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

Kokil Jaidka

kokil.jaidka@gmail.com

National University of Singapore https://orcid.org/0000-0002-8127-1157

Yphtach Lelkes

University of Pennsylvania

Subhayan Mukerjee

National University of Singapore

Harshit Aneja

National University of Singapore

Social Sciences - Article

Keywords:

Posted Date: June 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9347980/v1

Additional Declarations: There is **NO** Competing Interest.

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;

Abstract

We tested whether reducing the visibility of political elites in social media feeds can improve belief accuracy. In a preregistered field experiment, a custom mobile app directly manipulated participants' X (formerly Twitter) timelines, randomly assigning users to a control condition, *Elite Blocking* (hiding posts authored by political elites), or *Elite Scrubbing* (additionally removing posts mentioning elites). This design isolates elite-authored content from the broader stream of elite-referential discourse while preserving users' follow networks and ordinary scrolling experience. Both interventions decreased exposure to low-credibility and highly partisan material, with Elite Scrubbing producing larger reductions than Elite Blocking. Resharing of such content dropped substantially, indicating lower downstream diffusion. These feed-level shifts translated into significantly lower misinformation susceptibility and improved factual accuracy. Our findings illuminate how elite visibility is a consequential and actionable design lever for platform content moderation. Adjusting the prominence of elite-centered discourse can yield informational and epistemic benefits without broadly suppressing political content.

1 Introduction

Social media platforms are frequently blamed for accelerating misinformation. From the 2016 U.S. presidential election to Brexit and beyond, platforms such as Facebook, Twitter (now X), and YouTube are said to have reshaped the information ecosystem, enabling misleading claims to spread quickly and widely. Much of the existing research responding to this concern has treated misinformation as a problem of individual susceptibility, emphasizing traits such as partisanship [1], ignorance [2], or simple inattention [3], and interventions have followed suit. Media literacy programs, inoculation games, and corrective prompts seek to improve users' ability to distinguish truth from falsehood [4, 5]. While valuable, this bottom-up focus can miss a key feature of platform environments: who becomes salient and how attention is organized.

A complementary, top-down view is that a small number of hyperpartisan influencers and elite accounts produce and circulate a disproportionate share of low-quality and highly partisan content [6–8]. Elites are also central *subjects* of misinformation, with false or misleading claims often framed around their actions, reputations, and policy positions [9]. More broadly, elite-centered content helps coordinate attention: elite posts anchor agendas, cue partisan interpretations, and trigger cascades of replies, quote-tweets, and resharing that can amplify both accurate and misleading claims.

Yet the claim that elites "drive" misinformation is difficult to establish with observational data alone. Political content is often a small share of what users encounter in their feeds [10]. And elite-centered narratives frequently reach users indirectly through mentions, replies, and quote-tweets that recontextualize their messages as they travel across networks [6]. As a result, misinformation may co-occur with elite discourse without elites being the causal channel linking exposure to belief.

Algorithmic audits have documented how platforms systematically amplify political content [11, 12], but cannot isolate which exposure channels generate downstream epistemic effects. Most field experiments on social media provide causal evidence, yet typically shift many inputs at once—reducing platform use wholesale [13], manipulating feed ranking [14, 15], or altering broad content categories [16]—making it difficult to distinguish whether susceptibility is driven by *who* populates political discourse versus the wider stream of elite-referential talk that elites elicit. Unfollowing interventions isolate elite exposure more directly [17], but bundle it with network reconfiguration and downstream algorithmic learning. Establishing finer-grained causal effects has also become harder as platforms restrict data access [18]. Like recent platform-independent designs [14, 15], we sidestep this constraint using a custom client to manipulate live feeds without platform consent—but unlike prior work, we hold follow networks fixed and manipulate only elite *visibility*.

In this study, we tested whether elite-centered exposure causally shapes political beliefs in a preregistered field experiment (September–December 2024). A custom mobile X client manipulated consenting participants' timelines by muting certain accounts or keywords (Fig. 1). Participants were randomly assigned to an unaltered feed or one of two elite-suppression conditions, both preserving follow networks, non-political content, and ordinary scrolling experience. *Elite Blocking* removes elite-authored posts (elites as creators); *Elite Scrubbing* additionally removes elite-referential discourse—replies, quote-tweets, and commentary that mention elites

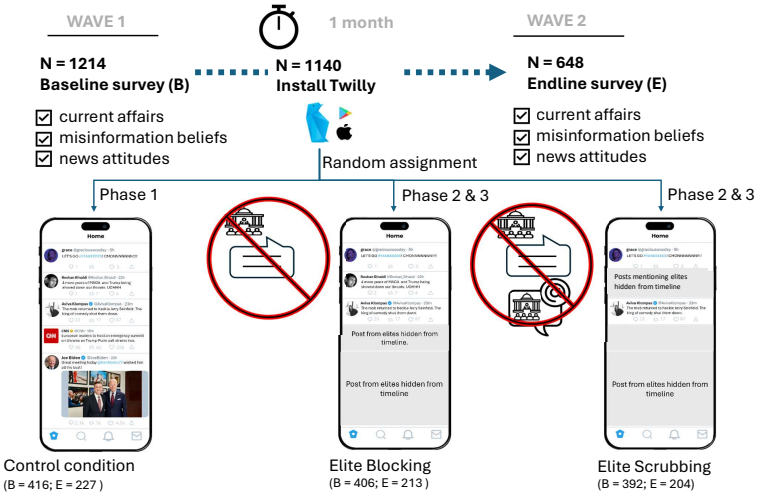

Fig. 1: Experimental design for the Twilly field study. Participants completed a baseline survey (Wave 1) measuring political knowledge, misinformation beliefs, and news attitudes, and then installed a custom X client. Over a one-month period, participants were randomly assigned to one of three conditions: a control group, an elite-blocking group in which posts from elite accounts were hidden, or an elite-scrubbing group in which both elite accounts and mentions were blocked. At the end of the study period, participants completed a follow-up survey (Wave 2) to assess changes in beliefs and knowledge.

