Social Media Toxicity and Mental Health*

George Beknazar-Yuzbashev[†] **Francesco Capozza**[‡]
Kylan Rutherford[§] **Mateusz Stalinski**[¶]

July 12, 2026

The most recent version is available here.

Abstract

We study whether content-level moderation that reduces exposure to social media toxicity, holding platform access constant, affects mental health. We deploy, for randomly assigned users, a cross-platform browser extension that hides posts and comments exceeding a prespecified toxicity threshold, and follow $N = 664$ desktop users over a six-week intervention across two waves – one during the 2024 U.S. election campaign and one outside of it. Contrary to basic intuition, filtering toxicity worsens mental health: treatment raises a standardized index of GAD-7 and PHQ-8 symptoms by 0.15σ, and the share of respondents above a moderate-symptom screening cutoff rises by about 12 pp. A QALY-based calibration implies a welfare cost of about $158 per treated participant. Heightened loneliness and a diminished relative moral self-view emerge as potential channels. We conclude that curbing toxic content can impose unintended psychological costs, even where it yields other social benefits.

JEL Classification: C93, D83, I10, I31, L82, L86, Z13

Keywords: mental health, toxic content, hate speech, moderation, social media, field experiment

*We would like to thank Vladimir Avetian, Tim Besley, Sonia Bhalotra, Luca Braghieri, Jon de Quidt, Ruben Enikolopov, Rafael Jiménez-Durán, John List, Luis Menendez, Elliot Motte, Andrew Oswald, Maria Petrova, Matthew Ridley, Chris Roth, and numerous seminar participants for helpful comments and suggestions. We thank Ceren Budak, Julia Kamin, and Jonathan Stray, with whom we built the shared experimental infrastructure. We thank Kevin Chen, Esha Parekh, Cody Prosperini, and Jessie Yu for their great research assistance. The field experiment was pre-registered in the AEA Registry under the number AEARCTR-0013987. The field trial was funded by Civic Health Project and Jigsaw LLC. We obtained the ethical approval to conduct this research from the Morningside IRB at Columbia University (protocol number IRB-AAAU8117). Moreover, the project is logged as externally approved by the Humanities and Social Sciences Research Ethics Committee at the University of Warwick under the reference number HSSREC 107/23-24.

[†]University of Chicago, The Normal Lab beknazar@uchicago.edu.
[‡]UB, IEB, and CESifo, francesco.capozza@ub.edu.
[§]New York University, kylan.rutherford@columbia.edu.
[¶]University of Warwick and CAGE, mateusz.stalinski@warwick.ac.uk.

1 Introduction

Social media platforms mediate a large share of information consumption and social interaction, and a growing body of causal evidence shows that they affect user well-being and mental health (Allcott et al., 2020; Braghieri et al., 2022; Allcott et al., 2022; Bursztyn et al., 2025; Allcott et al., 2025).[1] This work has focused predominantly on *access* – whether people use these platforms and how much. A distinct margin, the content users are exposed to conditional on access, has received less attention, despite offering a more targeted lever: reducing harmful exposure while preserving the benefits of platform access.

Much of the intuition behind content regulation reflects a view of social media as a form of media: toxicity, a commonly moderated type of harmful content, affects users through what it communicates, so filtering it should improve mental health.[2] Moreover, a feed scrubbed of toxic content can leave users with a more benign picture of the world – perceiving discrimination and hostility as less pervasive – and such perceptions are linked to better mental health. But the platforms are also social networks: they shape the social environments in which users participate, not merely the information they receive (Enikolopov et al., 2024). Toxicity is often woven into valued social participation. Even hostile exchanges signal that others are present and engaged – a sense of social presence whose removal could heighten loneliness – and others' bad behavior enables comparisons against which users view themselves favorably.[3] These forces pull in opposite directions, leaving the net effect on mental health unclear ex ante.

We study this trade-off in a six-week field experiment on Facebook, X, and Reddit. A browser extension filtered toxic posts and comments for randomly assigned participants while preserving their access to the platforms. We find that filtering toxicity *worsens* mental health, increasing symptoms of both anxiety and depression. The mechanism evidence points to the social function of the platforms: filtering increases loneliness, and the adverse effects

[1] Studying mental health is critical given that mental health disorders are the leading cause of disability worldwide (Friedrich, 2017; OECD, 2021), affecting roughly one in two individuals over the life course (OECD, 2019). Moreover, poor mental health and psychological distress hinder human capital accumulation (Currie and Stabile, 2006, 2009; Eisenberg et al., 2009), contribute to the misallocation of talent (de Quidt and Haushofer, 2016), and are associated with penalties in the labor market (Biasi et al., 2021).

[2] Toxicity is defined as "a rude, disrespectful, or unreasonable comment that is somewhat likely to make you leave a discussion or give up on sharing your perspective". This definition applies to classifiers such as Perspective API and Unitary's Detoxify, deployed in both industry moderation systems and academic research (e.g., Kim et al., 2021; Jiménez Durán et al., 2025; Beknazar-Yuzbashev et al., 2025; Motte, 2026).

[3] Loneliness is itself a well-established determinant of mental health: it prospectively predicts the onset and persistence of depression and anxiety (Jin et al., 2026), and recent population evidence from the *All of Us* program shows that it statistically mediates the link running from anxiety and depressive symptoms to suicidal ideation (Musacchio Schafer et al., 2026). Accordingly, loneliness has moved onto the public-health agenda as a first-order policy concern in its own right (Goldman et al., 2026).

are larger on platforms and in feeds where content is more strongly embedded in social ties and politically relevant communities.

We recruit desktop social media users via targeted ads posted on major social media platforms and through an online panel (CloudResearch) and ask them to install our extension. Participants are randomly assigned to (i) treatment conditions with toxicity hiding based on a Perspective toxicity score threshold of 0.33 or (ii) control conditions that do not hide any content (pure control) or hide a matched random fraction of content to equalize hiding rates (active control).[4] The Perspective classifier maps text to a continuous toxicity score – the predicted fraction of human annotators who would label a snippet as toxic – and is built on human-labeled training data widely used in academic and platform settings (Wulczyn et al., 2017; Muralikumar et al., 2023). We field two waves: one during the 2024 U.S. election campaign and one outside the campaign period. The study runs for six weeks until a post-treatment survey and continues for an additional six weeks to a follow-up survey; mental health outcomes (PHQ-8, GAD-7) are measured at baseline, post-treatment, and follow-up.

