670	-0.317	-0.093	0.079	0.172
≥ 3 weeks (n=1722)	-1.274	-1.741	-0.467	-0.068	0.037	0.105
≥ 4 weeks (n=1630)	-1.310	-1.603	-0.293	-0.214	-0.020	0.194
≥ 5 weeks (n=1470)	-0.990	-1.421	-0.431	-0.010	0.662	0.671

Table S117. Effect of treatment condition on content using progressively relaxed exclusion criteria (Right-Leaning content). The table displays the coefficients for each treatment (diversified extremity algorithm vs reverse chronological, DE - RC; engagement-based algorithm vs. reverse chronological, EB - RC) at different levels of exclusion criteria. Valid weeks were defined as the number of weeks participants logged in and filled out a survey. Bold values represent significant values at the $p < .05$.

Robustness across inclusion thresholds: descriptive_accuracy_democratic						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1815)	-2.719	-1.496	1.223	-0.720	0.182	0.902
≥ 2 weeks (n=1760)	-2.909	-1.651	1.258	-0.155	0.192	0.347
≥ 3 weeks (n=1722)	-2.982	-1.794	1.188	-0.131	0.146	0.276
≥ 4 weeks (n=1630)	-2.755	-1.608	1.147	-0.769	0.084	0.854
≥ 5 weeks (n=1470)	-2.676	-1.409	1.267	-0.887	0.017	0.904

Table S118. Effect of treatment condition on content using progressively relaxed exclusion criteria (Left-Leaning content). The table displays the coefficients for each treatment (diversified extremity algorithm vs reverse chronological, DE - RC; engagement-based algorithm vs. reverse chronological, EB - RC) at different levels of exclusion criteria. Valid weeks were defined as the number of weeks participants logged in and filled out a survey. Bold values represent significant values at the $p < .05$.

Robustness across inclusion thresholds: descriptive_accuracy_toxic						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1815)	-2.555	-3.638	-1.083	-0.352	-0.756	-0.404
≥ 2 weeks (n=1760)	-2.590	-3.775	-1.185	-0.345	-0.801	-0.456
≥ 3 weeks (n=1722)	-2.574	-3.928	-1.354	-0.319	-0.764	-0.445
≥ 4 weeks (n=1630)	-2.301	-3.478	-1.177	-0.175	-0.520	-0.346
≥ 5 weeks (n=1470)	-2.144	-3.661	-1.517	-0.703	-1.466	-0.763

Table S119. Effect of treatment condition on content using progressively relaxed exclusion criteria (toxic content). The table displays the coefficients for each treatment (diversified extremity algorithm vs reverse chronological, DE - RC; engagement-based algorithm vs. reverse chronological, EB - RC) at different levels of exclusion criteria. Valid weeks were defined as the number of weeks participants logged in and filled out a survey. Bold values represent significant values at the $p < .05$.

Robustness across inclusion thresholds: descriptive_accuracy_emotion						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1815)	-0.712	-1.134	-0.422	-2.730	-1.555	1.175
≥ 2 weeks (n=1760)	-0.728	-1.254	-0.526	-2.698	-1.550	1.148
≥ 3 weeks (n=1722)	-0.707	-1.329	-0.623	-2.737	-1.655	1.082
≥ 4 weeks (n=1630)	-0.499	-1.168	-0.669	-2.801	-1.655	1.146
≥ 5 weeks (n=1470)	-0.593	-1.043	-0.451	-3.346	-1.852	1.493

Table S120. Effect of treatment condition on content using progressively relaxed exclusion criteria (emotional content). The table displays the coefficients for each treatment (diversified extremity algorithm vs reverse chronological, DE - RC; engagement-based algorithm vs. reverse chronological, EB - RC) at different levels of exclusion criteria. Valid weeks were defined as the number of weeks participants logged in and filled out a survey. Bold values represent significant values at the $p < .05$.