(elites as objects of attention). If susceptibility is driven by elites as creators, Blocking should suffice; if by talk that mentions elites, Scrubbing should add leverage.

We estimated effects on preregistered outcomes spanning five domains: *political knowledge* and *misinformation susceptibility* (primary); *affective polarization*; *political interest and institutional trust*, and *media trust and news avoidance*. We report all preregistered outcomes regardless of significance (Fig. 5).

Both interventions sharply reduced retweeting of low-credibility (news domains rated unreliable by expert and crowdsourced evaluations; Mosleh and Rand 6, Robertson et al. 19) and highly partisan content (tweets with absolute ideology score > 0.3, inferred from network affiliation; Mukerjee et al. 10, Barberá 20) relative to the control, even as rates of liking remained stable. These behavioral shifts translated into improved epistemic outcomes: participants in both treatment conditions became less susceptible to misinformation and more accurate on factual knowledge. Together, the findings suggest that selectively suppressing elite-centered exposure can improve belief accuracy while preserving users' ordinary social-media experience, offering an alternative strategy to bottom-up interventions.

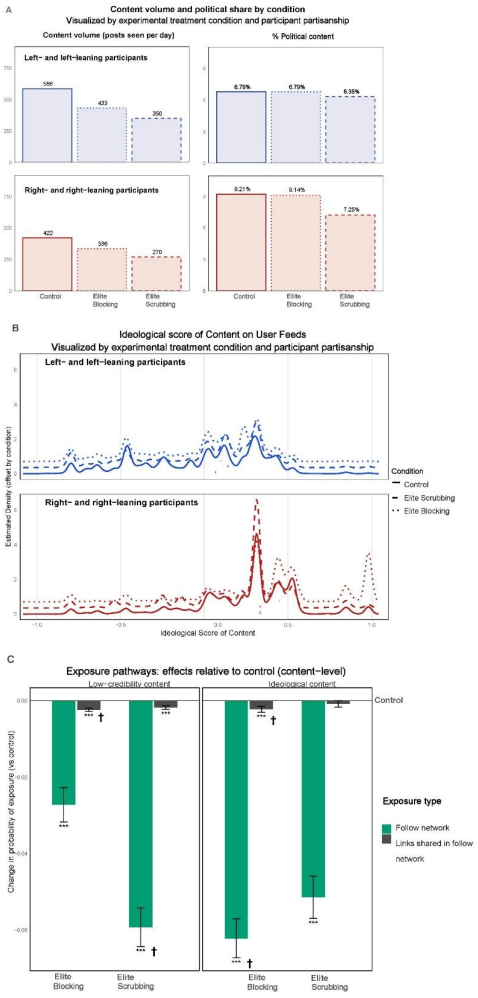

Fig. 2: Social media feed composition and exposure pathways across experimental conditions. Panel A shows the mean number of tweets seen per day and the share of political content per user, disaggregated by treatment condition and participant partisanship. **Panel B** shows the ideological distribution of political content in user feeds. Solid, dashed, and dotted lines correspond to the Control, Elite Scrubbing, and Elite Blocking conditions, respectively; curves are vertically offset for readability. The x-axis is shifted so that 0 corresponds to the political classification threshold (ideology score = 0.2). **Panel C** shows treatment effects on exposure to low-credibility and ideological content relative to control, disaggregated by provenance: content from followed accounts (follow network) versus links shared by followed accounts. Ideological content is defined as tweets with ideology scores beyond ±0.4. Error bars denote 95% confidence intervals. Stars indicate significance relative to the control condition; † indicates a significant pairwise contrast between Elite Scrubbing and Elite Blocking.

2 Results

2.1 Feed composition and exposure pathways

Before turning to outcomes, we verify what the interventions actually changed in participants' information diets. **Panel A** of Figure 2 shows that both interventions substantially reduced overall feed volume—among left-leaning participants, daily tweet volume fell from 586 (Control) to 433 under Elite Blocking and 350 under Elite Scrubbing—while leaving the share of political content largely unchanged (6.79%, 6.79%, and 6.35%, respectively). Volume reductions were similarly pronounced among right-leaning participants (422, 336, and 270 tweets/day), with a slightly larger drop in political share under Elite Scrubbing (9.21% to 7.25%). **Panel B** shows that the ideological distribution of political content was not meaningfully shifted across arms. For Democrats, no significant differences emerged across conditions (Kruskal-Wallis $p = .081$; all pairwise $p > .10$). For Republicans, the omnibus test reached significance ($p = .016$), driven by a small Scrubbing vs. Control difference ($p = .019$, $r = 0.29$), though absolute differences in mean ideology were modest (0.16–0.20) and cell sizes limited ($n = 52$–65 per condition). In short, the interventions thinned the feed without meaningfully tilting its partisan composition. **Panel C** shows that treatment effects are concentrated in first-degree exposure—posts from accounts participants directly follow—rather than in second-degree exposure via linked domains (i.e., external URLs embedded in followed accounts' posts). Both interventions significantly reduce partisan and low-credibility content from followed accounts (all $p < .001$), with Elite Scrubbing producing substantially larger reductions in low-credibility exposure than Elite Blocking ($\Delta = 3.2$ pp, $p < .001$), consistent with the role of elite-referential amplification. By contrast, changes in exposure through linked domains are comparatively small (all $|\Delta| < 0.24$ pp) and in several cases nonsignificant (Appendix I.1).