Using a difference-in-differences design with respondent and cohort-time fixed effects, we compare treated and control participants relative to baseline. Our main finding is that reducing exposure to toxic content *worsens* mental health. The standardized mental-health-symptoms index rises by 0.15σ (s.e. $= 0.05$, $p = 0.0030$). The anxiety and depression subindices increase by 0.13σ and 0.16σ respectively. On the raw clinical scales, GAD-7 rises by 0.7 points on the 21-point scale and the PHQ-8 rescaled to the full 27-point scale rises by 1.0 points. Expressed as an elasticity, a 1% reduction in shown toxic share is associated with approximately a 0.13% increase in GAD-7 symptoms and a 0.15% increase in PHQ-8 symptoms, evaluated at the control-group means. The deterioration is not confined to mean scores: the share of respondents crossing the threshold of "moderate" symptoms on any of the scales rises by 12.1 percentage points ($p = 0.0006$). Item-level decompositions show broad worsening across GAD-7 items and more concentrated increases on PHQ-8 items for sleep disturbance, depressed mood, and negative self-evaluation.

We translate the symptom movement into a welfare metric using a conservative, QALY-based calibration: it prices only net movement into worse PHQ-8 and GAD-7 severity bins and values the implied quality-of-life loss at respondents' own reported income, yielding a welfare cost of approximately $158 per treated participant over the treatment window.

To benchmark magnitudes, we compare our effect against two RCTs using the same or closely related clinical scales, both documenting mental-health improvements that our

[4]Participants in the treatment conditions were randomly assigned to one of two variants: one in which toxicity hiding was solely based on whether the Perspective score exceeded 0.33 and one that additionally screened content when the Alienation score exceeded 0.7.

deterioration mirrors. Angelucci et al. (2026) document a 0.29σ improvement in a composite PHQ-8/GAD-7 index at six months from an AI-powered cognitive behavioral therapy app among distressed women. Shreekumar and Vautrey (2024) find that randomized access to the Headspace mindfulness app reduces a composite anxiety-depression index by 0.46σ after four weeks. Our 0.15σ adverse effect runs in the opposite direction across both: about 52% of the Angelucci et al. benefit despite a treatment window more than four times shorter, and 33% of the Headspace improvement.[5]

The mechanism evidence points to social connection as the primary channel through which toxicity filtering worsens mental health. Toxic content on social media is not merely noxious media content: it is often embedded in social relationships, shared attention, and politically salient exchanges. Filtering it can therefore remove not only hostility but also information about what proximate or identity-relevant others are saying, fighting over, and treating as important. Among our mechanism outcomes, the strongest effect is on loneliness: treatment raises its index by 0.34σ ($p = 0.003$). A second channel is a deterioration in relative moral self-view: treatment changes the positive-self-view measure by -0.19σ ($p = 0.037$), consistent with the loss of a downward-comparison reference point supplied by others' toxic behavior, while the absolute moral superiority index does not move. Other channels, such as more benign social beliefs and compensatory behavioral adaptation, find no support.

The heterogeneity evidence sharpens this social-connection interpretation along two dimensions. The first is horizontal ties in the feed: controlling for platform use shares, the adverse effect on mental health is larger by 0.11σ for each one-standard-deviation increase in the share of feed impressions from peers ($p = 0.013$) – as opposed to broadcasts from big creators and entities. A toxic post from a socially proximate account carries different information than one from an impersonal source: it says something about what people in the user's social environment are discussing and how. Platform heterogeneity points in the same direction: the adverse ITT is 0.17σ for Facebook-heavy users and 0.21σ for Twitter-heavy users – platforms organized around interpersonal exchange – against only 0.01σ for Reddit-heavy users, where consumption is organized around topical communities rather than personal social connections.

The second dimension is political attachment. The ITT is 0.19σ for Democrats and 0.16σ for Republicans, against only 0.03σ for independents. Among pre-treatment political

[5]Two further RCTs offer useful benchmarks. Bhat et al. (2022), following two behavioral-activation psychotherapy trials in India four to five years on, find pooled depression reductions of 0.15σ (rising to 0.23σ for the more effective Healthy Activity Program, which cut moderate-depression incidence by 13 pp); our deterioration is a near-exact mirror image, with our 12.1 pp rise in threshold-crossing almost perfectly reflecting their 13 pp reduction. Francois et al. (2025), in the largest psychedelics RCT to date, find that a one-time ayahuasca treatment raises happiness and lowers distress by roughly 0.4σ at six months among 429 first-time users; our effect is roughly 38% of that benefit, again reversed.

and hostility measures, only the moral-disengagement index (MDI) – capturing a morally charged relation to the out-party, viewing it as threatening, evil, or lacking humanity – generates a robust interaction: a one-standard-deviation increase in MDI raises the ITT by 0.12σ ($p = 0.006$), while dehumanization, political polarization, violence support, and spiral-of-silence measures are all near zero. The relevant margin is therefore not generic partisan intensity but this moralized out-party hostility. Moreover, content evidence shows that most toxic political posts with a discernible affiliation removed by the extension are *party-congruent* with the user's own affiliation, suggesting that what is lost is not only hostile out-party speech but content connected to the user's own political side. Entered jointly, the peer-share and MDI interactions remain separately positive with similar magnitudes, indicating that horizontal social proximity and political self-relevance are separable dimensions of the same underlying mechanism: toxicity filtering is most costly when it hides content connected to the user's relevant social environment.

