Supplementary Materials for The Elite Effect: Blocking Political Leaders on X Decreases Misinformation Susceptibility and Low-credibility Content Sharing

Kokil Jaidka[1*], Yphtach Lelkes[2], Subhayan Mukerjee[1], Harshit Aneja[1]

[1*]Department of Communications and New Media & Centre for Trusted Internet and Community, National University of Singapore, 11 Computing Way, Singapore, 110097.
[2]Annenberg School for Communication, University of Pennsylvania, 3620 Walnut Street, Philadelphia, 19104, Pennsylvania, USA.

*Corresponding author(s). E-mail(s): jaidka@nus.edu.sg ;
Contributing authors: ylelkes@asc.upenn.edu; mukerjee@nus.edu.sg; contactaneja@gmail.com;

Contents

Appendix A Ethics

All participants provided informed consent prior to participation. Recruitment procedures, consent materials, and study protocols are described in the following sections. Participants were informed that their social media feeds would be modified for research purposes and that they could withdraw from the study at any time without penalty. Participants were also given the option to withdraw their data at any point prior to de-identification, including after completing study activities.

The experiment was conducted using a custom research application that delivered manipulated versions of participants' social media feeds. The study conditioned all analyses on participants meeting a minimum activity threshold within the experimental app, ensuring that treatment assignment corresponded to meaningful exposure to the intervention. Any social media use outside the experimental environment was neither manipulated nor logged and was disclosed to participants during consent. All behavioral and survey data were de-identified prior to analysis. Only de-identified data were accessible to the research team, and no personally identifying information was retained alongside outcome measures. Data were stored on secure servers and handled in accordance with institutional data protection guidelines.

The study protocol, consent procedures, and data collection methods were reviewed and approved by the Institutional Review Boards of the University of Pennsylvania (IRB # 832944) and the National University of Singapore (IRB # NUS-IRB-2021-431) . The research team retained full independence over study design, analysis, and reporting of results.

Appendix B Pre-registered Research Questions

In advance of data collection, we pre-registered all treatment procedures, outcome measures, compliance criteria, stopping rules, and statistical analyses on OSF. The pre-analysis plan specifies the experimental design, manipulation checks, operationalization of dependent variables, moderation tests, and multiple-comparison adjustments. The study was designed as a between-subject field experiment on X (Twitter), with block randomization based on baseline X usage and political partisanship. Compliance was defined as installing the Twilly app and using it at least once per week during the 28-day intervention period.

The pre-registered stopping criteria was 1,200 compliant post-survey completions. We stopped recruitment once 1,214 participants were assigned to one of the three study arms reported in this study.Among the 1,214 baseline respondents, 1,140 installed the Twilly app and thus entered the intervention phase; this group constitutes the baseline installation sample. The final analytic sample consists of participants who satisfied the pre-registered compliance criterion (installation plus minimum usage threshold) and completed the post-treatment survey.

Recruitment was discontinued earlier than the pre-registration for logistical and contextual reasons. First, our initial target of 1,200 compliant post-survey respondents assumed attrition consistent with prior beta-tests, and therefore required recruiting approximately 1,600 participants at baseline. Actual attrition exceeded expectations

(approximately 50%), a rate that could not have been anticipated at the time of pre-registration. Second, recruitment had been ongoing for seven weeks, and the conclusion of the U.S. election cycle constituted a natural stopping point. Because election outcomes substantially alter political discourse, feed composition, and political attitudes, extending recruitment beyond this point would have introduced a qualitatively different informational environment. Third, approximately half of recruited individuals did not meet eligibility criteria related to recent X usage, leading to a high bounce rate and diminishing marginal returns to continued outreach. Fourth, the number of 1200 post-survey responses was pre-registered for four study arms, while this study reports findings based on three treatment arms.

Last but not the least, the decision to discontinue recruitment was based on timing and feasibility constraints and was made without reference to treatment effects or outcome analyses. All primary analyses were conducted as preregistered.

We pre-registered the following research questions and hypotheses:

- **RQ1 (App usage):** What is the effect of (a) blocking all political content, (b) blocking political elites, and (c) demetrication on overall app usage (time spent, tweets viewed, likes, clicks)?
- **RQ2 (Sharing behaviors):** What is the effect of each treatment on specific sharing behaviors, including consuming, liking, retweeting, commenting on, and sharing partisan and non-partisan content, as well as engagement with low-credibility sources?
- **RQ3 (Affective polarization):** What is the effect of each treatment on affective polarization toward political in-groups and out-groups?
- **RQ4 (Perceived polarization):** What is the effect of each treatment on perceived political polarization?
- **RQ5 (Perceived discrimination):** What is the effect of each treatment on perceived political discrimination?
- **RQ6 (Political attitudes):** What is the effect of each treatment on political interest and political trust?
- **RQ7 (Political knowledge and misinformation):** What is the effect of each treatment on political knowledge, misinformation awareness, and belief in misinformation?
- **RQ8 (Media trust and news avoidance):** What is the effect of each treatment on media trust and news avoidance?
- **RQ9 (Moderation by partisan strength):** Do the effects of the treatments vary as a function of political partisanship strength?

Appendix C Technical design of Twilly

To enable the experimental manipulations and behavioral data collection described above, we developed a mobile platform named `Twilly`. The platform mimics a Twitter-like interface while allowing researchers to administer timeline treatments and record user interactions in a controlled environment. `Twilly` consists of three core components: a mobile app for participants, a web dashboard for researchers, and a backend system for coordinating user activity, treatment assignment, and data logging.

C.1 System Architecture

Figure C1 illustrates the architecture of the `Twilly` platform. The components are as follows:

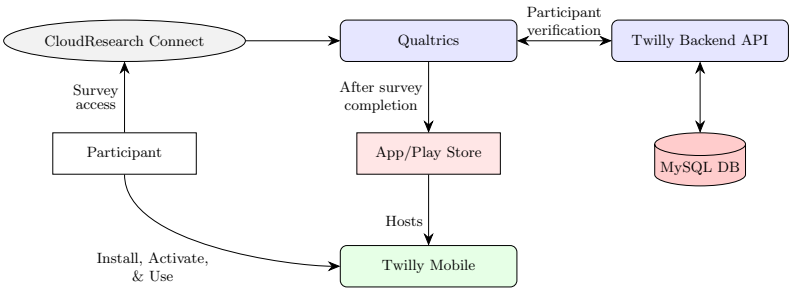


Fig. C1: System architecture of the Twilly platform.

1. **Mobile App:** The app provides participants with a familiar social media interface. It enables authentication, loads timelines, applies condition-specific modifications (e.g., muting specific usernames or keywords), and records interactions such as views, likes, and clicks.
2. **Backend Server:** The backend manages participant authentication, assigns experimental conditions, and stores incoming activity logs. It also communicates with the survey system to validate participants and update progress.
3. **Research Dashboard:** A web-based interface used by researchers to configure study parameters (e.g., treatment types, durations), monitor participation, and download usage logs.
4. **Database:** A secure relational database stores all experimental configurations, participant metadata, and user interaction logs.
5. **Local App Storage:** Logs are temporarily stored on the user's device before being uploaded to the backend, ensuring minimal data loss in the event of disconnection.
6. **External Services:** The platform integrates with external tools for participant recruitment and validation. Surveys are conducted using a third-party provider, and the app is distributed via standard app stores.

Appendix D Outcome Definitions

All survey-based outcomes are operationalized as the difference between endline and baseline values. That is, we compute the change in participants' responses across time to assess the effects of the intervention. Passive behavioral outcomes (e.g., app usage) are measured over the 28-day intervention period. Survey instruments with exact question wording are included in Appendix D.

D.1 App Usage

Operationalization: Measured using passive engagement logs from the Twilly app. Metrics include the number of:

- Timeline views (home and latest)
- Embedded media and video views
- Individual post views
- Profile page views
- Searches for content or users

D.2 Specific Sharing Behaviors

Measured using active engagement logs, that were broadly grouped into:

- Number of posts viewed
- Number of posts liked
- Number of posts retweeted

These are also comparable based on user demographics and content scores, as described in Figure 1, Appendix D, and Appendix E.

D.3 Content Classification: Political Relevance and Partisanship

We scored the content that participants encountered in their personalized feeds for its political relevance, partisanship, and credibility. Using these, we calculated the median score of the political relevance, partisanship, and credibility of users' *feed items* across different forms of engagement.

D.3.1 Lexicon Construction for Political Content

To identify politically relevant content, we adopted a semi-supervised lexicon-based classification approach, combining the scalability of unsupervised topic modeling with the precision of manual annotation, extending methods advocated in previous research [1, 2]. We collected the 758 annotated political keywords identified in prior work [3], based on large-scale topic models of historical social media posts. These were augmented with known political Twitter handles and a list of current and former U.S. Congressional representatives [4]. Altogether, our lexicon comprised 1,745 words.

D.3.2 Semantic Scoring with Sentence Transformers

We computed a continuous political relevance score using the `all-MiniLM-L6-v2` sentence transformer. The political keywords were divided into 18 chunks of approximately 100 words each, treating each chunk as a separate sentence input. Both the feed item and each keyword chunk were embedded into 384-dimensional vectors using mean pooling. Cosine similarity was computed between the feed item and each of the 18 keyword chunks, and the final political relevance score was calculated as the mean of these 18 similarity scores:

$$\text{PoliticalScore}_{\text{MiniLM}}(x) = \frac{1}{18}\sum_{j=1}^{18}\cos(\text{Embed}(x), \text{Embed}(\text{Keywords}_j))$$

This continuous measure allowed us to capture more nuanced expressions of political content that may have eluded keyword matching. The sentence-transformer approach complemented the keyword-based classification and enabled robustness checks across different thresholds of political relevance.

Examples of rows marked as relevant to politics are reported in the Table.

Table D1: Examples of tweets by political content score

Tweet Text	**Political content score**
RT moniquejolie: #USElections2024 #PresidentialElection2024 #GOTV Who will #PardonAssange if elected #POTUS in '24? GP: DrJillStei...	0.45
Donald J Trump KILLED THE BOARDER DEAL ... #ResistanceUnited #DemVoice1 #ProudBlue #KamalaInvestsInUS https://t.co/t0NY9lHdEN	0.45
RT RpsAgainstTrump: KamalaHarris It's okay to be a Republican and vote for VP Harris. Country over party	0.40
RT MLilyjo: FAKE MAGA ACCOUNTS Patriot_FreYa Celina_patriot ...	0.40
...by Working Families Party Councilmember NicForPhilly	0.35
It's not #Misogyny. #KomradeKamala is a low IQ pretender ... She is nothing more than a 4th term for BarackObama	0.35
Congresswoman Nancy Mace just officially entered into the Congressional record her messages from an unhinged CNN panel host ... This is the way. https://t.co/Olw5Uf8p7L	0.30
No part of Virginia should ever be left behind or forgotten. That's why I've worked to deploy high-speed internet to rural communities ...	0.25
Love when Republicans say "people are doing enormous damage by spreading inflammatory lies and misinformation" when by "people" they mean "the presidential ticket I intend to vote for"	0.25
pop crave is on a mission to make people hate chappell judging by the way they cherry pick her quotes ...	0.20
John Doyle actually said the term "Zionist Occupied Government" is just a joke on the internet ...	0.15
Mar '94 - Longtime artist Virginia Barratt of VNSMatrix asks – when will the game industry acknowledge & support female demographics? https://t.co/DFXb7vOKae	0.15

Tweet Text	Political content score
Recruiting update for 09/24/2024 Offers -2027 3 TE Luke Brewer	0.15
I'm sorry your family let you down	0.15
RT TomBrady: Big man on the move!!!	0.10
behold - trackball rotation!! (hence all the quaternion talk today) https://t.co/nfUGVfkicp	0.05
I put my heart and soul into my work, and have lost my mind in the process. – Vincent Van Gogh	0.05
Olive Garden: 0 https://t.co/HjGZAzZkyj	0.05
The only thing in the world that should be eternal is MOTHER! https://t.co/hSdszap4fM	0.00

D.4 Partisanship and Credibility

To characterize the political content in participants' feeds, we focused on two key dimensions: *partisanship* and *credibility*. These dimensions were operationalized using validated external datasets that score both individuals and media outlets based on their ideological leaning and source quality.

Partisanship scores.

We relied on prior work [4, 5] that infers the ideological orientation of over 1,200 X users—including elected officials, content creators, and prominent public figures—using patterns of network affiliation and audience engagement. For media domains, partisanship labels were taken from [6], which provide expert-coded ideological labels based on editorial slant and citation networks.

Credibility scores.

To assess information quality, we drew on datasets from [7] and [8], which assign credibility ratings to news domains based on expert evaluations and crowdsourced reliability metrics. These sources also identify domains with histories of misinformation or low editorial standards, enabling a binary classification of sources as high- or low-credibility.

Implementation.

Content Scoring and Political Exposure.

We applied partisanship and credibility scores to two principal types of content entities in the dataset: (1) the domains referenced in shared URLs, and (2) the X handles of users either authoring or mentioned in political posts.

Among the 197,214 domain mentions, 175,247 (88.87%) were scored based on available credibility or ideological scores, which accounted for 4,602 websites (30.64%) out of the 15,019 websites cited. Among the 359,146 X accounts, authored or mentioned

2,565,029 times, that users encountered across all their feeds, we could score 437,460 instances (17.06%) posted by 2,061 unique handles. This distribution reflects a skewed media environment in which a narrow subset of frequently cited sources — whether domains or user accounts — disproportionately shapes the political information landscape encountered by users. It also highlights the endemic nature of partisan exposure within the platform, driven primarily by the internal dynamics of user-generated content and interpersonal sharing, rather than by exogenous news sources.

In the control condition, **political content (citing political sources, or authored by political accounts) comprised approximately 17.71% of users' social media feeds**. In terms of ideological balance, **Democrats were exposed to predominantly cross-partisan content (63.93%)**, whereas **Republicans encountered more ideologically congruent material (64.50%)**. Low-credibility content remained relatively rare, accounting for 0.40% of political news overall, with slightly higher prevalence in Republican-aligned content (1.22%) than in Democrat-aligned content (0.65%).

D.5 Factual Knowledge

Participants evaluated whether a set of recent political events had occurred. Response options were: *Definitely didn't happen*, *Probably didn't happen*, *Probably did happen*, and *Definitely did happen*.

Responses were scored dichotomously: only exactly correct answers were considered accurate.

- For **true** statements, the correct answer was *Definitely did happen*.
- For **false** statements, the correct answer was *Definitely didn't happen*.

International knowledge and **U.S. national knowledge** were calculated separately as the number of correctly identified events in each domain.

D.5.1 International Political Knowledge

- Mexico elected its first female president, Claudia Sheinbaum. (T)
- The International Olympic Committee announced the suspension of Russia from the 2024 Paris Olympics. (T)
- The United Nations declared a global state of emergency due to climate-related disasters. (F)
- The EU and UK negotiated a new post-Brexit trade deal following extended disputes. (F)
- China launched its first crewed mission to Mars. (F)
- COP29 in Baku, Azerbaijan, set the goal to reduce global emissions by 2050 under a new agreement. (T)

D.5.2 National (U.S.) Political Knowledge

- Donald Trump announced plans to skip all remaining 2024 election debates. (T)
- A major cyberattack temporarily disabled several U.S. government websites. (F)

- President Joe Biden signed an executive order mandating nationwide electric vehicle incentives. (T)
- The U.S. hosted the 2024 NATO Summit in Washington, D.C., commemorating NATO's 75th anniversary. (T)
- Mass protests erupted after a federal ruling reversed several key environmental regulations. (F)
- California passed a statewide law banning all gasoline-powered car sales by 2035. (T)

D.6 Misinformation Susceptibility

Misinformation susceptibility was assessed using two modules: one based on text-only claims (fact-checked by AFP FactCheck and adapted from [9]), and one based on multimodal (headline + image) content.

D.6.1 Text-Based Claims

Participants rated the accuracy of widely circulated claims on a 4-point scale: *Definitely false*, *Probably false*, *Probably true*, *Definitely true*. Only the following responses were counted as correct:

- *Definitely false* for false claims
- *Definitely true* for true claims

Scores reflect the number of exactly correct responses, without partial credit. The items were:

- 10,000 illegal votes were counted in Arizona during the 2020 U.S. election. (F)
- Under President Biden, the U.S. borders are effectively gone and migrants are committing crimes at unprecedented levels. (F)
- The suspect in Donald Trump's assassination attempt was released from custody. (F)
- Pennsylvania election workers fraudulently filled out ballots in 2020. (F)
- Kamala Harris posed for a 2001 photo with Sean 'Diddy' Combs. (F)
- Violent crimes are at historic lows across the U.S. (T)
- A plot to assassinate Donald Trump was thwarted in 2024. (T)
- Prior to the 2016 presidential election, Donald Trump arranged a payment to an adult film star. (T)
- Donald Trump held a Bible upside-down in front of a church. (T)
- Masks and face coverings are not effective in preventing the spread of COVID-19. (F)

D.6.2 Multimodal Misinformation

Participants viewed five randomly assigned headline-image pairs (three false, two true). As an example:

Headline: Kamala Harris secures Democratic presidential nomination
Image: A photo of Rep. Ilhan Omar speaking at a podium.

This example was considered false.

For each item, they responded to six statements on a 5-point Likert scale from *Strongly disagree* to *Strongly agree*:

1. I am familiar with this topic.
2. I am familiar with this tweet/headline.
3. I think this news item is truthful.
4. I think this news item is fake.
5. If I saw this on social media, I would share it.
6. I am very confident about my rating.

Misinformation susceptibility was calculated by reverse-coding truth/fake responses:

- For false items: "This is truthful" (item 3) was reverse-coded.
- For true items: "This is fake" (item 4) was reverse-coded.

The overall misinformation score was computed by averaging all reverse-coded values. Higher values indicate better accuracy and lower susceptibility.