2.2 Content exposure and engagement

Figure 3 displays post-treatment differences in user behavior, disaggregated by usage type (Viewed, Favorited, Retweeted, Clicked) and treatment assignment. Effects are reported as percentage-point differences in the share of user actions relative to the Control group.

For **low-credibility content** (domain-rated; [6, 19]), both interventions reduced the share of viewed tweets classified as low-credibility. Elite Blocking decreased this share by 5.92 percentage points (95% CI: $[-8.51, -3.33]$, $p < .001$), and Elite Scrubbing by 5.75 percentage points (95% CI: $[-8.35, -3.14]$, $p < .001$). Favoriting behavior did not differ significantly across conditions (Elite Blocking: -1.74, $p = .227$; Elite Scrubbing: -0.59, $p = .690$). Retweeting of low-credibility content declined under Elite Blocking by 6.79 percentage points (95% CI: $[-10.98, -2.59]$, $p = .002$), while the reduction under Elite Scrubbing was smaller and not statistically significant (-3.44, 95% CI: $[-7.81, 0.92]$, $p = .122$). Click-based engagement also fell under both interventions, declining by 4.88 percentage points under Elite Blocking (95% CI: $[-7.53$,

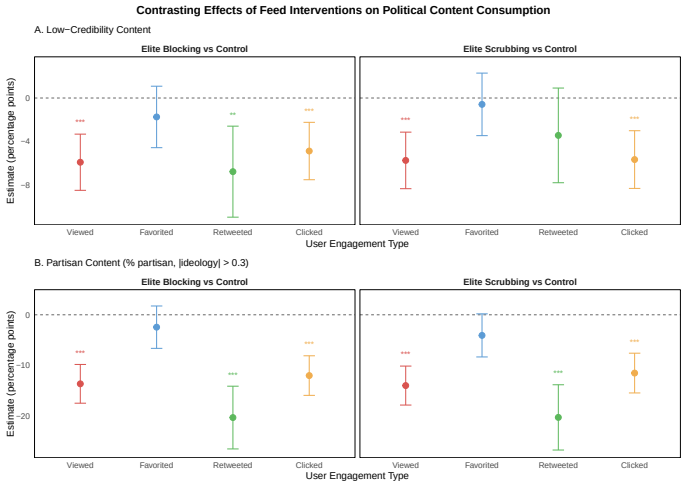

Fig. 3: Estimated effects of the two feed interventions (Elite Blocking and Elite Scrubbing) relative to the control, on the likelihood of consuming low-credibility or partisan content, stratified by user engagement type (Seen, Liked, Retweeted). Error bars indicate 95% confidence intervals. Significance levels: *p $< .05$, **p $< .01$, ***p $< .001$.

$-2.24]$, $p < .001$) and 5.67 percentage points under Elite Scrubbing (95% CI: $[-8.33$, $-3.01]$, $p < .001$).

For **partisan content** (tweets with absolute ideological score exceeding 0.3; [10, 20]), effects were substantially larger. The share of viewed tweets classified as partisan fell by 13.68 percentage points under Elite Blocking (95% CI: $[-17.52$, $-9.84]$, $p < .001$) and by 14.02 percentage points under Elite Scrubbing (95% CI: $[-17.88$, $-10.16]$, $p < .001$). Favoriting of partisan content was not significantly reduced under Elite Blocking (-2.45, $p = .252$), with a marginally significant reduction under Elite Scrubbing (-4.08, 95% CI: $[-8.35, 0.19]$, $p = .061$). Retweeting showed the sharpest declines: Elite Blocking reduced the share of retweeted partisan content by 20.36 percentage points (95% CI: $[-26.58$, $-14.15]$, $p < .001$), and Elite Scrubbing by 20.33 percentage points (95% CI: $[-26.80$, $-13.86]$, $p < .001$). Click-based engagement similarly declined, falling by 12.04 percentage points under Elite Blocking (95% CI: $[-15.96$, $-8.12]$, $p < .001$) and 11.55 percentage points under Elite Scrubbing (95% CI: $[-15.49$, $-7.61]$, $p < .001$).

Taken together, the interventions appear to have reduced both the exposure to and the amplification of low-credibility and highly partisan posts across multiple forms of engagement. Reductions were especially pronounced for retweeting and click-based exploration of partisan material, indicating that the interventions dampened both exposure and downstream amplification without inducing compensatory engagement through other interaction channels. We next examine whether these behavioral changes translate into shifts in belief accuracy and related attitudes.

