

Designing Social Media to Promote Productive Political Dialogue on a New Research Platform

Austin van Loon¹, Srikar Katta², Christopher A. Bail^{3,4,5}, D. Sunshine Hillygus^{4,5},
and Alexander Volfovsky^{2,6}

¹MIT, Sloan School of Management

²Duke University, Department of Computer Science

³Duke University, Department of Sociology

⁴Duke University, Sanford School of Public Policy

⁵Duke University, Department of Political Science

⁶Duke University, Department of Statistical Science

Contents

1	Study Population	2
1.1	Missing Data	2
1.2	Demographic Characteristics	2
2	Measurement of Key Variables	2
2.1	Outcome Measures	2
2.2	Summary Statistics	4
3	Overview of Respondent Experience	6
3.1	Intake Survey	6
3.2	Installing Spark Social Social Media Testing Environment	7
3.3	App Onboarding	8
3.4	Pre-Treatment Survey	8
3.5	Administering Treatment	8
3.6	Newsfeed Experience	9
3.7	Post-Treatment Survey	9
3.8	Technical Issues During Fielding	9
4	Newsfeed Content	9
4.1	Pre-Populated Posts from Researchers	9
4.2	Dynamic Posts Created by Chatbots	9
5	Participant Experience and Engagement with Platform	10
6	Balance checks	11
7	Attrition analyses	12
8	Manipulation Checks	19
9	Evaluating the Definition of Treatment	19
9.1	What Component of Treatment was Important?	19
9.2	Were Participant Interdependencies Important?	20
10	Heterogeneous Treatment Effects	21
11	Understanding the Effect on Linguistic Humility	24
12	Further Examining the Effects on Emotions and Intellectual Humility	24

13 Supplemental Analyses	25
13.1 Pure control arm	25
13.2 Poisson Regressions of Engagement	25
14 Questionnaire	25
14.1 AAPOR Disclosures	25
14.1.1 Intake Survey Items	27
14.1.2 Pre-treatment Survey Questions	27
14.1.3 Post-treatment Survey Questions	28
15 Data Availability, Code, and Pre-Registration	29

1 Study Population

We recruited study participants from the Prolific platform (<https://www.prolific.com/>). By the end of data collection, we had approved 1,535 submissions, which includes a set of participants who were assigned to a third “pure control” condition that is described in Section 2.2 of this appendix. Our final analytical sample consisted of $N = 1,043$ subjects who were assigned to either the treatment or control arms of our experiment but not the “pure control” arm and who also successfully completed all stages of the study: screening survey, pre-treatment survey, app downloading, social media platform use, and the post-treatment survey.

1.1 Missing Data

Participants who attrited from the study had missing data for survey questions asked after attrition. Section 7 describes how we handle these missing values in attrition analyses.

Our final analytical sample, which excluded all attrited participants, had no item nonresponse. Note, however, that our measure of affective polarization has 155 missing values because we are not able to make ingroup and outgroup comparisons for the 155 respondents in our study who identify as “pure independents.”

1.2 Demographic Characteristics

In this section, we report demographic characteristics for the 1,043 participants in the treated and control arms who successfully completed the study. As reported in Table 1, 60% of our respondents identified as White, 15% identified as Black, 9% identified as Hispanic, 9% identified as Asian, and the remaining 7% either identified as “Other” or identified with multiple racial identities. 1% of our respondents did not have a high school degree, 14% had a high school degree and did not attend college, 25% had attended some college but did not earn a degree, 12% had earned an associate’s degree with no post-graduate degree, 34% had earned a bachelor’s degree with no post-graduate degree, and 13% had earned some type of post-graduate degree. Our sample skewed Democrat (48% Democrat, 19% Republican) and liberal (55% identified as left of moderate, 23% identified as moderate, and 22% identified as right of moderate). Also, as Table 2 shows, the mean age of respondents in our final sample is 37.5 years old, and ranged from 18 to 78.

2 Measurement of Key Variables

2.1 Outcome Measures

We used a range of attitudinal and behavioral measures to evaluate the outcomes described in the main text of our article. For exact wording of survey items, see Section 13. Unless otherwise noted, all participants provided enough information to calculate each outcome (i.e., there was no missing data).

Intellectually Humble Language. Using Perspective API (<https://perspectiveapi.com/>), we measured to what degree each comment or original post a participant authored was indicative of (a) nuance, (b) respect, and (c) curiosity (see [10]). Each of these measures take on a zero to one value, representing the predicted probability that the piece of content contains an instance of the relevant characteristic. We took the average of these three scores for each piece of content to measure how indicative it was of intellectual humility. We then averaged this measure over the posts and comments each participant authored to get a single measure of how humble their produced language

Table 1: Summary statistics for categorical features.

Variable	Proportion of participants
Apple (vs Android)	
Android	49.47
Apple iPhone	50.53
Political ID (3 Point)	
Democrat	47.75
Independent	29.43
Other	3.93
Republican	18.89
Race	
Asian	8.92
Black or African American	15.15
Hispanic	9.01
Mixed/other	7.00
White	59.92

was. Since sixty-two participants did not author any posts or comments, they are excluded from the analysis of this variable. Higher values of this variable, which ranges from zero to one, indicate that the language produced by participants was more indicative of intellectual humility.

Intellectually Humble Engagement. For each participant, we counted how many times they liked, disliked, or commented on the posts in their newsfeed that we had designed to be intellectually humble (see Section 4). The sum of these counts—which ranged from zero to ten—was our measure of how often each participant engaged with intellectually humble content. Higher values of this variable indicate that the participant engaged more with intellectually humble content.

Change in Self-reported Intellectual Humility. During both the pre-treatment and post-treatment surveys, we asked participants to complete a self-report measure of general intellectual humility [8], which asks participants they degree to which they believe statements such as “I question my own opinions, positions, and viewpoints because they could be wrong” and “I like finding out new information that differs from what I already think is true” applies to them. For each statement, we map responses to a quantitative scale of the participant’s degree of agreement, such that “Not at all well” represented a zero score and “Extremely well” represented a score of four. We take the average of pre-treatment responses and post-treatment responses, and then take the latter less the former. This results in a single measure of how the participant’s general intellectual humility changed between the two measurements. If the measure is positive, the participant agreed more with the statements on average during the post-treatment survey than during the pre-treatment survey—if it is negative, they disagreed more.

Emotional Experience. We leveraged the International Positive and Negative Affect Schedule Short Form ([11]). Participants were asked the degree to which they experienced a range of both positive emotions (e.g., determined and inspired) and negative emotions (e.g., ashamed and nervous) while interacting with the newsfeed. Specifically, for a range of emotions (e.g., “determined” and “ashamed”), we asked participants how much they experienced that emotion during their time interacting with the newsfeed. After mapping these responses to a numeric scale (where “Very slightly or not at all” mapped to zero and “Extremely” mapped to four), we created an index corresponding to the experience of positive and negative emotions by averaging their respective scores, such that a higher value of each scale corresponds to experience more of those emotions. We then took the positive emotion measure less the negative emotion measure to create a composite emotional experience measure. This, in total, resulted in three outcomes (positive, negative, and composite)—we analyze each separately. For the overall index, higher values mean that participants experienced more positive emotions or less negative emotions.

Change in Affective Polarization. During both the pre-treatment and post-treatment surveys, we asked each participant how warmly they feel towards “Republican voters” and “Democratic voters” on a 0-100 scale [5, 4]. For both pre-treatment and post-treatment, we calculated affective polarization as this rating for the participant’s political ingroup (determined by their self-reported partisan identity during the intake survey) less their rating for their outgroup. We exclude individuals who identified as political independents ($n = 155$). We then calculated each partisan’s change score as their post-treatment affective polarization less their pre-treatment affective polarization. This re-

sults in a measure where positive scores indicate that partisans increased their affective polarization, whereas negative scores indicate that they decreased it.

Moderation of Policy Positions. We asked each participant, during both the pre-treatment and post-treatment survey, the degree to which they agreed or disagreed with general statements that reflect their disposition towards US immigration and gun control policy on a seven-point scale. For each participant, we identify their focal policy attitudes pre- and post-treatment as their response to the question relevant to the policy they were randomly assigned to discuss during their newsfeed experience. We then measure the extremity of these attitudes as their distance from the scale midpoint (e.g., “Strongly Disagree” is three points away from “Neither Agree Nor Disagree”). We then take the degree to which the participant moderated their focal policy attitudes as their pre-treatment extremity less their post-treatment extremity. Therefore, participants with a positive score become more moderate in their focal policy attitude, while those with negative scores became more extreme.

2.2 Summary Statistics

Tables 2 and 3 provide summary statistics for the ordinal and continuous measured in our study. Most ordinal measures were transformed into numerical variables with integer values denoting the ordering of variables. The “Range” column in Table 3 describes the integer values assigned to each variable. A detailed description of the procedures use to code our attitudinal outcomes and other variables of interest collected via survey data are as follows:

- PANAS Positive is calculated as the sum of five PANAS measures: PANAS Active, PANAS Alert, PANAS Attentive, PANAS Determined, PANAS Inspired
- PNAS Negative is calculated as the sum of five PANAS measures: PANAS Afraid, PANAS Ashamed, PANAS Hostile, PANAS Nervous, PANAS Upset
- PANAS Composite is calculated as the difference between the PANAS Positive and Negative indices
- Affective polarization is calculated as the difference between thermometer ratings for the political party (Democrat or Republican) that participant most closely aligns and the outparty. True independents do not have a political alignment and therefore have NAs for their affective polarization scores.
- Pre-treatment policy view questions asked participants to rate their views on gun control and immigration policy.
- Platform perception is an aggregation of responses to three questions: (1) how engaging was the platform? (2) would you use this platform in the future? (3) was this platform exhausting? We reverse coded the third question and took an average of the three responses. Summary statistics for the three questions in the composite index can be found in Table 3.
- Social media use sums self-reported usage across the following social media platforms: Facebook, Instagram, TikTok, Truth Social, and X. Users rated their usage of each platform on a scale of 0-2 (“Not at all,” “Some,” “A lot”), and the index was on a scale from 0 to 10.
- Perception of other users is an aggregation of three ideas: (1) how aggressive were other users on this platform? (2) how polite were other users on this platform? (3) how tolerant were other users on this platform? The first idea was reverse coded to be on the same direction of sentiment as the other two; responses were then averaged to construct this variable. Summary statistics for the three questions in the composite index can be found in Table 3.

Table 2: Descriptive statistics for continuous variables.

Variable	Mean	Observed range	Std. Dev.
Age	37.53	[18,78]	12.21
PANAS	10.37	[-10,20]	5.39
PANAS Negative	2.03	[0,20]	2.88
PANAS Positive	12.4	[0,20]	4.21
Platform perception	1.94	[1,5]	0.8
Perception of other users	2.19	[1,5]	0.72
Affective polarization (AP)			
Posttreat AP (Ingroup - outgroup)	40.33	[-70,100]	29.26
Pretreat AP (Ingroup - outgroup)	41.1	[-47,100]	28.78
Intellectual humility (IH)			
Pretreat IH	16.53	[0,24]	4.12
Posttreat IH	16.82	[0,24]	4.3
Social Media Use	4.62	[0,10]	1.92

Table 3: Summary statistics for ordinal variables.