It is important to interpret our content-level results in light of the literature on social media access and mental health. Allcott et al. (2020) find that deactivating Facebook for four weeks before the 2018 US midterms improved subjective well-being by 0.09σ, showing harmful effects of platform access. Relatedly, Allcott et al. (2025) find that deactivating Facebook before the 2020 US election improved an index of happiness, depression, and anxiety by roughly 0.06σ. Braghieri et al. (2022) report that Facebook's rollout across US college campuses (2004–2006) worsened mental health, measured by NCHA instruments, by 0.085σ. Our sizeable 0.15σ effect of toxicity hiding on mental health runs in the opposite direction, highlighting the importance of understanding access effects as *net* effects of various channels. For example, a strong impact of unfavorable social comparisons, a mechanism proposed by Braghieri et al. (2022), could render moderate negative access effects in the presence of countervailing effects such as those we document. We acknowledge that social media platforms have evolved considerably over the last decade and habitual users may differ from early adopters, limiting comparability of effect magnitudes. That said, context-dependence appears limited: our two waves, one during the 2024 U.S. election campaign and one outside it, yield similar effects, suggesting the estimate is unlikely to be an artifact of an election period.

To strengthen the interpretation of our main results, we conduct a battery of robustness checks. The mental-health-symptoms coefficient is stable across time horizon (adding the twelve-week follow-up), weighting (IPW for survey non-response), compliance window (restricting to respondents active for at least one or three weeks), recruitment channel (CloudResearch versus social-media ads), recruitment waves (during an election campaign or outside of it), questionnaire order, and classifier (toxicity-only versus alienation-augmented

arm). The four-arm design rules out generic feed disruption as an explanation: the random-hiding arm, which removes a matched fraction of non-toxic content, produces a null mental-health effect, while the toxicity-filtering ITT remains positive when random hiding serves as the comparison group, with a coefficient of 0.20σ (p=0.002). The adverse ITT persists dynamically: estimated with separate post-treatment and follow-up coefficients, the symptom index remains elevated at the twelve-week follow-up, inconsistent with a purely transitory adjustment response. Treated respondents were not more likely to report changes in their social media experience than those in the control group. This reduces the possibility of experimenter demand effects, as users did not understand the nature of the treatment. Finally, we find no differential attrition by treatment status – treatment assignment does not predict post-treatment survey completion or continued extension activity across any specification – and no shift toward mobile platform use that would attenuate the first stage.

Our findings do not imply that content moderation is futile or harmful. The harms of toxic content are real and well-established: hate speech online has been linked to a global increase in violence toward minorities (Council on Foreign Relations, 2019; Müller and Schwarz, 2021), and there is strong evidence that reducing the supply of hateful content can curb offline hate crimes (Jiménez Durán et al., 2025). Rather, our results call for a more granular, user-centered approach to how moderation is implemented. One promising alternative to outright hiding is friction-based flagging: platforms could label posts exceeding a toxicity threshold and leave the decision to view them to the user. Another follows from our mechanism evidence: because the cost concentrates where toxic content arrives through peer-level ties, platforms could moderate content from peers more lightly than content from influencers and institutions.

We contribute to the literature along four dimensions. First, we isolate a distinct content-level channel – exposure to toxic material – within the broader mental-health effects of social media. Unlike designs that jointly alter time online, content exposure, and social network exposure (Allcott et al., 2020; Braghieri et al., 2022; Allcott et al., 2025) or study the effects of algorithm personalization (Levy, 2021; Guess et al., 2023a; Mandile, 2025; Gauthier et al., 2026), our intervention directly manipulates toxicity exposure while holding platform access and other features of the recommendation algorithm constant. Our finding that filtering toxicity worsens rather than improves mental health shows that, for the viewing user, exposure to toxic content is not a net driver of the mental-health costs of social media: its removal does not alleviate them, and can itself impose a cost. This suggests that care should be taken in designing content moderation policies, as seemingly beneficial interventions may carry unintended psychological costs.

Second, we contribute to two related literatures: the effects of harmful content and its

moderation, and the effects of recommendation algorithms. We extend research on toxic content – which affects user engagement (Beknazar-Yuzbashev et al., 2025, 2024) and, when weaponized by politicians, voters' intentions (Motte, 2026) – and on the content moderation policies designed to curb it (Andres and Slivko, 2021; Jiménez Durán et al., 2025; Jiménez-Durán, 2023; Ahmad et al., 2024; Kominers and Shapiro, 2024; Madio and Quinn, 2024), by documenting that such content, and its removal, affect clinically validated depression and anxiety outcomes. More broadly, we add to a growing body of work on recommendation algorithms, which has studied the effects of overall personalization (e.g., Gauthier et al., 2026) and of specific alternative algorithms (Nyhan et al., 2023; Guess et al., 2023b; Piccardi et al., 2025; Milli et al., 2025; Allcott et al., 2026; Stray et al., 2026).

Third, we provide new evidence on the mechanisms linking content environments to mental health, identifying social connection – through loneliness (Enikolopov et al., 2024; Musacchio Schafer et al., 2026; Jin et al., 2026), horizontal ties, and political attachment – as the operative channel, with moral disengagement capturing the key political-attachment heterogeneity margin; these findings connect the mental-health literature to the political psychology of partisan identity (Graham et al., 2009; Rathje et al., 2021; Brady et al., 2017) and to the growing literature on how network structure shapes platform welfare effects (Burke and Kraut, 2016; Valkenburg and Peter, 2009).

Fourth, methodologically, we extend the growing toolkit of browser-extension experiments that directly manipulate on-platform content without platform collaboration (Aslett et al., 2022; Yu et al., 2024; Allcott et al., 2024; Farronato et al., 2024; Beknazar-Yuzbashev et al., 2025; Piccardi et al., 2025; Stray et al., 2026).

The paper is structured as follows. Section 2 discusses the experimental design. Section 3 discusses the empirical framework. Section 4 presents the main results. Section 5 examines mechanism and heterogeneity evidence. Section 6 reports robustness checks and alternative interpretations. Finally, Section 7 concludes.

2 Experimental Design

2.1 Intervention and Implementation

We used a Chromium-based desktop browser extension to introduce exogenous variation in exposure to toxic content on three social media platforms – Facebook, X, and Reddit – while leaving platform access otherwise unchanged.[6] The extension achieves that by modifying the

[6]Chromium-based browsers (e.g., Chrome, Edge, Opera) cover 87% of desktop browsing market share as of May 2026; https://gs.statcounter.com/browser-market-share/desktop/worldwide, accessed: 2026-06-06.

page's Document Object Model (DOM) to seamlessly hide targeted posts and comments. Figure 1 shows a toxic Facebook post being hidden, with the content below shifting up to close the gap. An analogous process was applied to tweets on X and posts on Reddit.