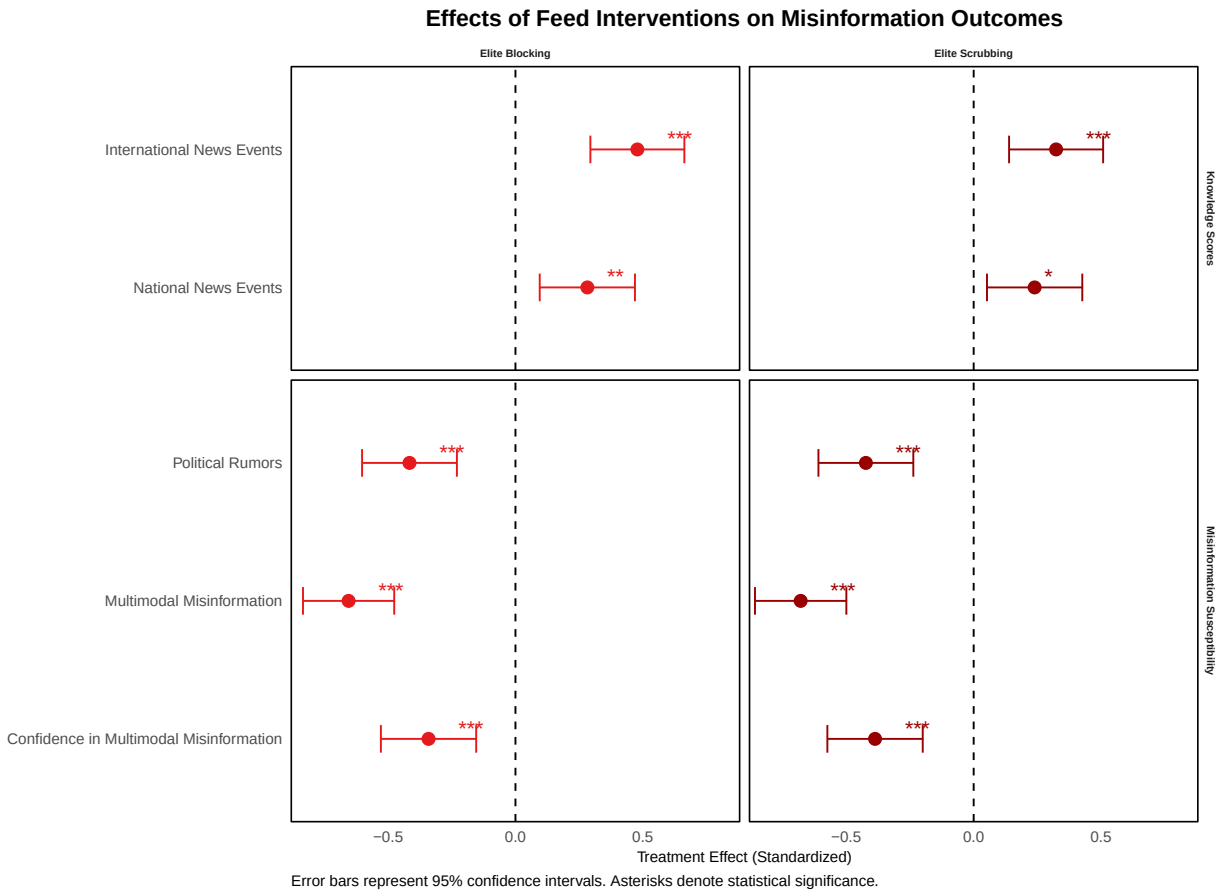

Fig. 4: Effects of feed interventions on the primary epistemic outcomes (Total vs. Complied). Estimated treatment effects on (a) knowledge scores and (b) misinformation susceptibility and confidence, shown separately for the full intent-to-treat sample (red) and for participants who complied with the intervention (black). Models include treatment indicators and covariates for gender, age, education, income and partisanship. Error bars represent 95% confidence intervals. Asterisks denote statistical significance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

2.3 Epistemic outcomes: knowledge and misinformation susceptibility

Figure 4 summarizes epistemic outcomes. Relative to control, both interventions increased factual knowledge and reduced susceptibility to misinformation. 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], $p = .003$), with comparable gains under Elite Scrubbing (international: 0.324 SD, 95% CI [0.139, 0.508], $p < .001$; national: 0.239 SD, 95% CI [0.052, 0.426], $p = .012$).

Misinformation susceptibility declined across formats. For political rumors, Elite Blocking reduced susceptibility by 0.416 SD (95% CI [−0.602, −0.230], $p < .001$) and Elite Scrubbing by 0.423 SD (95% CI [−0.609, −0.237], $p < .001$). Effects were largest for multimodal misinformation (Elite Blocking: −0.655 SD, 95% CI [−0.834, −0.476], $p < .001$; Elite Scrubbing: −0.679 SD, 95% CI [−0.859, −0.500], $p < .001$) and for confidence in misinformation judgments (Elite Blocking: −0.341 SD, 95% CI [−0.528, −0.154], $p < .001$; Elite Scrubbing: −0.387 SD, 95% CI [−0.574, −0.200], $p < .001$).

Robustness checks. All models adjust for demographic covariates. We further estimated instrumental-variable (2SLS) models using randomized treatment assignment as an instrument for overall feed exposure (Section K.6). The IV estimates identify local average treatment effects among compliers and confirm that exposure reductions improve knowledge and reduce misinformation belief, reinforcing the causal interpretation of the main findings under imperfect compliance (Table K33).

Results do not appear to be driven by differential attrition across experimental arms (Section K.2), and are also robust to the number of active days on the app (Section K.4) and recruitment timing (Section K.5).

2.4 Secondary outcomes: polarization, political attitudes, and media trust

Figure 5 presents effects of both interventions on all preregistered outcome variables. Neither Elite Blocking nor Elite Scrubbing significantly affected affective polarization, perceived partisan threat, political social identity, social distance, perceived polarization, or perceived political discrimination. Point estimates were generally small and confidence intervals consistently spanned zero. Similarly, the interventions produced no significant effects on political interest, institutional trust, media trust, news avoidance, or news skepticism. Full estimates, including complier-only analyses, are reported in SI Appendix, Sections H.3–H.6.

3 Discussion: The Infrastructure of Exposure

This study addressed a central question for both theory and practice: What epistemic shifts follow from curtailing exposure to elite and elite-centered discourse?

Across outcomes, reducing exposure to elite content was associated with lower overall app use but not with a proportional decline in political content encountered. Participants continued to see political material at similar rates, and political knowledge increased, indicating a reconfiguration of exposure rather than simple withdrawal. This pattern is consistent with work showing that structural exposure mechanisms shape information environments independently of self-selection [16, 21]. Our findings complement prior field experiments showing that feed re-ranking reduces exposure to political and untrustworthy content without necessarily improving news knowledge [17, 22]. The absence of effects on affective polarization aligns with interventions targeting algorithmic or elite cues [21, 23], though it contrasts with evidence that partisan feed re-ranking can reduce animosity over longer horizons [14]. Together, these results suggest that interventions targeting elite-centered exposure pathways operate differently from feed re-ranking alone.