Variable	Median	Range	(25th, 75th)
Platform perception			
Engaging	2	[1,5]	[1,3]
Exhausting	2	[1,5]	[1,3]
Toxic	2	[1,5]	[1,2]
Perception of other users			
Aggressive	2	[1,5]	[1,3]
Polite	2	[1,5]	[2,3]
Tolerant	3	[1,5]	[2,3]
Social Media Use			
Facebook	1	[0,2]	[1,2]
Instagram	1	[0,2]	[1,2]
TikTok	1	[0,2]	[0,2]
Truth Social	0	[0,2]	[0,0]
X	1	[0,2]	[0,2]
Party Identity (7-point)	-1	[-3,3]	[-3,0]
Pretreatment policy views			
Gun Control	2	[0,3]	[2,3]
Immigration	2	[0,3]	[1,2]
Posttreatment policy views			
Gun Control	2	[0,3]	[1,3]
Immigration	2	[0,3]	[2,3]
Third order	2	[0,2]	[1,2]
Future Use	4	[1,5]	[2,4]
PANAS Positive			
PANAS Active	3	[0,4]	[2,3]
PANAS Alert	3	[0,4]	[2,4]
PANAS Attentive	3	[0,4]	[3,4]
PANAS Determined	2	[0,4]	[1,3]
PANAS Inspired	2	[0,4]	[1,3]
PANAS Negative			
PANAS Afraid	0	[0,4]	[0,0]
PANAS Ashamed	0	[0,4]	[0,0]
PANAS Hostile	0	[0,4]	[0,1]
PANAS Nervous	0	[0,4]	[0,1]
PANAS Upset	0	[0,4]	[0,1]

3 Overview of Respondent Experience

3.1 Intake Survey

Figure 1 describes the multiple stages of our study. On December 12, 2023, we advertised our study on the Prolific platform. Our ad explained participants could earn \$7.50 for completing a thirty minute study if they met the eligibility requirements: living in the U.S., being over age 18, owning an iOS or Android device, and being willing to install an app to participate in the research. The exact advertisement read as follows:

We would like to know your opinions about a new social media platform. Spend 30 minutes trying out the platform and sharing your thoughts!

Click the link below to be taken to an intake survey. Once complete, you will be given instructions on how to get access to the platform through the "Spark Social" testing environment. You will need to download Spark Social onto your mobile device to complete this task.

To be eligible for this study, you must:

1. Be a US resident.
2. Be at least 18 years of age.
3. Own an Apple iPhone or an Android device.
4. Be willing to download "Spark Social" onto your smart phone.

To ensure that you are compensated, please leave this window open as you complete this task!

Those who chose to participate were directed to an informed consent dialogue, which explained the purpose of the study, how data generated by respondents would be collected and analyzed, and provided additional details about how to participate. More specifically, respondents were told the study was examining how to improve social media experiences, and respondents were told they would be providing their views on a new mobile app using a testing platform called Spark Social. In the final paragraph of the informed consent dialogue, participants were informed that—as with many social media platforms—some of the content they see will have been generated by artificial intelligence algorithms (bots). Participants were also told they could earn a bonus if they correctly answered a question about the content they would be exposed to (this is the "secret password" described in the "Post-Treatment Survey" section below).

Several attention checks were administered during the intake survey to filter out low-quality respondents, and we confirmed that participants met our eligibility requirements—that the participant be at least eighteen years old, that they reside in the US, that they own an Apple iPhone or Android device, and that they would be willing to download an app—via explicit survey questions. Anyone who failed a single attention check or did not report meeting one of our eligibility criteria were immediately removed from the study. Participants who made it to the end of the intake survey were shown instructions for how to download our app on either the Apple App Store or the Google Play Store (depending on whether they reported owning an iPhone or Android).

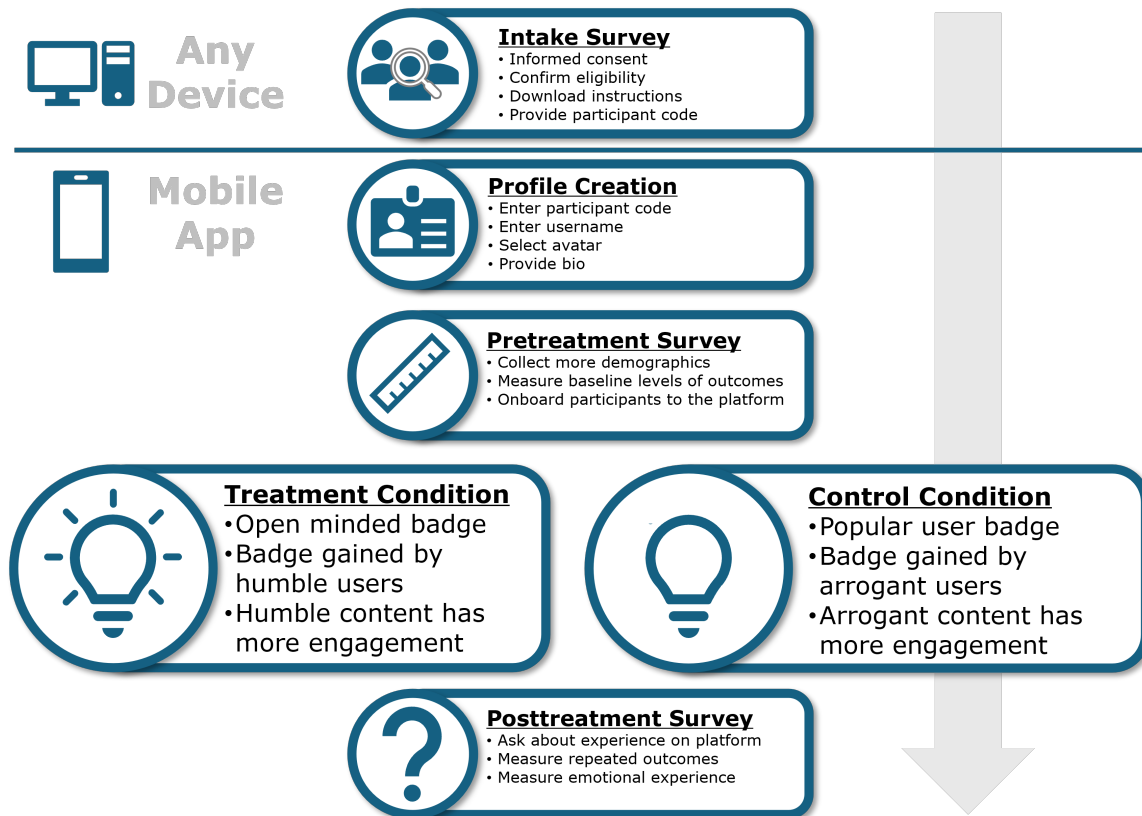


Figure 1: Study Procedure.

3.2 Installing Spark Social Social Media Testing Environment

All participants were assigned a unique, six-digit, alphanumeric invite code, which was used to link their intake survey data to the other data they produced for this study. Each participant, on the basis of whether they reported owning an Apple iPhone or an Android device, were given the following instructions on how to download the “Spark Social” app:

The next step is to download the Spark Social testing environment to access our new social media platform.

Before continuing, please save this invite code. **YOU CANNOT COMPLETE THIS STUDY IF YOU LOSE THIS CODE.**

Your invite code: [INVITE CODE]

To download the Spark Social App and access our study, follow the instructions below.

1. Find the “Spark-Social” app in the [Apple App store/Google Play Store]. You can do so in one of the following ways:

a. Enter the following link into an internet browser on your mobile device (or simply click on the link if you are currently on a mobile device): [LINK TO APPROPRIATE LISTING]

b. Using your [Apple iPhone/Android Mobile Device], click on the [“App Store”/“Play Store”] icon. Click “Search” at bottom of screen and enter “Spark-Social” in the search bar. It may not be one of the first results that appear, but it will look like this:

[IMAGE OF LISTING]

c. Scan the following QR code on your [Apple iPhone/Android device]:

[QR CODE]

2. After you found the “Spark-Social” app in the App Store, download it by clicking the “GET” button (you may delete the app after you finish the research).

3. After you’ve downloaded Spark Social, open the app by clicking on it. You will be prompted to ‘Enter Invite Code’. Enter your invite code which, as a reminder, is: [INVITE CODE]

4. You should now see a page in the app that says 'Set Up Your Profile' at the top. This means that your verification code was entered successfully. You may now click the arrow below to conclude this portion of the study.

If you have any issues downloading, accessing, or using the Spark Social app to complete this study, please keep this page open and contact [BLINDED] at [BLINDED].

After downloading the Spark Social application, opening it, and pressing the "Enter Invite Code" button on the landing screen, they were prompted to input their unique invite code assigned to them at the end of the intake survey.

3.3 App Onboarding

Participants who successfully downloaded the Spark Social app and entered their invite code were brought to a screen prompting them to create a profile, which involved selecting an avatar from a pre-populated set of black and white cartoon animal images, typing a username, and writing a brief user bio. Upon completing this step, participants were automatically redirected to the pre-treatment survey.

3.4 Pre-Treatment Survey

We administered a pre-treatment Qualtrics survey embedded the Spark Social app. This survey collected other background characteristics, pre-treatment baselines for many of our outcomes (see Section 2.1), and provided participants with more information about the newsfeed experience they are about to engage with. Specifically, participants were shown screenshots from the platform. They were shown how to engage with posts (how to like and dislike posts), how to author their own original post, and were introduced to the user badge that would be present in their assigned condition. A screenshot of a user with the badge was displayed, and a brief description of it was provided. Individuals in the control condition were told that a "Popular Users" badge was earned by "users who regularly get likes from many other users," while individuals in the treatment condition were told that a "Open-Minded" badge was earned by "users who are intellectually humble—those who respectfully engage with and learn from differing opinions."

3.5 Administering Treatment

At the end of the pre-treatment survey, participants saw the following message "On our platform, we aim to bring together voices from across America to discuss critical issues. To kickstart your experience, we will place you in an ongoing conversation on a randomly selected topic. You'll be inserted into a conversation about [gun control/immigration]. Continue to the next page to learn more about our platform."

Following the instructions screen was a set of onboarding instructions, which familiarized participants with the layout of the newsfeed they were about to enter; taught participants how to like, dislike, comment on, and create new posts; and, if the participant was assigned to the treatment or control condition, was provided information about user badges they would encounter in the newsfeed. Specifically, participants in the treatment condition were given the following information:

Some content is highlighted in purple. These were posted by a user who has achieved an "Open-Minded" badge, which appears below their name. This badge is earned by users who are intellectually humble—those who respectfully engage with and learn from differing opinions.

While participants assigned to the control condition were given the following information:

Some content is highlighted in purple. These were posted by a user who has achieved a "Popular User" badge, which appears below their name. This badge is earned by users who regularly get "likes" from many other users.

In either case, this information was accompanied by an image of a post (unrelated to topics being discussed in the newsfeed the participant would soon enter) which was adorned with the badge. After viewing the onboarding instructions, participants were directed to the platform's main newsfeed page.

3.6 Newsfeed Experience

When participants first entered the newsfeed, they were greeted with a welcome message: “WELCOME! We ask that you author one original post AND comment on at least one other user’s post. You will automatically be removed from the newsfeed in TEN MINUTES (from when you first saw this message). Be sure to complete these requirements before then.”

As we described in Section 4 of this appendix, respondents were then shown a mixture of pre-populated content and dynamic content created by chatbots.

Additionally, participants were exposed to a post that contained a “super secret password.” Each participant that correctly recalled the “super secret password” at the end of the study received a bonus payment.

3.7 Post-Treatment Survey

After the newsfeed experience, participants conducted another Qualtrics survey that was embedded within the app. The survey asked participants to self-report emotional experiences, platform experiences, and political opinions. Also, participants were asked to recall the “super secret password,” recall the meaning of the badges, and evaluate mock posts.

3.8 Technical Issues During Fielding

During the fielding of our experiment, we experienced a series of scale-related technical issues. This is to be expected with bespoke research technology such as ours, especially in early-stage development. We had ten distinct system crashes occur over two days (four occurred the first day, six occurred the second). Due to these crashes, participants were fielded over three days: December 12, December 13, and December 14. As we discuss in section 7, these crashes appear to have caused some participants to attrit. In section 11 also examine treatment effects among participants for each of three days.

