Supplementary Information Guide

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

William J. Brady*, Meriel Doyle, Abdo Elnakouri, Eli Finkel, Joshua Conrad Jackson, Nour Kteily, Victoria Parker, Curtis Puryear, Trevor Spelman, Jacob Teeny & Mark Torres

Kellogg School of Management, Northwestern University, Chicago, IL
Department of Psychology, Northwestern University, Chicago, IL
Booth School of Business, University of Chicago, Chicago, IL

* Corresponding author: William J. Brady (william.brady@kellogg.northwestern.edu)

Table of Contents

Supplementary Information Overview

This Supplementary Information (SI) file accompanies the main manuscript and provides detailed documentation of participant procedures, platform details, algorithm development, pilot testing, and comprehensive statistical results for all research questions. Below we summarize each major section.

Section 1: Participant procedure details

Contains participant instructions for the study, demographic and screening materials, all survey measures used in the intake and weekly surveys, procedures for identifying non-legitimate sign-ups, and descriptive statistics for the final participant pool.

Section 2: Bluesky Platform details

Provides background on Bluesky Social, including platform features, user interface details, and the rationale for selecting Bluesky as the study platform.

Section 3: Details of feed ranking algorithm development

Documents the full technical pipeline for building the three feed-ranking algorithms, including raw data collection from the Bluesky firehose, data filtering procedures, feature generation (on-platform behavior metrics, machine learning classification models, in-house IME classifiers, and super poster definitions), candidate generation, feed ranking and scoring, and feed postprocessing steps.

Section 4: Algorithm details

Provides technical details on the approximate nearest neighbor (ANN) approach used for suggesting posts and the ranking and scoring algorithms used across experimental conditions.

Section 5: Pre-pipeline pilot

Reports pilot testing conducted during Stage 1, demonstrating that the criteria used for the engagement-based and diversified extremity algorithms produced measurable differences in feed content composition.

Section 6: RQ1 supplementary information

Presents supplementary materials for Research Question 1, including examples of IME content classifications, descriptive statistics and correlations among content measures, estimates of baseline content proportions on Bluesky, model fit analyses, fully tabulated model results for both pre/post and base model specifications, topic modeling and content analyses (BERTopic, Named Entity Recognition, hashtags), plots of raw content proportions over the study period, polynomial time models, analyses of Bluesky's new user growth impact, comparisons of in- vs. out-of-network toxic content, further analyses of constructive dialogue classification, and re-classification of intergroup content.

Section 7: RQ2 supplementary information

Presents supplementary materials for Research Question 2, including attrition analyses, model fit analyses, fully tabulated results for pre/post and base models, effects of condition on raw perception variables (both pre/post and base model specifications), additional model specifications (including heterogeneity by degree of IME removal and controlling for baseline IME exposure and login frequency), and exploratory models for political sectarianism and attitude extremity.

Section 8: RQ3 supplementary information

Presents supplementary materials for Research Question 3, including descriptive statistics and correlations for engagement outcomes, conditional models, model fit comparisons, fully tabulated results for base, pre/post, and exposure model specifications, precision analysis of null engagement effects, and condition effects on general engagement.

Section 9: Additional attrition analyses and exclusion robustness tests

Reports robustness tests using progressively relaxed exclusion thresholds for all three research questions, demonstrating that substantive conclusions are robust to alternative exclusion criteria, and baseline individual difference analyses.

Titles and Legends for SI Figures and Tables

All SI figures and tables are contained in the accompanying Supplementary Information PDF file. Below we list titles and legends for key figures and tables.

Fig. S1. Screenshot of the Bluesky home screen (web version). Bluesky implements a user interface similar to X (formerly Twitter) that allows users to select custom feed-ranking algorithms.

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

Fig. S3. Overview of steps for recommending a post to a user. When a user logs into Bluesky, posts are selected and ranked according to the assigned experimental condition.

Table S1. Examples of "toxic" and "non-toxic" posts pulled from Bluesky's firehose as classified by Google's Perspective API.

Table S2. Examples of "constructive" and "non-constructive" posts pulled from Bluesky's firehose as classified by Google's Perspective API.

Table S3. Ratings of posts by both toxicity and constructiveness, demonstrating that filtering by toxicity and constructiveness allows strong moral opinions while removing personal attacks.

Table S4. Examples of "sociopolitical" and "non-sociopolitical" posts pulled from Bluesky's firehose.

Table S5. Validation metrics for in-house classification tools.

Table S6. Additional validation metrics for classification tools.

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

Table S8. Example feeds for the engagement versus diversified extremity conditions.

Table S9. Examples of posts for each label within the IME framework.

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

Figs. S11–S14. Correlation matrices for all content types measured in feeds across conditions and baseline.

Figs. S15–S22. Daily proportions of each content type (intergroup, moral, negative emotional, positive emotional, toxic, political, moral outrage, constructive) by condition compared to the baseline firehose estimate.

Table S11. Model fit comparison between the base model and the model with a pre-/post-election indicator for RQ1 outcomes.

Tables S12–S27. Fully tabulated model results for all RQ1 outcomes across pre/post and base model specifications.

Table S28. Top keywords in each of the top 5 topics uncovered by the BERTopic model.

Figs. S23–S27. Topic modeling, named entity recognition, and hashtag distribution figures across conditions.

Figs. S28–S35. Mean proportions of each content type in feeds by condition compared to a large sample from the Bluesky firehose.

Table S29. Exploration of various polynomial fit statistics.

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

Tables S30–S39. Tabulated model results from polynomial time models for all content types.

Figs. S37–S43. Figures related to new user growth impact, in- vs. out-network toxic content comparisons, constructive dialogue analyses, and intergroup content re-classification.

Tables S40–S83. Fully tabulated results for all RQ2 outcomes including attrition analyses, model fit, pre/post models, base models, raw perception variables, and additional model specifications.

Fig. S44. Mean attitude extremity over the 8-week study period by condition.

Fig. S45. Attitude extremity over the 8-week study period by condition and each issue.

Tables S84–S87. Tabulated results for exploratory models on political sectarianism and attitude extremity.

Fig. S46. Smooth toxic engagement trends by condition.

Tables S88–S107. Fully tabulated results for all RQ3 outcomes including model fit comparisons, base models, pre/post models, exposure models, and precision analysis of null engagement effects.

Tables S108–S139. Robustness tests using progressively relaxed exclusion thresholds for RQ1, RQ2, and RQ3, plus baseline individual difference analyses.