Robustness across inclusion thresholds: prescriptive_norm_accuracy						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1815)	-1.816	-4.991	-3.175	-1.839	-4.566	-2.727
≥ 2 weeks (n=1760)	-1.871	-4.949	-3.078	-1.824	-4.572	-2.748
≥ 3 weeks (n=1722)	-1.773	-4.975	-3.202	-1.832	-4.615	-2.783
≥ 4 weeks (n=1630)	-1.841	-5.014	-3.173	-1.768	-4.519	-2.751
≥ 5 weeks (n=1470)	-1.773	-5.196	-3.422	-1.840	-4.972	-3.132

Table S121. Effect of treatment condition on prescriptive norm accuracy using progressively relaxed exclusion criteria. The table displays the coefficients for each treatment (diversified extremity algorithm vs reverse chronological, DE - RC; engagement-based algorithm vs. reverse chronological, EB - RC) at different levels of exclusion criteria. Valid weeks were defined as the number of weeks participants logged in and filled out a survey. Bold values represent significant values at the $p < .05$.

Robustness across inclusion thresholds: Ingroup dislike						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1714)	-2.760	-2.889	-0.129	-2.734	-3.705	-0.971
≥ 2 weeks (n=1671)	-2.769	-3.078	-0.309	-2.742	-3.875	-1.132
≥ 3 weeks (n=1642)	-2.646	-3.014	-0.367	-2.717	-3.858	-1.141
≥ 4 weeks (n=1585)	-2.836	-3.323	-0.487	-2.447	-3.992	-1.545
≥ 5 weeks (n=1470)	-2.690	-3.068	-0.378	-2.094	-3.548	-1.455

Table S122. Effect of treatment condition on enjoyment of ingroup dislike using progressively relaxed exclusion criteria. The table displays the coefficients for each treatment (diversified extremity algorithm vs reverse chronological, DE - RC; engagement-based algorithm vs. reverse chronological, EB - RC) at different levels of exclusion criteria. Valid weeks were defined as the number of weeks participants logged in and filled out a survey. Bold values represent significant values at the $p < .05$.

Robustness across inclusion thresholds: Outgroup dislike						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1714)	-2.097	-3.130	-1.033	-3.539	-4.212	-0.673
≥ 2 weeks (n=1671)	-2.081	-3.472	-1.391	-3.555	-4.562	-1.007
≥ 3 weeks (n=1642)	-1.832	-3.311	-1.479	-3.281	-4.448	-1.167
≥ 4 weeks (n=1585)	-1.924	-3.309	-1.385	-3.342	-4.522	-1.181
≥ 5 weeks (n=1470)	-1.361	-2.851	-1.491	-2.856	-4.336	-1.480

Table S123. Effect of treatment condition on outgroup dislike using progressively relaxed exclusion criteria. The table displays the coefficients for each treatment (diversified extremity algorithm vs reverse chronological, DE - RC; engagement-based algorithm vs. reverse chronological, EB - RC) at different levels of exclusion criteria. Valid weeks were defined as the number of weeks participants logged in and filled out a survey. Bold values represent significant values at the $p < .05$.

Robustness across inclusion thresholds: enjoy_bs						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1815)	-0.110	0.018	0.128	-0.080	-0.638	-0.558
≥ 2 weeks (n=1760)	-0.126	-0.006	0.121	-0.078	-0.627	-0.549
≥ 3 weeks (n=1722)	-0.131	-0.012	0.119	-0.072	-0.617	-0.544
≥ 4 weeks (n=1630)	-0.161	-0.102	0.059	-0.075	-0.604	-0.529
≥ 5 weeks (n=1470)	-0.129	-0.099	0.031	-0.030	-0.584	-0.554

Table S124. Effect of treatment condition on enjoyment of Bluesky using progressively relaxed exclusion criteria. The table displays the coefficients for each treatment (diversified extremity algorithm vs reverse chronological, DE - RC; engagement-based algorithm vs. reverse chronological, EB - RC) at different levels of exclusion criteria. Valid weeks were defined as the number of weeks participants logged in and filled out a survey. Bold values represent significant values at the $p < .05$.