Belief in misinformation: was calculated as the sum of responses to item 6 ("I am very confident about my rating") across all five items. Higher scores reflect greater confidence in judgment.

D.7 Attitudinal Outcomes

We measured several secondary attitudinal outcomes capturing polarization, media trust, and political engagement. Unless otherwise noted, Likert items were measured on 5-point scales ranging from *Strongly disagree* (1) to *Strongly agree* (5).

Affective Polarization.

Measured along four complementary dimensions:

- *Feeling thermometer:* Difference between in-party and out-party warmth ratings (0 = very cold, 100 = very warm).
- *Perceived partisan threat:* Agreement with "The other political party is so misguided that they threaten the nation's well-being."
- *Social distance:* Average comfort interacting with out-party members in relational contexts (e.g., neighbor, colleague, in-law) (1 = extremely uncomfortable, 5 = extremely comfortable).
- *Political social identity:* Composite of nine items measuring identification with one's political party (e.g., "When I speak about the party, I usually say 'we' instead of 'they'.").

Internal consistency was assessed for multi-item measures. Political social identity ($\alpha = .773$) and social distance ($\alpha = .839$) demonstrated acceptable reliability. The feeling thermometer is operationalized as a difference score; reliability is reflected in the correlation between in-party and out-party ratings ($r = .765$). Perceived partisan threat is measured with a single item.

Analyses of polarization outcomes were restricted to participants with non-neutral partisan identification (partisan strength $\neq 0$ on a $[-3, 3]$ scale).

Perceived Polarization and Discrimination.

Perceived polarization was measured using agreement with "More and more Democrats and Republicans have extreme views these days." Perceived discrimination was measured with "How often do you feel that your political beliefs lead to unfair treatment?" Internal consistency was modest ($\alpha = .683$ for polarization; $\alpha = .646$ for discrimination).

Media Trust and News Avoidance.

Media trust was measured using items such as "I trust mainstream news sources to provide accurate information." News avoidance was measured with items such as "I find myself actively trying to avoid news these days." Higher scores indicate greater trust or avoidance, respectively. Reliability was acceptable (media trust $\alpha = .704$; news avoidance $\alpha = .744$; test–retest $r = .704$ and $r = .744$, respectively).

Political Interest and Institutional Trust.

Political interest ("I am interested in politics") and institutional trust ("The government can be trusted to do what is right") were measured using single 5-point Likert items and therefore do not have internal consistency estimates.

D.8 Covariates

Covariates were measured in the baseline survey and selected to account for key demographic, attitudinal, and behavioral differences among participants. Specifically, we controlled for:

- **Demographics:** Gender (male, female, other), age (categorized as 18–24, 25–34, 35–44, 45–54, 55–64, and 65+), education level, household income bracket, and race/ethnicity (non-Hispanic White, Hispanic, Black, Asian, and Other).
- **Political orientation:** Political ideology on a 6-point scale (strong Democrat to strong Republican, including leaners). True neutrals and independents were assigned NA.
- **Number of days active on the app:** Total number of distinct days a participant actively used the Twilly app during the 28-day intervention period, as recorded in passive usage logs. This variable accounts for variation in exposure to the experimental treatment.
- **Data collection phase:** To account for temporal differences in exposure and randomization rollout, fixed effects for the experimental phase (rollout batch) were included.

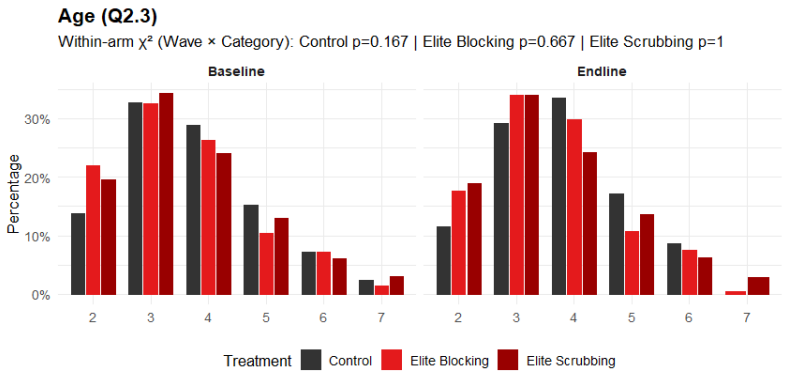
Appendix E Sample, Balance, Attrition, and Compliance

E.1 Sampling

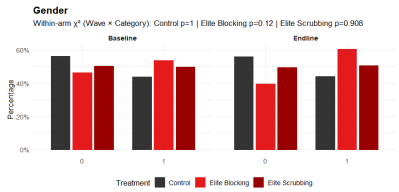
Participants were recruited from the U.S. population via CloudResearch. All participants—treatment and control—were drawn from a common pool of individuals willing to install and use the app, ensuring comparability in treatment exposure, attrition risk, and awareness of app-based monitoring. This choice mitigates bias that would arise from comparing to an external, non-intervened population and minimizes demand effects or placebo differences.

The inclusion criteria required that participants were active users of X (formerly Twitter), defined as having posted on X at least once in the past year. This criterion screened out participants with private accounts, as the date of their last post is then no longer accessible through the X API. After validation using the X API and manual checks to remove duplicates or inauthentic usernames, a final eligible sample of $1,214$ participants was identified, who installed the Twilly app and activated it using a unique activation code.

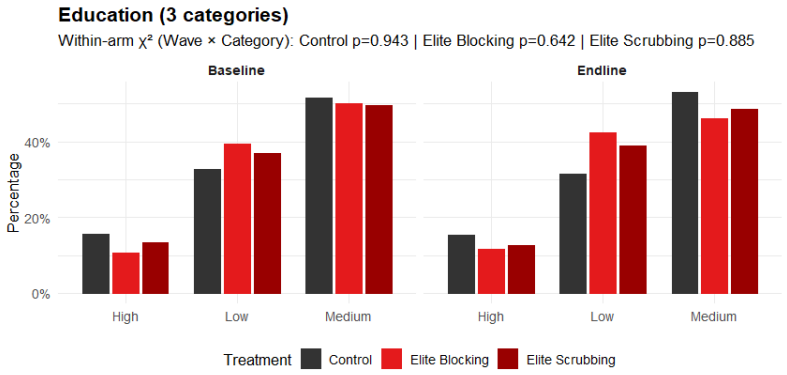
Figure E2 reports the participant demographics across the treatment arms and survey waves. Table E2 reports the percentages reported in Figure E2.



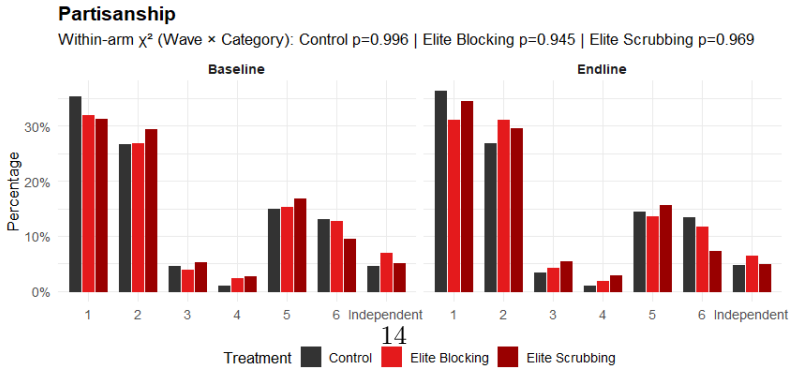
(a) Age



(b) Gender



(c) Education



(d) Ideology

Fig. E2: Demographic balance across the study waves (baseline and endline).

Table E2: Baseline and endline demographic distributions by treatment arm.

Variable	Treatment	Category	Baseline n (%)	Endline n (%)	Δ 95% CI	χ^2(df)	p
Education	Control	High	65 (15.6%)	32 (15.3%)	−0.003 [−0.060, 0.066]	0.12(2)	.943
		Low	136 (32.7%)	66 (31.6%)	−0.011 [−0.070, 0.092]		
		Medium	215 (51.7%)	111 (53.1%)	+0.014 [−0.101, 0.072]		
	Elite Blocking	High	44 (10.7%)	25 (11.6%)	+0.009 [−0.065, 0.046]	0.09(2)	.642
		Low	162 (39.3%)	91 (42.3%)	+0.030 [−0.115, 0.055]		
		Medium	206 (50.0%)	99 (46.0%)	−0.040 [−0.046, 0.125]		
	Elite Scrubbing	High	54 (13.5%)	26 (12.6%)	−0.009 [−0.052, 0.069]	0.24(2)	.885
		Low	147 (36.9%)	80 (38.8%)	+0.019 [−0.105, 0.066]		
		Medium	198 (49.6%)	100 (48.5%)	−0.011 [−0.077, 0.099]		
Income	Control	High	45 (10.8%)	28 (13.4%)	+0.026 [−0.084, 0.033]	1.31(2)	.519
		Low	133 (32.0%)	70 (33.5%)	+0.015 [−0.097, 0.067]		
		Medium	238 (57.2%)	111 (53.1%)	−0.041 [−0.045, 0.127]		
	Elite Blocking	High	49 (11.8%)	24 (11.2%)	−0.007 [−0.049, 0.063]	2.06(2)	.356
		Low	125 (30.2%)	77 (35.8%)	+0.056 [−0.138, 0.025]		
		Medium	240 (58.0%)	114 (53.0%)	−0.050 [−0.036, 0.135]		
	Elite Scrubbing	High	41 (10.3%)	20 (9.7%)	−0.006 [−0.048, 0.060]	0.54(2)	.764
		Low	138 (34.6%)	66 (32.0%)	−0.026 [−0.057, 0.108]		
		Medium	220 (55.1%)	120 (58.3%)	+0.031 [−0.118, 0.056]		
Gender	Control	Female	234 (56.2%)	117 (56.0%)	−0.003 [−0.083, 0.088]	0.00(1)	1.00
		Male	182 (43.8%)	92 (44.0%)	+0.003 [−0.088, 0.083]		
	Elite Blocking	Female	192 (46.4%)	85 (39.5%)	−0.068 [−0.016, 0.153]	2.42(1)	.120
		Male	222 (53.6%)	130 (60.5%)	+0.068 [−0.153, 0.016]		
	Elite Scrubbing	Female	201 (50.4%)	102 (49.5%)	−0.009 [−0.079, 0.096]	0.01(1)	.908
		Male	198 (49.6%)	104 (50.5%)	+0.009 [−0.096, 0.079]		
Partisanship	Control	1	147 (35.3%)	76 (36.4%)	+0.010 [−0.094, 0.073]	0.60(6)	.996
		2	111 (26.7%)	56 (26.8%)	+0.001 [−0.076, 0.074]		
		3	19 (4.6%)	7 (3.4%)	−0.012 [−0.023, 0.047]		
		4	4 (1.0%)	2 (1.0%)	−0.000 [−0.016, 0.016]		
		5	62 (14.9%)	30 (14.4%)	−0.006 [−0.057, 0.068]		
		6	54 (13.0%)	28 (13.4%)	+0.004 [−0.064, 0.056]		
		Independent	19 (4.6%)	10 (4.8%)	+0.002 [−0.040, 0.035]		
	Elite Blocking	1	132 (31.9%)	67 (31.2%)	−0.007 [−0.073, 0.087]	1.70(6)	.945
		2	111 (26.8%)	67 (31.2%)	+0.044 [−0.122, 0.035]		
		3	16 (3.9%)	9 (4.2%)	+0.003 [−0.039, 0.033]		
		4	10 (2.4%)	4 (1.9%)	−0.006 [−0.021, 0.032]		
		5	63 (15.2%)	29 (13.5%)	−0.017 [−0.044, 0.078]		
		6	53 (12.8%)	25 (11.6%)	−0.012 [−0.045, 0.069]		
		Independent	29 (7.0%)	14 (6.5%)	−0.005 [−0.040, 0.050]		
	Elite Scrubbing	1	125 (31.3%)	71 (34.5%)	+0.031 [−0.114, 0.052]	1.35(6)	.969
		2	117 (29.3%)	61 (29.6%)	+0.003 [−0.083, 0.077]		
		3	21 (5.3%)	11 (5.3%)	+0.001 [−0.039, 0.038]		
		4	11 (2.8%)	6 (2.9%)	+0.002 [−0.031, 0.028]		
		5	67 (16.8%)	32 (15.5%)	−0.013 [−0.053, 0.078]		
		6	38 (9.5%)	15 (7.3%)	−0.022 [−0.027, 0.072]		
		Independent	20 (5.0%)	10 (4.9%)	−0.002 [−0.036, 0.040]		

Notes: Baseline sample sizes: Control $N = 416$, Elite Blocking $N = 412$–414, Elite Scrubbing $N = 399$. Endline sample sizes: Control $N = 209$, Elite Blocking $N = 215$, Elite Scrubbing $N = 206$. $\Delta = \%_{\text{endline}} - \%_{\text{baseline}}$ with Wald-type 95% CI from a two-sample proportion test. χ^2 tests the within-arm association between wave (baseline vs. endline) and category distribution. Gender categories 0/1 correspond to Female/Male.

Table E3: Covariate balance before and after matching/weighting (Arm 1: FEP01A vs TRD158B). Standardized mean differences (SMD) are shown for the unmatched sample (Diff.Un) and matched/weighted sample (Diff.Adj). Variance ratios are shown for continuous covariates (V.Ratio.Un / V.Ratio.Adj); variance ratios are not reported for binary covariates (NA).

Construct	Covariate	Type	Diff.Un	Diff.Adj	VR Un	VR Adj
Demographics						
	Age: low	Binary	0.0814	0.0000		
	Age: mid	Binary	0.0727	0.00563		
	Age: high	Binary	0.00878	0.00563		
	Gender (binary)	Binary	0.0989	0.0000		
	Education: low	Binary	0.0663	0.0000		
	Education: mid	Binary	0.0168	0.0000		
	Education: high	Binary	0.0495	1.39×10^{-17}		
Political predispositions						
	Political ideology scale	Contin.	0.0479	0.0122	1.03	1.04
Propensity score						
	Estimated propensity score distance	Distance	0.319	0.00384	1.16	1.01

E.2 Balancing: Matching and Covariate Balance

To estimate the average treatment effect on the treated (ATT), we conducted pairwise comparisons between each intervention arm (FEP01A and FEP01D) and the reference arm (TRD158B) using propensity score full matching on baseline covariates.

For each contrast, we first restricted the baseline sample to respondents in the two relevant arms. Observations with missing or non-finite values on any matching covariate were excluded using complete-case filtering. The following covariates exhibited missingness and were therefore subject to listwise deletion in the matching stage: education category (edu_3), affective polarization, Bogardus social distance, social identity (socid), and image misinformation susceptibility.

Propensity scores were estimated using logistic regression, including baseline political ideology, affective polarization, social identity, political knowledge, misinformation susceptibility (text and image), and demographic covariates. Full matching with an ATT estimand was then applied. Education (tertiary variable) and gender (binary variable) were enforced as exact matching constraints.

Table E3 and Table E4 report standardized mean differences (SMDs) before and after matching for the Elite Blocking and Elite Scrubbing conditions versus the control condition. After matching, covariate balance improved substantially, with all covariates falling below the conventional threshold (absolute SMD < 0.10). Figure E3 visualizes the reduction in imbalance via Love plots.

E.3 Experimental Procedure

The study employed a two-wave within-subject design. Participants first completed a baseline survey (Wave 1) that included measures of misinformation susceptibility, affective polarization, and media trust. Upon completion, they were asked to install *Twilly*, a custom-built X (formerly Twitter) client that preserved core platform functionality while enabling experimental manipulation of users' timelines.

Table E4: Covariate balance before and after matching/weighting (Arm 2: FEP01D vs TRD158B). Standardized mean differences (SMD) are shown for the unmatched sample (Diff.Un) and matched/weighted sample (Diff.Adj). Variance ratios are shown for continuous covariates; variance ratios are not reported for binary covariates (NA).

Construct	Covariate	Type	Diff.Un	Diff.Adj	VR Un	VR Adj
Demographics						
	Age: low	Binary	0.0589	0.0102		
	Age: mid	Binary	0.0518	0.00596		
	Age: high	Binary	0.00710	0.00422		
	Gender (binary)	Binary	0.0612	0.0000		
	Education: low	Binary	0.0420	0.0000		
	Education: mid	Binary	0.0206	5.55×10^{-17}		
	Education: high	Binary	0.0214	2.78×10^{-17}		
Political predispositions						
	Political ideology scale	Contin.	0.00303	0.00713	1.11	1.01
Propensity score						
	Estimated propensity score distance	Distance	0.195	0.0172	1.31	1.01

Random assignment occurred immediately after the baseline survey, through a unique activation code linked to the participant's assigned condition. Each participant remained in the study for 28 days following activation of the Twilly app.

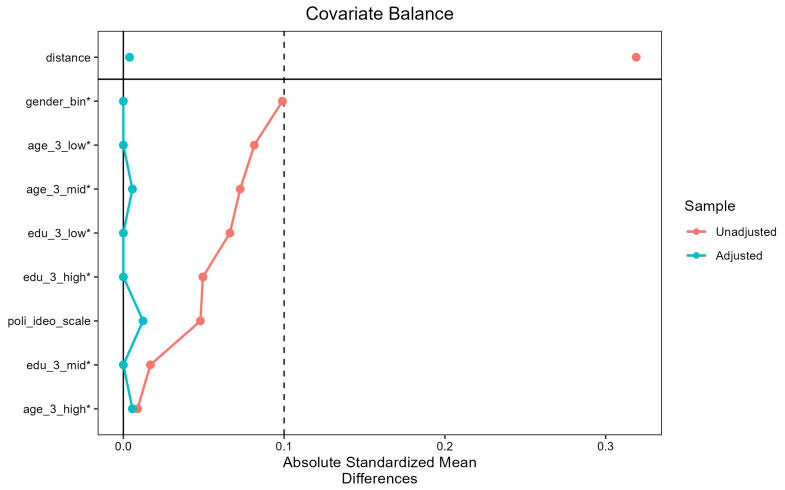
Treatment rollout. Treatments were deployed in two recruitment phases between 18 September and 18 November. We pre-registered that to mitigate attrition risks, data collection would be staggered and continue until the pre-registered sample size was met. In the first phase, participants were randomly assigned to either the control condition or a separate pre-registered treatment condition not analyzed in the current study. In the second phase, new participants were randomly assigned to one of the two focal treatment conditions examined here.