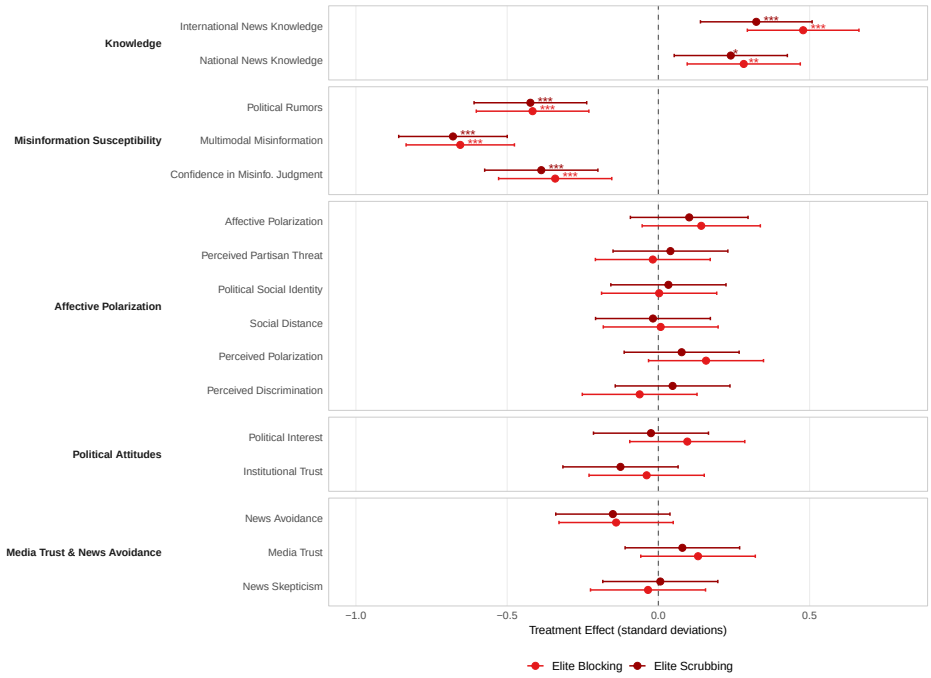

Fig. 5: Effects of feed interventions on all pre-registered outcomes. Estimated treatment effects of Elite Blocking and Elite Scrubbing relative to the control condition, estimated using Eq. 2. Outcomes are grouped by domain: knowledge and misinformation susceptibility (primary) and affective polarization, political attitudes, and media trust (secondary). All outcomes are standardized into units of SD. Horizontal lines represent 95% confidence intervals. Asterisks denote statistical significance (* $p < .05$, ** $p < .01$, *** $p < .001$).

These findings extend work showing that elite accounts contribute disproportionately to misinformation [6, 9] and that elite visibility amplifies emotionally charged and polarizing narratives [24, 25]. In our setting, most participants did not recall following many major political elites directly; however, elite cues likely entered users' feeds indirectly through retweets, mentions, and posts recommended by X's algorithm. In this context, blocking elite-attributed content and elite-related keywords altered elite *visibility* rather than elite affiliation. This pattern supports prior evidence of the role of elites in spreading political misinformation [6, 21, 24], yet clarifies that they do not gatekeep all political content. Reducing elite exposure therefore reshaped the informational environment without requiring users to disengage from politics altogether. A potential alternative explanation is algorithmic confounding: diminished engagement with elite content may have led X's recommendation system to downrank political material independently of the intervention. This account is unlikely. Filtering occurred entirely on the client side, invisible to the platform, and participants' networks and historic interaction patterns remained unchanged.Moreover, the overall share of political content in users' feeds remained stable across conditions, suggesting that reducing

elite visibility reconfigured exposure pathways rather than eliminating political content altogether, consistent with prior evidence that social media environments are not typically characterized by complete ideological isolation [26, 27].

Taken together, the results suggest that exposure volume constitutes an important intermediate pathway through which elite-targeted feed interventions influence political knowledge and misinformation susceptibility. These findings reframe misinformation as a problem of exposure architecture rather than individual belief or motivation. Developing scalable designs that reshape exposure pathways, rather than relying on individual-level corrections, remains a central challenge for mitigating elite-driven misinformation.

Several limitations qualify our conclusions. First, the intervention targeted a specific set of elite accounts and elite-related keywords, which does not exhaust the broader universe of politically salient actors and narratives circulating on the platform. Second, compliance with the intervention was incomplete, and treatment effects are therefore conservative estimates of the impact of reduced elite exposure. Third, the study window captures short-run behavioral responses; longer-term dynamics, including adaptation by users or by the platform's recommendation system, remain uncertain. Finally, because the filtering operated at the client level, the intervention does not speak directly to how platform-side ranking or moderation policies could be applied to disrupt users' social media feeds, thereby affecting the effects we report here.

Methodologically, the study contributes to an emerging class of platform-independent field experiments that operate without platform collaboration. Such approaches are essential for sustaining cumulative causal research in an era of limited API access and shifting moderation policies [15]. Our findings therefore speak not only to elite-driven misinformation but also to the feasibility of studying exposure architectures without platform consent.

4 Methods

The study was pre-registered[1] in 2022 and updated in September 2024. It was reviewed and approved by the Institutional Review Boards of the National University of Singapore and the University of Pennsylvania.

