Impacts of Normsy.ai Interventions: Co-attitudinal and Counter-attitudinal content

Prepared by Kylan Rutherford

Civic Health Project

June 2026

Normsy

Executive Summary

Our prior lab study tested the impact of Normsy interventions on reducing the level of agreement that a sample of social media users express with a sample of toxic social media content, which was matched to each user's ideological lean. Our second study expands on this design by also exposing participants to counter-attitudinal posts. We analyze the same central question: can a single well-crafted reply comment reduce viewers' support for toxic social media content? The convergent answer is *yes, under the right conditions.*

The headline finding: Normsy interventions reliably **reduce copartisan agreement with toxic content**. When viewers encounter a toxic post that aligns with their own political identity, a Normsy intervention measurably reduces their agreement. For Democratic viewers of left-leaning toxic content, the effect reaches conventional significance ($b = -0.22$, $p = .050$). When Republican co-attitudinal data from both studies are pooled, the effect is also significant ($b = -0.135$, $p = .037$).

A second, underappreciated finding: Normsy interventions also **moderate outpartisan rejection**. Viewers who already oppose a toxic post's political lean — and who would be most likely to respond with oppositional toxicity — become more measured when an intervention is present. This effect is statistically significant for Democrats ($b = +0.203$, $p = .018$). The effect is in the same direction and magnitude for Republicans, but is not statistically significant (largely due to a smaller sample size). This convergence at both ends of the agreement scale is a signature of reduced affective polarization, not merely content moderation.

We also compared AI-drafted and human-written interventions. **AI-drafted interventions outperform human-written ones** on reducing agreement with toxic posts ($b = -0.13$, $p = .038$). However, human-written interventions have a complementary advantage on emotional tone, reducing negative affect more than AI. The optimal model — on which Normsy now operates — is **AI-drafted, human-reviewed**: combining AI's precision with human judgment and accountability.

The Challenge: Partisan Amplification and the Limits of Platform Moderation

The reach of political content on social media is amplified by polarization. Partisan networks elevate content they agree with through positive engagement (likes, shares, follows), boosting content from their own side. But they also elevate content they find disagreeable through negative engagement (angry comments and hate-sharing). The result is a social media ecosystem that surfaces the content one side loves and the other hates, rather than the best of what users have to offer. Toxicity, clickbait, misinformation, and attacks on democratic norms are not what most users — or social media platforms — want (Milli et al., 2025), but it is what rises to the top of feeds (Germano et al., 2026).

Platform-level moderation addresses this only partially and inconsistently. Content removal is reactive, labels often go unnoticed, and algorithmic suppression operates in the background without producing the kind of friction that might change how a viewer processes what they see (Pennycook et al., 2021). None of these tools put a voice *in the thread* alongside the toxic post.

Counter-speech — placing a direct, substantive reply alongside harmful content — has been shown to improve a variety of outcomes on social media. It can shift descriptive norms ("others do not endorse this"), provide factual reframing, and introduce cognitive friction that promotes more deliberative processing (Cialdini, 2003; Kozyreva et al., 2022). Empirically, the evidence base for counter-speech has been growing: field studies have shown reductions in racist Twitter behavior following reply interventions (Munger, 2017), and computational work has documented norm-shifting effects from well-designed replies in online communities (Garland et al., 2022).

The open questions have been: *Does it work at the post level, in a controlled environment? Does AI authorship help or hurt? And does it work differently for partisan viewers who already agree versus already disagree with the claims embedded in toxic content?* Normsy has now run two independent laboratory studies to answer these questions.

The Normsy Approach

Normsy enters the social media conversation by identifying toxic content[1] and responding to it using intervention strategies specifically designed to bridge partisan divides and reduce toxicity. Normsy equips *real social media users* with tools to intervene and improve discussions. These resources include scalable toxicity identification, AI-generated context summaries, recommendations for how to reply, and impact monitoring. The comments generated by Normsy are not platform labels or system flags, but substantive replies drafted by an AI tool grounded in proven depolarization strategies.

Every Normsy deployment includes a trained human reviewer before the comment is posted — a feature, not a constraint. AI drafts are reviewed, edited where needed, and approved by a human operator. The human reviewer ensures that deployed content is contextually appropriate, factually accurate, and consistent with community norms. It also provides a clear chain of accountability that fully automated systems cannot offer. This model is compatible with the oversight expectations of platforms and regulators alike, and it is the standard to which we hold all deployments.