Across both phases, a randomization script ensured balance across experimental arms based on baseline characteristics. Participants were assigned to one of three conditions for a 3-week intervention:

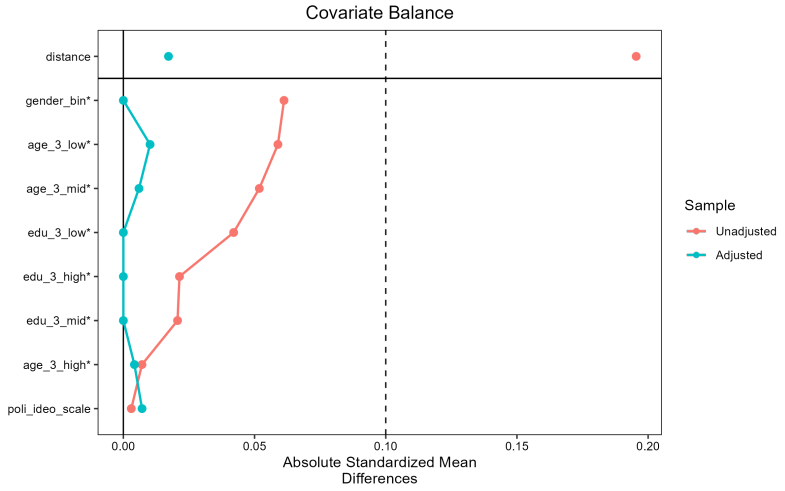
- **Control:** The feed remained unaltered.
- **Elite Blocking:** Posts from political elite accounts were hidden.
- **Elite Scrubbing:** Both posts from political elites and posts mentioning them were removed.

After the intervention period, participants completed the endline survey (Wave 2), which repeated the same outcome measures as the baseline.

Screening process. Twitter user validation was conducted automatically as part of the screening survey. Out of over 2000 who attempted the survey, about 50% were screened out or returned the project. 1,214 participants were ultimately deemed eligible after multiple validation steps. These included verifying usernames via the Twitter API, removing duplicate submissions, and excluding accounts that had not posted within the past year. Initially, private accounts were also rejected, though this restriction was later relaxed to improve inclusivity.



(a) Elite Blocking vs Control



(b) Elite Scrubbing vs Control

Fig. E3: Covariate balance before and after propensity score full matching for each intervention arm relative to the control condition. Points represent standardized mean differences (SMDs) for baseline covariates before and after matching. Vertical reference lines indicate the conventional balance threshold of $|\text{SMD}| = 0.10$.

Participants who submitted high-profile or clearly inauthentic usernames (e.g., `iamsrk`, `elonmusk`) were manually flagged and excluded during post-survey validation. Recruitment began using both CloudResearch Connect and its MTurk Toolkit; however, the MTurk Toolkit yielded a high proportion of invalid or recycled Twitter IDs—over 80% of submissions were unusable due to duplication or impersonation. In

contrast, no such issues were observed with CloudResearch Connect. Consequently, all remaining recruitment was conducted exclusively through Connect.

At the end of the study, participants were debriefed regarding their treatment condition and offered the option to withdraw their data. No participants opted to do so.

E.4 Compliance

We assessed behavioral compliance using logged app usage, defining compliance as using the Twilly app for at least nine days during the study period among the endline respondents. Behavioral compliance was high in all conditions (Elite Blocking: 194/213 = 91.1%; Elite Scrubbing: 181/204 = 88.7%; Control: 227/227 = 100%). Compared with the control group, both Elite Blocking ($\chi^2(1) = 13.42$, p¡.001) and Elite Scrubbing ($\chi^2(1) = 17.62$, p¡.001) showed significantly lower compliance rates. In total, 602 participants met this behavioral compliance criterion.

E.5 Attrition

Table E5 reports the recruitment attrition at every stage of the experiment. Attrition during recruitment and eligibility screening was substantial but broadly similar across conditions, suggesting no strong evidence of differential selection into the experimental sample.

Table E5: Retention summary by study stages

Stage	Total			
Total survey attempts	3638			
Stage	Control	Blocking	Scrubbing	**Total**
Completed the baseline survey	416	406	392	**1214**
Installed the app	327	414	399	**1140**
Daily app users	Mean = 143.62, Median = 128.50			
Eligible for endline survey invitation (used app $\geq$ 7 days)	245	230	214	**689**
Completed endline survey	227	210	198	**648**
Behavioral compliance (used app $\geq$ 9 days)	227	194	181	**602**
Self-reported compliance	95	93	72	**260**

Survey attrition between baseline and endline was 48.1% (Elite Blocking), 48.4% (Elite Scrubbing), and 49.8% (Control), with no significant differential attrition across arms ($\chi^2(2) = 0.27$, $p = .874$). Such levels of attrition have been reported in other longitudinal experiments, especially those involving secondary app use [10–12]. Lee (2009)

trimming bounds and Manski (1990) extreme-value bounds confirm that treatment effects are robust to potential attrition bias (see Section K).

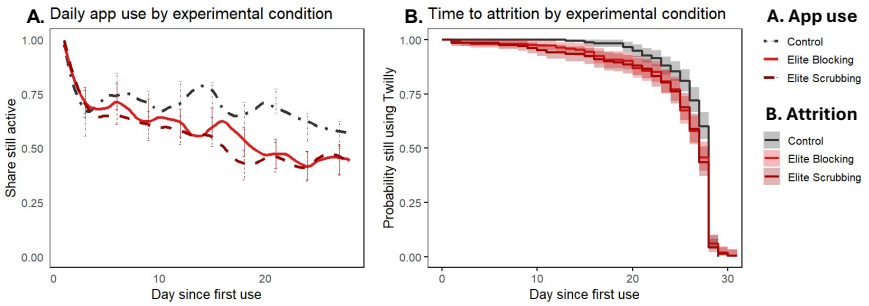


Fig. E4: **App usage and retention over time by experimental condition:** Panel A plots the share of participants still using the Twilly app on each day since first use. Panel B shows Kaplan–Meier survival curves for time to permanent discontinuation. Error bars and Shaded bands represent 95% confidence intervals. Control is shown in black, Elite Blocking in light red, and Elite Scrubbing in dark red.

Appendix F User engagement characteristics

User interactions on the Twilly app were passively recorded to assess exposure and engagement with political content.

F.1 Active and passive engagement

Engagement signals were grouped into three categories based on their type and intensity:

- **Passive engagement:** Signals of content exposure that do not involve contribution, amplification, or visible interaction. These include instances where users intentionally browse or view content but do not engage in ways that alter visibility or provide feedback (e.g., liking or retweeting). Examples include:
  - Viewing the home timeline
  - Exploring content on the Explore page
  - Viewing tweet details
  - Viewing user profiles or tweets
  - Watching videos or images
  - Performing broadcast queries and search timeline navigation
- **Active engagement:** Signals of contribution, amplification, or evaluative feedback that make a user's interaction visible to others or to the platform's algorithms. These actions reflect deliberate participation—such as endorsing, sharing, or saving content—and typically leave a digital trace. Examples include:

– **Favoriting behavior:** Lower-effort active engagement signals such as liking and bookmarking posts
– **Sharing behavior:** Resharing posts

The frequency distribution of user actions logged per treatment condition, across all content, partisan content, and low-credibility content are reported in Table F6.

Table F6: Engagement by treatment and content type

Content type	Treatment	Seen	Favorited	Retweeted
By post (total counts)				
All content	Control	607,183	53,332	4,171
All content	Elite Blocking	455,977	33,456	1,072
All content	Elite Scrubbing	389,425	16,383	1,960
Partisan content	Control	578,979	52,327	4,067
Partisan content	Elite Blocking	444,467	32,884	1,038
Partisan content	Elite Scrubbing	380,674	16,250	1,941
Low-credibility	Control	23,402	1,428	988
Low-credibility	Elite Blocking	17,335	640	23
Low-credibility	Elite Scrubbing	15,183	216	191
By user (median per participant)				
All content	Control	2,106	544	246
All content	Elite Blocking	1,842	468	200
All content	Elite Scrubbing	1,732	410	150
Partisan content	Control	1,015	261	94
Partisan content	Elite Blocking	919	223	77
Partisan content	Elite Scrubbing	865	199	66
Low-credibility	Control	676	148	25
Low-credibility	Elite Blocking	519	88	7
Low-credibility	Elite Scrubbing	468	74	8

F.2 Sessions

Participants were active on the app for a median of 18 days (range: 1-29, SD = 7.8). While participants self-reported spending a median of 1.45 hours per day on the app, observed activity patterns were lower: users in the Control condition were active for a median of 20 days (SD = 6.0) with session durations ranging from 0.67 seconds to 2.25 hours. Usage in the treatment groups was slightly lower but comparable, with median active days of 16 (SD = 6.3) and 15.5 (SD = 6.5), and time ranges of 0.84 seconds to 48.9 minutes and 1.18 seconds to 25.5 minutes in the Elite Blocking and Elite Scrubbing conditions, respectively.

To define user sessions, we first identified different kinds of user actions. The Twilly app logged 35 user actions, of which 21 were performed by at least 10% of unique users. The remaining actions were aggregated into an "Other" category, resulting in 22 final user actions. Next, sequences of these user actions were used to define user sessions. We computed inter-activity time gaps and applied a two-component Gaussian Mixture Model [13]. The inferred thresholds varied by condition, ranging from 209 to

239 seconds. We define a new session when the time between two user actions exceeds the threshold. Sessions that span across calendar days (UTC) were rare (fewer than 5) and excluded from our session-level timing statistics. We computed daily engagement metrics per user, focusing on the number of active days and the median time spent per day. A non-trivial proportion of users exhibited minimal engagement, with a median daily time below 10 seconds—often limited to passive behaviors such as viewing the home timeline.

We computed daily session statistics metrics per user, focusing on the number of active days and the median time spent per day. A non-trivial proportion of users exhibited minimal engagement, with a median daily time below 10 seconds—often limited to passive behaviors such as viewing the home timeline. Summary session statistics for each experimental group are reported in Table F7.

Table F7: Session and engagement statistics by experimental condition

Metric	Control	Elite blocking	Elite scrubbing
Median daily sessions	209	233	239
Median daily active users	215	185	184
Active days (median, SD)	20 (6.0)	16 (6.3)	15.5 (6.5)
Range of daily activity times	0.67s–2.25h	0.84s–48.9m	1.18s–25.5m

Appendix G Summary statistics

Condition-wise summary statistics for the baseline survey are reported in Table G8.

Appendix H All Preregistered Results

H.1 App usage (time spent, tweets viewed, likes, clicks) (RQ1)

To assess whether feed interventions altered overall app usage, we modeled user-level counts of actions using a negative binomial regression with interactions between treatment assignment and usage type (Viewed, Favorited, Clicked, Retweeted). Table H9 (Figure H5) reports marginal differences in expected counts relative to the Control group.

Overall usage declined under both interventions, with particularly pronounced reductions for active behaviors. Elite Scrubbing reduced viewed tweets by 530.5 actions per user (95% CI: [-1041.6, -19.3], $p = .042$), while Elite Blocking produced a negative but noisier estimate that was not statistically distinguishable from zero ($p = .125$). Both interventions reduced favoriting (Elite Blocking: -83.1, 95% CI: [-146.7, -19.6], $p = .010$; Elite Scrubbing: -174.8, 95% CI: [-229.9, -119.8], $p < .001$) and click-based exploration (Elite Blocking: -137.2, 95% CI: [-244.7, -29.7], $p = .012$; Elite Scrubbing: -300.0, 95% CI: [-393.1, -206.9], $p < .001$). Retweeting also declined, significantly so under Elite Blocking (-30.1, 95% CI: [-42.8, -17.4], $p < .001$), with a smaller and marginal reduction under Elite Scrubbing ($p = .057$).

Table G8: Baseline summary statistics for each outcome variable by treatment group. Cells report mean, median (standard deviation).

Variable	Elite Blocking	Elite Scrubbing	Control
Knowledge (RQ7)			
International news knowledge	10.27, 11.00 (5.41)	9.79, 9.00 (5.07)	8.56, 8.00 (5.65)
National news knowledge	7.41, 7.00 (2.51)	7.14, 7.00 (2.45)	6.56, 7.00 (3.00)
Misinformation susceptibility (RQ7)			
Political rumors	14.30, 14.00 (2.19)	14.50, 15.00 (2.13)	12.90, 14.00 (4.38)
Multimodal misinformation	3.06, 3.00 (0.19)	3.03, 3.00 (0.19)	2.54, 2.60 (0.53)
Confidence in misinformation assessment	19.10, 19.00 (3.75)	19.10, 19.00 (4.04)	18.10, 18.00 (3.85)
Affective polarization (RQ3–RQ5)			
Affective polarization (feeling thermometer)	43.6, 40 (28.8)	44.1, 50 (31.6)	49.1, 50 (29.3)
Perceived partisan threat	−0.05, −0.11 (0.88)	−0.04, −0.11 (0.90)	0.02, −0.11 (0.90)
Social identity	2.87, 2.88 (0.92)	2.92, 2.88 (0.88)	2.88, 2.88 (0.91)
Social distance	3.18, 3.00 (1.12)	3.10, 3.00 (1.17)	3.09, 3.00 (1.19)
Perceived polarization	2.62, 3.00 (0.96)	2.64, 2.50 (0.96)	2.61, 3.00 (1.01)
Political attitudes (RQ6)			
Political interest	3.65, 4.00 (1.14)	3.79, 4.00 (1.05)	3.95, 4.00 (0.99)
Institutional trust	2.18, 2.00 (0.79)	2.16, 2.00 (0.81)	2.17, 2.00 (0.82)
Media trust & news avoidance (RQ8)			
News avoidance	0.12, 0.11 (0.95)	0.12, 0.11 (0.92)	−0.13, −0.17 (0.92)
Media trust	0.03, 0.13 (0.88)	0.03, 0.12 (0.90)	−0.00, 0.12 (0.89)
News skepticism	−0.06, −0.13 (0.84)	0.00, 0.14 (0.85)	0.04, 0.11 (0.80)

Taken together, the interventions reduced absolute activity on the app, with the largest and most consistent declines occurring for interactive engagement (favoriting and clicking) rather than passive viewing. This pattern aligns with the session-level metrics, which similarly show contraction in engagement intensity.

Table H9: Marginal Differences in Absolute Usage Counts Relative to Control. p_{BH}: Benjamini–Hochberg adjusted across 8 tests.

Usage Type	**Contrast**	**Estimate (95% CI)**	p	p_{BH}
Viewed	Elite Blocking vs Control	−408.9 [−931.6, 113.8]	0.125	0.125
	Elite Scrubbing vs Control	−530.5 [−1041.6, −19.3]	0.042	0.056
Favorited	Elite Blocking vs Control	−83.1 [−146.7, −19.6]	0.010	0.019
	Elite Scrubbing vs Control	−174.8 [−229.9, −119.8]	$< .001$	$< .001$
Clicked	Elite Blocking vs Control	−137.2 [−244.7, −29.7]	0.012	0.019
	Elite Scrubbing vs Control	−300.0 [−393.1, −206.9]	$< .001$	$< .001$
Retweeted	Elite Blocking vs Control	−30.1 [−42.8, −17.4]	$< .001$	$< .001$
	Elite Scrubbing vs Control	−14.8 [−30.1, 0.4]	0.057	0.065

Table H10: Effects of Feed Interventions on Political Content by Usage Type (Percentage Points)

Outcome	Usage Type	Contrast	Estimate (95% CI)	p-value
Low-Credibility	Viewed	Elite Blocking vs Control	−5.92 [−8.51, −3.33]	< .001
		Elite Scrubbing vs Control	−5.75 [−8.35, −3.14]	< .001
	Favorited	Elite Blocking vs Control	−1.74 [−4.58, 1.09]	.227
		Elite Scrubbing vs Control	−0.59 [−3.47, 2.29]	.690
	Retweeted	Elite Blocking vs Control	−6.79 [−10.98, −2.59]	.002
		Elite Scrubbing vs Control	−3.44 [−7.81, 0.92]	.122
	Clicked	Elite Blocking vs Control	−4.88 [−7.53, −2.24]	< .001
		Elite Scrubbing vs Control	−5.67 [−8.33, −3.01]	< .001
Partisan Content (|ideology| > 0.3)	Viewed	Elite Blocking vs Control	−13.68 [−17.52, −9.84]	< .001
		Elite Scrubbing vs Control	−14.02 [−17.88, −10.16]	< .001
	Favorited	Elite Blocking vs Control	−2.45 [−6.65, 1.75]	.252
		Elite Scrubbing vs Control	−4.08 [−8.35, 0.19]	.061
	Retweeted	Elite Blocking vs Control	−20.36 [−26.58, −14.15]	< .001
		Elite Scrubbing vs Control	−20.33 [−26.80, −13.86]	< .001
	Clicked	Elite Blocking vs Control	−12.04 [−15.96, −8.12]	< .001
		Elite Scrubbing vs Control	−11.55 [−15.49, −7.61]	< .001

H.2 Sharing behaviors (RQ2)

Table H10 reports percentage-point differences in user interaction with low-credibility and partisan political content across engagement types (Viewed, Favorited, Retweeted, Clicked), relative to the Control group. These results have been discussed in the main text (Figure 2).