4.1 Recruitment

Participants were recruited through CloudResearch between September 18, 2024 and November 17, 2024 (see Figure 1) in three phases over a four week period: the first randomly assigned participants to the control group or an unrelated treatment not analyzed here, while the second and third assigned new participants to either the Elite Blocking or Elite Scrubbing condition. Eligibility criteria required participants to be US citizens and active X users, defined as having posted at least once within the past year. We summarize the characteristics of our sample in SI Appendix, Section E. Over 2200 participants attempted the study, of which 1,214 passed our screening checks

[1] https://osf.io/pn8f4

(reported in the SI Appendix) and completed the baseline survey, and 1,214 installed the app. This group is our "baseline sample."

The baseline sample (**N = 1,214**) was 50.0% male (N = 607). The sample skewed younger than the U.S. population, with over half of participants aged 18–34 (52.0%), and relatively few aged 55 or older (9.1%). The sample leaned Democratic, with 64.7% identifying as far left, left, or slightly left on a 7-point political partisanship scale. Participant demographics are reported in the SI Appendix, Table E2. Upon qualifying for participation based on recent X use and completing the baseline survey, participants were instructed to install our custom application with a personalized activation code corresponding to their randomly assigned treatment condition, and use it in place of the default X app on their devices for the duration of the study.

Table 1: Sample Demographics from the baseline survey

Variable	Category	Count (%)
Gender	Female	591 (48.7%)
	Male	607 (50.0%)
	Others	16 (1.3%)
Age	18–24 years	224 (18.5%)
	25–34 years	407 (33.5%)
	35–44 years	319 (26.3%)
	45–54 years	154 (12.7%)
	55–64 years	82 (6.8%)
	65+ years	28 (2.3%)
Education	No formal education	132 (10.9%)
	High school graduate	304 (25.1%)
	Some college	148 (12.2%)
	Associate degree	461 (38.0%)
	Bachelor's degree	128 (10.5%)
	Master's degree	33 (2.7%)
	Doctorate or professional degree	8 (0.7%)
Household Income	Less than $20,000	113 (9.3%)
	$20,000–$44,999	279 (23.0%)
	$45,000–$139,999	625 (51.5%)
	$140,000–$149,999	65 (5.4%)
	$150,000–$199,999	77 (6.3%)
	More than $200,000	55 (4.5%)
Political Ideology	Far left	397 (32.7%)
	Left	334 (27.5%)
	Slightly left	55 (4.5%)
	Center	24 (2.0%)
	Slightly right	191 (15.7%)
	Right / Far right	145 (11.9%)
	Independent / No response	68 (5.6%)

689 participants with at least 7 active days as per the 3-week monitoring mark were invited to complete the survey, of which **648** participants completed the post-survey following the 4-week study period and comprise our "endline sample." .

We monitored behavioral compliance by tracking daily app usage (Figure E4). Pairwise comparisons between treatment arms revealed no statistically significant differences; full results are reported in SI Appendix, Section E.

Table 2: 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

4.2 Experimental Design

Participants were randomly assigned to one of three experimental conditions after completing a pre-survey measuring their current affairs knowledge and misinformation beliefs. All participants then installed and used the Twilly app - a custom-built X client - for one month. The app allowed us to implement timeline-level interventions while preserving the native user experience.

In the **elite-blocking** condition, posts authored by elite accounts were hidden from participants' timelines. Elites were defined as elected officials, content creators, and prominent public figures whose ideological orientation had been inferred from patterns of network affiliation and audience engagement, following the classifications in [20] and [10], with updates to reflect prominent political voices at the time of data collection. In the **elite-scrubbing** condition, both posts from elites and posts that mentioned elites were hidden from the timeline, thereby removing both direct and indirect exposure. The **control** condition featured an unaltered X experience with no filtering applied. The full list of filtered elite accounts is provided in our Open Science Framework (OSF) repository.

Participants entered the study in waves. In the first recruitment wave, participants were randomized to the Control condition (and to an unrelated manipulation not analyzed here). In subsequent waves, newly recruited participants were randomized to Elite Blocking or Elite Scrubbing. Thus, the two elite-suppression arms were not recruited contemporaneously with the full Control sample. We compare baseline attitudes and demographics in SI Appendix, Section E. This design raises the possibility that time-varying shocks or cohort differences drive estimated effects. We address this in four complementary ways (Section K.5).

Manipulation check. All participants completed a post-survey assessing whether they noticed any changes in their social media feeds. Across conditions, awareness of any modification was low.

A majority of participants in the control group (64%) reported no change in their feed content, compared with 47% in the treatment groups. About 28% of control

participants and 30% of treated participants said they were unsure whether anything had changed. Only 8% in the control group and 23% in the treatment groups indicated that they noticed a difference.

When asked specifically about posts from political accounts they followed, the pattern was similar: 37% of treated participants said they saw no change, 38% were unsure, and 25% reported that such posts were missing from their feeds. Overall, participants across all conditions showed limited awareness of the feed alterations implemented through Twilly.

We further examined the extent of filtering to contextualize these perceptions. 208 participants in the elite-blocking condition had a median of 58 political tweets filtered out of their timelines, while 206 participants in the elite-scrubbing condition had 918 tweets filtered.

4.3 Analysis

We implemented a preregistered set of linear and mixed-effects regression models to estimate the causal effects of timeline interventions - *elite-blocking*, *elite-scrubbing*, and *control* - on key outcomes such as misinformation susceptibility, media trust, political disengagement, and polarization-related attitudes. Fixed effects were included for treatment conditions and baseline covariates, while random intercepts were included for participants to account for repeated measures and individual heterogeneity.

4.4 Analysis

We implemented a preregistered set of linear regression models to estimate the causal effects of timeline interventions—*elite-blocking*, *elite-scrubbing*, and *control*—on key outcomes such as misinformation susceptibility, media trust, political disengagement, and polarization-related attitudes.

