normsy
CIVIC HEALTH PROJECT

EXECUTIVE SUMMARY

Normsy.ai Impact Study

Civic Health Project · May 2025

Study Purpose

The Normsy.ai Impact Study tested whether prosocial, norm-engaging ("Normsy") comments reduce bystander support for toxic political posts on social media. The findings are intended to assess and guide interventions to improve online discourse.

Design

Participants: 1,021 politically active X users (535 Republicans, 486 Democrats).

Procedure: Each participant completed an online survey, in which they viewed four real examples of toxic political posts, each with several real comments.

Experimental conditions: Participants were randomly split into two groups.

- **Normsy group:** Viewed each post with one added Normsy comment.
- **Control group:** Viewed the same posts, but without any Normsy comments.

The posts and comments were otherwise identical across both groups.

Measures: We measured post engagement intentions (i.e., likes, follows, shares), agreement with toxic posts, perceived harmfulness of the toxic posts, and broader attitudes toward tolerance.

This off-platform design allowed us to isolate the causal impact of Normsy comments on bystander reactions to toxic political content.

Key Findings

- **Reduced Engagement:** Normsy comments decreased likes by 4%, decreased follows by 5%, and increased unfollows by 3%. No change in commenting, resharing, or time spent on thread.
- **Reduced Support:** Participants agreed less with toxic posts and rated them as more harmful when Normsy comments were present.
- **No Change in Broader Attitudes:** A one-time exposure did not shift general beliefs about tolerance or bridging divides.
- **Consistent Effects:** The intervention yielded similar impacts across parties and toxic content types. It yielded stronger effects when participants recalled seeing at least one Normsy comment and perceived the comment as human-written.

Limitations

This study tested the psychological impact of human-written, prosocial comments – not the current output of Normsy.ai's language model. The results also reflect the effect of these comments in an experimental setting, not real-world social media. Additional research would build confidence in Normsy's impact directly on social media platforms and across more content.

Take-aways and Recommendations

Normsy comments have the potential to meaningfully reduce bystander support for toxic political content. For greater impact, we recommend encouraging early, visible, and frequent posting of Normsy comments, especially on high-profile toxic posts. Moreover, authentic and community-driven participation may help sustain trust and long-term effectiveness.

More information: For details, please see our slide deck or full design document. All survey measures, data, and analysis code will be available at OSF/Github (soon). Questions about this research project can be directed to the author and scientific consultant, Catherine Garton (cgarton@stanford.edu).

BRIEF REPORT · 5 PAGES

Normsy.ai Impact Study

Civic Health Project · May 2025

Project Overview

Purpose: The Normsy.ai Impact Study was designed to evaluate whether prosocial, norm-engaging comments – referred to as "Normsy comments" – can reduce bystander support for toxic political content on social media. Rather than testing current language model (LLM) outputs or comparing different Normsy optimization strategies, this proof-of-concept study focused on the psychological impact of human-crafted comments that exemplify prosocial engagement. The goal was to inform future development and deployment of Normsy.ai tools and to support similar interventions that promote healthier online discourse.

Research question: Does the presence of Normsy comments on harmful social media posts "detoxify" bystander beliefs and behaviors, compared to exposure to toxic posts without such comments?

Outcome metrics: To assess the potential "detoxifying" effect of Normsy comments, the study measured three categories of outcomes:

- **Post Engagement:** Participants' intentions to like, follow/unfollow, reshare, comment, and spend time on the original toxic thread.
- **Post Support:** Agreement with the toxic post and perceptions of its helpfulness versus harmfulness.
- **Personal Beliefs:** Participants' perceptions of online tolerance norms and their personal attitudes toward tolerance.

Study Design

Overview: We conducted an online survey experiment from March 24 to April 1, 2025. Participants viewed images of four real social media posts, which were classified as toxic political content and accompanied by three to four comments, all presented in Qualtrics. After each post – and again at the end of the survey – we measured key

outcomes. This off-platform experimental design enabled us to isolate the causal effect of Normsy comments on participant responses. (See Figure 1 for a chronological overview of the study.)

Sample: We recruited politically active X users through the Bovitz-Forthright online panel, a participant recruitment service. To meet our eligibility criteria, each participant had to:

- Report checking X at least once per week,
- Have at least 25% of their social media feed comprised of political content, and
- Successfully pass an attention check to ensure data quality.

The final sample included 1,021 users (535 Republicans, 486 Democrats). Most reported checking X multiple times daily and an average of 62% political content in their feeds. Respondents also showed high levels of partisan animosity and sectarianism.

Experimental conditions: Participants were randomly assigned to one of two groups.

- **Normsy group:** Saw each post with one added Normsy comment.
- **Control group:** Saw the same posts, but without any Normsy comments.

Apart from the presence or absence of a Normsy comment, the survey experience was identical for both groups. Random assignment ensured the groups were comparable and minimized potential confounding factors.

Outcome metrics: After each post, participants completed assessments related to Post Engagement and Post Support. Post Engagement included binary ("yes" vs. "no") intentions to "like," follow, unfollow, reshare, and comment on the original toxic post, as well as their inclination to spend additional time on the thread. Post Support included their agreement with the toxic post and their perceptions of its helpfulness or harmfulness, on 1–7 Likert scales.