H.3 Affective polarization (RQ3-RQ5)

Table H11 (Figure H6) reports treatment effects on multiple dimensions of political polarization. Across both the full sample (intent-to-treat) and the complied-only sample, none of the estimated effects reach conventional levels of statistical significance. Point estimates are generally small in magnitude and confidence intervals are wide, consistently spanning zero.

For affective polarization and perceived polarization, estimates are positive but imprecisely estimated in both treatment conditions. Effects on social identity, social distance, and perceived political discrimination are near zero and statistically indistinguishable from the Control group. The complied-only estimates mirror the intent-to-treat pattern, with slightly larger point estimates in some cases but similarly overlapping confidence intervals.

Taken together, these results provide no evidence that removing elite-authored or elite-referential political content produced measurable short-term changes in affective or perceived dimensions of political polarization.

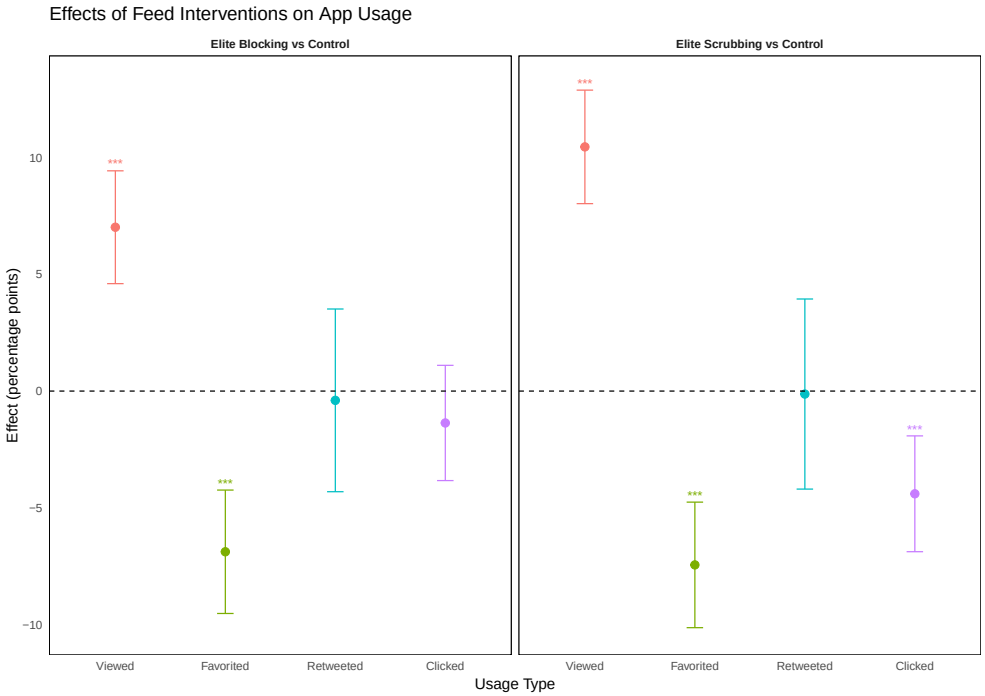


Fig. H5: Effects of Feed Interventions on Absolute Usage Counts.

H.4 Political Interest and Political Trust (RQ6)

Table H12 (Figure H7) reports treatment effects on political interest and institutional trust. Across both the full sample and the complied-only sample, none of the estimated effects reach conventional levels of statistical significance. Point estimates for political interest are small in magnitude, with Elite Blocking showing a modest positive association and Elite Scrubbing showing near-zero or slightly negative effects. For institutional trust, estimates are negative under both interventions, particularly under Elite Scrubbing, but confidence intervals consistently include zero. Overall, the results provide no evidence that suppressing elite-authored or elite-referential political content produced measurable short-term changes in political interest or trust in government.

H.5 Political Knowledge and Misinformation (RQ7)

The main findings related to the effects on political knowledge and misinformation, reported in Figure 3, are presented in Table H13.

Knowledge. Elite Blocking increased international news knowledge by 0.479 SD (95% CI [0.294, 0.663], $p < .001$) and national news knowledge by 0.282 SD (95% CI [0.095, 0.469], $p = .003$). Elite Scrubbing increased international news knowledge by 0.324 SD (95% CI [0.139, 0.508], $p < .001$) and national news knowledge by 0.239 SD (95% CI [0.052, 0.426], $p = .012$).

Table H11: Effects of Feed Interventions on Polarization and Intergroup Attitudes

Outcome	Treatment	Estimate (95% CI)	p
All participants (N = 648)			
Affective Polarization	Elite Blocking	0.142 [−0.054, 0.337]	0.155
Affective Polarization	Elite Scrubbing	0.102 [−0.093, 0.296]	0.304
Perceived Partisan Threat	Elite Blocking	−0.018 [−0.208, 0.171]	0.848
Perceived Partisan Threat	Elite Scrubbing	0.040 [−0.150, 0.230]	0.680
Social Identity	Elite Blocking	0.002 [−0.188, 0.193]	0.980
Social Identity	Elite Scrubbing	0.033 [−0.157, 0.224]	0.732
Social Distance	Elite Blocking	0.008 [−0.182, 0.197]	0.937
Social Distance	Elite Scrubbing	−0.018 [−0.208, 0.172]	0.853
Perceived Polarization	Elite Blocking	0.157 [−0.033, 0.348]	0.104
Perceived Polarization	Elite Scrubbing	0.077 [−0.113, 0.267]	0.427
Perceived Political Discrimination	Elite Blocking	−0.062 [−0.252, 0.128]	0.522
Perceived Political Discrimination	Elite Scrubbing	0.047 [−0.143, 0.237]	0.627
Complied (N = 602)			
Affective Polarization	Elite Blocking	0.128 [−0.073, 0.329]	0.212
Affective Polarization	Elite Scrubbing	0.156 [−0.046, 0.357]	0.129
Perceived Partisan Threat	Elite Blocking	−0.022 [−0.215, 0.171]	0.825
Perceived Partisan Threat	Elite Scrubbing	−0.011 [−0.206, 0.183]	0.909
Social Identity	Elite Blocking	−0.026 [−0.220, 0.168]	0.792
Social Identity	Elite Scrubbing	0.009 [−0.186, 0.205]	0.927
Social Distance	Elite Blocking	−0.000 [−0.193, 0.192]	0.998
Social Distance	Elite Scrubbing	−0.030 [−0.224, 0.164]	0.762
Perceived Polarization	Elite Blocking	0.135 [−0.057, 0.328]	0.169
Perceived Polarization	Elite Scrubbing	0.100 [−0.094, 0.294]	0.310
Perceived Political Discrimination	Elite Blocking	−0.089 [−0.282, 0.105]	0.368
Perceived Political Discrimination	Elite Scrubbing	0.062 [−0.133, 0.257]	0.532

Awareness (Misinformation susceptibility). Elite Blocking reduced susceptibility to political rumors by 0.416 SD (95% CI [−0.602, −0.230], $p < .001$), and for multimodal misinformation by 0.655 SD (95% CI [−0.834, −0.476], $p < .001$). Elite Scrubbing produced comparable reductions: 0.423 SD for political rumors (95% CI [−0.609, −0.237], and $p < .001$), and by 0.679 SD for multimodal misinformation (95% CI [−0.859, −0.500], $p < .001$).

Belief (Confidence in judgments). Elite Blocking reduced confidence in multimodal misinformation judgments by 0.341 SD (95% CI [−0.528, −0.154], $p < .001$), while Elite Scrubbing reduced confidence in multimodal by 0.387 SD (95% CI [−0.574, −0.200], $p < .001$).

Overall, the interventions increase factual knowledge while simultaneously reducing susceptibility to political and multimodal misinformation, with larger effects under Elite Scrubbing.

H.6 Media Trust and News Avoidance (RQ8)

Table H14 (Figure H8) reports treatment effects on news avoidance, media trust, and news skepticism. Across both the full sample (intent-to-treat) and the complied-only

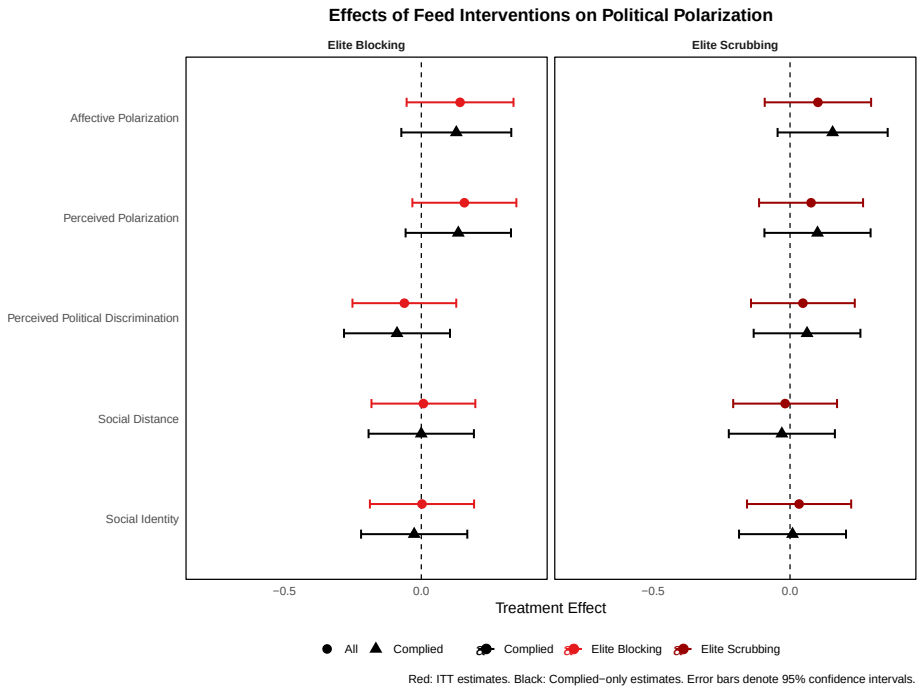


Fig. H6: Effects of Feed Interventions on Affective Polarization.

sample, none of the estimated effects reach conventional levels of statistical significance. Point estimates for news avoidance are negative under both interventions, suggesting modest reductions, but confidence intervals include zero. Effects on media trust are positive but imprecisely estimated, while estimates for news skepticism are small and near zero in magnitude.

The complied-only analyses closely mirror the intent-to-treat results, with slightly larger point estimates in some cases but similarly wide confidence intervals that span zero. Overall, the interventions do not appear to produce measurable short-term changes in these downstream news-related attitudes.

H.7 Partisan Differences (RQ9)

Table H15 and Figure H9 report heterogeneous treatment effects by partisan group across five preregistered knowledge and misinformation outcomes. Across outcomes, the direction of treatment effects is largely consistent across partisan groups. For knowledge measures (International and National News Events), both partisan groups exhibit positive treatment effects, with somewhat larger gains under Elite Blocking. For misinformation outcomes—including Political Rumors, Multimodal Misinformation, and Confidence in Multimodal Misinformation—effects are negative for both partisan groups, indicating reduced susceptibility under treatment. Importantly, differences between partisan groups are modest in magnitude and not consistently

Table H12: Effects of Feed Interventions on Political Content by Usage Type (Percentage Points). p_{BH}: Benjamini–Hochberg adjusted across 16 tests.

Outcome	Usage Type	Contrast	Estimate (95% CI)	p	p_{BH}
Low-Credibility	Viewed	Elite Blocking vs Control	−5.92 [−8.51, −3.33]	< .001	< .001
	Viewed	Elite Scrubbing vs Control	−5.75 [−8.35, −3.14]	< .001	< .001
	Favorited	Elite Blocking vs Control	−1.74 [−4.58, 1.09]	0.227	0.259
	Favorited	Elite Scrubbing vs Control	−0.59 [−3.47, 2.29]	0.690	0.690
	Retweeted	Elite Blocking vs Control	−6.79 [−10.98, −2.59]	0.002	0.003
	Retweeted	Elite Scrubbing vs Control	−3.44 [−7.81, 0.92]	0.122	0.150
	Clicked	Elite Blocking vs Control	−4.88 [−7.53, −2.24]	< .001	< .001
	Clicked	Elite Scrubbing vs Control	−5.67 [−8.33, −3.01]	< .001	< .001
Partisan Content (|ideology| > 0.3)	Viewed	Elite Blocking vs Control	−13.68 [−17.52, −9.84]	< .001	< .001
	Viewed	Elite Scrubbing vs Control	−14.02 [−17.88, −10.16]	< .001	< .001
	Favorited	Elite Blocking vs Control	−2.45 [−6.65, 1.75]	0.252	0.269
	Favorited	Elite Scrubbing vs Control	−4.08 [−8.35, 0.19]	0.061	0.081
	Retweeted	Elite Blocking vs Control	−20.36 [−26.58, −14.15]	< .001	< .001
	Retweeted	Elite Scrubbing vs Control	−20.33 [−26.80, −13.86]	< .001	< .001
	Clicked	Elite Blocking vs Control	−12.04 [−15.96, −8.12]	< .001	< .001
	Clicked	Elite Scrubbing vs Control	−11.55 [−15.49, −7.61]	< .001	< .001

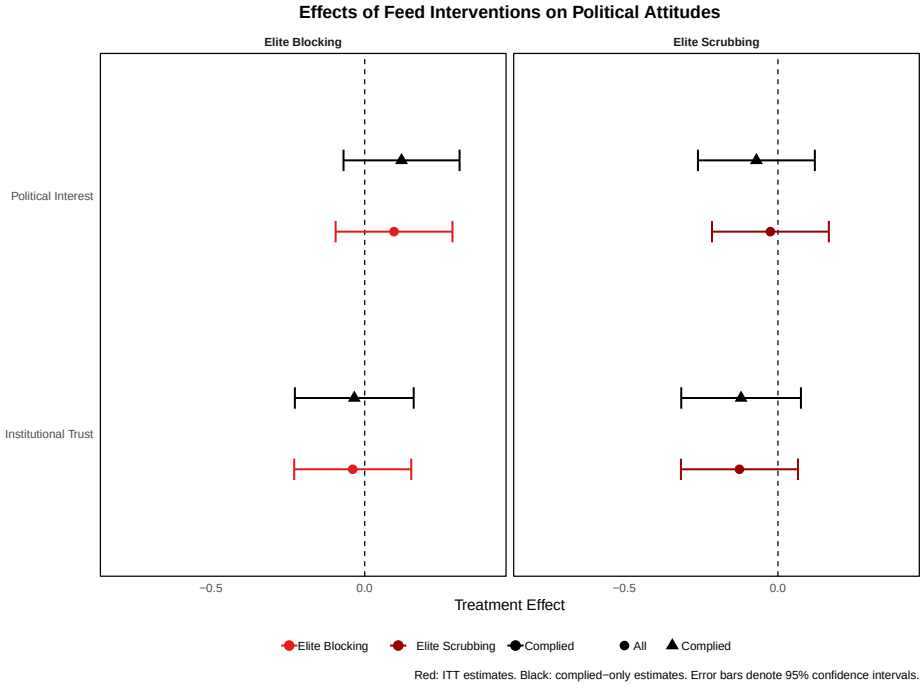


Fig. H7: Effects of Feed Interventions on Political Interest and Political Trust.

Table H13: Regression results for epistemic outcomes (Intent-to-Treat; N = 648). Estimates are OLS coefficients with 95% confidence intervals. Models include treatment indicators and covariates for gender, age, education, income, and political ideology. Asterisks denote significance levels (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ Benjamini–Hochberg adjusted across 20 treatment tests.)