Behavioral Engagement Model. To assess how the interventions affected political content exposure within the app, we estimated the following linear regression model:

$$\begin{aligned}\bar{Y}_{ig} = \beta_0 + \beta_1 \text{Treatment}_i + \beta_2 \text{EngagementType}_g \\ + \ \beta_3(\text{Treatment}_i \times \text{EngagementType}_g) + \varepsilon_{ig}\end{aligned} \tag{1}$$

where $\bar{Y}_{ig}$ is the mean proportion of political content consumed by user i in engagement context g. EngagementType_g is a categorical variable representing how the tweet was encountered: *Viewed* (home timeline, latest timeline, explore page, or search results), *Favorited* (liked or bookmarked), *Retweeted*, or *Clicked* (tweet detail or user profile pages), with *Viewed* as the reference category. Treatment indicators correspond to the experimental conditions *Elite Blocking* and *Elite Scrubbing*, with the *Control* condition as the reference category. Interaction terms between treatment and engagement type test whether intervention effects vary by how participants encountered political content. The unit of observation is one user–engagement-type cell, and standard errors are estimated via OLS.

Survey-Based Outcomes Model. To estimate the effects of our timeline interventions on belief correction, misinformation susceptibility, media trust, and other attitudinal outcomes, we used a linear regression framework with the following specification:

$$Y_i = \beta_0 + \beta_1 \text{EliteBlock}_i + \beta_2 \text{EliteScrub}_i + \gamma X_i + \varepsilon_i \tag{2}$$

Here, Y_i represents the outcome variable of interest for participant i, such as change in belief accuracy, news sharing intention, or trust in media. X_i includes baseline covariates: pre-treatment scores, demographics (age, gender, political affiliation), and cognitive traits (e.g., cognitive reflection scores). We report standardized coefficients, post-hoc pairwise comparisons, and 95% confidence intervals. Treatment indicators capture assignment to the *Elite Blocking* and *Elite Scrubbing* conditions, with the control group serving as the reference category.

4.5 Outcomes

We pre-registered multiple outcome domains capturing informational and attitudinal consequences of elite-centered exposure. All outcomes were measured in both the baseline and endline surveys; treatment effects are estimated on pre–post difference scores, standardized into units of within-condition SD unless otherwise noted. Full variable construction details and survey instruments are provided in SI Appendix, Section D.

Primary outcomes

Our primary outcomes directly test the core theoretical claim that elite-centered exposure shapes belief accuracy.

Political knowledge. Participants evaluated whether a series of recent political events had occurred, selecting from *Definitely didn't happen* to *Definitely did happen* (4-point scale). Responses were scored by their distance from factual truth; absolute deviations were summed and reverse-coded so that higher values indicate greater accuracy. Separate indices were computed for *international* events (e.g., foreign elections, climate accords) and *U.S. national* events (e.g., executive actions, legislation). Full item lists are in SI Appendix, Section D.

Misinformation susceptibility. Susceptibility was assessed through two modules. *Text-based misinformation* presented 10 political statements (true and false) rated on a 4-point truthfulness scale; deviations from factual truth were summed, with higher totals indicating greater susceptibility. *Multimodal misinformation* presented five headline–image pairs, including out-of-context pairings (adapted from [28]), rated on perceived truth, familiarity, and willingness to share.

Confidence in misinformation assessment. After each multimodal item, participants reported how confident they were in their judgment (5-point Likert scale). Higher scores indicate greater confidence.

Secondary outcomes

Secondary outcomes were pre-registered alongside the primary outcomes and are reported in full in the Supplementary Information (SI Appendix, Section H), regardless of statistical significance. We summarize their construction here so that the full scope of the pre-registered outcome space is transparent; Figure 5 presents effects on all pre-registered outcomes.

Affective polarization. Measured along four dimensions: (i) a *feeling thermometer* (difference between in-party and out-party warmth, 0–100 scale); (ii) *perceived partisan threat* (agreement that the opposing party "threatens the nation's well-being"); (iii) *social distance* (comfort interacting with out-party members across relational contexts; $\alpha = .839$); and (iv) *political social identity* (nine-item composite measuring party identification; $\alpha = .773$). Analyses were restricted to participants with non-neutral partisan identification.

Perceived polarization and discrimination. *Perceived polarization*: agreement that "more and more Democrats and Republicans have extreme views." *Perceived discrimination*: frequency of feeling that one's political beliefs lead to unfair treatment.

Political interest and institutional trust. Single 5-point Likert items measuring interest in politics and trust that "the government can be trusted to do what is right."

Media trust and news avoidance. *Media trust*: agreement that mainstream news sources provide accurate information ($\alpha = .704$). *News avoidance*: actively trying to avoid news ($\alpha = .744$). *News skepticism*: tendency to question the accuracy of news reports.

Supplementary information. Supplementary analyses and data are reported in the online materials and are available on the Open Science Framework.

Acknowledgments. This research was supported by the Ministry of Education, Singapore, through its MOE AcRF Tier 3 Grant (MOE- MOET32022-0001), and the Department of Communications and New Media at the National University of Singapore through a start-up grant. We are grateful to Jeremiah Yee, Marc Riven Reyes Herrera, Insyirah Mujtahid, and Cai Yang for their research assistance. We are grateful to Dr. Pablo Barbera and to Dr. Sandra González-Bailón, as well as the members of the Centre for Information Networks and Democracy at the Annenberg School for Communication, University of Pennsylvania for their invaluable feedback.

Declarations.

- Funding: This research was supported by the Ministry of Education, Singapore, through its MOE AcRF Tier 3 Grant (MOE-MOET32022-0001), and by the Department of Communications and New Media at the National University of Singapore through a start-up grant.
- Conflict of interest: We declare that none of the authors has competing financial or non-financial interests as defined by Nature Portfolio.
- Ethics approval: The study was reviewed and approved by the Institutional Review Boards of the National University of Singapore (NUS-IRB-2021-431) and the University of Pennsylvania (832944).