Robustness across inclusion thresholds: enjoy_content						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1815)	-0.092	-0.043	0.049	-0.578	-0.561	0.018
≥ 2 weeks (n=1760)	-0.091	-0.041	0.050	-0.570	-0.552	0.019
≥ 3 weeks (n=1722)	-0.090	-0.037	0.053	-0.560	-0.541	0.019
≥ 4 weeks (n=1630)	-0.091	-0.036	0.055	-0.549	-0.528	0.021
≥ 5 weeks (n=1470)	-0.082	-0.048	0.034	-0.587	-0.059	0.528

Table S125. Effect of treatment condition on enjoyment of content using progressively relaxed exclusion criteria. The table displays the coefficients for each treatment (diversified extremity algorithm vs reverse chronological, DE - RC; engagement-based algorithm vs. reverse chronological, EB - RC) at different levels of exclusion criteria. Valid weeks were defined as the number of weeks participants logged in and filled out a survey. Bold values represent significant values at the $p < .05$.

Robustness across inclusion thresholds: percent_posts_dem						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1815)	-2.742	-1.511	1.231	-0.689	0.203	0.892
≥ 2 weeks (n=1760)	-2.932	-1.664	1.267	-0.125	0.209	0.334
≥ 3 weeks (n=1722)	-3.008	-1.808	1.200	-0.096	0.165	0.261
≥ 4 weeks (n=1630)	-2.779	-1.624	1.155	-0.737	0.106	0.843
≥ 5 weeks (n=1470)	-2.694	-1.426	1.268	-0.863	0.040	0.903

Table S126. Effect of condition X pre/post election period on descriptive norm perception (percentage) of democratic posts using progressively relaxed exclusion criteria.

Robustness across inclusion thresholds: percent_posts_rep						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1815)	-1.268	-1.558	-0.291	-0.117	0.078	0.195
≥ 2 weeks (n=1760)	-1.346	-1.666	-0.320	-0.079	0.087	0.166
≥ 3 weeks (n=1722)	-1.266	-1.737	-0.471	-0.052	0.046	0.098
≥ 4 weeks (n=1630)	-1.303	-1.599	-0.296	-0.199	-0.010	0.189
≥ 5 weeks (n=1470)	-0.984	-1.416	-0.431	0.002	0.672	0.670

Table S127. Effect of condition X pre/post election period on descriptive norm perception (percentage) of republican posts using progressively relaxed exclusion criteria.

Robustness across inclusion thresholds: percent_posts_pol						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1815)	-3.389	-2.681	0.708	-1.493	-0.728	0.765
≥ 2 weeks (n=1760)	-3.366	-2.682	0.684	-1.458	-0.721	0.737
≥ 3 weeks (n=1722)	-3.359	-2.728	0.631	-1.433	-0.717	0.715
≥ 4 weeks (n=1630)	-3.374	-2.797	0.577	-1.553	-0.761	0.791
≥ 5 weeks (n=1470)	-3.195	-3.109	0.086	-1.821	-1.024	0.797

Table S128. Effect of condition X pre/post election period on descriptive norm perception (percentage) of political posts using progressively relaxed exclusion criteria.

Robustness across inclusion thresholds: percent_posts_emo						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1815)	-0.721	-1.140	-0.419	-2.732	-1.556	1.176
≥ 2 weeks (n=1760)	-0.737	-1.259	-0.522	-2.700	-1.552	1.149
≥ 3 weeks (n=1722)	-0.717	-1.335	-0.618	-2.740	-1.656	1.083
≥ 4 weeks (n=1630)	-0.508	-1.175	-0.666	-2.803	-1.657	1.147
≥ 5 weeks (n=1470)	-0.600	-1.050	-0.450	-3.348	-1.854	1.494

Table S129. Effect of condition X pre/post election period on descriptive norm perception (percentage) of emotional posts using progressively relaxed exclusion criteria.