Outcome	Term	Estimate	CI Lower	CI Upper	p	Stars
	International News Knowledge					
	(Intercept)	-0.271	-0.638	0.097	0.148	
	Elite Blocking	0.479	0.294	0.663	$< .001$	***
	Elite Scrubbing	0.324	0.139	0.508	$< .001$	***
	Male	0.051	-0.103	0.206	0.514	
	Age	-0.014	-0.080	0.051	0.667	
	Education	0.051	-0.012	0.114	0.112	
	Income	0.017	-0.054	0.087	0.644	
	Ideology	-0.065	-0.108	-0.022	0.003	**
	National News Knowledge					
	(Intercept)	-0.336	-0.708	0.035	0.076	
	Elite Blocking	0.282	0.095	0.469	0.003	**
	Elite Scrubbing	0.239	0.052	0.426	0.012	*
	Male	-0.001	-0.158	0.155	0.989	
	Age	0.061	-0.005	0.128	0.071	
	Education	0.046	-0.018	0.110	0.159	
	Income	-0.007	-0.078	0.065	0.851	
	Ideology	-0.068	-0.112	-0.024	0.002	**
	Political Rumors					
	(Intercept)	0.235	-0.135	0.605	0.213	
	Elite Blocking	-0.416	-0.602	-0.230	$< .001$	***
	Elite Scrubbing	-0.423	-0.609	-0.237	$< .001$	***
	Male	-0.021	-0.177	0.135	0.790	
	Age	0.005	-0.061	0.071	0.886	
	Education	-0.026	-0.090	0.037	0.420	
	Income	0.007	-0.064	0.079	0.838	
	Ideology	0.034	-0.009	0.078	0.122	
	Multimodal Misinformation					
	(Intercept)	0.577	0.221	0.933	0.002	**
	Elite Blocking	-0.655	-0.834	-0.476	$< .001$	***
	Elite Scrubbing	-0.679	-0.859	-0.500	$< .001$	***
	Male	-0.159	-0.309	-0.009	0.038	*
	Age	-0.016	-0.080	0.047	0.612	
	Education	0.002	-0.059	0.064	0.937	
	Income	-0.030	-0.099	0.039	0.391	
	Ideology	0.029	-0.013	0.071	0.182	
	Confidence in Misinformation					
	(Intercept)	0.346	-0.026	0.718	0.069	
	Elite Blocking	-0.341	-0.528	-0.154	$< .001$	***
	Elite Scrubbing	-0.387	-0.574	-0.200	$< .001$	***
	Male	0.144	-0.012	0.301	0.071	
	Age	-0.030	-0.096	0.037	0.383	
	Education	0.005	-0.059	0.069	0.889	
	Income	-0.040	-0.111	0.032	0.278	
	Ideology	0.009	-0.035	0.053	0.685	

Table H14: Effects of Feed Interventions on Political Attitudes

Outcome	Treatment	Estimate (95% CI)	p
All participants (N = 648)			
News Avoidance	Elite Blocking	−0.140 [−0.329, 0.049]	0.146
News Avoidance	Elite Scrubbing	−0.151 [−0.339, 0.038]	0.118
Media Trust	Elite Blocking	0.131 [−0.058, 0.320]	0.174
Media Trust	Elite Scrubbing	0.079 [−0.110, 0.269]	0.411
News Skepticism	Elite Blocking	−0.034 [−0.224, 0.156]	0.723
News Skepticism	Elite Scrubbing	0.006 [−0.184, 0.196]	0.948
Complied (N = 602)			
News Avoidance	Elite Blocking	−0.151 [−0.345, 0.043]	0.126
News Avoidance	Elite Scrubbing	−0.163 [−0.358, 0.032]	0.102
Media Trust	Elite Blocking	0.117 [−0.076, 0.311]	0.235
Media Trust	Elite Scrubbing	0.126 [−0.069, 0.321]	0.206
News Skepticism	Elite Blocking	−0.060 [−0.255, 0.136]	0.550
News Skepticism	Elite Scrubbing	0.045 [−0.152, 0.241]	0.657

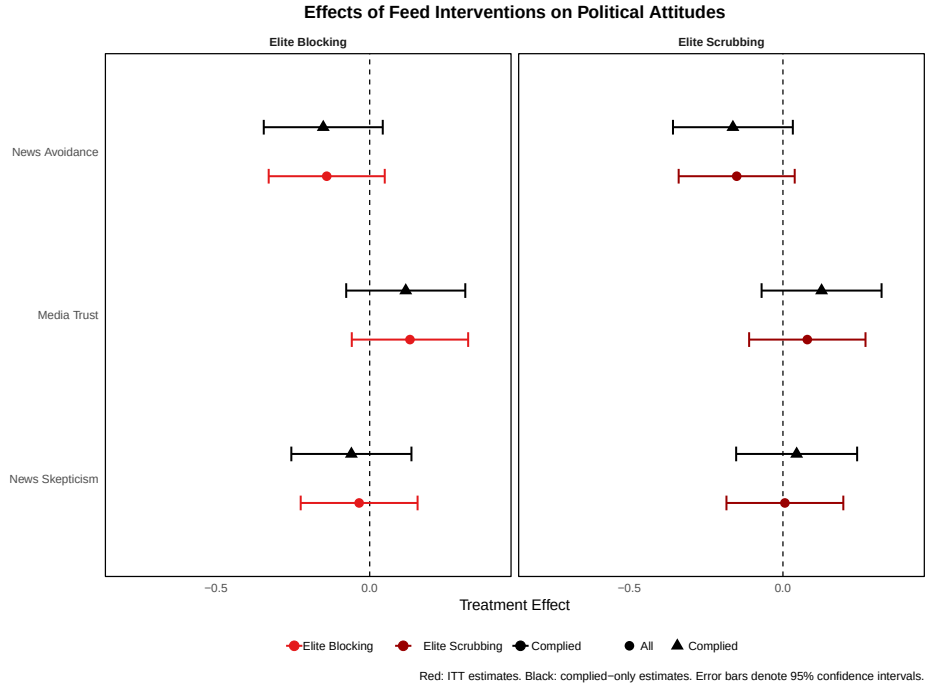


Fig. H8: Effects of Feed Interventions on Political Attitudes.

statistically distinguishable from one another. While point estimates occasionally vary across Democrats and Republicans, the overall pattern suggests broadly similar treatment responses rather than strong partisan divergence. Taken together, the

interventions appear to improve knowledge and reduce misinformation susceptibility in both partisan groups, with limited evidence of substantial partisan moderation.

Table H15: Heterogeneous Treatment Effects by Partisanship (Pre-registered RQ9)

Outcome	Treatment	Estimate	SE	95% CI	p
International News Events					
	Elite Blocking	0.295	0.115	[0.070, 0.520]	0.010
	Elite Scrubbing	0.197	0.114	[-0.026, 0.421]	0.084
	Elite Blocking	0.913	0.179	[0.563, 1.263]	<0.001
	Elite Scrubbing	0.614	0.184	[0.255, 0.974]	<0.001
National News Events					
	Elite Blocking	0.161	0.116	[-0.066, 0.388]	0.164
	Elite Scrubbing	0.153	0.115	[-0.072, 0.379]	0.183
	Elite Blocking	0.568	0.180	[0.215, 0.921]	0.002
	Elite Scrubbing	0.427	0.185	[0.064, 0.791]	0.021
Political Rumors					
	Elite Blocking	-0.255	0.115	[-0.482, -0.029]	0.027
	Elite Scrubbing	-0.310	0.115	[-0.535, -0.085]	0.007
	Elite Blocking	-0.671	0.179	[-1.023, -0.320]	<0.001
	Elite Scrubbing	-0.656	0.184	[-1.018, -0.295]	<0.001
Multimodal Misinformation					
	Elite Blocking	-0.617	0.111	[-0.835, -0.400]	<0.001
	Elite Scrubbing	-0.712	0.110	[-0.927, -0.496]	<0.001
	Elite Blocking	-0.577	0.172	[-0.914, -0.239]	<0.001
	Elite Scrubbing	-0.519	0.178	[-0.868, -0.171]	0.004
Confidence in Multimodal Misinformation					
	Elite Blocking	-0.421	0.116	[-0.649, -0.194]	<0.001
	Elite Scrubbing	-0.398	0.116	[-0.625, -0.172]	<0.001
	Elite Blocking	-0.195	0.181	[-0.549, 0.160]	0.282
	Elite Scrubbing	-0.373	0.186	[-0.738, -0.009]	0.045

Appendix I Auxiliary Findings

I.1 Feed analysis

Most participants reported following few elite accounts: 71.6% indicated that they did not follow any right-leaning elite accounts, and 47.7% reported that they did not follow any left-leaning elite accounts. Among those who did follow elites, left-leaning elites were more common than right-leaning ones: 31.8% of participants followed three or more left-leaning elites, compared to only 3.8% for right-leaning elites. These measures were based on a small, predefined list of prominent political elites and therefore provide an indicative rather than comprehensive measure of users' elite-following behavior.

In contrast, the feed-based exposure analysis, reported in Figure 2 in the main text, provides higher recall by enabling us to measure the volume of political content that users actually encountered in their timelines, including content originating from

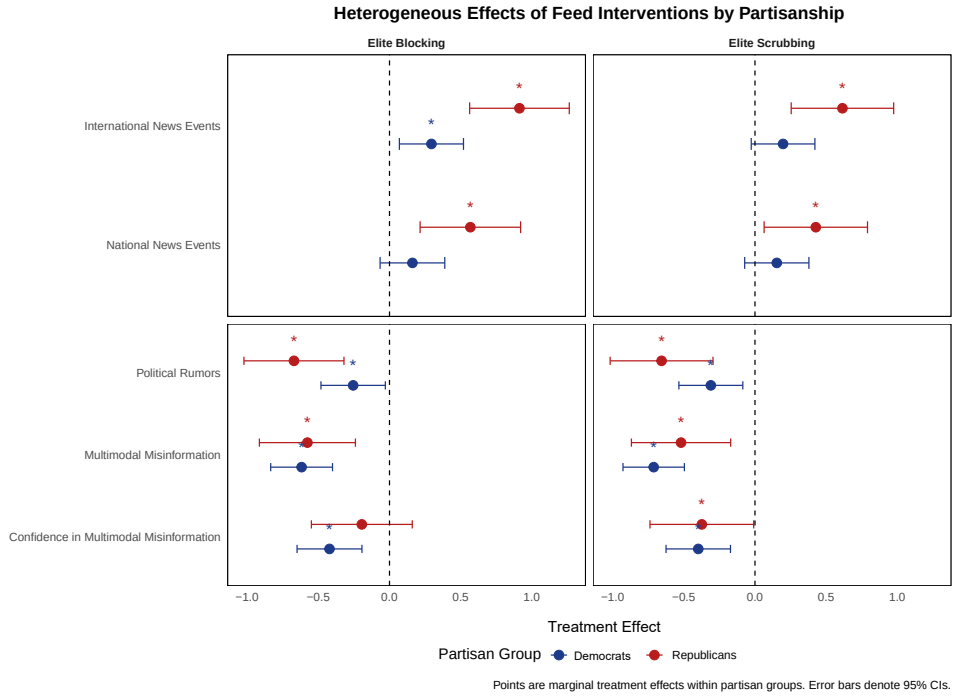


Fig. H9: Heterogeneous treatment effects by partisan group. Points represent marginal effects within partisan subgroups; error bars denote 95% confidence intervals. Positive values indicate increases in knowledge scores, while negative values indicate reductions in misinformation susceptibility.

accounts they do not directly follow. Among **left- and left-leaning participants (417 participants, or 64.35% of the total)**, the Control group saw on average 586 tweets/day, of which 6.79% were political. Under Elite Blocking, volume dropped to 433/day (6.79% political), and under Elite Scrubbing to 350/day (6.35% political). Among **right- and right-leaning participants (164 participants, 25.30% of the total)**, the Control group averaged seeing 422 tweets/day (9.21% political), compared to 336/day (9.14%) under Elite Blocking and 270/day (7.25%) under Elite Scrubbing. In short, the interventions thinned the feed without substantially altering the overall balance of political content.

To assess whether the interventions shifted the ideological composition of content beyond volume changes, we compared the distribution of user-level mean ideology scores across conditions (Table I16). For Democrats, no significant differences emerged across conditions (Kruskal-Wallis $p = .081$; all pairwise $p > .10$; all KS $D < 0.14$). For Republicans, the omnibus test was significant ($p = .016$), driven by a difference between Elite Scrubbing and Control ($p = .019$, $r = 0.29$). Post-level breakdowns suggest this is a mechanical consequence of which content each intervention removes: Elite Blocking disproportionately reduced centrist and left-leaning content in Republican feeds (82.9% centrist in Control vs. 79.9% in Blocking), while Elite Scrubbing

increased the centrist share (88.3%) by broadly thinning right-leaning material as well. However, user-level sample sizes for Republicans are modest (n = 52–65 per condition), and the absolute differences in mean ideology are small (0.157–0.201), so these shifts should be interpreted with caution.

Panel C of Figure 2 in the main text, breaks down exposure by provenance—content from followed accounts versus links shared by followed accounts—with full estimates reported in Table I17. Treatment effects are concentrated in first-degree exposure from followed accounts. Both interventions significantly reduce partisan and low-credibility content from followed accounts (all $p < .0001$), with Elite Scrubbing producing substantially larger reductions in low-credibility exposure than Elite Blocking ($\Delta = 3.2$ pp, $p < .0001$), consistent with the hypothesized elite-referential amplification channel. By contrast, changes through linked domains are comparatively small (all $|\Delta| < 0.24$ pp) and in some cases nonsignificant. Overall, the interventions primarily reshape content flowing through users' immediate follow networks rather than indirect exposure pathways.

Table I16: User-level ideological composition across conditions. Mean ideology score per user, compared across treatment conditions within each partisanship group. Kruskal-Wallis omnibus test and pairwise Wilcoxon rank-sum tests (Holm-corrected) are reported alongside rank biserial effect sizes (r) and Kolmogorov-Smirnov D statistics.

Partisanship	Contrast	n per arm	Wilcoxon p	Rank biserial r	KS D	KS p
Democrats (Kruskal-Wallis $\chi^2 = 5.03$, $p = .081$)						
	Control – Blocking	65 / 58	.69	0.03 [−0.10, 0.16]	0.06	.95
	Control – Scrubbing	65 / 52	.17	−0.11 [−0.24, 0.01]	0.12	.20
	Blocking – Scrubbing	58 / 52	.10	0.14 [0.01, 0.27]	0.14	.13
Republicans (Kruskal-Wallis $\chi^2 = 8.23$, $p = .016$)						
	Control – Blocking	65 / 58	.12	0.20 [−0.00, 0.39]	0.29	.008
	Control – Scrubbing	65 / 52	.019	0.29 [0.09, 0.47]	0.28	.017
	Blocking – Scrubbing	58 / 52	.30	−0.12 [−0.32, 0.10]	0.27	.025

I.2 Twitter behavioral covariates

We conducted a series of supplementary analyses to examine (a) whether the main treatment effects are robust to controlling for the number of tweets filtered by the intervention, (b) whether filtered tweets mediate the treatment effects via an instrumental variable (IV) approach, and (c) whether pre-app political engagement moderates the treatment effects.

I.2.1 Filtered tweets as a covariate

Table I18 reports the main treatment effects after adding the number of filtered tweets as a covariate. The pattern of results is virtually unchanged from the primary analysis:

Table I17: Exposure pathway effects relative to control (content-level). Estimates represent changes in probability of exposure. Stars indicate significance: * $p < .05$, ** $p < .01$, *** $p < .001$.

Pathway	Contrast	Estimate	SE	95% CI	Sig.
Partisan content					
Follow network	Control – Elite Blocking	−0.062	0.002	[−0.068, −0.057]	***
	Control – Elite Scrubbing	−0.051	0.002	[−0.057, −0.046]	***
	Blocking – Scrubbing	0.011	0.003	[0.004, 0.017]	***
Linked domains	Control – Elite Blocking	−0.002	0.000	[−0.003, −0.001]	***
	Control – Elite Scrubbing	−0.001	0.000	[−0.002, 0.000]	
	Blocking – Scrubbing	0.001	0.000	[0.000, 0.002]	**
Low-credibility content					
Follow network	Control – Elite Blocking	−0.027	0.002	[−0.032, −0.022]	***
	Control – Elite Scrubbing	−0.059	0.002	[−0.065, −0.054]	***
	Blocking – Scrubbing	−0.032	0.003	[−0.038, −0.026]	***
Linked domains	Control – Elite Blocking	−0.002	0.000	[−0.003, −0.002]	***
	Control – Elite Scrubbing	−0.002	0.000	[−0.002, −0.001]	***
	Blocking – Scrubbing	0.001	0.000	[0.000, 0.001]	*

both Elite Blocking (FEP01A) and Elite Scrubbing (FEP01D) significantly improved knowledge scores and reduced misinformation susceptibility across all five outcomes (all $p < .05$). The robustness of these effects to the inclusion of the filtering covariate suggests that the treatment effects are not confounded by differential content filtering.

Next we used an IV approach to test whether the number of filtered tweets mediates the treatment effects on misinformation outcomes, with treatment assignment as the instrument and filtered tweet count as the endogenous variable. The first stage confirmed that the instruments are strong ($F = 143.41$, $p < .001$), though the relationship is driven almost entirely by the Elite Scrubbing condition ($\beta = 1.24$, $p < .001$), with no significant first-stage effect for Elite Blocking ($\beta = 0.11$, $p = .180$).

Table I19 reports the second-stage (2SLS) estimates. The number of filtered tweets significantly predicted reductions in political rumors ($\beta = -0.209$, $p = .003$), multimodal misinformation susceptibility ($\beta = -0.339$, $p < .001$), and confidence in misinformation ($\beta = -0.205$, $p = .005$). Wu–Hausman tests confirmed that the IV correction is warranted for these three outcomes ($p < .05$), indicating that OLS estimates of the filtering–outcome relationship are biased. The IV estimates were not significant for the two knowledge score outcomes, suggesting that content filtering operates primarily on misinformation susceptibility rather than factual knowledge.

I.2.2 Moderation by pre-app political engagement

We tested whether two indicators of pre-app political engagement moderated treatment effects: (a) the proportion of political tweets posted by participants in the three

Table I18: Treatment effects with filtered tweet count as covariate. Standardized coefficients from OLS regression with demographic controls. Asterisks denote significance levels (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$), Benjamini–Hochberg adjusted across 20 treatment tests.

Outcome	Treatment	β	SE	95% CI	p
Knowledge scores					
International News Events	Elite Blocking	0.481	0.094	[0.296, 0.666]	$< .001^{***}$
	Elite Scrubbing	0.354	0.110	[0.138, 0.571]	$.001^{**}$
National News Events	Elite Blocking	0.281	0.095	[0.094, 0.469]	$.003^{**}$
	Elite Scrubbing	0.230	0.112	[0.011, 0.449]	$.040^{*}$
Misinformation susceptibility					
Political Rumors	Elite Blocking	−0.413	0.095	[−0.600, −0.227]	$< .001^{***}$
	Elite Scrubbing	−0.394	0.111	[−0.612, −0.176]	$< .001^{***}$
Multimodal Misinfo	Elite Blocking	−0.652	0.091	[−0.832, −0.473]	$< .001^{***}$
	Elite Scrubbing	−0.647	0.107	[−0.857, −0.437]	$< .001^{***}$
Confidence in Misinfo	Elite Blocking	−0.346	0.095	[−0.534, −0.159]	$< .001^{***}$
	Elite Scrubbing	−0.448	0.112	[−0.667, −0.228]	$< .001^{***}$

Table I19: Instrumental variable (2SLS) estimates. Endogenous variable: filtered tweet count. Instrument: treatment assignment. Asterisks denote significance levels (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$), Benjamini–Hochberg adjusted across 20 treatment tests.

Outcome	β_{2SLS}	SE	95% CI	p	Wu–Hausman p
International News Events	0.104	0.071	[−0.034, 0.243]	.140	.128
National News Events	0.102	0.071	[−0.036, 0.240]	.148	.264
Political Rumors	−0.209	0.071	[−0.349, −0.070]	$.003^{**}$	$.029^{*}$
Multimodal Misinfo	−0.339	0.072	[−0.480, −0.199]	$< .001^{***}$	$< .001^{***}$
Confidence in Misinfo	−0.205	0.072	[−0.347, −0.064]	$.005^{**}$	$.003^{**}$

months before the study (operationalized as the percentage of tweets with a political content score ≥ 0.15, see Appendix D.3.2), and (b) the total number of tweets posted in the same period (June–August 2024). Each moderator was interacted with treatment assignment in separate regressions for all five outcomes, yielding 10 interaction tests (Table I20). After applying Benjamini–Hochberg correction, no interaction remained significant, suggesting that the treatment effects are broadly uniform across levels of prior political engagement on Twitter.