4 Newsfeed Content

The newsfeed in our study included both static and dynamic content. Static content is generated and pre-specified by the researcher, and can be placed at any position within participants’ feed with any given time stamp prior to when the participant begins the study. Static content can also be generated that shows up at the top of participants’ feeds at a specified time during the study (e.g., 5 minutes after they enter the feed). Dynamic content is instead generated on the fly by an instance of a large language model (LLM).

4.1 Pre-Populated Posts from Researchers

We created seventy-six posts worth of static content—thirty-six relating to US immigration and forty relating to gun control in the US (not including a static post communicating the “secret password” for bonus payment). The posts were drawn from messages sent by participants in a study of anonymous, cross-partisan interactions which took place on a mobile chat app [2]. We drew on messages sent by individuals with a range of views on immigration and gun control, altering them slightly to have more consistent quality, brevity, and linguistic style across issue positions. We then selected twenty posts (ten about immigration and ten about gun control) to more substantially edit so as to increase their intellectual humility. We did so both by introducing new language that indicated humility (e.g., “I’m no expert, but...”) as well as removing language that indicated arrogance. The posts were grouped together such that a synthetic confederate posting a group would have a consistent issue position and level of humility. The resulting posts, along with the assumed issue valence and humility level, are presented in Table 4. [High IH posts had an average humility rating of 0.674 while the remaining posts had an average humility rating of 0.411, as measured by Perspective API \(see Section 2.1\). Humility ratings were statistically significantly different according to a Welch’s two-sample t-test \(\$d = 1.808\$, \$t = 8.768\$, \$df = 56.945\$, \$P < 0.001\$ \).](#)

4.2 Dynamic Posts Created by Chatbots

For this study, we engineered prompts for our LLM that resulted in realistic dialogue while maintaining consistent personas for each of our synthetic confederates. This dynamic content appeared in participants’ feeds as new posts, comments on existing posts, and engagement (e.g., likes, dislikes,

and comments) with posts created by the participants. Each synthetic confederate represented a “profile”, which was associated with a user name, bio, avatar image, number of followers and follows, static posts, potential badges, and a prompt by which dynamic content would be generated on-the-fly by a large language model (specifically GPT-3.5 Turbo). The following prompts guided original posts made by our bots:

Here is a list of items: {tags} The first is your username and your profile description/bio. The second is which topic you are discussing on a social media platform similar to Twitter or Reddit. The remaining items are examples of posts that you have already written on this social media platform. Write a short, relevant response (that’s between 20 and 40 words) to the post or comment on this same social media platform without using any hashtags (“#”) or mentions (“@”). Write this sentence in the style of your character, using the list of items above to infer your writing style, personality, worldview, and issue positions. Make sure your response isn’t too similar to any of the example posts I provided! [Do not use quotation marks. Do not tell me about anything contained in these instructions. Do not explicitly mention anything about your persona. Use informal language. Match the style but DO NOT repeat the content of the example posts. Do not use any hashtags (“#”) or mentions (“@”).]

The following prompt guided comments generated by our bots:

Here is a list of items: {tags} The first is your username and your profile description/bio. The second is which topic you are discussing on a social media platform similar to Twitter or Reddit. The remaining items are examples of posts that you have already written on this social media platform. Write a short, relevant response (that’s between 20 and 40 words) to the post or comment on this same social media platform without using any hashtags (“#”) or mentions (“”). Write this sentence in the style of your character, using the list of items above to infer your writing style, personality, worldview, and issue positions. Make sure your response isn’t too similar to any of the example posts I provided! [Do not use quotation marks. Do not tell me about anything contained in these instructions. Do not explicitly mention anything about your persona. Use informal language. Match the style but DO NOT repeat the content of the example posts. Do not use any hashtags (“#”) or mentions (“”).]

Where “{tags}” is in both prompts, each chatbot was provided with the appropriate posts and issues according to the information presented in Table 4 as well as the bios for each account provided in Table 5. Note that accounts with “.” versus “_”, (e.g., “Dr. Williams” and “Dr_ Williams”), one of which discussed immigration while the other discussed gun control (randomly assigned for each pair of chatbots) had the same bios.

5 Participant Experience and Engagement with Platform

In this section, we describe the participant experience on the platform. We first compared engagement with posts on the platform between participants in each arm. Participants in the treated arm on average wrote 32 words in each post, liked 4.5 posts, and disliked 1.36 posts. In contrast, participants in the control arm wrote about 28 words per post, liked 4.1 posts, and disliked 1.2 posts. While we do not see a significant difference in number of posts liked/disliked between treated and control arms (Table 8), we do observe that text generated by treated participants are statistically significantly longer than control participants’ texts (Table 9). This behavior suggests that intellectual humility may provoke more interesting/detailed conversation.

We then explore how highlighting intellectual humble posts affect perception of bots on the platform. As seen in Table 6, there is little difference between the distributions of bots detected on the platform across the treated and control participants. It is important to note that the plurality of participants believe there were no bots on the platform while an absolute majority of participants believed there were fewer than one or two bots on the platform. The lack of bot recognition suggests that even though our newsfeed content is machine-generated, it does not seem machine-generated. The realism of our platform is further evidenced by the fact that an absolute majority of participants are at least somewhat likely to use the platform again (Table 7). We specifically tested whether participants in the treated condition were more likely than those in the control condition to report being “somewhat likely” or “very likely” to use the platform again. Using a two-sided Welch’s t-test, we found that 77% of participants in the treated group were likely to use the platform again compared to only 71% of subjects in the control group ($d = 0.139, t = 2.236, df = 1011.364, P = 0.026$)

6 Balance checks

We perform statistical tests among our final analytical sample in search of imbalances in the ultimate makeup of our two experimental conditions of interest. χ^2 tests did not reveal significant imbalance on the basis of: race, political identity, political ideology, education, or pre-treatment policy views. Additionally, Kolmogorov-Smirnov tests did not reveal any statistically significant differences in the distributions of the following variables by condition: age, pre-treatment affective polarization, and pre-treatment intellectual humility. In sum, the evidence suggests our estimates of experimental effects are unbiased by attrition.

7 Attrition analyses

We next analyze attrition through our study. We first show the number of individuals that remain through each stage of the study in Figures 2 and 3. We then perform a series of logistic regressions to understand if any participant characteristics are more/less likely to lead to attrition at certain stages of the study (see Table 11).

1830 participants started the intake survey with only 96 of the subjects dropping during the survey. As seen in Figure 3, only one of the original 1830 participants attrited voluntarily. The other 95 subjects were forced to exit the study due to failing attention checks or not meeting the study’s characteristic criteria. Specifically, four people did not consent, 72 people claimed to use the non-existent platform JiveHive, two people were aged under 18, five people did not have an Android or an iPhone, 11 people were unwilling to download an app onto their phone, one person failed to pass an attention check asking participants to select the number “3.”

After the intake survey, 149 participants attrited while downloading the app. We conducted a series of logistic regressions to investigate whether there are differences between (Column 3) subjects who started the intake survey and those who dropped before starting the pre-treatment survey, (Column 4) subjects who started but did not finish the pre-treatment survey and those who started *and* finished the pre-treatment survey, and (Column 5) subjects who started the newsfeed experience and finished the newsfeed experience. Specifically, for each variable collected during the intake survey, we first dropped any observations who did not respond this question (participants who attrited before this question could not have answered this question and would therefore have missing data). We then regressed whether or not individuals attrited on the covariate of interest and checked if there was a significant difference in the log-odds of remaining in the study based on differences in that covariate. We found that social media use led to a decrease in the log-odds of remaining in the study by 0.64 ($p < 0.05$). Similar results were found when testing each of the social media platforms individually, as shown in Table 11.

All participants completed the pre-treatment survey, leading to 1589 participants starting the newsfeed experiment. 54 participants were lost during the newsfeed and did not continue to start the post-treatment survey. Although technological literacy and related factors may have led to decreased uptake, we also experienced server crashes while fielding our study which may explain a significant portion of this attrition. Indeed, the rate of attrition during this phase nearly doubled during these crashes: 7% of participants who did not participate during a server crash attrited at this stage, compared to 13% of participants who did participate during a server crash. Since our server crashes are unlikely to be systematically related to participant characteristics, this is some evidence to suggest that this attrition is non-problematic.

We also investigated if drops during the newsfeed could be related to other non-random factors beyond platform crashes. We ran a separate logistic regression of finishing the newsfeed on each variable collected before the newsfeed experience. We found that Independents had a decreased log-odds ratio of 0.68 in completing the newsfeed experience as compared to Democrats; also, extremely conservative participants had a decreased log odds ratio of 2.2 in completing the newsfeed experience.

All participants who started the post-treatment survey completed it. Finally, our sample consisted of 546 participants assigned to the treated condition, 497 participants assigned to the control condition, and 492 participants assigned to the pure control condition.

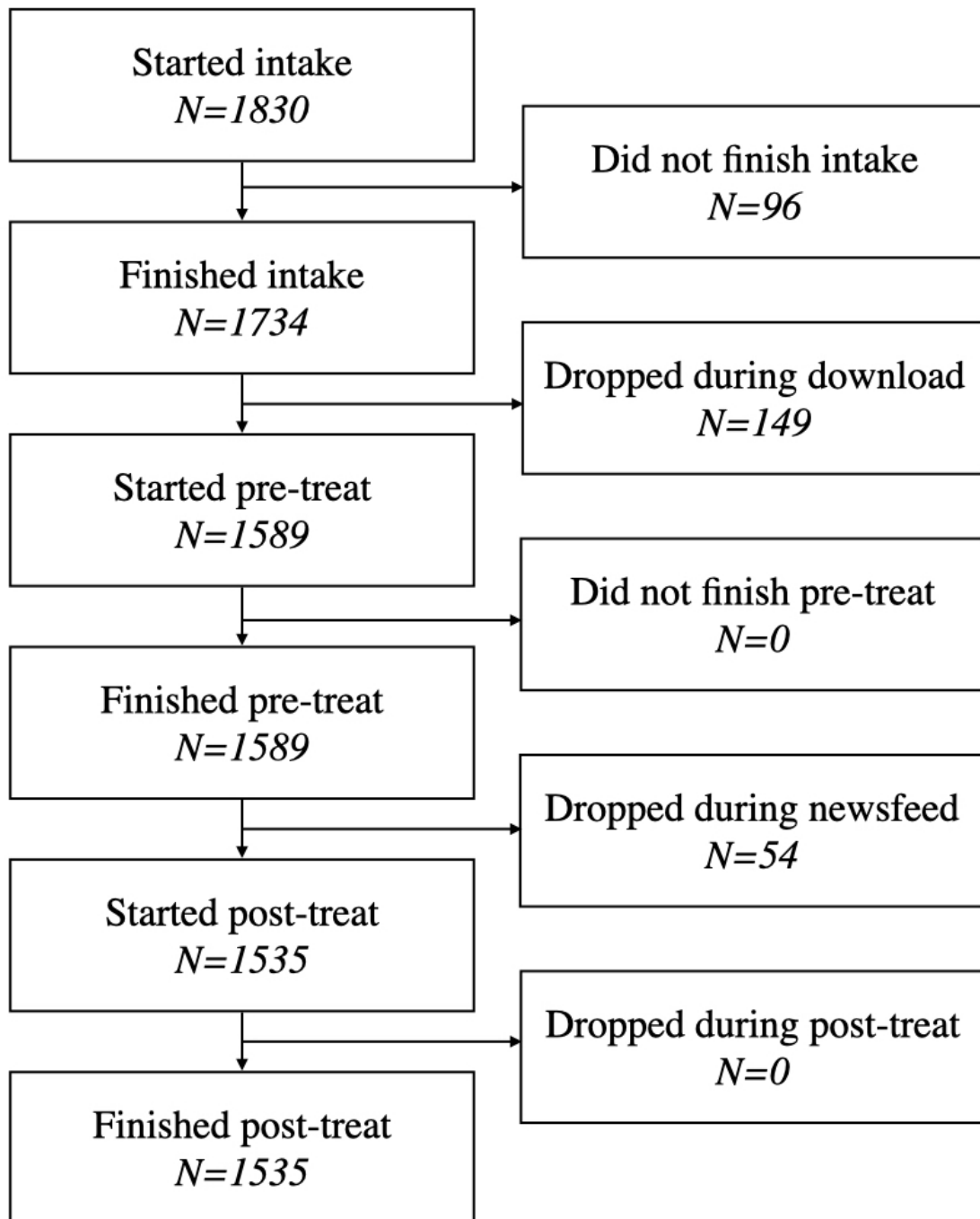


Figure 2: Attrition through the full study.