At the very end of the survey, participants responded via 1–100 scales to questions about:

- Their perceptions of online tolerance norms.
- Personal beliefs about tolerance and bridging political divides.

Complete survey measures are available on OSF/Github (soon).

Toxic political posts and comments — Posts: Participants were shown political posts from accounts that matched their own political affiliation, reflecting the typical partisan online experience. The research team curated eight toxic posts per political side, covering four key themes identified in prior research: out-party misperceptions, dehumanization, violence, and anti-institutional sentiment. Each participant viewed one post from each theme, with posts randomly assigned to ensure balance across the sample. We selected posts to be recent, broadly understandable (not requiring specialized knowledge), and "evergreen." When possible, posts referenced groups (e.g., "liberals," "MAGAs") rather than specific high-profile individuals. All usernames were blurred.

Comments: To simulate a realistic online environment, each post was displayed with three real, supportive comments from the original thread. In the Normsy condition, an additional researcher-generated Normsy comment – designed by CHP staff to model prosocial, ingroup norm engagement – was included. To minimize the influence of order effects, we randomized the position of the Normsy comment (either 1st or 3rd).

Ethics: All participants were fully informed about the study and provided their consent before taking part. They were notified that they would view real X content and could withdraw at any time. After completing the survey, participants were debriefed about the study's purpose and reassured that most toxic political content online is generated by a small, highly active minority. Participants were invited to leave feedback, and none reported negative experiences. Overall, the survey received high ratings, averaging 4.7 out of 5 stars on the Forthright platform.

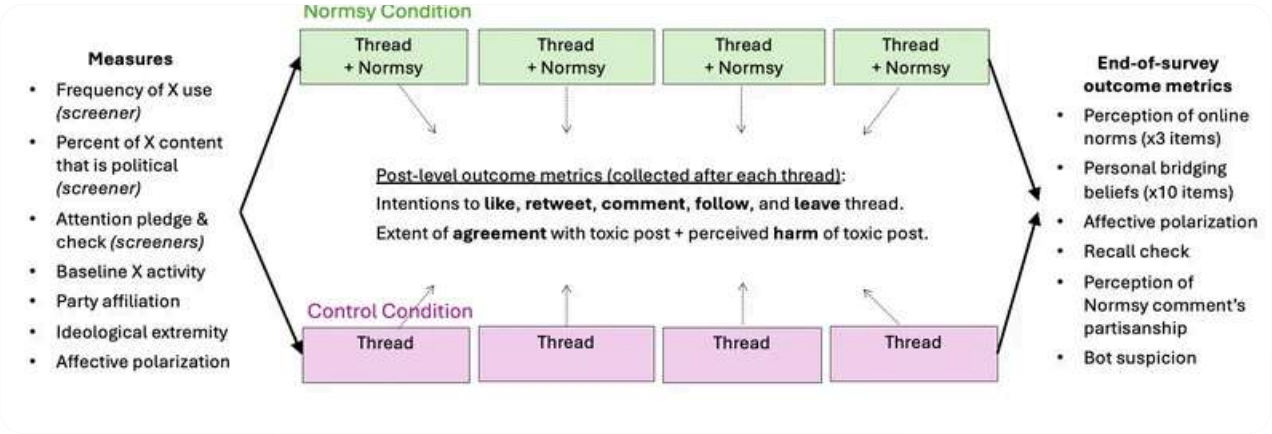


Figure 1. This diagram summarizes the structure of our study and the chronological flow of the participant experience.

Analyses

We conducted our analyses using R (version 4.2.2). Reproducible code is publicly available on OSF/Github (soon).

For post-level outcomes (such as Post Engagement and Post Support), we used generalized linear or logistic mixed-effects regression models to examine the effect of condition. These models included random intercepts to account for the fact that each participant rated multiple posts, and each post was rated by multiple participants. We also controlled for pre-treatment variables, including X usage, gender, party affiliation, ideological extremity, partisan animosity, and sectarianism. For outcomes measured only once (e.g., personal attitudes toward tolerance), we used simple linear regression with the same set of control variables. Additionally, we ran interaction models to explore whether the effect of Normsy comments varied by political party, whether participants noticed the Normsy comment, and whether they believed the comment was written by a human or a bot.

Key Results

Exposure to Normsy comments statistically significantly reduced Post Engagement and Post Support for toxic political content. We believe these effects could be meaningful at scale. Normsy comments did not shift broader attitudes about tolerance or online norms after this one-time exposure.

Post Engagement: Relative to the control condition, the presence of Normsy comments led to a 4% absolute reduction in "likes" for toxic posts (from 39% in the control group to 35% in the Normsy group; OR = .74, p = .03). Figure 2 shows the impact of Normsy comments on "liking" behavior, for both Democrat and Republican participants. Additionally, "follows" of the toxic poster declined by 5% (from 26% to 21%; OR = .64, p = .01), and "unfollows" increased by 3% (from 12% to 15%; OR = 1.5, p = .02). The Normsy condition did not change the rate of commenting, resharing, or time spent on the thread.