Appendix J Results with weighted estimates

To evaluate the sensitivity of the intent-to-treat (ITT) results to residual covariate imbalance, we re-estimated treatment effects using inverse-probability weighting. Balance diagnostics indicate that standardized mean differences were reduced below conventional thresholds following weighting (see Figure E2).

Table I20: Moderation of treatment effects by pre-app political engagement. Only interaction terms shown. p_{BH}: Benjamini–Hochberg adjusted p-values across 10 tests. No interaction survived correction.

Outcome	Moderator	β	SE	95% CI	p	p_{BH}
Political tweet proportion						
Intl News	Block. × pol. tw.	0.249	0.208	[−0.16, 0.66]	.232	.580
Natl News	Block. × pol. tw.	−0.020	0.211	[−0.43, 0.39]	.924	.924
Pol. Rumors	Block. × pol. tw.	−0.066	0.210	[−0.48, 0.35]	.753	.924
Multimod.	Block. × pol. tw.	0.187	0.202	[−0.21, 0.58]	.356	.712
Conf. Mis.	Block. × pol. tw.	0.205	0.211	[−0.21, 0.62]	.333	.712
Pre-app tweet volume						
Intl News	Block. × tw. vol.	−0.536	0.456	[−1.43, 0.36]	.240	.580
Natl News	Block. × tw. vol.	−0.426	0.463	[−1.33, 0.48]	.358	.712
Pol. Rumors	Block. × tw. vol.	0.262	0.461	[−0.64, 1.17]	.570	.814
Multimod.	Block. × tw. vol.	−0.771	0.442	[−1.64, 0.10]	.082	.410
Conf. Mis.	Block. × tw. vol.	1.067	0.461	[0.16, 1.97]	.021	.209

Knowledge outcomes

The knowledge effects observed in the ITT models are sensitive to covariate adjustment and lose statistical robustness after weighting. In the unweighted ITT models, both interventions were associated with significant increases in knowledge, as reported in the main findings. However, after weighting, these knowledge effects attenuated and were no longer statistically distinguishable from zero. For international knowledge, Elite Blocking (b = 2.26, 95% CI [0.720, 3.79], p_{adj} = .005) and Elite Scrubbing (b = 1.15, 95% CI [-0.051, 2.35], p_{adj} = .076) show reduced precision, with confidence intervals crossing zero in the latter case. For national knowledge, Elite Blocking (b = 1.20, 95% CI [0.150, 2.25], p_{adj} = .025) and Elite Scrubbing (b = 0.672, 95% CI [-0.155, 1.50], p_{adj} = .111) similarly weaken, with Scrubbing no longer significant. Overall, once covariate imbalance is addressed, the evidence for knowledge gains becomes inconsistent and statistically fragile.

Misinformation outcomes

In contrast, the reductions in misinformation susceptibility remain robust. In the ITT models, Elite Blocking and Elite Scrubbing significantly reduced political rumor susceptibility, as reported in the main findings. These patterns persist under weighting. Political rumor susceptibility remains significantly reduced (Blocking: b = -1.54, 95% CI [-2.49, -0.595], p_{adj} = .004; Scrubbing: b = -1.24, 95% CI [-2.02, -0.454], p_{adj} = .003). Multimodal misinformation remains significantly lower (Blocking: b = -0.402, 95% CI [-0.529, -0.274], p_{adj} ¡ .001; Scrubbing: b = -0.383, 95% CI [-0.535, -0.230], p_{adj} ¡ .001). Confidence in misinformation is likewise reduced (Blocking: b = -1.36, 95% CI [-2.22, -0.492], p_{adj} = .004; Scrubbing: b = -1.90, 95% CI [-2.71, -1.09], p_{adj} ¡ .001).

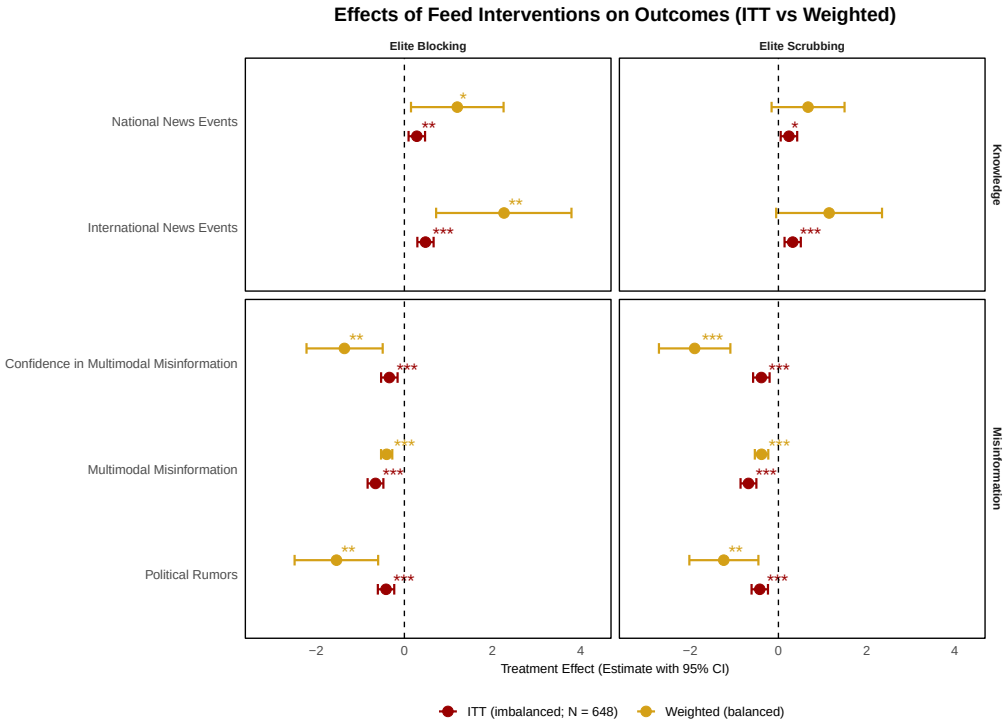


Fig. J10: Weighted average treatment effects on the treated (ATT) for each outcome and intervention arm relative to the control condition. Points denote point estimates and horizontal bars indicate 95% confidence intervals. Estimates are obtained from weighted regressions on matched samples. p-values are adjusted for multiple testing using the Benjamini–Hochberg false discovery rate procedure.

Appendix K Robustness checks

K.1 Validation of feed scores

To validate our findings, we compared our primary measures of ideology and credibility against three independent sources: Media Bias/Fact Check (MBFC) [14], AllSides [15].

Because AllSides indexes outlets by name rather than domain, we used a SerpAPI-based name-to-domain lookup to bridge AllSides entries to our domain key. MBFC bias ratings, which include numeric scores on a $[-10, 10]$ scale, were rescaled to $[-1, 1]$; AllSides categorical ratings were mapped as Left $= -1$, Lean Left $= -0.5$, Center $= 0$, Lean Right $= 0.5$, Right $= 1$. For credibility, low credibility was coded as 1 if the MBFC credibility field contained "Low" or the outlet was categorized as "fake-news" or "conspiracy,".

Table J21: Estimated treatment effects for Arm 1 (FEP01A vs TRD158B). Estimates are reported with 95% confidence intervals and both nominal and multiple-testing-adjusted p-values (p.adj).

Domain	Outcome	N	Est.	SE	95% CI	p / p.adj
Knowledge	International News Events	424	2.26	0.784	[0.720, 3.79]	0.00399 / 0.00499
Knowledge	National News Events	424	1.20	0.537	[0.150, 2.25]	0.0251 / 0.0251
Susceptibility	Political Rumors	424	-1.54	0.484	[-2.49, -0.595]	0.00142 / 0.00353
Susceptibility	Multimodal Misinformation	424	-0.402	0.0652	[-0.529, -0.274]	7.39e-10 / 3.69e-09
Susceptibility	Confidence in Multimodal Misinformation	424	-1.36	0.442	[-2.22, -0.492]	0.00212 / 0.00353

Table J22: Estimated treatment effects for Arm 2 (FEP01D vs TRD158B). Estimates are reported with 95% confidence intervals and both nominal and multiple-testing-adjusted p-values (p.adj).

Domain	Outcome	N	Est.	SE	95% CI	p / p.adj
Knowledge	International News Events	415	1.15	0.614	[-0.051, 2.35]	0.0606 / 0.0758
Knowledge	National News Events	415	0.672	0.422	[-0.155, 1.50]	0.111 / 0.111
Susceptibility	Political Rumors	415	-1.24	0.399	[-2.02, -0.454]	0.00194 / 0.00323
Susceptibility	Multimodal Misinformation	414	-0.383	0.0779	[-0.535, -0.230]	8.79e-07 / 4.39e-06
Susceptibility	Confidence in Multimodal Misinformation	415	-1.90	0.414	[-2.71, -1.09]	4.47e-06 / 1.12e-05

Domain-level validation.

We matched 19,037 unique reference domains against the external sources. Coverage varied across the three validation sources: MBFC provided labels for 10% of domains and AllSides for less than 1%. Pairwise consistency was evaluated using Spearman's ρ for ideology and Cohen's κ for credibility across all available overlaps between each source and our measure, rather than restricting analyses to the small intersection of items with complete labels.

As seen in Table K23, across all available overlaps, correlations ranged from $\rho = 0.66$ to 0.87, indicating moderate to strong directional agreement among the independent coding sources. MBFC, with the largest independent coverage at 10%, showed moderate-to-strong correlation ($r = 0.700$), with the higher MAE likely reflecting differences in how categorical labels are mapped to a continuous scale. AllSides showed strong correlations (r¿0.83) despite limited coverage. These results suggest that despite

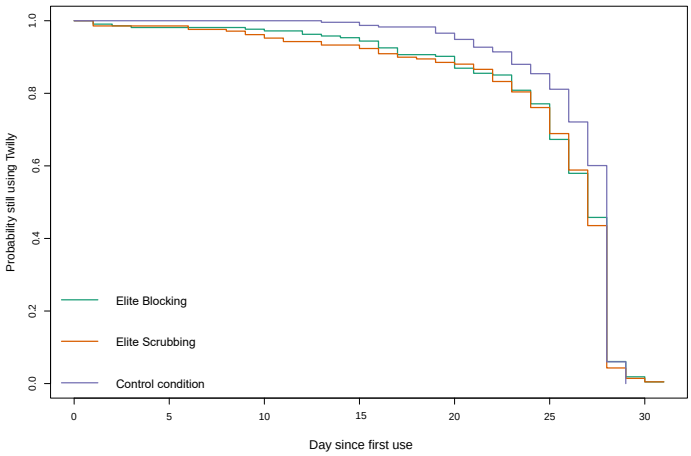


Fig. J11: **Time to permanent discontinuation by experimental condition.** Kaplan–Meier survival curves showing the probability that participants remain active in Twilly as a function of days since first use, separately for each treatment group. Discontinuation is defined as permanent cessation of use within the 30-day intervention window.

Table K23: Domain-level ideology validation against external sources.

Source	n overlap	Coverage (%)	Pearson r	Spearman ρ	MAE
MBFC	1,889	10.0	0.700	0.656	0.267
AllSides	64	0.3	0.887	0.870	0.343

Table K24: Domain-level low-credibility validation against external sources.

Source	n overlap	Coverage (%)	Agreement (%)	κ	TP	FP	FN	TN
MBFC (credibility)	2,272	11.9	98.6	0.920	206	26	6	2,034
MBFC (category)	18,980	99.7	99.1	0.715	210	26	138	18,606

varying domain coverage, the different bias estimation approaches capture a consistent underlying ideological dimension.

For low credibility (Table K24), MBFC's credibility field achieved excellent agreement ($\kappa = 0.920$) with near-zero false negatives, indicating that our scores reliably flag the same outlets as low credibility. The MBFC category-based measure—flagging outlets listed under "fake-news" or "conspiracy"—covered nearly all reference domains and showed strong agreement ($\kappa = 0.715$), with most disagreements arising from false

negatives: outlets our measure flags as low credibility that MBFC does not categorize under those two labels.

Handle-level validation.

For handle-level validation, we bridged X handles to domains using a compiled handle-to-domain mapping, covering 1,385 of 4,078 reference handles, and then applied domain-level external scores. Tables K25 and K26 report the results.

Handle-level ideology correlations were moderate ($r \approx 0.34$–0.48), reflecting both the indirectness of the domain bridge and the fact that many handles represent individual accounts whose ideology may diverge from their affiliated outlet. Coverage was limited to 3–4% of handles with ideology scores.

In contrast, handle-level credibility validation was strong. The domain bridge provided roughly one-third coverage, and agreement with MBFC ($\kappa = 0.948$–0.958) was excellent. This suggests that credibility, unlike ideology, transfers reliably from the outlet domain to associated handles.

Table K25: Handle-level ideology validation against external sources (via domain bridge).

Source	n overlap	Coverage (%)	Pearson r	Spearman ρ	MAE
AllSides	35	2.7	0.482	0.533	0.751
MBFC	41	3.2	0.376	0.331	0.448

Table K26: Handle-level low-credibility validation against external sources (via domain bridge).

Source	n overlap	Coverage (%)	Agreement (%)	κ	TP	FP	FN	TN
MBFC (credibility)	1,313	32.2	99.7	0.958	48	3	1	1,261
MBFC (category)	1,318	32.3	99.6	0.948	47	3	2	1,266

Summary.

Our ideology scores show strong convergent validity at the domain level, with correlations ranging from $\rho = 0.66$ to 0.87 across independently constructed scales. Our low-credibility classifications are well-validated at both the domain and handle levels, with Cohen's κ exceeding 0.92 against MBFC for domains and exceeding 0.95 for handles. Handle-level ideology validation remains limited by sparse external coverage and the inherent looseness of the handle-to-domain mapping, though the moderate correlations observed are consistent with expectations given the indirectness of the comparison.

K.2 Treatment effects on user attrition

To assess whether feed-filtering interventions affected user retention, we analyzed time to permanent discontinuation using a Cox proportional hazards regression, with experimental condition as the sole predictor and the control group as the reference category. Time was measured in days since each participant's first recorded use of the application. Figure J11 shows the corresponding Kaplan-Meier survival curves.

The model indicates modest differences in discontinuation rates across experimental arms. Participants in the Elite Scrubbing condition exhibited about 23% higher hazard of permanent discontinuation relative to the control condition (HR = 1.23, 95% CI [1.02, 1.48], p = .032). Participants in the Elite Blocking condition showed about 15% higher hazard of permanent discontinuation relative to the control, but this difference was not statistically significant (HR = 1.15, 95% CI [0.95, 1.39], p = .14). The overall likelihood ratio test for treatment effects was not significant (χ^2 = 4.91, p = .09).

Substantively, these results indicate that users in the intervention conditions, particularly the Elite Scrubbing condition, discontinued use slightly earlier than users in the control group. However, the magnitude of these differences is small, and the survival curves diverge primarily in the later portion of the intervention window. Importantly, the main epistemic effects reported in the study emerge despite these modest differences in attrition, suggesting that treatment effects are not driven by selective retention of more engaged users but occur in the presence of slightly higher interface friction under the intervention conditions.

K.3 Attrition Bounds Analysis

As reported in the main text, survey attrition did not differ significantly across arms. To further assess whether outcome missingness could bias treatment-effect estimates, we computed Manski [16] extreme-value bounds and Lee [17] trimming bounds using the `attrition` package [18]. Manski bounds replace missing outcomes with the best- and worst-case values in the observed range, yielding conservative intervals that, as expected, include zero for all outcomes (Table K27).

Lee trimming bounds provide tighter identification by trimming the group with higher response rates to equalize completion across arms. Because attrition was nearly identical across conditions, trimming affected very few observations, and the resulting bounds are narrow. Importantly, for all five outcomes under both treatment arms, the Lee bounds exclude zero (Table K28): bounds for international and national news knowledge are entirely positive, indicating treatment-group advantages, while bounds for political rumors susceptibility, multimodal misinformation susceptibility, and confidence in misinformation are entirely negative, indicating treatment-group reductions. These results suggest that the main treatment effects reported in the paper are robust to potential attrition bias.

K.4 Sensitivity to minimum active-days threshold

Because exposure to the client-side feed intervention depended on actual app use, treatment assignment did not imply uniform exposure across participants. To assess

Table K27: Manski Extreme-Value Bounds on Treatment Effects

Outcome	Treatment	Lower	Upper	CI Lower	CI Upper
International News Knowledge	Blocking	−20.68	23.34	−22.22	24.78
	Scrubbing	−21.16	23.00	−22.70	24.45
National News Knowledge	Blocking	−12.70	13.72	−13.61	14.58
	Scrubbing	−12.80	13.69	−13.72	14.56
Political Rumors	Blocking	−12.99	11.47	−13.77	12.33
	Scrubbing	−13.03	11.51	−13.82	12.38
Multimodal Misinformation	Blocking	−2.34	1.87	−2.47	2.01
	Scrubbing	−2.35	1.88	−2.48	2.03
Confidence in Misinformation	Blocking	−17.84	16.40	−18.88	17.51
	Scrubbing	−18.01	16.33	−19.05	17.47

Note. Manski bounds assume missing outcomes can take any value within the observed range. 95% CIs reported.