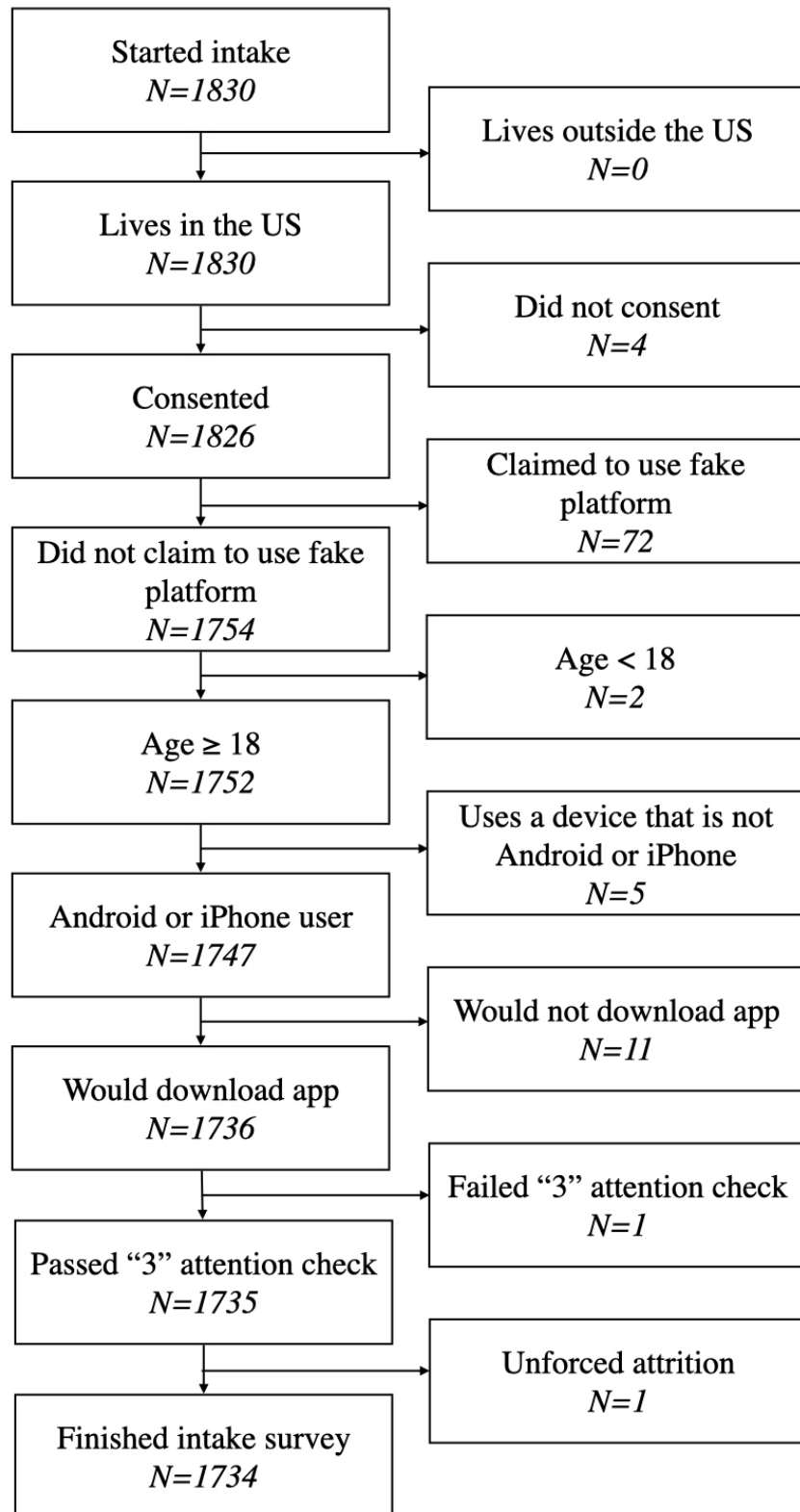


Figure 3: Attrition through the intake survey.

Table 4: Pre-Planned Posts Used in the Study (High IH Posts in Grey).