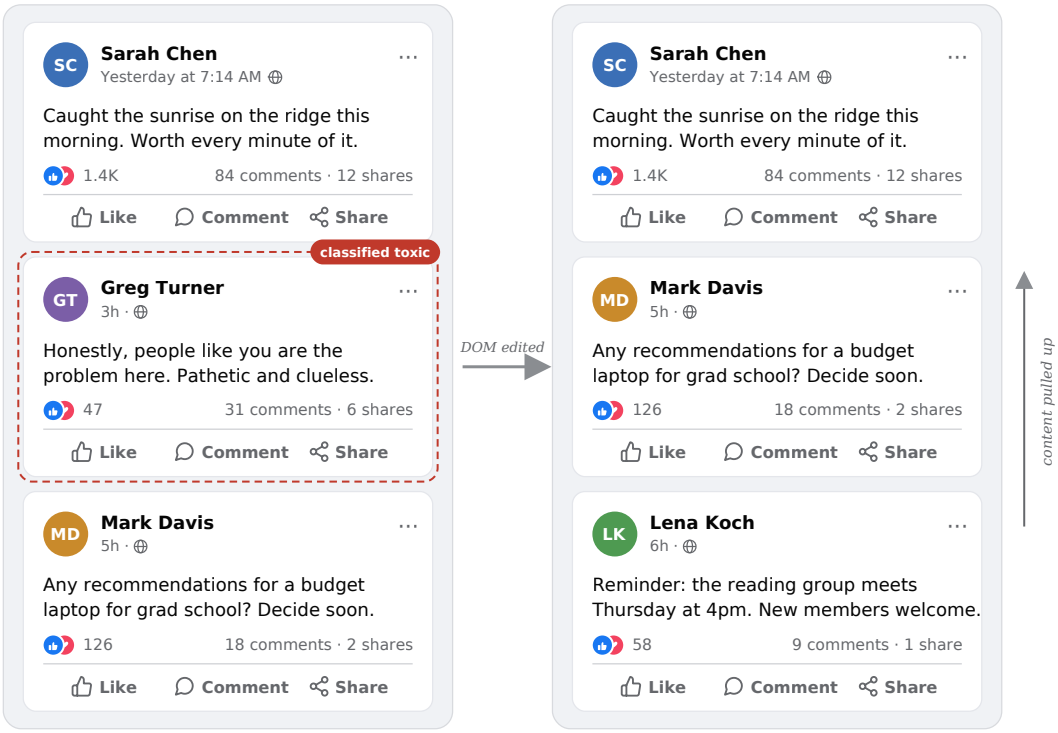


FIGURE 1: HOW THE EXTENSION HIDES TOXIC POSTS

Notes: The figure illustrates the post-hiding mechanism on a stylized Facebook feed. *Left:* the feed as delivered by the platform; one post is flagged as toxic by the extension's classifier (outlined in red and labeled here for exposition only). *Right:* the extension hides the flagged post's container from the page's DOM, and the posts below shift up to fill the vacated space, so that no gap or placeholder remains visible to the participant. The same procedure applies to posts on X and Reddit. The names and texts presented here are fictional.

Comments are hidden in the same way on X and Reddit, whereas on Facebook a hidden comment is instead replaced in place with the italicized placeholder "Deleted comment."[7] Two platform-specific details reflect how each site structures conversations. On Reddit, replies are nested beneath the comment they answer, so hiding a comment also hides its replies. On X, where a single post can span several linked tweets (a "thread"), the original poster's own tweets are exempt within a conversation, so that such a thread is not fragmented; replies left by other users are hidden when toxic. The extension acts only on these three platforms and only on feed posts and comments; advertisements, direct messages, the participant's own content, and anything off-platform are left untouched.

[7] At the time of the experiment, the technology that hid Facebook comments was unstable in testing, and thus we opted for a more reliable moderation tool.

Hiding is fast: in Beknazar-Yuzbashev et al. (2025), which deploys a similar browser extension, the median hiding time is about 400 milliseconds. Because the extension classifies content in parallel, and platforms load most content off-screen, this speed means hiding typically completes before an item reaches the screen, rendering the intervention effectively seamless. An additional experiment there shows that neither longer hiding times nor the hiding process itself affects user experience. On X and Reddit, where feasible, the present extension goes further: newly loaded content is held invisible for a brief interval (to allow the hiding decision to finalize), after which it appears only if it was not flagged.[8] Despite these efforts, we do not rely on speed or seamlessness for identification. Instead, we hold user experience constant through a random-hiding condition, described in Section 2.2.

2.2 Treatment Arms and Randomization

After installing the extension, participants were randomized at the individual level in two stages: first to treatment (60%) or control (40%), then evenly into two conditions within each arm. The resulting four conditions – two treatment (30% each) and two control (20% each) – are summarized in Figure 2.

All conditions score text content with Jigsaw's Perspective API, which assigns each post or comment a toxicity score equal to the predicted share of human annotators who would rate it toxic – that is, rude, disrespectful, or likely to make someone leave a conversation. Our threshold of 0.33 thus flags content that at least a third of annotators would label toxic. The two treatment conditions differ only in the hiding rule: one hides any content above this threshold; the other additionally hides content scoring above 0.7 on Alienation, an experimental attribute of the Perspective API.[9]

The two controls differ in whether anything is hidden. The no-hiding control hides nothing, only recording and scoring content. The random-hiding control instead hides content at random rather than by toxicity, with the rate calibrated to approximate the treatment arms. Comparing treatment with random hiding isolates the effect of hiding toxic content, holding constant both the amount hidden and any possible effects of the hiding process itself.

Our headline comparison pools the two treatment conditions and the two control conditions, contrasting any toxicity hiding with either control. We pre-registered this pooling,

[8] As content loads, the extension immediately makes each new post or comment invisible – while preserving the space it occupies, so the page does not jump – and the toxicity classifier then evaluates it; the item is revealed if it is not flagged and kept hidden if it is. This was not feasible on Facebook.