Normsy's approach is also platform-agnostic. While Normsy has thus far focused on X (Twitter), our interventions require only that a comment can be placed in a reply thread. Any platform with that capability is a potential deployment surface. We are preparing to scale Normsy for additional platforms, such as TikTok, Twitch, and Reddit.

Prior Validation

Prior to this study, an earlier lab study was designed and fielded by the Civic Health Project research team in early 2025, in partnership with researchers at Stanford. One of the key findings from that study is that Normsy interventions significantly reduced agreement with the post and increased perceived harmfulness of the original post across the full sample. When subset to parties (Republican and Democrats), the effects are statistically significant for Democrats, but not for Republicans. Importantly, all participants in this study were shown **co-attitudinal** content, meaning that all of the original posts aligned with their expressed partisan identity.

A more extensive summary of the design and results from the first study can be found online at https://normsy.ai/labs.

[1]By *toxic content*, we mean social media posts that dehumanize individuals or groups, normalize anti-democratic conduct, or weaponize partisan outrage in ways that degrade the quality of public discourse. Critically, toxicity in this sense is not the exclusive province of any political orientation: posts can be toxic regardless of the direction from which they lean.

Lab Experiment: Co-attitudinal and Counter-attitudinal Effects

This lab experiment was designed and fielded by the Civic Health Project in collaboration with researchers at NYU's Center for Social Media, AI, and Politics, with data collected in April 2026.

Design. A sample of $N = 1{,}007$ active social media users each viewed seven simulated social media posts, one at a time. For each post, three binary treatments were independently randomized: (1) whether a Normsy intervention comment was present in the thread, (2) whether the intervention was AI-drafted or human-written, and (3) whether an "Automated Account" disclosure label was shown. This within-person design yielded 7,049 post-level observations.

Posts. The post bank comprised 20 posts spanning diverse toxic content types: dehumanizing rhetoric, conspiracy theories, partisan outrage, and immigration-related hostility. Posts varied in partisan lean (left, right, or ambiguous), and post-level outcomes were analyzed with post fixed effects to control for baseline differences in content.

Outcomes. Agreement with the post (1–7 scale, lower is better), entertainingness, negative and positive affect, and perceived informativeness were measured post-by-post. Perceived acceptability of toxic online behavior (a four-item norm index) was measured at survey completion.

Sample composition. The analytic sample was majority-Democratic (50.2%), with Republicans comprising 32.4% and true Independents 17.4% (using a folded three-category party measure where leaners are assigned to their leaned party).

Key Finding 1: Reducing Copartisan Support for Toxicity

The strongest case for Normsy interventions is when the person most likely to endorse a toxic post — someone whose political identity aligns with the post's framing — encounters an intervention and becomes less agreeable. These are co-attitudinal exposures: a Republican viewer seeing a right-leaning toxic post, or a Democrat seeing a left-leaning one. They may also be some of the most consequential exposures, because co-attitudinal viewers are the ones most likely to amplify and endorse the content.

Democratic viewers. Study 1 established this effect using a between-subjects design in which all exposures were co-attitudinal: Democratic viewers of left-leaning toxic posts showed a significant reduction in agreement ($b = -0.30$, $p = .003$, $d = 0.15$, $N = 476$). Study 2 confirmed this in a within-person design (Democrats: $b = -0.215$, $SE = 0.109$, $p = .050$,

$N = 506$; pooled across both studies: $b = -0.163$, $SE = 0.085$, $p = .055$).

Republican viewers. Study 1 found a consistent but individually non-significant effect for Republicans ($b = -0.12$, $SE = 0.09$, $p = .21$), and Study 2 replicated the same directional pattern ($b = -0.114$, $SE = 0.113$, $p = .31$). When the Republican co-attitudinal data from both studies are combined using inverse-variance weighting, the pooled estimate reaches significance ($b = -0.135$, $SE = 0.065$, $p = .037$; Figure 1). The directional consistency across two independent studies with distinct designs strengthens confidence in this finding.