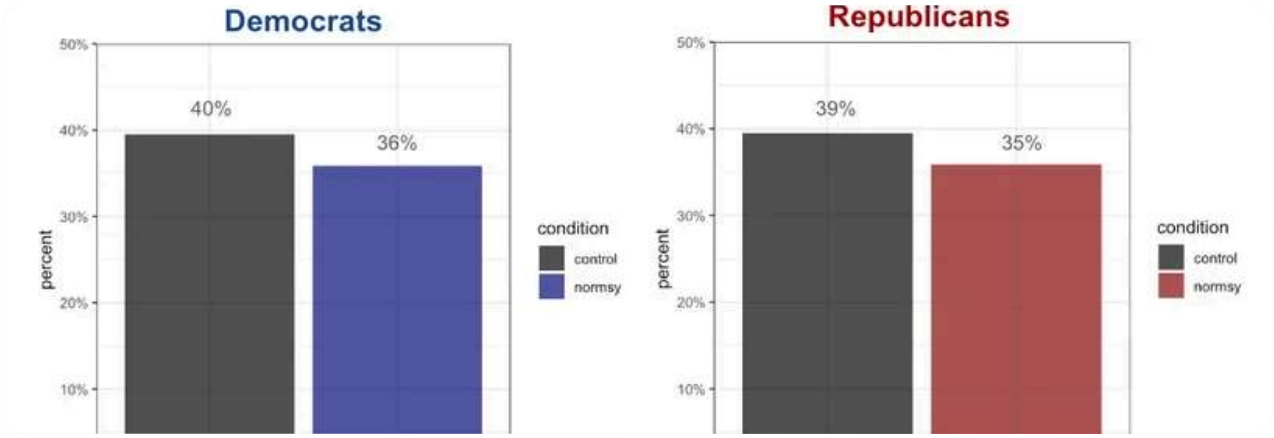


Figure 2. The percentage of toxic posts that were "Liked" by participants, depending on exposure to a Normsy comment.

Post Support: Participants agreed less with toxic posts (M_{diff} = -.21, Cohen's d = .10, p < .001) and rated them as more harmful (M_{diff} = .20, Cohen's d = .14, p = .004) when Normsy comments were present.

Personal Beliefs: A single exposure to Normsy comments did not significantly change participants' broader beliefs about respect, tolerance, or bridging political divides.

For whom did the intervention work best? Consistent results.

Please note: the following analyses are exploratory. Although informative, they should be interpreted with caution, because the study does not provide controlled comparisons between these subgroups.

- **Political party:** The reduction in "likes," follows, unfollows, and post agreement were similar for both Democrats and Republicans (interaction terms were not significant, all ps > .23). For perceptions of harm, the intervention yielded stronger

results among Democrats (Democrat M_{diff} = .32, Republican M_{diff} = .11, p = .04). Figures of all results split by party are in the slide deck.

- **Comment recall:** The intervention had a stronger impact for participants who explicitly remembered seeing a Normsy comment (which included 58% of those in the Normsy group).
- **Human vs. bot perception:** The effect was also stronger when participants believed the comment was written by a human (84% perceived it as human-posted).
- **Toxic content types:** The positive effect of Normsy comments was consistent across all types of toxic content tested, including posts about out-party misperceptions, dehumanization, violence, and anti-institutional sentiment.
- **Other factors:** The intervention worked similarly regardless of participants' level of polarization, baseline social media activity, or whether they thought the commenter was from their own party or the opposing party.

Limitations

There are some limitations to the generalizability of these findings. The study's off-platform design allows for strong experimental control, but at the expense of ecological validity, meaning results may not fully reflect real-world social media behavior. Additionally, the study included only 16 real X posts, so it remains unknown how Normsy comments would perform across a broader set of posts. Lastly, due to survey constraints, the images used in this experiment involved a higher ratio of Normsy to toxic comments (1:3) than is typical online, unless prosocial commenting is significantly scaled up.

Importantly, the Normsy comments tested in this study were human-written, not generated by AI. Therefore, while the findings highlight the psychological benefits of prosocial, norm-engaging comments, they do not directly validate the current output of Normsy.ai. However, this approach likely sets a conservative baseline, as future AI-generated comments could be better optimized for impact.

Take-Aways & Recommendations

We conclude that Normsy comments can meaningfully reduce bystander post engagement and support for toxic political content, in controlled settings. Moreover, Normsy visibility was a key factor determining the effectiveness of the intervention. That said, a single exposure did not shift broader attitudes, indicating that repeated and prominent interventions may be necessary for lasting change.

To maximize impact:

- **Increase visibility:** Post Normsy comments early, encourage upvotes/likes, and use multiple commenters per thread.
- **Target strategically:** Focus efforts on a few high-profile toxic posts each day rather than spreading interventions too thin.
- **Foster authenticity:** Promote organic, community-driven participation to maintain trust and long-term effectiveness.

These strategies may amplify the positive effects of Normsy comments and support healthier online discourse.

For more information or to discuss collaboration, please contact:

Kristin Hansen, Executive Director of Civic Health Project
Catherine Garton, Scientific Research Consultant