[9] Alienation is an experimental Perspective attribute developed by Google Jigsaw; Schmer-Galunder et al. (2024), who benchmark it among six prosocial dimensions, define it as "language that refers to an opposing individual or group as less human, inferior, and/or not fitting in within the norms of a social group or society."

and the data support it: both the random-versus-no-hiding difference and the difference between the two treatment rules are null – see Section 6.1. In the same section, we report the variant-specific comparisons for robustness.

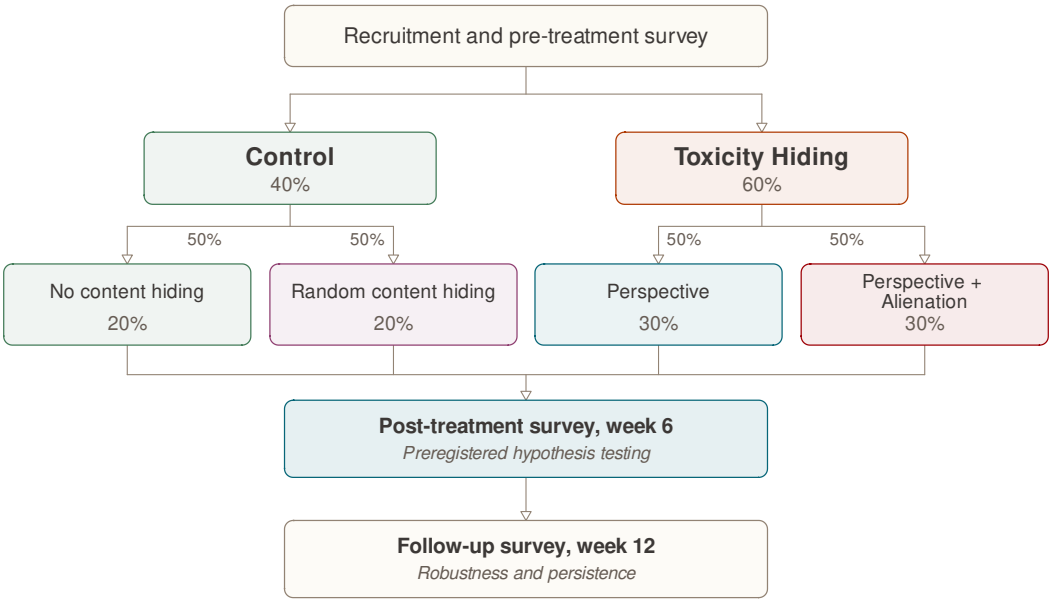


FIGURE 2: EXPERIMENTAL FLOW

2.3 Study Flow

We recruited participants through ads on Facebook and X and through the CloudResearch panel, restricting eligibility to English-speaking U.S. residents on a Google Chrome desktop browser, the environment in which the extension operates. To limit selection into the study on toxicity preferences, recruitment materials described the study only as a paid opportunity to improve one's social media experience and did not disclose the extension's functionality (see examples of recruitment ads in the Online Appendix).

Figure 2 summarizes the study flow. After installing the extension, participants completed the pre-treatment survey and were randomized to experimental conditions. Six weeks later they completed the post-treatment survey, our primary pre-registered measurement, which we compare against the pre-treatment baseline. To assess whether the effect persists, we re-surveyed participants at twelve weeks (the follow-up survey). Due to higher expected attrition, we intended it as a suggestive robustness check. Participants were remunerated through a mixture of lottery entries for Amazon gift cards of up to $500 and flat payments

for survey completion.[10] We ran two waves of identical design – the first from 16 July to 5 November 2024, spanning the U.S. presidential election campaign, the second from 8 May to 6 September 2025, outside any campaign – to test whether effects are specific to a high-salience political period. In total, $N = 1394$ participants enrolled.

This experiment (preregistered as AEARCTR-0013987) was one component of a larger research platform: the same extension, randomization, recruitment, and questionnaire simultaneously supported a separately preregistered political-attitudes study(AEARCTR-0013997) run by an overlapping but distinct set of investigators. A single questionnaire, fielded at all three surveys, collected both the mental-health module we analyze here and the political module analyzed in that paper; the two trials preregistered distinct primary-outcome families *ex ante* – clinical symptom scales here, political attitudes there – cross-reference each other in their registry entries, and are intended for separate publication. We analyze only the mental-health outcomes; the political items enter this paper solely as pre-treatment heterogeneity covariates. Because the combined instrument was administered identically across arms and is orthogonal to treatment assignment, any effect of its length on reported symptoms is common to both arms and nets out of our within-individual estimates – shifting the level of symptoms but not the intent-to-treat effect. The order of the mental-health and political modules was, moreover, cross-randomized, so it too is orthogonal to treatment; we confirm directly that it does not affect our estimates (Section 6).

2.4 Sample Definition

Our starting sample comprises the $N = 1394$ participants who completed the pre-treatment survey and installed the extension. The arms are well balanced on pre-treatment characteristics: for each of the fourteen covariates we examine – demographics (age, sex, race, education), party affiliation, recruitment channel, media-use intensity and per-platform desktop shares, and baseline anxiety and depression – an F-test of equality across the four arms fails to reject at the 5% level, as expected under random assignment (balance table in the Online Appendix).

Estimating treatment effects additionally requires a post-treatment outcome and actual exposure to the intervention. The primary analysis sample therefore retains participants who completed the post-treatment survey and recorded extension activity on a supported platform during the study; this yields $N = 664$. Attrition into this sample is unrelated to

[10]For installing the extension and completing the pre-treatment survey, all participants were entered into a lottery for Amazon gift cards of up to $500. In addition, participants recruited via social media ads were paid $10 for completing both the post-treatment and follow-up surveys, while participants recruited via CloudResearch were paid $5 for each of the pre-treatment, post-treatment, and follow-up surveys.

treatment, and we probe sensitivity to alternative activity requirements in Section 6.

2.5 Design Tradeoffs

Our design has no pre-treatment baseline period for extension activity – the intervention begins as soon as participants complete the baseline survey and are randomized. This protects our survey outcomes: a baseline observation window would have added weeks and raised attrition, whereas starting immediately let us keep the study short and the intervention strong, preserving power on the mental-health effect. The cost is that, without baseline behavioral data, the study is not powered for engagement outcomes and cannot support within-person analysis of behavior. We therefore do not study engagement here; its response to hiding toxic content is already documented in Beknazar-Yuzbashev et al. (2025), whose design supports such within-person analysis.