Table K28: Lee Trimming Bounds on Treatment Effects

Outcome	Treatment	Lower Bound	Upper Bound
International News Knowledge	Blocking	1.94	2.50
	Scrubbing	1.32	1.79
National News Knowledge	Blocking	0.62	1.03
	Scrubbing	0.56	0.91
Political Rumors	Blocking	−1.52	−1.17
	Scrubbing	−1.55	−1.22
Multimodal Misinformation	Blocking	−0.50	−0.41
	Scrubbing	−0.49	−0.44
Confidence in Misinformation	Blocking	−1.66	−0.92
	Scrubbing	−1.89	−1.31

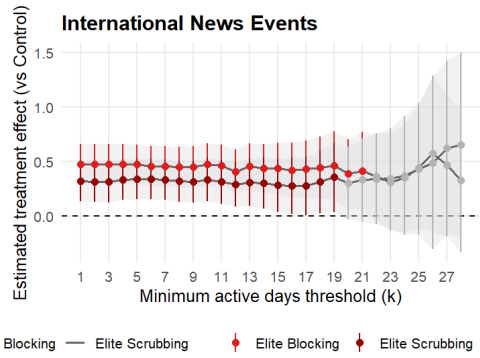
Note. Lee bounds trim the better-responding group to equalize response rates. All bounds exclude zero, indicating treatment effects are robust to attrition.

whether the primary ITT results are sensitive to differential engagement, we re-estimated the main outcome models after progressively restricting the analysis sample to participants with at least k active days of app use, for all observed values of k. Each model follows the primary specification, with standardized outcomes, treatment indicators, and the full set of pre-treatment covariates. This threshold-based analysis evaluates whether estimated treatment effects are robust to increasingly stringent minimum usage requirements.

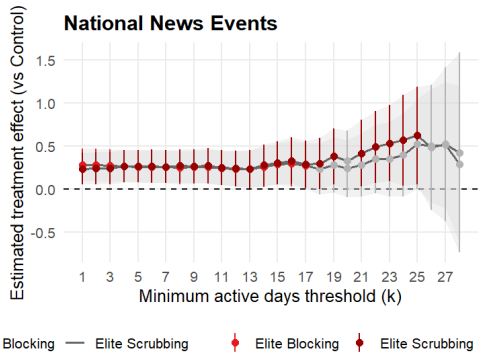
Figure K12 presents the resulting threshold-based sensitivity plots. Across outcomes, estimates are stable in sign and broadly consistent in magnitude across a wide range of active-days thresholds, with uncertainty increasing at both low exposure levels and the most restrictive thresholds. In particular, statistical significance is attenuated when exposure is minimal or when sample sizes become small, reflecting reduced precision rather than systematic changes in effect direction. Knowledge outcomes are generally positive for both interventions, with Elite Blocking exhibiting more frequent statistical significance.

As expected, susceptibility to political rumors, multimodal misinformation, and confidence in multimodal misinformation outcomes are predominantly negative across

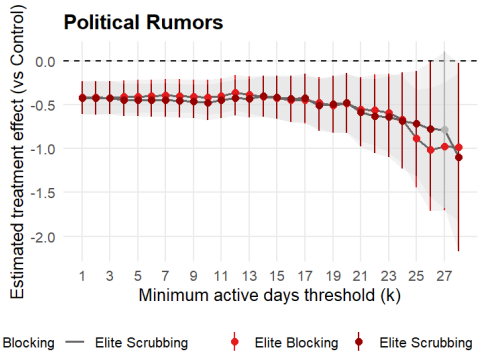
thresholds for both interventions. Overall, the results indicate that the primary conclusions are not driven by any single choice of minimum usage threshold and are robust to reasonable variation in sample restrictions based on participant engagement.



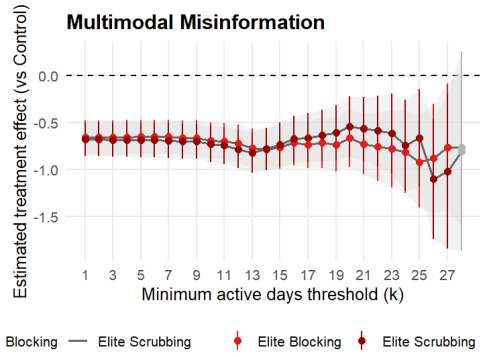
(a) International News Events



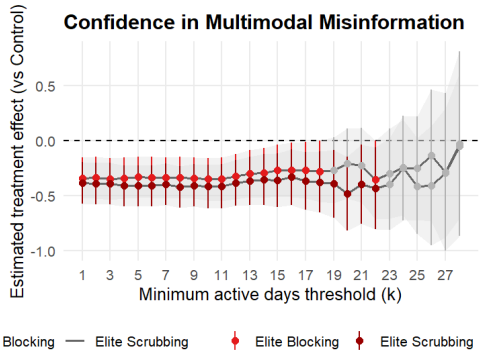
(b) National News Events



(c) Political Rumors



(d) Multimodal Misinformation



(e) Confidence in Multimodal Misinformation

Fig. K12: Sensitivity of treatment effects to minimum active-days thresholds. Each point shows the ITT estimate from the primary regression re-estimated on the subsample of participants with at least k active days of app use. Vertical bars indicate treatment-specific 95% confidence intervals. Points are colored by treatment when statistically significant at $p < 0.05$ and shown in gray otherwise. The dashed horizontal line denotes the null effect.

Table K29: Treatment Effects on Temporally Overlapping Cohort (±3 Days)

Outcome	Treatment	b	95% CI	p	
Panel A: All participants (N = 334)					
International News Knowledge	Elite Blocking	0.607	[0.265, 0.949]	<.001	***
	Elite Scrubbing	0.374	[0.033, 0.715]	.032	*
National News Knowledge	Elite Blocking	0.029	[−0.326, 0.384]	.873	
	Elite Scrubbing	−0.080	[−0.434, 0.274]	.657	
Political Rumors	Elite Blocking	−0.613	[−0.958, −0.268]	<.001	***
	Elite Scrubbing	−0.599	[−0.943, −0.254]	<.001	***
Multimodal Misinformation	Elite Blocking	−0.505	[−0.853, −0.157]	.005	**
	Elite Scrubbing	−0.585	[−0.932, −0.238]	.001	**
Confidence in Misinformation	Elite Blocking	−0.656	[−1.004, −0.308]	<.001	***
	Elite Scrubbing	−0.633	[−0.980, −0.286]	<.001	***
Panel B: Complied participants (N = 305)					
International News Knowledge	Elite Blocking	0.567	[0.228, 0.907]	.001	**
	Elite Scrubbing	0.331	[−0.009, 0.671]	.056	
National News Knowledge	Elite Blocking	−0.011	[−0.376, 0.354]	.952	
	Elite Scrubbing	−0.089	[−0.454, 0.277]	.633	
Political Rumors	Elite Blocking	−0.587	[−0.941, −0.233]	.001	**
	Elite Scrubbing	−0.586	[−0.941, −0.232]	.001	**
Multimodal Misinformation	Elite Blocking	−0.507	[−0.863, −0.150]	.005	**
	Elite Scrubbing	−0.620	[−0.977, −0.263]	<.001	***
Confidence in Misinformation	Elite Blocking	−0.663	[−1.015, −0.312]	<.001	***
	Elite Scrubbing	−0.639	[−0.991, −0.287]	<.001	***

Note. OLS estimates on standardized outcomes. Sample restricted to participants whose first day of app activity fell within ±3 days of at least one participant in a different treatment arm. Compliance defined as ≥9 active days. All models include covariates for gender, age, education, income, and political ideology. * $p < .05$, ** $p < .01$, *** $p < .001$.

K.5 Sensitivity to recruitment timing

Because participants were recruited in staggered waves—control in wave 1 (September–October 2024) and treatment arms in waves 2–3 (October–November 2024)—the U.S. general election (November 5, 2024) fell within the treatment window for later-recruited participants but not for most control participants. We conduct four complementary checks to assess whether this temporal non-overlap drives the observed treatment effects.

Start-date overlap cohort (±3 days).

We first re-estimate the main models restricting to participants whose first day of app activity fell within ±3 days of at least one participant in a different treatment arm (ITT $N = 334$; Complied $N = 305$). Across misinformation outcomes, the overlap-cohort estimates (Figure K14, Table K29) closely track the full-sample patterns. Political rumor susceptibility remains strongly negative in both arms (ITT: Elite Blocking $b = -0.61$, 95% CI $[-0.96, -0.27]$, $p < .001$; Elite Scrubbing $b = -0.60$, 95% CI $[-0.94, -0.25]$, $p < .001$), as do multimodal misinformation susceptibility (Elite Blocking $b = -0.50$, $p = .005$; Elite Scrubbing $b = -0.59$, $p = .001$) and confidence in misinformation judgments (Elite Blocking $b = -0.66$, $p < .001$; Elite Scrubbing

Table K30: **Calendar-week fixed effects robustness check.** Comparison of baseline (Eq. 2) and calendar-week FE specifications. All models include demographic covariates. Outcomes are standardized.

		Baseline (Eq. 2)		Calendar-Week FE		
Outcome	Treatment	b	p	b [95% CI]	p	Δb
Panel A: ITT ($N = 644$)						
International News	Elite Blocking	$+0.48^{***}$	<.001	$+0.98$ $[+0.24, +1.71]^{**}$	.009	+0.50
	Elite Scrubbing	$+0.32^{***}$	<.001	$+0.85$ $[+0.12, +1.58]^{*}$	.023	+0.53
National News	Elite Blocking	$+0.28^{**}$	.003	$+0.52$ $[-0.23, +1.26]$	.172	+0.24
	Elite Scrubbing	$+0.24^{*}$	.012	$+0.50$ $[-0.24, +1.24]$	.185	+0.26
Political Rumors	Elite Blocking	-0.41^{***}	<.001	-1.32 $[-2.05, -0.58]^{***}$	<.001	−0.90
	Elite Scrubbing	-0.42^{***}	<.001	-1.35 $[-2.08, -0.62]^{***}$	<.001	−0.93
Multimodal Misinfo	Elite Blocking	-0.66^{***}	<.001	-0.77 $[-1.49, -0.04]^{*}$	.038	−0.11
	Elite Scrubbing	-0.67^{***}	<.001	-0.77 $[-1.50, -0.05]^{*}$	.036	−0.10
Confidence in Misinfo	Elite Blocking	-0.33^{***}	<.001	-0.98 $[-1.72, -0.24]^{**}$	.009	−0.64
	Elite Scrubbing	-0.38^{***}	<.001	-1.05 $[-1.79, -0.32]^{**}$	.005	−0.67
Panel B: Complied ($N = 602$)						
International News	Elite Blocking	$+0.45^{***}$	<.001	$+0.97$ $[+0.23, +1.72]^{*}$	.010	+0.52
	Elite Scrubbing	$+0.32^{**}$	.001	$+0.83$ $[+0.09, +1.58]^{*}$	.028	+0.52
National News	Elite Blocking	$+0.26^{**}$	.009	$+0.49$ $[-0.27, +1.25]$	.207	+0.23
	Elite Scrubbing	$+0.27^{**}$	.007	$+0.50$ $[-0.26, +1.26]$	.198	+0.23
Political Rumors	Elite Blocking	-0.40^{***}	<.001	-1.32 $[-2.07, -0.57]^{***}$	<.001	−0.92
	Elite Scrubbing	-0.46^{***}	<.001	-1.37 $[-2.12, -0.63]^{***}$	<.001	−0.91
Multimodal Misinfo	Elite Blocking	-0.66^{***}	<.001	-0.77 $[-1.50, -0.03]^{*}$	.040	−0.11
	Elite Scrubbing	-0.70^{***}	<.001	-0.80 $[-1.53, -0.07]^{*}$	.033	−0.10
Confidence in Misinfo	Elite Blocking	-0.34^{***}	<.001	-0.97 $[-1.71, -0.23]^{*}$	.011	−0.63
	Elite Scrubbing	-0.41^{***}	<.001	-1.02 $[-1.76, -0.28]^{**}$	.007	−0.61

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. Δb = change in point estimate when adding calendar-week FE.

$b = -0.63$, $p < .001$). International news knowledge effects remain significant for Elite Blocking ($b = +0.61$, $p < .001$) and reach significance for Elite Scrubbing in the ITT sample ($b = +0.37$, $p = .032$). National news knowledge effects are not significant in this restricted sample, consistent with the weaker full-sample effects for this outcome.

Calendar-week fixed effects.

A more direct test adds ISO-calendar-week-of-start-date fixed effects to the main specification (Eq. 2), absorbing any time-varying shocks—including the election—that affected all participants who began the study in the same week. As shown in

Table K31: **Treatment × start-date interaction test.** Interaction coefficients from models including Treatment × Start Day (continuous, days since earliest participant). A significant interaction would indicate that treatment effects vary with recruitment timing. Joint F-test assesses both interaction terms simultaneously.

		Interaction Term			Joint F-test	
Outcome	Treatment	b	95% CI	p	F	p
Panel A: ITT ($N = 644$)						
International News	Elite Blocking	+0.05	[−0.04, +0.14]	.315	0.80	.449
	Elite Scrubbing	+0.04	[−0.05, +0.13]	.370		
National News	Elite Blocking	−0.01	[−0.10, +0.08]	.798	0.12	.884
	Elite Scrubbing	−0.01	[−0.10, +0.08]	.843		
Political Rumors	Elite Blocking	−0.11*	[−0.20, −0.02]	.018	2.82	.060
	Elite Scrubbing	−0.11*	[−0.20, −0.02]	.019		
Multimodal Misinfo	Elite Blocking	+0.01	[−0.07, +0.10]	.755	0.67	.513
	Elite Scrubbing	+0.02	[−0.07, +0.11]	.649		
Confidence in Misinfo	Elite Blocking	−0.07	[−0.16, +0.02]	.125	1.38	.253
	Elite Scrubbing	−0.07	[−0.17, +0.02]	.109		
Panel B: Complied ($N = 602$)						
International News	Elite Blocking	+0.05	[−0.04, +0.14]	.295	0.66	.516
	Elite Scrubbing	+0.04	[−0.05, +0.13]	.331		
National News	Elite Blocking	−0.01	[−0.10, +0.08]	.818	0.06	.942
	Elite Scrubbing	−0.01	[−0.10, +0.08]	.848		
Political Rumors	Elite Blocking	−0.11*	[−0.20, −0.02]	.015	3.13	.045*
	Elite Scrubbing	−0.11*	[−0.20, −0.02]	.013		
Multimodal Misinfo	Elite Blocking	+0.01	[−0.08, +0.10]	.771	0.55	.578
	Elite Scrubbing	+0.02	[−0.07, +0.11]	.669		
Confidence in Misinfo	Elite Blocking	−0.07	[−0.16, +0.02]	.143	1.35	.260
	Elite Scrubbing	−0.07	[−0.16, +0.02]	.120		

*$p < .05$. Non-significant joint F-tests indicate no evidence that effects vary with timing.

For political rumors, the negative interaction direction implies effects are *stronger* for later recruits.

Figure K13 (Table K30), treatment effects on misinformation outcomes remain statistically significant and in the same direction across all three measures. For political rumor susceptibility, the calendar-week FE estimates are $b = -1.32$ ($p < .001$) for Elite Blocking and $b = -1.35$ ($p < .001$) for Elite Scrubbing, compared with baseline estimates of $b = -0.41$ and $b = -0.42$, respectively. For multimodal misinformation susceptibility, both arms remain significant (Elite Blocking $b = -0.77$, $p = .038$; Elite Scrubbing $b = -0.77$, $p = .036$), as do confidence in misinformation judgments (Elite Blocking $b = -0.98$, $p = .009$; Elite Scrubbing $b = -1.05$, $p = .005$). Point estimates are larger in magnitude with calendar-week FE, suggesting that if anything, temporal confounding attenuated rather than inflated the baseline misinformation effects.

Table K32: **Treatment effects in strict window-overlap cohort (≥14 days).** Restricted to participants whose full 28-day treatment window overlaps by at least 14 days with at least one participant in a different arm. All models include demographic covariates. Outcomes are standardized.

Outcome	Treatment	b	95% CI	p
Panel A: ITT ($N = 550$)				
International News	Elite Blocking	$+0.40^{***}$	$[+0.20, +0.60]$	$<.001$
	Elite Scrubbing	$+0.33^{**}$	$[+0.13, +0.53]$	.001
National News	Elite Blocking	$+0.31^{**}$	$[+0.11, +0.52]$	.002
	Elite Scrubbing	$+0.26^{*}$	$[+0.06, +0.46]$	.012
Political Rumors	Elite Blocking	-0.37^{***}	$[-0.58, -0.17]$	$<.001$
	Elite Scrubbing	-0.41^{***}	$[-0.61, -0.21]$	$<.001$
Multimodal Misinfo	Elite Blocking	-0.61^{***}	$[-0.81, -0.42]$	$<.001$
	Elite Scrubbing	-0.67^{***}	$[-0.86, -0.48]$	$<.001$
Confidence in Misinfo	Elite Blocking	-0.41^{***}	$[-0.61, -0.21]$	$<.001$
	Elite Scrubbing	-0.44^{***}	$[-0.64, -0.24]$	$<.001$
Panel B: Complied ($N = 518$)				
International News	Elite Blocking	$+0.36^{***}$	$[+0.16, +0.57]$	$<.001$
	Elite Scrubbing	$+0.30^{**}$	$[+0.09, +0.51]$	.005
National News	Elite Blocking	$+0.28^{**}$	$[+0.07, +0.49]$	.009
	Elite Scrubbing	$+0.27^{*}$	$[+0.06, +0.48]$	.012
Political Rumors	Elite Blocking	-0.36^{***}	$[-0.57, -0.15]$	$<.001$
	Elite Scrubbing	-0.41^{***}	$[-0.62, -0.20]$	$<.001$
Multimodal Misinfo	Elite Blocking	-0.61^{***}	$[-0.82, -0.41]$	$<.001$
	Elite Scrubbing	-0.70^{***}	$[-0.90, -0.49]$	$<.001$
Confidence in Misinfo	Elite Blocking	-0.43^{***}	$[-0.64, -0.22]$	$<.001$
	Elite Scrubbing	-0.46^{***}	$[-0.67, -0.26]$	$<.001$

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

The wider confidence intervals reflect the additional degrees of freedom consumed by the week indicators. International news knowledge remains significant (Elite Blocking $b = +0.98$, $p = .009$; Elite Scrubbing $b = +0.85$, $p = .023$), while national news knowledge becomes non-significant ($p = .172$ and $p = .185$), consistent with the pattern observed in the start-date overlap cohort.