Robustness across inclusion thresholds: percent_toxic						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1815)	-2.555	-3.638	-1.083	-0.351	-0.755	-0.404
≥ 2 weeks (n=1760)	-2.589	-3.774	-1.185	-0.344	-0.800	-0.456
≥ 3 weeks (n=1722)	-2.574	-3.928	-1.354	-0.318	-0.763	-0.446
≥ 4 weeks (n=1630)	-2.300	-3.477	-1.177	-0.174	-0.520	-0.346
≥ 5 weeks (n=1470)	-2.144	-3.661	-1.517	-0.702	-1.465	-0.763

Table S130. Effect of condition X pre/post election period on descriptive norm perception (percentage) of toxic posts using progressively relaxed exclusion criteria.

Robustness across inclusion thresholds: others_appropriate						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1815)	-1.802	-4.982	-3.180	-1.871	-4.587	-2.716
≥ 2 weeks (n=1760)	-1.857	-4.941	-3.084	-1.856	-4.590	-2.735
≥ 3 weeks (n=1722)	-1.758	-4.967	-3.209	-1.867	-4.634	-2.767
≥ 4 weeks (n=1630)	-1.825	-5.003	-3.178	-1.802	-4.541	-2.740
≥ 5 weeks (n=1470)	-1.763	-5.186	-3.423	-1.869	-4.999	-3.130

Table S131. Effect of condition X pre/post election period on descriptive norm perception (percentage) of ... posts using progressively relaxed exclusion criteria.

Robustness across inclusion thresholds: self_appropriate						
Entries are β coefficients; bold indicates $p < .05$.						
Inclusion threshold	Pre-election			Post-election		
	DE – RC	EB – RC	EB – DE	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1815)	-0.497	-4.154	-3.657	-0.695	-3.940	-3.245
≥ 2 weeks (n=1760)	-0.485	-4.132	-3.647	-0.686	-3.971	-3.285
≥ 3 weeks (n=1722)	-0.339	-4.125	-3.786	-0.690	-4.031	-3.341
≥ 4 weeks (n=1630)	-0.606	-4.382	-3.777	-0.459	-3.805	-3.346
≥ 5 weeks (n=1470)	-0.227	-4.157	-3.930	-0.507	-4.171	-3.664

Table S132. Effect of condition X pre/post election period on descriptive norm perception (percentage) of ... posts using progressively relaxed exclusion criteria.

9.1.3 RQ3 results

RQ3 Robustness (Base Model): Any intergroup engagement			
Entries are β (logit); bold indicates $p < .05$.			
Inclusion threshold	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1824)	0.155	-0.114	-0.270
≥ 2 weeks (n=1766)	0.137	-0.115	-0.252
≥ 3 weeks (n=1728)	0.158	-0.090	-0.248
≥ 4 weeks (n=1636)	0.137	-0.088	-0.225
≥ 5 weeks (n=1475)	0.134	0.012	-0.121

Table S133: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period, using progressively relaxed exclusion criteria.

RQ3 Robustness (Base Model): Any moral engagement			
Entries are β (logit); bold indicates $p < .05$.			
Inclusion threshold	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1824)	-0.149	-0.069	0.080
≥ 2 weeks (n=1766)	-0.134	-0.047	0.087
≥ 3 weeks (n=1728)	-0.132	-0.039	0.093
≥ 4 weeks (n=1636)	-0.158	-0.035	0.122
≥ 5 weeks (n=1475)	-0.154	0.023	0.177

Table S134: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period, using progressively relaxed exclusion criteria.

RQ3 Robustness (Base Model): Any valence_negative engagement			
Entries are β (logit); bold indicates $p < .05$.			
Inclusion threshold	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1824)	-0.183	-0.237	-0.053
≥ 2 weeks (n=1766)	-0.195	-0.230	-0.035
≥ 3 weeks (n=1728)	-0.194	-0.222	-0.027
≥ 4 weeks (n=1636)	-0.126	-0.126	0.000
≥ 5 weeks (n=1475)	-0.117	-0.053	0.063

Table S135: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period, using progressively relaxed exclusion criteria.

RQ3 Robustness (Base Model): Any valence_positive engagement			
Entries are β (logit); bold indicates $p < .05$.			
Inclusion threshold	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1824)	-0.112	-0.163	-0.051
≥ 2 weeks (n=1766)	-0.147	-0.154	-0.007
≥ 3 weeks (n=1728)	-0.134	-0.143	-0.010
≥ 4 weeks (n=1636)	-0.077	-0.059	0.018
≥ 5 weeks (n=1475)	-0.075	-0.002	0.074

Table S136: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period, using progressively relaxed exclusion criteria.