- Consent for participation and publication: Participants were provided with detailed information about the study procedures and the types of data being collected, excluding the specific nature of the treatment condition to preserve experimental integrity. They provided informed consent for participation and for their data to be used in research publications. Participants were free to withdraw from the study at any time and receive partial reimbursement. Following the post-treatment survey, all participants were debriefed regarding the experimental manipulation and invited to retract their data if desired. No participants chose to withdraw after debriefing.
- Code and data availability: X (formerly Twitter) data was collected in accordance with the platform's API terms of use. The code and preprocessed data necessary to reproduce the results reported in this paper are available on OSF.
- Author contribution:
  K.J.: Conceptualization, Methodology, Software, Data curation, Investigation, Writing - Original draft preparation, Visualization, Supervision.
  Y.L.: Conceptualization, Methodology, Investigation, Writing - Original draft preparation, Visualization, Supervision.
  S.M.: Conceptualization, Investigation, Data curation, Writing - Original draft preparation, Visualization, Supervision.
  H.A.: Conceptualization, Methodology, Software, Data curation, Investigation, Writing - Original draft preparation.

References

[1] Osmundsen, M., Bor, A., Vahlstrup, P.B., Bechmann, A., Petersen, M.B.: Partisan polarization is the primary psychological motivation behind political fake news sharing on twitter. American political science review **115**(3), 999–1015 (2021)

[2] Flynn, D.J., Nyhan, B., Reifler, J.: The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. Political psychology **38**, 127–150 (2017)

[3] Pennycook, G., Rand, D.G.: Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. Cognition **188**, 39–50 (2019)

[4] Bode, L., Vraga, E.K.: See something, say something: Correction of global health misinformation on social media. Health communication **33**(9), 1131–1140 (2018)

[5] Arechar, A.A., Allen, J., Berinsky, A.J., Cole, R., Epstein, Z., Garimella, K., Gully, A., Lu, J.G., Ross, R.M., Stagnaro, M.N., *et al.*: Understanding and combatting misinformation across 16 countries on six continents. Nature Human Behaviour **7**(9), 1502–1513 (2023)

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

[7] Lasser, J., Aroyehun, S.T., Simchon, A., Carrella, F., Garcia, D., Lewandowsky, S.: Social media sharing of low-quality news sources by political elites. PNAS nexus **1**(4), 186 (2022)

[8] Nielsen, R.: Forget technology–politicians pose the gravest misinformation threat. The Financial Times, 15–15 (2024)

[9] Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., Lazer, D.: Fake news on twitter during the 2016 us presidential election. Science **363**(6425), 374–378 (2019)

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

[11] 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)

[12] Gauthier, G., Hodler, R., Widmer, P., Zhuravskaya, E.: The political effects of x's feed algorithm. Nature (2026) https://doi.org/10.1038/s41586-026-10098-2

[13] 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)

[14] Piccardi, T., Saveski, M., Jia, C., Hancock, J., Tsai, J.L., Bernstein, M.S.: Reranking partisan animosity in algorithmic social media feeds alters affective polarization. Science **390**(6776), 5584 (2025)

[15] Allen, J., Tucker, J.A.: Platform-independent experiments on social media. Science **390**(6776), 883–884 (2025)

[16] Guess, A.M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., *et al.*: Reshares on social media amplify political news but do not detectably affect beliefs or opinions. Science **381**(6656), 404–408 (2023)

[17] Rathje, S., Pretus, C., He, J.K., Harjani, T., Roozenbeek, J., Gray, K., Linden, S., Van Bavel, J.J.: Unfollowing hyperpartisan social media influencers durably reduces out-party animosity. Preprint at https://osf. io/acbwg/download (2024)

[18] Wagner, M.W.: Independence by permission. Science **381**(6656), 388–391 (2023)

[19] 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)

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

[21] Bail, C.A., Argyle, L.P., Brown, T.W., Bumpus, J.P., Chen, H., Hunzaker, M.F., Lee, J., Mann, M., Merhout, F., Volfovsky, A.: Exposure to opposing views on social media can increase political polarization. Proceedings of the National Academy of Sciences **115**(37), 9216–9221 (2018)

[22] Guess, A.M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., *et al.*: How do social media feed algorithms affect attitudes and behavior in an election campaign? Science **381**(6656), 398–404 (2023)

[23] González-Bailón, S., Lazer, D., Barberá, P., Zhang, M., Allcott, H., Brown, T., Crespo-Tenorio, A., Freelon, D., Gentzkow, M., Guess, A.M., *et al.*: Asymmetric ideological segregation in exposure to political news on facebook. Science **381**(6656), 392–398 (2023)

[24] Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. science **359**(6380), 1146–1151 (2018)

[25] Brady, W.J., Wills, J.A., Jost, J.T., Tucker, J.A., Van Bavel, J.J.: Emotion shapes the diffusion of moralized content in social networks. Proceedings of the National Academy of Sciences **114**(28), 7313–7318 (2017)

[26] Eady, G., Nagler, J., Guess, A., Zilinsky, J., Tucker, J.A.: How many people live in political bubbles on social media? evidence from linked survey and twitter data. Sage Open **9**(1), 2158244019832705 (2019)

[27] Tucker, J.A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., Nyhan, B.: Social media, political polarization, and political disinformation: A review of the scientific literature. Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018) (2018)

[28] Qi, P., Yan, Z., Hsu, W., Lee, M.L.: Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13052–13062 (2024)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TwillyEchoChamberEffectssupp.pdf