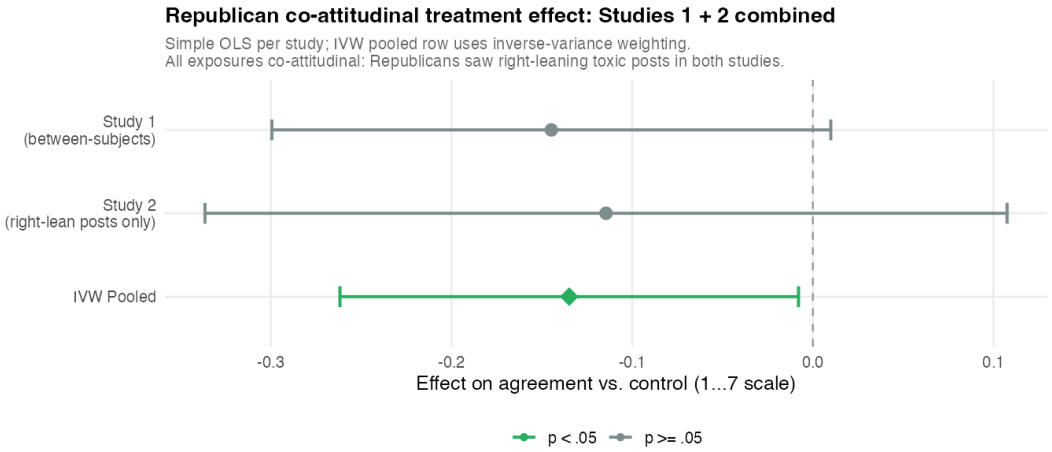


Figure 1: Republican co-attitudinal treatment effect: individual-study estimates and inverse-variance weighted (IVW) pooled estimate. All exposures co-attitudinal (Republicans saw right-leaning toxic posts in both studies). Individual estimates use simple OLS; pooled estimate uses inverse-variance weighting. Green = $p < .05$; grey = $p \geq .05$.

Key Finding 2: Moderating Outpartisan Rejection

The second finding is less intuitive but equally important for discourse health: interventions also change how *outpartisans* respond to toxic content. Viewers who oppose a post's political lean already disagree with it, but the degree of disagreement matters. Outright, maximum-intensity rejection ("strongly disagree") is a signature of affective polarization: a reflexive partisan response that bypasses the content itself.

In Study 2, counter-attitudinal viewers exposed to an intervention showed higher agreement scores than counter-attitudinal viewers in the control condition (pooled: $b = +0.203$, $SE = 0.086$, $p = .018$). This is a movement *toward the middle* of the scale, from extreme rejection to moderate disagreement or neutrality, not a movement toward endorsement. The outpartisan viewer still disagrees with the post; they simply disagree less reflexively and intensely.

Figure 2 shows both co- and counter-attitudinal patterns side by side.

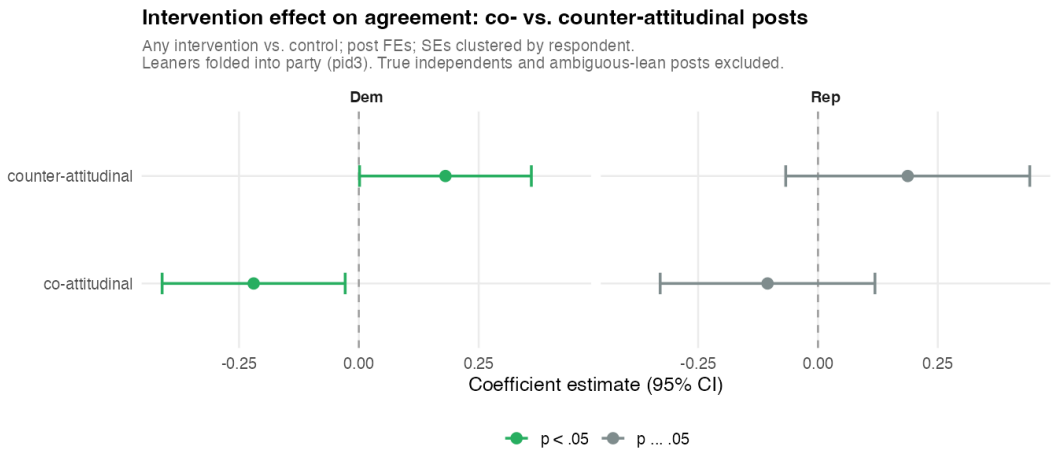


Figure 2: Intervention effect on agreement by attitudinal direction (Study 2). Co-attitudinal: the post's partisan lean matches the viewer's party. Counter-attitudinal: the post's lean opposes the viewer's party. Leaners folded into party; true Independents and ambiguous-lean posts excluded. Post fixed effects; standard errors clustered by respondent. Green points indicate $p < .05$.