RQ3 Robustness (Base Model): Any sociopolitical engagement

Entries are β (logit); bold indicates $p < .05$.

Inclusion threshold	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1824)	-0.175	-0.204	-0.029
≥ 2 weeks (n=1766)	-0.211	-0.205	0.006
≥ 3 weeks (n=1728)	-0.190	-0.198	-0.008
≥ 4 weeks (n=1636)	-0.128	-0.120	0.008
≥ 5 weeks (n=1475)	-0.116	-0.045	0.071

Table S137: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period, using progressively relaxed exclusion criteria.

RQ3 Robustness (Base Model): Any toxic engagement			
Entries are β (logit); bold indicates $p < .05$.			
Inclusion threshold	DE – RC	EB – RC	EB – DE
≥ 1 weeks (n=1824)	-0.242	-0.243	-0.001
≥ 2 weeks (n=1766)	-0.245	-0.224	0.021
≥ 3 weeks (n=1728)	-0.227	-0.201	0.027
≥ 4 weeks (n=1636)	-0.202	-0.143	0.058
≥ 5 weeks (n=1475)	-0.271	-0.179	0.092

Table S138: Odds ratios from logistic mixed-effects regression models with Reverse Chronological (RC) as the reference group, averaged across the entire study period, using progressively relaxed exclusion criteria.

9.2 Baseline individual difference analysis

Selective attrition by baseline extremity. To assess whether attrition could bias estimated treatment effects via selective dropout, we modeled retention in the analytic sample as a function of experimental condition, baseline attitude extremity (mean across five items), baseline ideological extremity (absolute distance from the midpoint category “Neither” on a seven-point ideology scale ranging from “Extremely liberal” to “Extremely conservative”), and their interactions (Table S139). Baseline attitude extremity was associated with higher retention in both the engagement-based and reverse-chronological conditions ($ORs = 1.29$ and 1.47 , $ps = .021$ and $< .001$), whereas it was unrelated to retention in the diversified extremity condition ($OR = 0.98$, $p = .911$). Accordingly, the relationship between attitude extremity and retention differed between the reverse-chronological and diversified extremity conditions ($OR = 1.50$, $p = .025$). Importantly, however, the reverse-chronological condition did not differ significantly from the engagement-based condition in the relationship between baseline attitude extremity ($p = .389$) or ideological extremity ($p = .716$) and retention, despite the reverse-chronological condition exhibiting the highest descriptive attrition rate. We also observed that ideologically extreme participants were somewhat more likely to be retained in the diversified extremity condition relative to the engagement-based condition ($OR = 1.46$, $p = .035$). Because participants higher in baseline extremity are plausibly more likely to consume and engage with IME-type content, these patterns make it unlikely that observed differences in IME exposure across conditions are attributable to selective dropout; if anything, the reverse-chronological condition retained more attitude-extreme participants, which would bias against finding lower IME exposure in that condition.

Exploring Selective attrition by baseline extremity			
Baseline predictor	Condition / comparison	Odds ratio [95% CI]	p
A. Effect of baseline extremity on retention (Odds Ratios, 95% CI)			
Attitude extremity (z)	EB	1.29 [1.04, 1.60]	0.021
Attitude extremity (z)	DE	0.98 [0.74, 1.31]	0.911
Attitude extremity (z)	RC	1.47 [1.19, 1.82]	< .001
Ideology extremity (z)	EB	0.87 [0.69, 1.09]	0.232
Ideology extremity (z)	DE	1.27 [0.97, 1.66]	0.079
Ideology extremity (z)	RC	0.92 [0.73, 1.16]	0.506
B. Differences in slopes between conditions			
Attitude extremity (z)	DE – EB	0.76 [0.53, 1.09]	0.138
Attitude extremity (z)	RC – EB	1.14 [0.84, 1.55]	0.389
Attitude extremity (z)	RC – DE	1.50 [1.05, 2.13]	0.025
Ideology extremity (z)	DE – EB	1.46 [1.03, 2.07]	0.035
Ideology extremity (z)	RC – EB	1.06 [0.77, 1.47]	0.716
Ideology extremity (z)	RC – DE	0.73 [0.51, 1.04]	0.078
Note. Odds ratios (ORs) reflect the change in odds of retention for a 1 SD increase in the baseline predictor. Panel A reports within-condition slopes; Panel B reports differences in slopes between conditions.			