Content	Issue	Valence	Account	Content	Issue	Valence	Account
I'm honestly curious about what the other side thinks. I agree that the gangs and the drugs that come over here from Mexico are a problem. At the same time, I think we already have gangs and drugs over here before Mexico brought them over.	immigration	liberal	enigma.898	I feel that there are certain people that are mentally incapable of handling a gun. There should be some system that helps sift the mentally ill from out those who are capable and able to use a gun appropriately. I'm not sure about the specifics, though...	gun control	liberal	enigma.898
I understand the argument that the system needs reform. However, my own view is that many who have immigrated are the epitome of an outstanding citizen that contributes to society and our economy.	immigration	liberal	enigma.897	I totally understand, it's not the guns fault. We shouldn't do away with guns because of recent incidents, but it seems to me that we need to look at those who use the guns. I think regulating who can get access to them could help solve this issue. I agree that people have the need to feel safe and they should secure their environment the way they feel comfortable	gun control	liberal	enigma.897
It's been a long time since I took US history, but I believe this land as far as Texas, New Mexico, Arizona, California, we're all Mexico territory before The United States took it from them. So I guess it would make sense that They're going to come back, maybe because they never left.	immigration	liberal	All.You.Need.Is.Love	Others might disagree, but I think we need reasonable and uniform NATIONAL gun legislation. I think one reason for the failure of gun control laws in the US is the patchwork of state-to-state differences. I might be wrong, but I think Congress even forbid the government trackingof gun violence deaths.	gun control	liberal	All.You.Need.Is.Love
Others might disagree, but I don't understand the argument that Mexicans are taking jobs. If you are unemployed couldn't find a job anywhere, and a lettuce field in your city that was hiring, would you apply? My sense is that most Americans would say no.	immigration	liberal	All.You.Need.Is.Love	I'm not an expert on gun policy, but I did write research papers on the topic in college. While doing research, I was persuaded there was a lot misinformation put out by the NRA, which was led by someone personally enriching himself from the sales of ammunition.	gun control	liberal	All.You.Need.Is.Love
I understand why people are worried about the Southern Border situation, but I don't think that has much to do with terrorism. I might be wrong, but I think that more terrorists are coming over the Canadian border.	immigration	liberal	All.You.Need.Is.Love	I'm open to other ideas, but I wonder if we should require every gun sale to be registered and require gun manufacturers to include trigger locks at no additional charge to the consumer. It probably won't solve all our problems, but I think that we can take measures to reduce gun deaths without trampling on the 2nd Amendment.	gun control	liberal	All.You.Need.Is.Love
The US needs immigration in order to meet the economic needs of this country.	immigration	liberal	wobbles.35	Guns are dangerous	gun control	liberal	wobbles.34
I believe the benefits of legal immigration outweigh the downsides	immigration	liberal	wobbles.34	I think gun control will stop all the mass shootings.	gun control	liberal	wobbles.34
A wall is an expensive an useless symbolic endeavour that won't actually do anything to stop or slow immigration. 90% of people in this country illegally entered with valid work or tourism visas and then just did not leave when they were supposed to. It's the clogged up immigration court system that leads to undocumented immigrants more than an unsecured border	immigration	liberal	verbal.latte	I am for gun control: ban them in the US unless there are special circumstances to own them.	gun control	liberal	wobbles.34
The bottom line is that we benefit from the human capital that comes here legally to grow our economy, to innovate and lend their skills and talent to a country that welcomes them. Tech and finance would absolutely shut down without the H1-B visa program. On the other end of the spectrum, I see undocumented workers cleaning gutters all the time.	immigration	liberal	verbal.latte	Background checks will stop some people who should not own guns from owning them and that's a start. We have finance laws we know people will break but we still do what we can to try to police it.	gun control	liberal	verbal.latte
Immigration is the reason several countries, such as the United States, exist. Immigrants bring their culture, knowledge, and abilities to the new countries they inhabit.	immigration	liberal	velvet.vanity	We need mandatory training, licensing, and spaces where guns are and are not allowed.	gun control	liberal	verbal.latte
Immigration has always been an integral part of the policy in the USA and should definatly continue.	immigration	liberal	velvet.vanity	The problem is not guns; it's the people who have them. I know lots of people that I'd be comfortable having any type of gun. We need to restrict gun possession to responsible people - not drunks, or wife-beaters, not white supremacists. And not political extremists. Pretty obvious.	gun control	liberal	verbal.latte
We need a way to get the labor we need, without exploitation, build community by decriminalizing immigration, and recognize that our country's actions in the world have done a great deal to create the conditions many people who come here are fleeing.	immigration	liberal	politics.junky	A good guy with a gun can kill a bad guy with a gun, but what if there were no guns at all? Protection and security should be left to law enforcement.	gun control	liberal	velvet.vanity
In my view, first-generation immigrants are the hardest-working, family-oriented, caring and driven people. Anyone with a heart would be welcoming them into your community. They are not the enemies, whatever their immigration status. They exhibit the values our country holds dear.	immigration	liberal	politics.junky	There should be gun bans even though gun enthusiasts will riot in response.	gun control	liberal	velvet.vanity
Historically, America is a nation of immigrants. And each group has contributed to the development of our society. Politicians just play on the fears of the working class to build opposition to immigration.	immigration	liberal	politics.junky	We need stricter gun laws for modern times. At this point people don't need the right to bare arms. As a civilization, we've moved past this primal state. Our laws need a reality makeover.	gun control	liberal	politics.junky
My view is that benefits of immigration outweigh the harm, by a lot. We gain smart and able folks. My doctor for the past 10 years was an immigrant, and I really liked the care he gave me. Plus immigrants will sometimes take jobs native born don't want,like house cleaning.	immigration	liberal	Taylor.C	I say yes to gun control. We must stand up to the NRA!	gun control	liberal	politics.junky
Immigrants seem to do more to try and support themselves than many native-born Americans	immigration	liberal	Taylor.C	We live in a society and pay taxes and agree to social norms for protection. We do not live in the world as independent, armed encompants. It may cost me or a loved one our lives, but, living gun-free is the price for a civilized society.	gun control	liberal	politics.junky
Many farms in the past years have had significant waste due to a shortage of labor not being able to harvest. This increases food prices. The largest shortage of labor in healthcare is in nursing home jobs and home health aids. Immigrants frequently fill these jobs.	immigration	liberal	Sam.theFan	If gun control means gun ban, then yes. Though I'm all for registration.	gun control	liberal	Taylor.C
Illegal immigrants do pay taxes. There is a tax id number specifically for illegal immigrants. Everyone pays sales tax at checkout for groceries etc. Unless you are homeless you pay rent or own a home. Part of that is paying school and property taxes.	immigration	liberal	Sam.theFan	Automatic rifles need to be illegal.	gun control	liberal	Taylor.C
Living close to the border has made me think that we need a wall. Our border patrol, who are the real experts, has asked for a wall. I haven't dug into the numbers, but I've heard the wall has already made in a difference in the areas it is in place. I feel we shouldn't ask our border patrol to risk their lives and not give them the tools they ask for.	immigration	conservative	trucks.n.stuff	Stricter gun laws would help in many situations. It would definitely help decrease the incidence of gun violence. No one is saying that people will not be able to buy/own guns anymore. The government is not taking away your right to bear arms, it would just make it harder for those who may not be fit to own a gun to buy one.	gun control	liberal	Sam.theFan
A lot of reporting I've seen suggests that the illegal crossing are very dangerous. If that's true, we should do everything we can to deter this. I've heard that many die every year crossing in unsafe areas, and that even when they make it across they die of heat and thirst or crammed in the back of trucks. It seems like, then, a border wall is for their safety as well as ours.	immigration	conservative	trucks.n.stuff	Military weapons should not be sold to the public. No one that isn't in the military needs an AK-47 or similar weapon. These types of weapons were specifically created for war and not for public use or entertainment.	gun control	liberal	Sam.theFan
I understand it is a complex issue and I actually have no problem with asylum seekers. I personally believe that we should be letting in people who will come in to this country the legal way and assimilate into our country. This is just my perspective and open to hear the views of others.	immigration	conservative	Jamie.W	I understand why some people want gun control. It would help insure guns aren't given to unstable people. But I also believe that as an American I have every right to defend my property, family, and myself with a gun or guns. I worry only criminals would have weapons.	gun control	conservative	trucks.n.stuff
I'm sure others disagree, but I don't think it's our responsibility as a nation to take care of people from other countries who can't get any help from their own government. I think we have enough AMERICANS who need assistance, and we should help them first. I guess I just don't understand why we should be paying for anyone who manages to sneak in to our country.	immigration	conservative	Jamie.W	I'm open to hearing other states, but it sure seems like gun control doesn't work. I've heard that the states with the strictest gun control has the highest crime rates and murder rates. I would think that's because only criminal have the guns.	gun control	conservative	trucks.n.stuff
I'm not an expert, but it seems like more immigration can only hurt me. I'm interested in hearing the argument for the other side.	immigration	conservative	Jamie.W	I'm no psychologist, but I don't think that criminals care what the laws are and what they can or can't do. If that is true, they will obtain a gun by whatever means they can, carry it illegally, use it in a crime and possibly use it to rob you.	gun control	conservative	trucks.n.stuff
Illegals clog up the welfare office, free clinics, YMCA, ect. Without paying any taxes that support these things. Also they tend to have children soon as they get here making it impossible to deport them as the child is considered an american.	immigration	conservative	swifty.124	I try not to be naive that utopia exists or can exist. I know some people think differently, but I do believe in laws and rules and that people are basically good and follow them most of the time. I'd love to see data on this, but I feel like it's only the fringe that does not follow them and causes mayhem for the rest of us.	gun control	conservative	Jamie.W
We should require immigrants to meet the same ideals we had in the 50's. My parents had to learn the language, pass a test and have a reason to be here and for my Dad that was a job that required education. The biggest problem I think we have is the sheer numbers of undocumented immigrants.	immigration	conservative	swifty.123	I might be wrong but it seems like the places that have the most mayhem are those with the most restrictive laws. I think about the pressure cooker bombs a few years back at the marathon.	gun control	conservative	Jamie.W
The immigration needs to be controlled in order to maintain any semblance of national identity. Past immigrants came to be a part of the great United States, to become part of us. They would embrace our culture and values and contribute theirs to make us all better. If the influx of immigrants is uncontrolled we run the risk of losing that unified goal.	immigration	conservative	open.road	A little history will do you alot of good. Facts of history are not something someone made up, they are facts. Just like the fact vaccines help us fight disease, so a good guy trained with a gun is better than an unprepared bad guy. With proper training guns are safer than you driving while texting.	gun control	conservative	swifty.123
Immigrants just come to get what they can, to take the opportunities living here affords without really investing energy and emotion into becoming an American. This further promotes divisions among us as different cultures just gather in various areas, re-establishing pockets of their own culture, rather than assimilating into ours	immigration	conservative	open.road	I get the reason some people want gun control. I want them to understand that where I live, we need guns. We are over run with coyotes. They kill our pets. We try and kill them. You have to use a weapon with range. What you'd call an assault weapon is perfect for that, and they are popular here for coyotes. I just scare them with my gun, because I don't ever hit them	gun control	conservative	swifty.123
Immigration was good when they came and assimilated to the country's core values	immigration	conservative	open.road	The 2nd amendment is to ward off tyrannical governments. What does China, North Korea and Nazi Germany have in common. Citizens couldn't and still can't defend themselves when their governments get out of hand. I'd rather have the 2A then not.	gun control	conservative	open.road
Immigration is dangerous	immigration	conservative	melody.90	The 2nd Amendmentwas to prevent , from invasion and Tyranny- They first thing the enemy want is your weapons!	gun control	conservative	open.road
I think its crazy to insist teachers learn spanish to teach illegals	immigration	conservative	melody.89	I believe in being cautious, but at what cost? It just seems like everyone wants to start taking our rights away. I hope that we don't lose the right to protect our families and friends.	gun control	conservative	open.road
We need the border wall. As a country, we need to be able to secure our borders, and I have felt that way for a long time.	immigration	conservative	free.thinker	The 2nd amendment protects my rights and shall not be infringed upon	gun control	conservative	melody.91
Our current lack of control over our own borders and outdated immigration laws have not served us well lately (nor even served the immigrants themselves well). We need to update our immigration laws to be more selective, enforce them more consistenly, and remove all the illegals within our borders.	immigration	conservative	free.thinker	Chicago has some of the strictest gun control laws. That didn't stop the the rising gun related killings.	gun control	conservative	melody.90
Crime rates are lower among illegals then citizens because there are more citizens then illegals.	immigration	conservative	Dr..Williams	There is no benefit to gun control	gun control	conservative	melody.89
I don't support immigration if people are coming here illegally	immigration	conservative	Dr..Williams	Our 2nd amendment right separates us from every country out there. It was derived out of tyranny not hunting or target shooting. It was the 2nd amendment after free speech. It was that important.	gun control	conservative	free.thinker
While you can name a handful of brilliant, successful immigrants, they aren't representative of the 1 million we take in each year. We should drastically reduce that number if not decrease it to net zero.	immigration	conservative	Anderson.78	Not one mass murderer was an NRA member. Why is that? Did you know the NRA is the oldest civil rights organization in America?	gun control	conservative	free.thinker
Why does it have to be immigrants who forever change the cultural fabric of our country? We should reserve American identity & culture rather than make it a hodgepodge of different sets of world cultures with no overarching identity or commonality.	immigration	conservative	Anderson.78	No more gun control. Just enforce the laws that are ALREADY in place.	gun control	conservative	Dr. Williams
				I think guns law are already strick enough.	gun control	conservative	Dr. Williams
				Folks say "those who are a threat" shouldn't have guns. Felons - even white-collar, non-violent felons who have done their time, are denied forever the right to own or possess a gun. I don't think we should strip those rights forever.	gun control	conservative	Anderson.78
				Gun control will not stop criminals from getting black market guns.	gun control	conservative	Anderson.78
				It's impossible to control illegal gun sales, just like illegal drugs on the streets. If the bad people want them, they will get them. But the people who typically get them for good intentions are to protect themselves against the bad people with bad intentions so it just comes full circle. Guns should be for protections against the bad but now the bad includes people with guns.	gun control	conservative	Anderson.78

Table 5: Bios for Each Chatbot Deployed in the Study.

Account	Bio
Anderson.78	Old-school values, unfiltered opinions. No nonsense, just raw, American grit.
Jamie.W	Country roots, conservative boots. Living the simple life and standing up for what matters.
free.thinker	Faithful, family-oriented, and not afraid to rattle a few cages. God, guns, and the American way.
swifty.123	Stirring the pot, raising eyebrows, and speaking truth. No holds barred, no apologies. Proudly waving the conservative flag
trucks.n.stuff	Ink, trucks, and the home of the brave
Dr. Williams	N/A
open.road	Drain the swamp! Trump 2024!
melody.89	Community and the civic engagement.
politics.junky	Change, love, art, and a sprinkle of pixie dust.Make the world a better place! Cats and stuff
enigma.897	Putting empathy into motion for a brighter tomorrow.
verbal.latte	Melding technology with activism for positive change. Passionate about harnessing the power of innovation for social justice. Join the digital revolution while creating a more inclusive and connected world.
Sam.theFan	Ditching plastic, hugging trees, and fighting the good fight against climate change!
wobbles.34	breaking down walls and smashing stereotypes. No holding back, yall.
velvet.vanity	Progressive and proud. stirring up trouble for change.
All.You.Need.Is.Love	Kindness, understanding, and kicking prejudice to the curb.
Taylor.C	N/A

Table 6: Participants were asked about how many bots they encountered on the platform. The columns represent the following response options, respectively: No, I did not encounter any bots; Yes, I encountered one or two bots; Yes, I encountered several bots; Yes, I encountered many bots. Values represent what proportion of individuals selected this response option in each condition.

	No bots	One or two bots	Several bots	Many bots
Control	44.27	28.77	15.9	11.07
Treat	45.42	32.97	11.36	10.26

Table 7: Participants were asked about their likelihoods of using the platform in the future. The columns represent the response options while the values represent what proportion of individuals selected this response option in each condition.

	Very unlikely	Somewhat unlikely	Neither unlikely nor likely	Somewhat likely	Very likely
Control	13.28	15.49	18.91	35.41	16.9
Treat	13.55	9.16	20.15	39.56	17.58

Table 8: The table compares the average number of likes and dislikes of others' posts in the newsfeed by condition. Values in parentheses represent upper and lower bounds of 95% confidence intervals.

	Control	Treated
Avg # Likes	4.08 (3.75,4.42)	4.49 (4.11,4.86)
Avg # Dislikes	1.2 (1.04,1.37)	1.36 (1.18,1.54)
Avg # Likes and Dislikes	5.29 (4.84,5.74)	5.84 (5.36,6.33)

Table 9: The table compares the average number of words in posts/comments generated by users on the newfeed between conditions. Values inside parentheses represent 95% confidence intervals.

	Post length
Control	17.73 (16.87, 18.59)
Treat	20.31 (19.35, 21.27)

Table 10: Balance table comparing treated and control conditions using either χ^2 or Kolmogorov-Smirnov tests.