Treatment × start-date interaction.

We also test directly whether treatment effect magnitudes vary with recruitment timing by interacting the treatment indicators with a continuous start-date variable (days since the earliest participant's start date). If the election or other temporal shocks

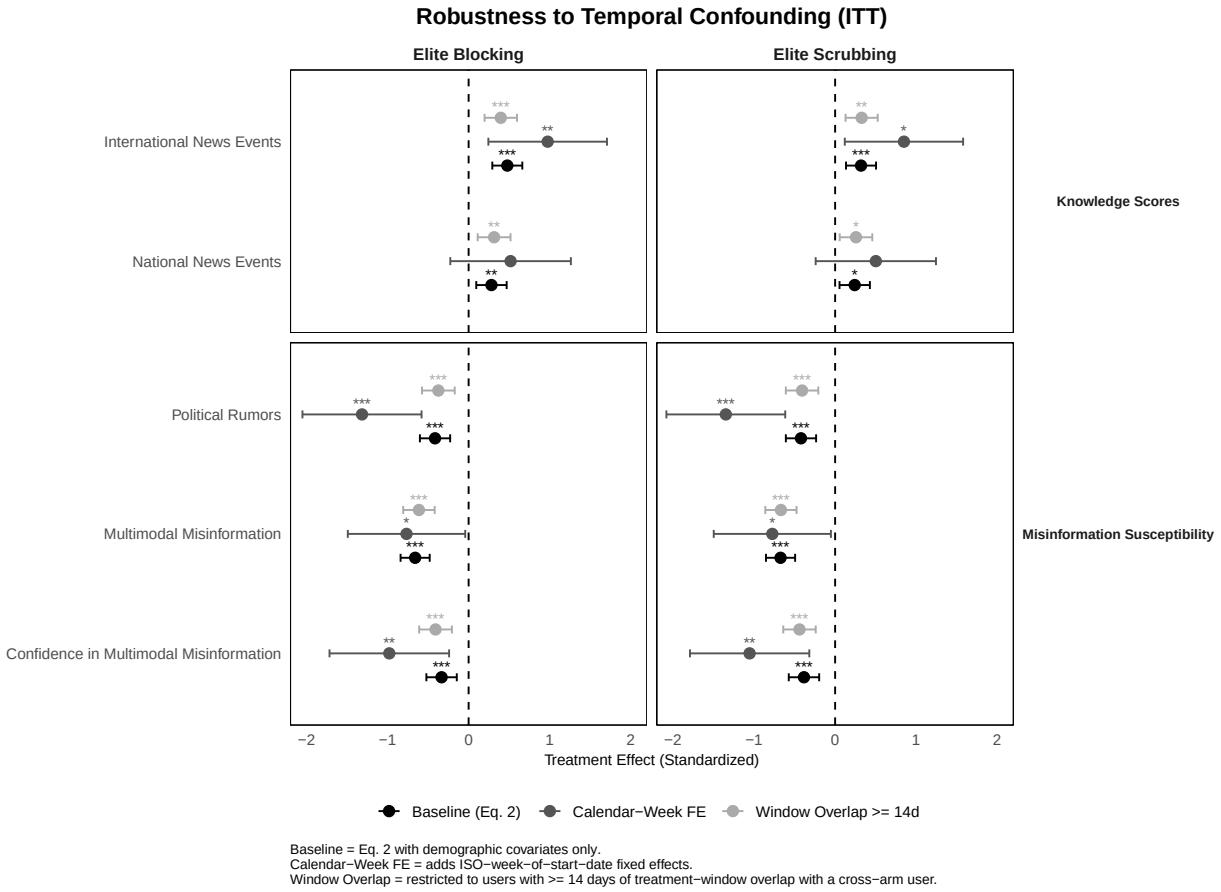


Fig. K13: **Robustness of treatment effects to temporal confounding (ITT, $N = 644$).** Comparison of three specifications: Baseline (Eq. 2 with demographic covariates), Calendar-Week FE (adding ISO-week-of-start-date fixed effects), and Window Overlap (restricted to participants with ≥14 days of treatment-window overlap with a cross-arm participant, $N = 550$). Error bars indicate 95% confidence intervals. Asterisks denote statistical significance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

inflated the treatment effects for later-recruited participants, we would expect significant interactions. Joint F-tests for the interaction terms are non-significant for knowledge outcomes (international news $p = .449$; national news $p = .884$), multimodal misinformation ($p = .513$), and confidence in misinformation ($p = .253$). For political rumor susceptibility, the joint F-test is marginal ($p = .060$), with negative individual interaction coefficients ($b = -0.11$, $p = .018$ for both arms), suggesting that political rumor effects may be *stronger* for later-recruited participants (Table K31).

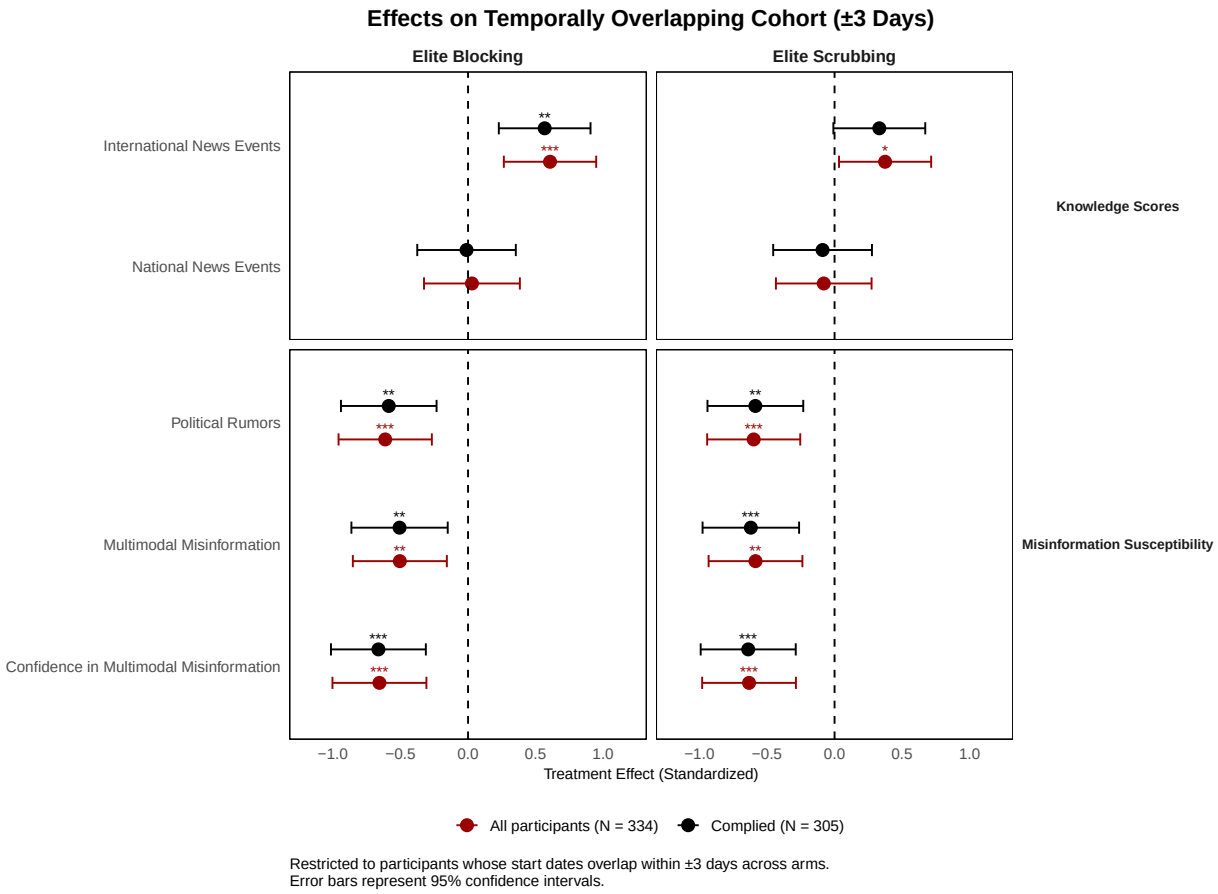


Fig. K14: **Robustness of treatment effects in temporally overlapping cohort.** Estimated effects of Elite Blocking and Elite Scrubbing on knowledge scores and misinformation susceptibility, restricted to participants whose first day of app activity fell within ±3 days of at least one participant in a different treatment arm (ITT $N = 334$; Complied $N = 305$). All specifications include demographic covariates (gender, age, education, income, and political ideology). Error bars indicate 95% confidence intervals. Asterisks denote statistical significance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

Strict treatment-window overlap (≥14 days).

Finally, we apply a stricter temporal restriction that accounts not just for start-date proximity but for overlap in the full 28-day treatment window. Each participant's treatment window is defined as the 28 days beginning on their first active day. We retain only participants whose treatment window overlaps by at least 14 days with at least one participant in a different arm (ITT $N = 550$; Complied $N = 518$). As seen in Figure K13 (Table K32), all misinformation outcomes remain highly significant: political rumor susceptibility (Elite Blocking $b = -0.37$, $p < .001$; Elite Scrubbing

$b = -0.41$, $p < .001$), multimodal misinformation (Elite Blocking $b = -0.61$, $p < .001$; Elite Scrubbing $b = -0.67$, $p < .001$), and confidence in misinformation (Elite Blocking $b = -0.41$, $p < .001$; Elite Scrubbing $b = -0.44$, $p < .001$). International news knowledge remains significant for both arms (Elite Blocking $b = +0.40$, $p < .001$; Elite Scrubbing $b = +0.33$, $p = .001$), and notably, national news knowledge also remains significant in this larger overlap sample (Elite Blocking $b = +0.31$, $p = .002$; Elite Scrubbing $b = +0.26$, $p = .012$).

Direction-of-bias argument.

The 2024 election resulted in a Republican victory, an outcome associated with increased partisan misinformation circulation on social media platforms. If the post-election information environment increased baseline misinformation susceptibility for all participants, then treatment-group participants who experienced the election during their treatment window would face a *harder* test of the intervention's effectiveness. The finding that treated participants nonetheless showed *reduced* misinformation susceptibility relative to controls is therefore conservative: any election-driven increase in misinformation exposure would bias against finding the treatment effects we observe. However, marginal evidence of treatment effect heterogeneity by start date for political rumor susceptibility (joint F-test $p = .060$) suggests the possibility that the intervention is more effective in higher-misinformation environments. For this outcome, the calendar-week fixed effects specification provides the most direct control for temporal confounding, and the treatment effects remain highly significant under that specification ($p < .001$ for both arms).

Taken together, these checks indicate that the primary misinformation findings are robust to temporal confounding from staggered recruitment.

K.6 Instrumental-variable (IV) analysis

Assessment of IV assumptions.

To assess the robustness of the main results under imperfect compliance, we estimated two-stage least squares (2SLS) models using randomized treatment assignment as an instrument for behavioral exposure measures. In two-arm specifications comparing each treatment to the control group, treatment assignment strongly predicted overall feed exposure (total number of tweets seen) in both the Blocking and Scrubbing conditions (first-stage F statistics ranging from approximately 14 to 26, all p¡.001), satisfying the relevance condition.

Models instrumenting self-reported compliance did not meet the relevance condition (first-stage $F < 4$, p¿.05) and are therefore not emphasized. Wu–Hausman tests rejected exogeneity of behavioral exposure measures across outcomes (p¡.01), indicating that ordinary least squares estimates are biased and supporting the use of an IV estimator.

Results from the IV analysis.

Table K33 reports 2SLS estimates that instrument overall feed exposure with randomized treatment assignment. These models identify Local Average Treatment Effects

(LATEs) among compliers—participants whose exposure volume was shifted by the intervention. The strong first stage confirms that treatment assignment exogenously reduces exposure among this subgroup.

Across outcomes, the IV estimates are consistent in direction with the intention-to-treat results. Among compliers, reductions in exposure are associated with improvements in international and national political knowledge and reductions in belief in political rumors, multimodal misinformation, and confidence in misinformation. Effect magnitudes are larger than the corresponding ITT estimates, as expected under partial compliance, indicating that the main results are not driven by noncompliance or attenuation bias.

Table K33: Instrumental-variable decomposition of treatment effects using total exposure as mediator. π_1 denotes the first-stage effect of treatment on exposure; β_1 denotes the second-stage effect of exposure on the outcome; θ_1 denotes the reduced-form treatment effect. Indirect effects are computed as $\pi_1 \times \beta_1$, and residual (direct) effects as $\theta_1 - (\pi_1 \times \beta_1)$. In just-identified IV (one instrument, one endogenous regressor), $\theta_1 = \pi_1 \times \beta_1$ by construction, so the residual direct effect is approximately zero up to numerical precision.

Outcome	Arm	π_1	β_1	θ_1	Indirect	Direct	SE(indirect)	p_{indirect}
Confidence in misinformation	FEP01A	−0.419	0.788	−0.330	−0.330	0.000	0.157	0.035
Confidence in misinformation	FEP01D	−0.521	0.752	−0.392	−0.392	0.000	0.142	0.006
Multimodal misinformation	FEP01A	−0.419	1.580	−0.661	−0.661	0.000	0.268	0.014
Multimodal misinformation	FEP01D	−0.519	1.292	−0.670	−0.670	0.000	0.206	0.001
Political rumors	FEP01A	−0.419	0.989	−0.414	−0.414	0.000	0.182	0.023
Political rumors	FEP01D	−0.521	0.815	−0.424	−0.424	0.000	0.151	0.005
International knowledge	FEP01A	−0.419	−1.122	0.470	0.470	0.000	0.204	0.021
International knowledge	FEP01D	−0.521	−0.615	0.321	0.321	0.000	0.134	0.017
National knowledge	FEP01A	−0.419	−0.664	0.278	0.278	0.000	0.150	0.063
National knowledge	FEP01D	−0.521	−0.468	0.244	0.244	0.000	0.122	0.045

References

[1] Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., Ungar, L.: Beyond binary labels: political ideology prediction of twitter users. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (volume 1: Long Papers), pp. 729–740 (2017)

[2] Jaidka, K., Mukerjee, S., Lelkes, Y.: Silenced on social media: the gatekeeping functions of shadowbans in the american twitterverse. Journal of Communication **73**(2), 163–178 (2023)

[3] Mukerjee, S., Yang, T., Peng, Y.: Metrics in action: how social media metrics shape news production on facebook. Journal of Communication **73**(3), 260–272 (2023)

[4] Mukerjee, S., Jaidka, K., Lelkes, Y.: The political landscape of the us twitterverse. Political Communication **39**(5), 565–588 (2022)

[5] Barberá, P.: Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. Political analysis **23**(1), 76–91 (2015)

[6] Hohenberg, B., Menchen-Trevino, E., Casas, A., Wojcieszak, M.: A list of over 5000 US news domains and their social media accounts. Zenodo (2021). https://doi.org/10.5281/zenodo.7651047 . https://doi.org/10.5281/zenodo.7651047

[7] Robertson, R.E., Jiang, S., Joseph, K., Friedland, L., Lazer, D., Wilson, C.: Auditing partisan audience bias within google search. Proceedings of the ACM on Human-Computer Interaction **2**(CSCW), 1–22 (2018)

[8] Mosleh, M., Rand, D.G.: Measuring exposure to misinformation from political elites on twitter. Nature Communications **13**(1), 7144 (2022)

[9] Allcott, H., Gentzkow, M., Mason, W., Wilkins, A., Barberá, P., Brown, T., Cisneros, J.C., Crespo-Tenorio, A., Dimmery, D., Freelon, D., *et al.*: The effects of facebook and instagram on the 2020 election: A deactivation experiment. Proceedings of the National Academy of Sciences **121**(21), 2321584121 (2024)

[10] Wang, S., Huang, S., Zhou, A., Metaxa, D.: Lower quantity, higher quality: Auditing news content and user perceptions on twitter/x algorithmic versus chronological timelines. Proceedings of the ACM on human-computer interaction **8**(CSCW2), 1–25 (2024)

[11] Kovacs, G., Wu, Z., Bernstein, M.S.: Rotating online behavior change interventions increases effectiveness but also increases attrition. Proceedings of the ACM on Human-Computer Interaction **2**(CSCW), 1–25 (2018)

[12] Paredes, P., Gilad-Bachrach, R., Czerwinski, M., Roseway, A., Rowan, K., Hernandez, J.: Poptherapy: Coping with stress through pop-culture. In: Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare, pp. 109–117 (2014)

[13] Halfaker, A., Keyes, O., Kluver, D., Thebault-Spieker, J., Nguyen, T., Grandprey-Shores, K., Uduwage, A., Warncke-Wang, M.: User session identification based on strong regularities in inter-activity time. In: Proceedings of the 24th International Conference on World Wide Web, pp. 410–418 (2015)

[14] Media Bias/Fact Check: Media Bias/Fact Check. Accessed: 2026-03-22 (2025). https://mediabiasfactcheck.com

[15] AllSides: AllSides Media Bias Ratings. Accessed: 2026-03-22 (2025). https://www.allsides.com/media-bias/ratings

[16] Manski, C.F.: Nonparametric bounds on treatment effects. American Economic Review **80**(2), 319–323 (1990)

[17] Lee, D.S.: Training, wages, and sample selection: Estimating sharp bounds on treatment effects. Review of Economic Studies **76**(3), 1071–1102 (2009)

[18] Coppock, A.: attrition: An R Package for Analyzing Experiments with Attrition (2019). https://github.com/acoppock/attrition