Table S139. Exploring selective attrition by baseline extremity. Odds ratios (ORs) from logistic regression models predicting retention in the analytic sample as a function of experimental condition, baseline attitude extremity (z-scored; mean across five items), baseline ideological extremity (z-scored; absolute distance from “Neither” on a seven-point ideology scale), and their interactions. Panel A reports within-condition associations between baseline extremity and retention. Panel B reports differences in these associations between conditions (interaction contrasts). ORs > 1 indicate greater odds of retention with higher baseline extremity.

Supplementary Information References

1. Lees, A. *et al.* A New Generation of Perspective API: Efficient Multilingual Character-level Transformers. Preprint at <https://doi.org/10.48550/arXiv.2202.11176> (2022).
2. Anil, R. *et al.* PaLM 2 Technical Report. Preprint at <https://doi.org/10.48550/arXiv.2305.10403> (2023).
3. About the API - Attributes and Languages.
https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US.
4. OpenAI Platform. <https://platform.openai.com>.
5. meta-llama/Llama-2-7b · Hugging Face. <https://huggingface.co/meta-llama/Llama-2-7b>.
6. Hopp, F. R., Fisher, J. T., Cornell, D., Huskey, R. & Weber, R. The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behav Res* **53**, 232–246 (2021).
7. Shaver, P., Schwartz, J., Kirson, D. & O'Connor, C. Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology* **52**, 1061–1086 (1987).
8. Rajadesingan, A., Budak, C. & Resnick, P. Political Discussion is Abundant in Non-political Subreddits (and Less Toxic). Preprint at <https://doi.org/10.48550/arXiv.2104.09560> (2021).
9. Baly, R., Martino, G. D. S., Glass, J. & Nakov, P. We Can Detect Your Bias: Predicting the Political Ideology of News Articles. Preprint at <https://doi.org/10.48550/arXiv.2010.05338> (2020).
10. Twitter's Recommendation Algorithm.
https://blog.x.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm.
11. facebookresearch/faiss. Meta Research (2024).

12. spotify/annoy. Spotify (2024).
13. Hellwig, N. C., Fehle, J., Bink, M., Schmidt, T. & Wolff, C. Exploring Twitter discourse with BERTopic: topic modeling of tweets related to the major German parties during the 2021 German federal election. *Int J Speech Technol* **27**, 901–921 (2024).
14. Egger, R. & Yu, J. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Front Sociol* **7**, 886498 (2022).
15. sentence-transformers/all-MiniLM-L6-v2 · Hugging Face.
<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2> (2024).
16. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at <https://doi.org/10.48550/arXiv.1802.03426> (2020).
17. Stewart, G. & Al-Khassaweneh, M. An Implementation of the HDBSCAN* Clustering Algorithm. *Applied Sciences* **12**, 2405 (2022).
18. Egger, R. & Yu, J. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Front Sociol* **7**, 886498 (2022).
19. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. Preprint at <https://doi.org/10.48550/arXiv.2203.05794> (2022).
20. Stats for Bluesky by Jaz (jaz.bsky.social). <https://bsky.jazco.dev/stats>.
21. Ha, A. Bluesky grows to 9M+ users. *TechCrunch*
<https://techcrunch.com/2024/09/07/bluesky-grows-to-9m-users/> (2024).
22. Perez, S. Bluesky continues to soar, adding 2M more new users in a matter of days. *TechCrunch*
<https://techcrunch.com/2024/09/03/bluesky-continues-to-soar-adding-2m-more-new-users-in-a-matter-of-days/> (2024).