Among *co-attitudinal* viewers (left panel), the intervention effect is concentrated at the top of the scale: AI interventions reduce the share of "Strongly agree" (7) responses from roughly 21% in control to 15%, with corresponding mass redistributing toward the middle. Among *counter-attitudinal* viewers (right panel), the pattern is reversed. Posts without intervention comments produce the highest rate of "Strongly disagree" (1) responses (˜30%). Human interventions reduce this to ˜25%, with the displaced mass accumulating at "Neither" (4), which rises from 27% to 33%. The movement is not from disagreement to agreement (most viewers still opposes the post) but from maximum-intensity rejection to a more measured neutral-to-negative response.

The de-polarization interpretation. The results we observe for reduced copartisan support and moderated outpartisan rejection together suggest a convergence effect. Both the partisan amplifiers and the partisan rejectors become more measured when a Normsy comment is in the thread. This is consistent with the theoretical mechanism of cognitive friction: the presence of a substantive reply prompts more deliberative processing before forming a judgment (Kozyreva et al., 2022), pulling responses away from the emotional extremes in both directions.

Key Finding 3: AI-Drafted Interventions and the Human-in-the-Loop Advantage

Within the intervention arm of Study 2, AI-authored comments reduced agreement significantly more than human-authored ones ($b = -0.130$, $SE = 0.063$, $p = .038$; $N = 1{,}007$). This held across the post bank. We are cautious about overstating these results given the limited number of posts analyzed, and have future research planned to confirm.

Human-written interventions produced a greater reduction in negative affect compared to AI interventions ($b = -0.029$, $SE = 0.014$, $p = .042$). Viewers who saw a human-written comment reported lower rates of negative emotions relative to those who saw an AI-generated comment. The warmth and interpersonal attunement of human phrasing appears to dampen the emotional valence of the exchange, even when it does not move agreement as far.

These findings point directly to the Normsy operating model. AI drafting captures the precision and scalability advantages that drive the agreement-reduction finding. Human review captures the contextual judgment, warmth, and accountability that AI alone cannot guarantee. The human-in-the-loop is not a compromise, but a crucial feature of quality deployments.

Conclusion

Two independent lab studies have now validated the core Normsy finding: a single, well-crafted Normsy comment in a toxic thread reduces copartisan amplification of that content and moderates outpartisan rejection, pulling discourse toward the middle from both ends. When Republican co-attitudinal data from both studies are combined, the intervention effect is significant for Republican viewers as well, confirming that the finding generalizes across the partisan spectrum.

AI-drafted interventions outperform human-written ones on agreement reduction. Human review ensures that deployed content is accurate, contextually appropriate, and accountable. The combination is not a compromise between scale and quality, but the design that produces both.

Normsy is not a silver bullet. It does not work on all post types or situations, and it does not replace platform moderation. But we know that it *can* change how people who encounter toxic posts think and feel about what they just read. At scale, that is a meaningful contribution to discourse health.

References

Robert B Cialdini. Crafting normative messages to protect the environment. *Current Directions in Psychological Science*, 12(4):105–109, 2003.

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. Impact and dynamics of hate and counter speech online. *EPJ Data Science*, 11(1):1–24, 2022.

Fabrizio Germano, Vicenç Gómez, and Francesco Sobbrio. Ranking for engagement: How social media algorithms fuel misinformation and polarization. *Journal of Public Economics*, 255:105589, 2026.

Anastasia Kozyreva, Philipp Lorenz-Spreen, Stefan M Herzog, Ullrich K H Ecker, Stephan Lewandowsky, and Ralph Hertwig. Toolbox of interventions against online misinformation and manipulation. *Nature Human Behaviour*, 6(12):1168–1180, 2022.

Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D Dragan. Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS Nexus*, 4(3):pgaf062, 2025.

Kevin Munger. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39(3):629–649, 2017.

Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855):590–595, 2021.