Second, because the extension operates only on desktop devices, we recruited users who browse social media primarily there, which maximizes the share of their consumption it can screen. The resulting sample is a selected segment of the population of social media users. We examine its representativeness in Section 3.5.

3 Data, Outcomes, and Estimands

This section discusses the data, outcome measures, and empirical framework. Our approach is based on our preregistration and any deviations or extensions are disclosed in a dedicated section of the Online Appendix.

3.1 Primary Outcomes

The primary outcome is based on two psychometric scales widely used in clinical practice to screen for anxiety and depressive disorders: the Generalized Anxiety Disorder scale 7 (GAD-7) and the Patient Health Questionnaire 8 (PHQ-8) (Spitzer et al., 2006; Kroenke et al., 2009). Each instrument asks respondents to indicate, on a four-point scale from "not at all" to "nearly every day," the frequency with which they experienced a set of symptoms over the preceding two weeks. The exact wording of each question is provided in the Online Appendix. Responses are summed to yield a non-negative integer score: GAD-7 ranges from 0 to 21 and PHQ-8 from 0 to 24. Higher values denote higher incidence of mental health symptoms and therefore worse mental health.

Because clinical screening thresholds and population norms are conventionally expressed in terms of the nine-item PHQ-9, we additionally construct a *PHQ-9 proxy* by rescaling the

PHQ-8 sum by 9/8, placing the depression measure on the same 0–27 range as PHQ-9. With integer PHQ-8 scores, the standard moderate-symptom cutoff of PHQ-9 $\geq$ 10 is equivalent to PHQ-8 $\geq$ 9 (Kroenke et al., 2009; Franco et al., 2025).

Our headline outcome is a standardized composite index of mental health symptoms, defined as the simple average of the z-scorized PHQ-8 and GAD-7 items and normalized to mean zero and unit standard deviation using the control pre-treatment distribution.[11] All primary outcomes are elicited in each of the three surveys (pre-treatment, post-treatment, and follow-up).

3.2 Secondary Outcomes

We organize the secondary outcomes into four groups. Two groups are predicted ex ante to be *negative mediators* – channels through which toxicity filtering could worsen mental health: moral self-view and social experiences. The third group – social beliefs – is predicted ex ante to be a *positive mediator* – a channel through which filtering could improve mental health. Lastly, we report behavioral adaptation based on willingness to follow toxic accounts and substitution to other social media.

Moral self-view. The moral-self-view group comprises the absolute moral superiority index and the relative moral superiority measure. The theoretical motivation is that a persistently hostile social-media feed may allow users to derive utility from implicit comparison: others behave toxically, but they do not. If this sense of relative moral rectitude has hedonic value (Tappin and McKay, 2017; Taylor and Brown, 1988), filtering toxicity removes the reference point that sustains it – a form of downward social comparison (Wills, 1981) or self-affirming contrast (Steele, 1988) – and could thereby worsen mental health. We operationalize this channel using two complementary measures. The absolute moral superiority index is constructed from five items of the Marlowe–Crowne Social Desirability Scale, an instrument designed to measure the propensity to present oneself favorably; high scores capture a latent sense of moral virtue. The relative moral superiority measure directly elicits respondents' assessment of how many more positive qualities they have relative to others.

[11] The separately preregistered political-attitudes study has 4 preregistered primary outcomes that are outside this paper's estimand. Even if these outcomes are placed in the same Benjamini–Hochberg (Benjamini and Hochberg, 1995) family as the mental-health-symptoms index, the main ITT inference is unchanged: with 0.0030 as the unadjusted p-value, the largest possible q-value in the resulting 5-outcome family is 0.0149, below 0.05. The Online Appendix reports broader FDR calculations that also include secondary outcomes.

Social experiences. The social-experience group comprises anger, social-media fear of missing out (FOMO), and loneliness.[12] The theoretical motivation is that toxic content, despite its adversarial character, is embedded in social exchanges that provide a sense of shared participation and social presence. Even when experienced passively, hostile exchanges signal that others are engaged in the same arena, creating a sense of common discourse. Filtering such content could strip away this ambient connection, heightening feelings of isolation and apprehension about missing consequential exchanges, thereby worsening mental health. To capture this channel, we administer two items from the validated instrument widely used in large surveys, UCLA Loneliness Scale (UCLA-3) (Hughes et al., 2004). We complement the loneliness measure with two related instruments. First, a single question elicits respondents' fear of missing out (FOMO) – the apprehension that, with toxic content hidden, consequential social and political exchanges are occurring beyond their view (Bursztyn et al., 2025). Second, we measure anger directly through a qualitative survey question, to assess whether toxicity filtering reduces affective arousal rather than social connection.

Social beliefs. The social-beliefs group comprises perceived discrimination, respondent-relevant discrimination, and perceived altruism. The theoretical motivation runs in the opposite direction. Toxic content on social media can target, among others, gender, race, sexual orientation, and age, and its removal may shift respondents' beliefs about how prevalent discriminatory behavior is in everyday life. More benign beliefs about the social environment could in turn improve mental health (Pascoe and Richman, 2009). We measure perceived discrimination by presenting six vignette scenarios spanning these domains and eliciting beliefs about the frequency with which each event occurs.[13] We also report respondent-relevant discrimination, which restricts attention to the subset of scenarios that apply to each respondent's own demographic group. Lastly, we elicit respondents' beliefs about the fraction of a bonus payment that participants in an unrelated study donated to Save the Children USA. This belief question is not susceptible to self-presentational motives and therefore provides a direct test of beliefs about others' prosocial behavior.

Behavioral adaptation. We also study whether participants seek out toxic content when given the opportunity, or shift consumption from the treated platforms toward other social media we did not target. First, in the post-treatment and follow-up surveys, each respondent is cross-randomized to one of two conditions. In the neutral condition, three X accounts posting content on varied innocuous topics are offered for following. In the toxic condition,

[12] Social-experience outcomes were added to the survey after the post-treatment survey of Wave 1.