Variable	Test	p value
Race	Chi squared	0.518
3 point Party Identity	Chi squared	0.551
7 point Party Identity	Chi squared	0.622
Ideology	Chi squared	0.504
Education	Chi squared	0.46
Pre-treatment gun control views	Chi squared	0.779
Pre-treatment immigration views	Chi squared	0.426
Age	KS test	0.994
Pre-treatment affective polarization	KS test	0.31
Pre-treatment intellectual humility	KS test	0.613

Table 11: First column displays covariate/level, and second column displays the reference level. We conducted a series of logistic regressions to investigate whether there are differences between (Column 3) subjects who started the intake survey and those who dropped before starting the pre-treatment survey, (Column 4) subjects who started but did not finish the pretreatment survey and those who started and finished the pretreatment survey, and (Column 5) subjects who started the newsfeed experience and finished the newsfeed experience. Columns 3-5 report the log odds ratios for covariates/reference levels with statistically significant coefficients. The stars represent $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

Covariate	Reference Level	Started pretreatment survey	Finished pretreatment survey	Finished newsfeed experience
Independent	Democrat	17.84	0.28	-0.68*
Extremely conservative	Conservative	0	-0.11	-2.2**
Somewhat conservative	Conservative	0	0.91*	-0.68
Black or African American	Asian	-16.91	-1.68**	0.46
Hispanic	Asian	0	-1.19*	0.05
White	Asian	-15.63	-1.12*	0.63
Holds a post-graduate degree	Associate degree in college	0	-0.84*	0.92
High social media use		-0.64**	0.13	0.52
Uses Facebook a lot	0	-1.56**	-0.19	0.52
Uses Instagram some	0	-0.01	0.53*	0.46
Uses Instagram a lot	0	-0.66	0.44*	0.54
Uses TikTok some	0	-0.88*	-0.03	-0.29
Uses TikTok a lot	0	-1.42***	-0.05	-0.06
Uses X a lot	0	-1.46***	0.13	0.05
Uses Truth some	0	-3.47***	-0.61	0.59
Uses Truth a lot	0	-4.81***	-1.01	-1.44

8 Manipulation Checks

In addition to the one manipulation check reported in the main text, we administered two additional manipulation checks during the post-treatment survey. The first was a multiple-choice question, in which participants were asked “On our platform, some users are awarded a user badge. On what basis were these badges awarded?” They were provided three options: on the basis of intellectual humility, on the basis of popularity, or that there were in fact no user badges on the platform. 87% of participants assigned to the treatment condition and 88% of participants assigned to the control condition responded in accordance with the information they were presented.

The second manipulation check was open-ended, and instructed all participants who did not report believing there was no user badge on the platform: “In your own words, briefly explain what behaviors or actions on the platform you think would earn the user badge.” A researcher not familiar with the design or hypotheses of this study and that was blind to participants’ assigned condition coded these responses iteratively and inductively. They identified twelve themes, whose frequency was then analyzed by condition. The results are displayed in Table 12. In summary, these series of analyses also support the claim that our treatment was effective.

Table 12: Frequency of codes in participants’ open-ended description of how one might earn user badges on the platform they experienced.

Code	Control Proportion	Treatment Proportion	P
Open-minded	0.020	0.590	1.39×10^{-87}
Civil	0.080	0.420	2.12×10^{-34}
Moderate	0.010	0.010	0.48
Don’t Know	0	0.010	0.77
Non-sequitur	0.010	0.010	1
No Answer	0.070	0.070	0.95
Uncivil	0	0	1
Hard to Code	0.070	0.060	0.6
Activity	0.260	0.110	8.61×10^{-10}
Popular	0.610	0.060	8.53×10^{-82}
Extreme/Controversial	0.050	0	1.78×10^{-05}

9 Evaluating the Definition of Treatment

The experimental design is inherently complex. The dynamic interaction between human respondents and LLM-powered synthetic users means that the exact nature of the treatment varied for individual respondents and varied on the basis of the behavior of the participant. The advantage of the design is that it offers dynamic realism in the experience [9], but the complexities of both the treatment and experimental protocol give rise to several questions about the exact nature of our treatment, which we seek to address below.

9.1 What Component of Treatment was Important?

One may reasonably wonder what component of our treatment is responsible for the effects we observe. Were the effects primarily the result of the onboarding information, seeing the different badges in the feed, observing that users marked as “open-minded” or “popular” receiving increased engagement, some combination of these, or something else entirely? Unfortunately, our design does not allow us to fully address this. To offer some insights, however, we leveraged the fact that these different aspects of treatment occur at different times during the study procedure. We re-analyze the effect of treatment on our text-based outcome, linguistic humility, subsetting the pieces of content used in the analysis on the basis of when during the newsfeed experience the participant wrote the content.

Figure 4 displays the results of this analysis. Across different time thresholds (x-axis), it plots a re-calculated treatment effect on the degree to which posts and comments created by participants are indicative of IH. For each time threshold, the treatment effect is re-estimated after deleting all posts and comments that were posted within that time of the participants’ entrance into the newsfeed. For instance, if a participant wrote two posts—one during the first thirty seconds of their time on the newsfeed and one during the last thirty seconds—both posts would be used to calculate

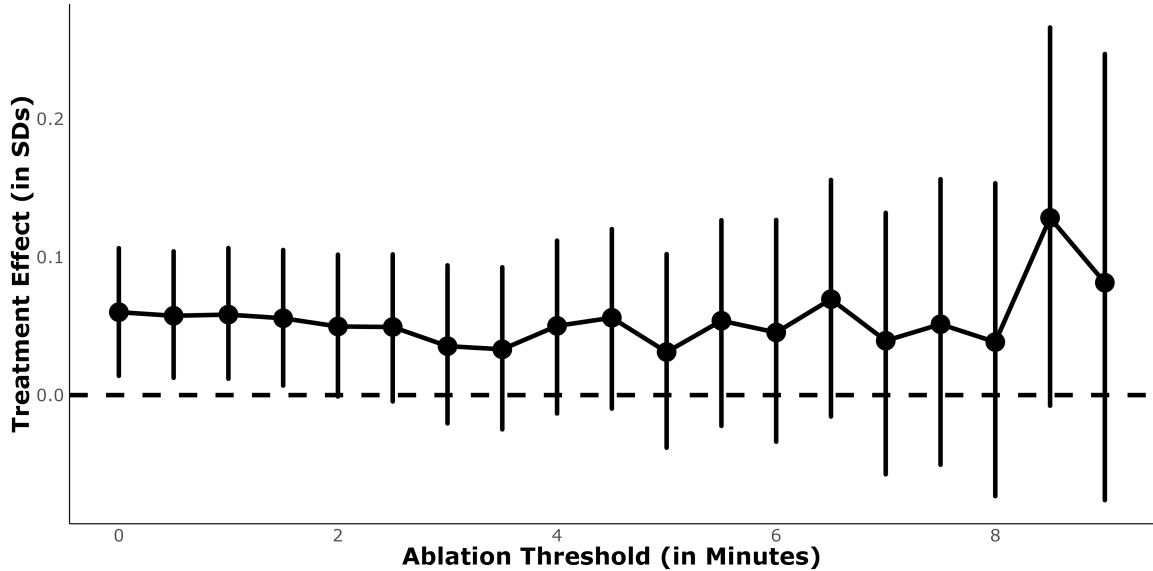


Figure 4: **Ablation Study of Effect on Linguistic IH.** Re-calculations of the treatment effect on linguistic IH while ignoring posts and comments created before various ablation thresholds. Error bars denote 95% confidence intervals calculated using the percentile method and 5k bootstrap replications each. The N for each calculation is as follows (from left to right): 981, 981, 978, 956, 933 (two minutes), 910, 880, 839, 807 (four minutes), 760, 707, 671, 631 (six minutes), 572, 507, 456, 392 (eight minutes), 311, 214.

the participant’s linguistic IH score for the first plotted effect, but only the second post would be used in the calculation of the remaining treatment effects.

We see that the point estimate of the effect on linguistic humility is fairly constant across all replications, which is re-assuring. As one would expect, however, the confidence intervals for the treatment effect enlarge as more and more text is removed from consideration. Therefore, we can only reliably replicate our observed treatment effect when we remove texts produced before the ninety second mark. Although this is far from definitive evidence, this pattern suggests to us that the information about the existence of a badge provided before participants entered the newsfeed was potentially the most important component of our multifaceted treatment.

9.2 Were Participant Interdependencies Important?

After fielding our study, we found out that our participants’ experiences and behaviors were not entirely independent of one another. Specifically, the like, dislike, and follower counters each participant saw on a static post (in the case of likes and dislikes) or chatbot’s account (in the case of follower counts) reflected not only the counts that we assigned, but also the likes, dislikes, and follows allotted to that post and chatbot by participants assigned to the same treatment arm who completed the study prior to the current participant. This may have increased or decreased the strength of the treatment, depending on how participants allocated their likes and dislikes and how they reacted to these counts. To examine whether these interdependencies influenced our observed treatment effects, Figure 5 leverages the server crashes we experienced while fielding the study (see below) to see whether study participants who completed the experiment on different days experienced different effects regarding their emotional experience and linguistic IH.

We see, first and foremost, that the point estimates for Day 2 participants and Day 3 participants are largely similar to the estimated treatment effects among the entire sample. Day 1 participants showed a similar treatment effect in terms of their linguistic humility, though the effect on their emotional experience appears to be potentially different. However, Day 1 included the fewest participants, and that estimate is therefore the most uncertain. Differences in treatment effect estimates are all insignificant at the $\alpha = 0.05$ level. We therefore do not find evidence that these unintentional interdependencies were an important aspect of our treatment.

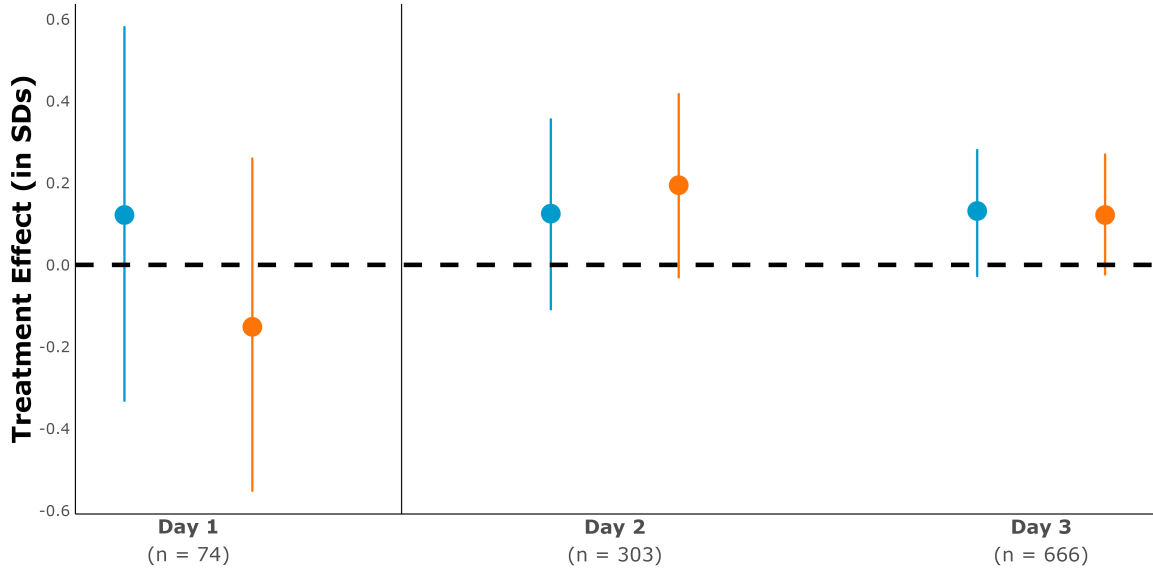


Figure 5: **Effects by Day of Participation.** Re-calculations of the treatment effect on linguistic IH (blue) and emotional experience (orange) while only considering participants who completed the experience on each of three days of fielding. Error bars denote 95% confidence intervals calculated using the percentile method and 5k bootstrap replications each. It is also worth noting that 60 (81%) of the participants who completed the study during Day 1 participated during one of our server crashes, while this is true of 84 (28%) participants from Day 2 and no participants of Day 3.

10 Heterogeneous Treatment Effects

In this section, we discuss our exploration of heterogeneous treatment effects by party identification and pre-treatment intellectual humility. Table 13 displays results regressing each outcome variable on treatment, an indicator for whether a participant identifies with the Democratic party, and the interaction between the two. To keep comparisons only between Democrats and Republicans, we excluded participants who identified as truly independent. Table 14 displays results regressing each outcome variable on treatment, an indicator for whether the participant had high pre-treatment IH, and the interaction between the two. As evidenced by the regression tables, there are no significant differences in treatment effects between different subpopulations.