[13] The wording of the vignettes is provided in the Online Appendix, alongside all other survey instruments.

the same three neutral accounts are presented alongside a fourth account posting politically toxic content. Following Bursztyn et al. (2019), we use an Item Count procedure to elicit the number of accounts respondents wish to follow, reducing social-desirability bias in the reporting of preferences for toxic content. Second, we use data from Beknazar-Yuzbashev et al. (2025) to assess whether any potential decline in mental health could be explained by increased engagement with social media platforms where we did not intervene.

3.3 Pre-Treatment Political Covariates

The mechanism analysis in Section 5 splits the ITT by political characteristics measured in the pre-treatment survey: party affiliation and a battery of political-hostility indices.[14]

Our central measure of political hostility is the moral disengagement index (MDI), which averages agreement with three statements about the respondent's political out-party – that it is "a serious threat to the United States and its people," that it is "not just worse for politics" but "downright evil," and that many of its members "lack the traits to be considered fully human" – each rated on a seven-point agreement scale adapted from Kalmoe and Mason (2022). High values capture a morally charged relation to the political out-party: viewing it as threatening, evil, or less than fully human.

We use four other indices as comparison dimensions in Section 5.3. The dehumanization index combines two mind-perception ratings of the extent to which out-party members can experience sensations and engage in reasoning, the "ascent of man" blatant-dehumanization slider (Kteily et al., 2015), and a demonic–saintly rating of the average out-party member, signed so that higher values denote more dehumanizing views. The political polarization index combines the in-party-minus-out-party differences in feeling-thermometer ratings and in reported comfort with having friends from each party (Iyengar et al., 2012). The support-for-political-violence index averages four items on the acceptability of sending threatening messages to out-party leaders, harassing ordinary out-party members online, and using violence to advance political goals or in response to electoral defeat (Kalmoe and Mason, 2022). The spiral-of-silence index combines the frequency of talking about politics with five items on how comfortable the respondent would feel voicing honest opinions on contentious political issues in a social media discussion, signed so that higher values denote greater self-silencing (Noelle-Neumann, 1974).

[14]The pre-treatment survey carries political outcomes for a separately preregistered political-attitudes study (see Section 2.3). In the present paper these measures enter exclusively as pre-treatment covariates to study heterogeneity of treatment effects on mental health.

3.4 Platform Use and Feed Composition

Post-randomization measurement. The mechanism analysis requires covariates that proxy four pre-treatment features of a user's social-media environment: platform use, toxicity of supplied content, author composition of the feed, and partisan orientation of toxic content. We do not observe these features at baseline: the extension's logs begin at intervention start (Section 2.5). We therefore construct proxies from the extension's post-randomization activity logs, which record platform activity under the same logging rule in the treated and control arms. A post-treatment variable is a credible proxy for a pre-treatment covariate only to the extent that assignment does not shift it: conditioning heterogeneity estimates on a treatment-shifted covariate would mix pre-treatment differences with endogenous treatment response. The remainder of this subsection defines the four measures and documents why this concern is limited in this setting.

Two design features limit the scope for treatment-induced movement in these measures. First, content-based measures use *supplied*, pre-hiding feed posts: hidden posts are recorded like any other impression supplied by the platform, so measurement is identical across arms. Furthermore, for feed posts, unlike comments, delivery does not depend on which items the user chooses to open and is instead determined solely by the platform. Second, measures that treatment could shift quickly are computed from an early window: the first qualifying days after treatment start, before users or ranking algorithms are likely to respond to an unannounced, visually seamless filter. This choice trades off post-randomization stability against sampling precision. Two measures are too sparse for early-window construction: feed composition requires posts with resolvable authors, and the political classification uses only toxic posts, so both are computed over the full logging window. For these variables, we give measure-specific reasons below why treatment-induced movement is unlikely and test its observable implications.

The Online Appendix details each construction, verifies balance, documents the outcome-blind diagnostics used to choose the day-count, coverage, and window criteria, and shows that the heterogeneity estimates are insensitive to each construction choice.

Platform mix. The platform mix is the share of a user's logged platform time spent on Facebook, Reddit, and X, averaged over the early window. We measure platform allocation early because it is the margin most directly exposed to treatment: it reflects user behavior, toxicity hiding changes the content shown in the treated feed, and toxicity filtering potentially may induce differential effects across platforms.

Toxicity of supplied content. We define supplied-content toxicity as the daily share of main-feed posts served to the user that score above the 0.33 Perspective threshold, likewise averaged over the early window.

Peer share. The main feed-composition covariate, the *peer share*, is the share of a user's supplied post impressions authored by *horizontal* accounts – personal accounts of socially proximate peers – as opposed to *vertical* accounts such as pages, public figures, and other broadcast-style sources. Authors are classified using three author-level signals – reach across study participants, typical engagement, and verified status – with computational details in the Online Appendix. The logs resolve author identity only on Facebook and X, so the measure covers those two platforms; each day's two platform shares are combined using weights fixed at the user's early platform mix. Because posts with resolvable authors are a far thinner stream than the time and impression logs behind the other measures, the daily shares are averaged over the full six-week logging window rather than the early one. The longer window widens the scope for treatment to move the measure. In practice, however, the peer share is anchored in who the user follows and befriends – a margin the author-blind filter does not directly target and one that turns over far more slowly than browsing behavior.[15]

Toxic political content. To characterize the partisan orientation of toxic content, we classify toxic-post text on all three platforms with a large language model, assigning each political post to the partisan side to which it is congenial; hostility toward one side is coded as congenial to the other, and mixed or unclear cases are labeled ambiguous. Like the peer share, the classification covers posts only and uses the full logging window: its denominator, the thin set of toxic posts, is too sparse to estimate early. The Online Appendix details the labeling scheme and verifies arm balance in the resulting measure. Section 5.3 uses these labels to measure the party congruence of the toxic content each user encounters.

The mechanism analysis uses a common sample in every column of Table 2: the 431 users with a full-window peer share, pre-treatment MDI and baseline symptoms, and early platform-mix and toxic-post-share measures that clear their coverage gates (criteria in the Online Appendix).