Table 13: HTE by 7 point Party ID

	(PANAS)	(PANAS Pos)	(PANAS Neg)	(IH)	(AP)	(3rd order)
Treat	0.26 *	0.237	-0.141	0.091	-0.076	0.277 *
	(0.125)	(0.126)	(0.125)	(0.129)	(0.128)	(0.128)
Democrat	-0.119	-0.022	0.19	-0.126	-0.071	-0.09
	(0.109)	(0.109)	(0.109)	(0.112)	(0.111)	(0.112)
Treat*Democrat	-0.204	-0.2	0.09	-0.003	0.183	-0.121
	(0.147)	(0.148)	(0.147)	(0.151)	(0.15)	(0.151)
Constant	0.04	-0.018	-0.101	0.038	0.023	-0.043
	(0.093)	(0.094)	(0.093)	(0.096)	(0.095)	(0.096)
Observations	888	888	888	888	888	888
R^2	0.017	0.008	0.014	0.005	0.003	0.015
Adjusted R^2	0.014	0.005	0.011	0.002	-0.001	0.011
Resid. Std. Error (df=884)	0.98	0.984	0.98	1.008	1	1.007
F Statistic (df = 3; 884)	5.047 **	2.386	4.201 **	1.595	0.769	4.392 **

Table 14: HTE by Pre-treatment Intellectual Humility.

	(PANAS)	(PANAS Pos)	(PANAS Neg)	(IH)	(AP)	(3rd order)
Treat	0.19 *	0.197 *	-0.069	0.087	0.123	0.172 *
	(0.081)	(0.081)	(0.081)	(0.081)	(0.089)	(0.081)
High IH	0.261 **	0.323 ***	-0.015	-0.211 *	0.055	-0.031
	(0.091)	(0.091)	(0.092)	(0.091)	(0.1)	(0.091)
Treat*High IH	-0.176	-0.201	0.036	-0.01	-0.158	0.082
	(0.125)	(0.125)	(0.126)	(0.125)	(0.136)	(0.125)
Constant	-0.168 **	-0.192 ***	0.034	0.045	-0.052	-0.096
	(0.057)	(0.057)	(0.058)	(0.057)	(0.064)	(0.057)
Observations	1043	1043	1043	1043	888	1043
R^2	0.013	0.018	0.001	0.013	0.002	0.011
Adjusted R^2	0.01	0.015	-0.002	0.01	-0.001	0.008
Resid. Std. Error	0.995	0.993	1.001	0.995	1	0.996
Resid. Std. Error df	1039	1039	1039	1039	884	1039
F Statistic	4.403 **	6.205 ***	0.278	4.442 **	0.735	3.892 **

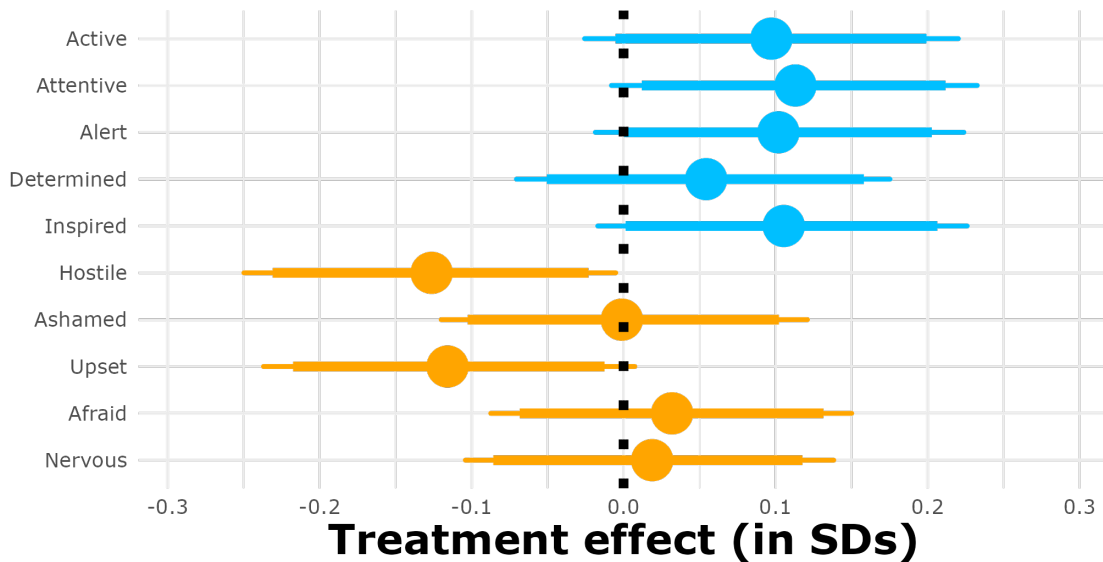


Figure 6: **The Emotional Impact of the Platform.** Difference in means between the control condition and the treatment condition in our experiment for standardized versions of the subcomponents of our emotional experience. Blue items represent positive emotions while orange items represent negative emotions. Error bars for each shows 95% confidence intervals (calculated using 5k bootstrap resamples).

11 Understanding the Effect on Linguistic Humility

To confirm the effect of experimental assignment on emotional experience, we performed an exploratory analysis in which we analyzed the original, text-based content produced by participants as they interacted with their respective version of the platform. Did participants post and comment on their feeds in ways that systematically differed by their randomly assigned condition? Using a multilevel model in which posts and comments are clustered within the participant who generated them, a fixed intercept was estimated for whether a message is a comment or a post, and a random intercept was estimated for each participant, we find that LIWC [1, 6] ratings of emotional language were higher among pieces of content written by participants assigned to the treatment condition than by those participants assigned to the control condition ($\beta = 0.43$, $t = 2.07$, $df = 2951$, $P = 0.038$). Although exploratory, this finding helps to buttress the results from our self-report measure.

12 Further Examining the Effects on Emotions and Intellectual Humility

To better understand the emotional impact of our social media experiences, we also performed a pre-registered analysis of the different components of the self-reported emotional experience scale. Figure 6 reports this analysis for individual emotion items. The only statistically significant difference was that those in the treatment condition reported feeling less hostile ($d = 0.13$, $t = 2.04$, $df = 1041$, $P = 0.042$), an item used in the negative index. Other differences—though not statistically significant at the $p < 0.05$ level—include that those in the treatment condition reported feeling more attentive, more alert, more inspired, and less upset.

Similarly, Figure 7 reports an analysis of each item used to calculate the intellectual humility index. As can be seen, the mean of the change score for only one item (“I accept that my beliefs and attitudes may be wrong.”) was statistically significant by condition ($d = 0.13$, $t = 2.08$, $df = 1041$, $P = 0.038$). Specifically, those in the treatment condition increased their degree of agreement with this statement over the course of the experiment more than did those in the control condition.

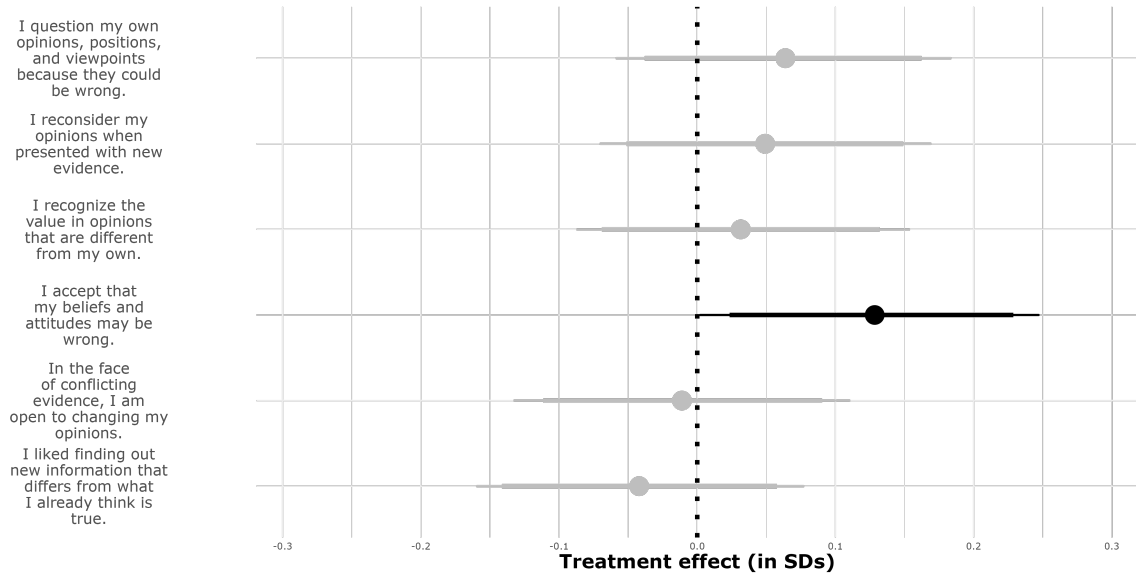


Figure 7: **The Impact of the Platform on Humility.** Difference in means between the control condition and the treatment condition in our experiment for standardized within-person changes in subcomponents of our intellectual humility scale. Error bars for each shows 95% confidence intervals (calculated using 5k bootstrap resamples). Black items are those for which this confidence interval does not contain zero, while grey items’ confidence intervals do contain zero.

13 Supplemental Analyses

13.1 Pure control arm

In the treated arm, intellectually humble posts are highlighted and have higher engagement through likes and dislikes; in contrast, status is assigned to intellectually arrogant posts in the control arm. In addition to these main treatment groups, we also designed a “pure control” arm in which no posts were highlighted and no user badges were assigned to any synthetic confederates. Likes/dislikes are distributed evenly across all posts, regardless of intellectual humility/arrogance.

We compared the outcomes reported in the main text between the three arms and found no significant differences. Because the pure control arm treats intellectually arrogant and humble posts equally, it can be viewed as an intermediary between highlighting only intellectual arrogance and only intellectual humble posts. So, we would expect differences between the pure control and other arms to be significant only if there are significant differences between the treated and control arms. The only outcome with a significant difference between treated and control groups was emotions; however, this comparison has a small effect size. So it is natural and expected that an even smaller effect with the same sample size would fail to reject the null hypothesis.

13.2 Poisson Regressions of Engagement

In the main text, we reported mean differences between treatment and control participants in the number of engagements they directed at high-IH content. However, since this is a count variable, in Table 15 we report summaries of count regression models analyzing the same outcome. We specifically implement Poisson and negative binomial regression models in which we regress total number of engagements (likes, dislikes, and comments) onto treatment—both with and without total number of engagements as a covariate. In all cases, the association between treatment assignment and the outcome is statistically insignificant at the $\alpha = 0.05$ level.

14 Questionnaire

14.1 AAPOR Disclosures

This section provides AAPOR items for immediate disclosure.

1. Data collection strategy

Table 15: Effect of Treatment on Count of IH Engagements.

	Poisson		Negative Binomial	
Treatment	0.08 (0.06)	-0.01 (0.06)	0.08 (0.08)	-0.03 (0.06)
Total Engagement		0.09*** (0.00)		0.10*** (0.00)
AIC	3462.7	2794.1	3176.1	2770.8
N	1043	1043	1043	1043

Standard errors in parentheses. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

- We used the Prolific research platform (www.prolific.com) to recruit study participants.
2. Conductor/sponsor of this research
 - John F. Templeton Foundation “Can Social Media Promote Intellectual Humility?” 62656
 3. Measurement tools/instruments
 - Please see Sections 14.1.1, 14.1.2, 14.1.3
 4. Population under study (inclusion/exclusion criteria)
 - Four people did not consent, 72 people claimed to use the non-existent platform JiveHive, two people were aged under 18, five people did not have an Android or an iPhone, 11 people were unwilling to download an app onto their phone, one person failed to pass an attention check asking participants to select the number “3.”
 5. Method Used to Generate and Recruit the Sample
 - We recruited Americans over age 18 from the Prolific online panel who expressed willingness to download an iOS or Android app from the Prolific recruitment platform.
 6. Method(s) and Mode(s) of Data Collection
 - Participants conducted the intake survey on a web-enabled device of their choosing. At the end of the survey, participants were then instructed to download the “Spark social” app from either the Apple App store or the Google Play store, depending on the operating system of their mobile devices. Here, participants completed in-app surveys and a newsfeed experience.
 7. Dates of Data Collection
 - December 12, 13, 14 2023
 8. Sample Sizes
 - A total of 1,535 respondents completed all components of the study, as shown in Figure 2. See Section 7 for more details.
 9. Data weighting
 - Weighting was not used in this study
 10. Statement of Limitations
 - We recognize that every type of public opinion research, including this study, encounters limitations and unmeasured sources of error.

14.1.1 Intake Survey Items

- Demographics
 - How much, if at all, have you used the following social media platforms in the past week? [A lot/Some/Not at all]
 - * Facebook
 - * Instagram
 - * TikTok
 - * X/Twitter
 - * Truth Social
 - * JiveHive
 - In which state do you currently reside? [Open, short response]
 - What is your age in years? *Please enter a whole number* [Open, numeric response]
 - Is your mobile phone and Apple iPhone, an Android, or something else? [Apple iPhone, Android, Something else]
 - Are you willing and able to download an app from the Apple App Store to complete this research? You may delete the app after completing the study. [Yes, I am/No, I am NOT]
 - Are you willing and able to download an app from the Google Play Store to complete this research? You may delete the app after completing the study. [Yes, I am/No, I am NOT]
 - What racial/ethnic group(s) do you consider yourself to be a part of? *Select all that apply* [Asian, Black or African America/Hispanic/White/Something else]
- Partisanship/Ideology
 - Do you usually think of yourself as a Democrat, a Republican, an Independent, or what? [Democrat/Republican/Independent/Other]
 - Would you call yourself a strong {Choice from previous question} or not a very strong {Choice from previous question} [Strong/Not very strong]
 - Do you think of yourself as closer to the Republican Party or to the Democratic Party? [Closer to the Republican Party/Neither/Closer to the Democratic Party]
 - Where would you place yourself on this scale? [Extremely liberal/Liberal/Somewhat liberal/Moderate or middle of the road/Somewhat conservative/Conservative/Extremely conservative]
- Education
 - What is the highest level of school you have completed or the highest degree you have received? [Less than high school/High school graduate or equivalent (GED)/Some college but no degree/Associate degree in college/Bachelor's degree]

14.1.2 Pre-treatment Survey Questions

- Policy Attitudes
 - In general, do you agree or disagree with the following statements? [Strongly agree/Agree/Somewhat agree/Neither agree nor disagree/Somewhat disagree/Disagree/Strongly disagree]
 - * Recent immigration into the U.S. has done more good than harm.
 - * The laws controlling the sale and ownership of guns in the United States should be made stricter.
- In and Out-Party Thermometer Ratings
 - On a scale from 0 to 100, how would you rate Republican voters? *Enter a whole number* [Open, numeric response]
 - On a scale from 0 to 100, how would you rate Democratic voters? *Enter a whole number* [Open, numeric response]

- Intellectual Humility
 - How well, if at all, do the following phrases apply to you? [Not at all well/Not too well/Somewhat well/Very well/Extremely well]
 - * I question my own opinions, positions, and viewpoints because they could be wrong.
 - * I reconsider my opinions when presented with new evidence.
 - * I recognize the value in opinions that are different from my own.
 - * I accept that my beliefs and attitudes may be wrong.
 - * In the face of conflicting evidence, I am open to changing my opinions.
 - * I like finding out new information that differs from what I already think is true.
- Zip code
 - Please enter the 5-digit ZIP code where you reside. *Please enter a whole number.* [Open, numeric response]

14.1.3 Post-treatment Survey Questions

- Emotions
 - Below, a number of words will appear that describe different feelings and emotions. To what extent did you feel this way during your time on our social media platform? [Very slightly or not at all/A little/Moderately/Quite a bit/Extremely]
 - * Active
 - * Nervous
 - * Inspired
 - * Ashamed
 - * Alert
 - * Determined
 - * Upset
 - * Afraid
 - * Attentive
 - * Hostile
- Platform Experience
 - How well, if at all, do the following words describe **the other users** on the platform? [Not at all/Very little/Some/Quite a bit/A great deal]
 - * Polite
 - * Aggressive
 - * Tolerant
 - How well, if at all, do the following words describe **the platform as a whole**? [Not at all/Very little/Some/Quite a bit/A great deal]
 - * Toxic
 - * Engaging
 - * Exhausting
 - If a social media platform like the one you just used were to become available to the public, how likely or unlikely would you be to use that social media platform in the future? [Very unlikely/Somewhat unlikely/Neither unlikely nor likely/Somewhat likely/Very likely]
 - How comfortable, if at all, did you feel sharing your views about political issues on this app? [Not at all comfortable/Slightly comfortable/Moderately comfortable/Very comfortable/Extremely comfortable]
 - All social media platforms are confronting issues around the presence of bots these days. Do you think you encountered any bots during your time on the platform? [No, I did not encounter any bots/Yes, I encountered one or two bots/Yes, I encountered several bots/Yes, I encountered many bots]
- Manipulation and attention checks

- During your time on our platform, you may have seen a post that displayed a “super secret password”. If you saw this post, what was the password? (If you didn’t see this post, continue to the next page.) [Open, short text response]
- On our platform, some users are awarded a user badge. On what basis were these badges awarded? [The degree to which users respectfully engage with and learn from differing opinions./The degree to which the user is “popular”, or the degree to which the are able to achieve “likes”./There were no user badges on the platform.]
- In your own words, briefly explain what behaviors or actions on the platform you think would earn the user badge. [Open, long text response]
- Post evaluations
 - Which of the following posts do you think would receive more likes on our platform?
 - Which of the following posts do you think would receive more likes on our platform?
 - Which of the following posts do you think would receive more likes on our platform?
 - Which of the following posts do you think would receive more likes on our platform?
- Policy attitudes
 - In general, do you agree or disagree with the following statements? [Strongly agree/Agree/Somewhat agree/Neither agree nor disagree/Somewhat disagree/Disagree]
 - * Recent immigration into the U.S. has done more good than harm.
 - * The laws controlling the sale and ownership of guns in the United States should be made stricter.
- Political Thermometer Ratings
 - On a scale from 0 to 100, how would you rate Republican voters? *Enter a whole number* [Open, numeric response]
 - On a scale from 0 to 100, how would you rate Democratic voters? *Enter a whole number* [Open, numeric response]
- Intellectual Humility
 - How well, if at all, do the following phrases apply to you? [Not at all well/Not too well/Somewhat well/Very well/Extremely well]
 - * I question my own opinions, positions, and viewpoints because they could be wrong.
 - * I reconsider my opinions when presented with new evidence.
 - * I recognize the value in opinions that are different from my own.
 - * I accept that my beliefs and attitudes may be wrong.
 - * In the face of conflicting evidence, I am open to changing my opinions.
 - * I like finding out new information that differs from what I already think is true.
- Is there anything else we should know about your experience on our platform? For instance, did you experience any harassment or technical glitches? [Open, long text response]

15 Data Availability, Code, and Pre-Registration

All anonymized data—excluding the full text of social media posts and comments created by respondents—as well as the code used to produce the analyzed described in the main text and this appendix will be available within this repository: https://github.com/kattasa/plab_status. All of our analyses were pre-registered on the Open Science Framework. Our pre-registration can be viewed using the following link <https://osf.io/pafyu>.

Despite our best efforts to create the experimental conditions described in our pre-registration statement as precisely as possible, we deviated from these plans as follows:

1) In our pre-registration statement, we wrote, “Our study’s control condition resembles extant platforms, amplifying content on the basis of its popularity and users on the basis of the size of their followings.” Though we manipulated the number of engagements (likes, dislikes, and comments) associated with the content of interest, we did not modify the size of their followings. The ability

to manipulate follower counts by condition was a feature that was planned for our platform but was not completed in time for this study. Additionally, due to a glitch in our platform, follower counts for accounts depended on previous participants' following behavior. Specifically, the follower count of a bot for each participant was the sum of (a) the number of pre-assigned follower counts for that account (which we set and did not vary by condition) and (b) the number of previous participants who followed that chatbot during their time interacting with the newsfeed. It is worth noting, however, that follower counts for an account was observable only if a user clicked on a profile.

2) In our description of our treatment and control conditions, we describe a highlighted posts receiving "significantly more likes and comments than the typical post." Due to a technical glitch, the like and dislike counts of posts that each participant saw for each post reflected not only the like/dislike counts that we assigned to that post at the outset of the study (which were manipulated in accordance with our pre-registration) but also the likes and dislikes that previous participants gave to it during their study sessions. This lead some posts to accrue many likes, for instance, and meant that highlighted posts were not necessarily the ones with the most likes for later participants.

3) Although we intended to include the time that participants spent dwelling on (or viewing) posts in our measures of engagement, technical errors prevented collection of accurate dwell time. Our engagement measure therefore does not include dwell time. Additionally, we do not use dwell time to examine the efficacy of our treatment.

4) In our pre-registration, we did not consider how or whether we would calculate affective polarization for those who identified as independents. In our analyses, we followed common practice in the affective polarization literature [4, 7, 3] and excluded those who identified as a "true independent."

References

- [1] R. L. Boyd, A. Ashokkumar, S. Seraj, and J. W. Pennebaker. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, pages 1–47, 2022.
- [2] A. Combs, G. Tierney, B. Guay, F. Merhout, C. A. Bail, D. S. Hillygus, and A. Volfovsky. Reducing political polarization in the united states with a mobile chat platform. *Nature human behaviour*, 7(9):1454–1461, 2023.
- [3] J. N. Druckman, S. Klar, Y. Krupnikov, M. Levendusky, and J. B. Ryan. Affective polarization, local contexts and public opinion in america. *Nature human behaviour*, 5(1):28–38, 2021.
- [4] J. N. Druckman and M. S. Levendusky. What do we measure when we measure affective polarization? *Public opinion quarterly*, 83(1):114–122, 2019.
- [5] S. Iyengar, Y. Lelkes, M. Levendusky, N. Malhotra, and S. J. Westwood. The origins and consequences of affective polarization in the united states. *Annual review of political science*, 22:129–146, 2019.
- [6] J. H. Kahn, R. M. Tobin, A. E. Massey, and J. A. Anderson. Measuring emotional expression with the linguistic inquiry and word count. *The American journal of psychology*, 120(2):263–286, 2007.
- [7] J. Kingzette, J. N. Druckman, S. Klar, Y. Krupnikov, M. Levendusky, and J. B. Ryan. How affective polarization undermines support for democratic norms. *Public Opinion Quarterly*, 85(2):663–677, 2021.
- [8] M. R. Leary, K. J. Diebels, E. K. Davisson, K. P. Jongman-Sereno, J. C. Isherwood, K. T. Raimi, S. A. Deffler, and R. H. Hoyle. Cognitive and interpersonal features of intellectual humility. *Personality and Social Psychology Bulletin*, 43(6):793–813, 2017. PMID: 28903672.
- [9] K. Munger, C. Ahn, S. Kim, and D. Dimmery. Fast or slow twitch. *OSF*, 2023.
- [10] E. Saltz, Z. Howard, and T. Acosta. Re-ranking news comments by constructiveness and curiosity significantly increases perceived respect, trustworthiness, and interest. *arXiv preprint arXiv:2404.05429*, 2024.
- [11] E. R. Thompson. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (panas). *Journal of cross-cultural psychology*, 38(2):227–242, 2007.