[15]The Online Appendix tests the observable implications of this claim: a parallel experiment with pre-treatment author data finds that toxicity hiding leaves the feed's author base unchanged, the measured share is balanced across arms, an early-window version yields the same heterogeneity estimates, and the results are unchanged when varying other ingredients of the construction.

3.5 Descriptive Statistics

We summarize observable characteristics of the primary main sample of $N = 664$ participants used for the headline six-week ITT estimates and compare them against the population of U.S. social media users. Table 1 provides the full descriptive statistics. The respondents' characteristics, on most dimensions, sit inside the bounds Facebook, Reddit, and X span without cleanly resembling a single platform. To our knowledge, no other dataset documents PHQ-8 and GAD-7 scores for a population of social-media users, so to benchmark the pre-treatment scores in our sample (partly recruited through social media ads) we use the sample of CloudResearch respondents who completed the pre-treatment survey.

3.6 Empirical Framework

Setup. We index individuals by i and periods by $t \in \{0, 1\}$, the pre-treatment baseline and the six-week post-treatment survey. $D_i \in \{0, 1\}$ is assignment to the pooled toxicity-hiding arm, fixed at randomization. The regressor of interest, $\text{Treated}_{it} \equiv D_i \cdot \mathbf{1}\{t = 1\}$, is a treatment-by-post indicator equal to one for treated individuals in the post-treatment period and zero otherwise. Each individual carries an intervention start date $s(i)$; users sharing a start date form a start-date cohort, nested within the two recruitment waves. α_i is an individual fixed effect and $\gamma_{t \times s(i)}$ a cohort-time fixed effect. Y_{it} is the outcome variable, such as the standardized index of mental health.

Our estimand is the intent-to-treat (ITT) effect of assignment to toxicity hiding. Because treatment is randomly assigned, parallel trends holds in expectation. Individuals enrolled and began treatment on different days, so we condition on the start-date cohort through $\gamma_{t \times s(i)}$, which absorbs cohort-specific time effects and confines identification to comparisons of treated against never-treated users within the same cohort and period. Standard errors are clustered at i throughout.

Main specification. Our headline specification is

$$Y_{it} = \beta \, \text{Treated}_{it} + \alpha_i + \gamma_{t \times s(i)} + \varepsilon_{it}. \tag{1}$$

The coefficient β is the average ITT effect of assignment to toxicity hiding on Y_{it}. Following our pre-registration, it pools the two treatment and two control conditions into a single treated-vs-control contrast; we report the variant-specific comparisons in Section 6.

Table 1: Descriptive statistics

	Study sample		Benchmark			
Variable	Mean / Share	N	Facebook	Reddit	X	CR
Panel A. Demographics (vs. platform users)						
Age (years)	41.17	664	47	38	41	38.70
Male	45.6%	664	44%	55%	60%	51.3%
White	67.0%	664	60%	64%	53%	64.7%
Bachelor's degree or higher	52.0%	663	39%	53%	44%	52.8%
Income $\geq$ \$100k	23.9%	664	30%	42%	36%	26.6%
Panel B. Political affiliation (vs. platform users)						
Democrat	47.9%	664				48.2%
Republican	20.5%	664				22.2%
Independent	25.3%	664				24.7%
Unaffiliated / no answer	6.3%	664				4.8%
Democratic (incl. leaners)	69.5%	663	50%	60%	45%	67.8%
Panel C. Recruitment and baseline platform use (vs. site traffic)						
Cloud Research sample	66.3%	664				
Self-reported social-media use (min/day)	116.49	663				109.24
Share of Facebook use on desktop	61.0%	597	∼35–40%			63.2%
Share of Reddit use on desktop	66.7%	552		∼55–60%		69.7%
Share of X use on desktop	60.8%	479			∼40–45%	62.9%
Panel D. Baseline mental health (vs. Cloud Research users)						
Baseline PHQ-8 score	5.85	664				5.77
Baseline PHQ-9 proxy score (PHQ-8 x 9/8)	6.58	664				6.50
Baseline GAD-7 score	5.21	664				5.29
Baseline PHQ-8 $\geq$ 10	23.6%	664				23.4%
Baseline PHQ-9 proxy $\geq$ 10	27.4%	664				27.3%
Baseline GAD-7 $\geq$ 10	19.1%	664				19.2%
Either PHQ-9 proxy or GAD-7 $\geq$ 10	30.9%	664				31.1%

Notes: Study sample is the primary analysis sample ($n = 664$). The Facebook/Reddit/X columns give the demographic composition of each platform's U.S.-adult user base, recovered by inverting Pew Research Center (2025) shares of each group that uses the platform against U.S.-adult base rates (Bayes' rule within each demographic dimension); these are model-based estimates and the age figure uses band midpoints. The CR column reports the same statistics for the Cloud Research (CR) reference sample ($n = 598$): participants recruited into the study through the CloudResearch panel who completed the pre-treatment survey and installed the extension, whether or not they went on to record extension activity or complete the post-treatment survey (i.e., not conditioned on finishing the study). We include them as an external-validity benchmark: our analysis sample is recruited partly through social-media advertising, and, to our knowledge, no other dataset documents these variables for a population of social-media users, so the CR panel provides a same-instrument comparison drawn from a standard survey population (albeit who agreed to install the browser extension).

Heterogeneity. To examine heterogeneity along pre-specified characteristics, we interact the treatment-by-post indicator with a heterogeneity variable M_i:

$$Y_{it} = \beta_1 \,\text{Treated}_{it} + \beta_2 \,\text{Treated}_{it} \times M_i + \alpha_i + \gamma_{t \times s(i)} + \varepsilon_{it}. \tag{2}$$

Here β_2 is the differential ITT and β_1 the ITT in the reference group; M_i may be a binary indicator or a standardized continuous covariate. Because M_i is treatment-invariant and absorbed by α_i, reading β_2 as a differential treatment effect requires that, absent treatment, changes in the outcome not differ systematically across values of M_i within a cohort—parallel trends imposed at the subgroup level.

Heterogeneity by pre-treatment mental health requires a more saturated specification. Unlike the characteristics in equation (2), the splitting variable here is a lagged realization of the outcome itself: partitioning users on baseline symptoms sorts them partly on transitory shocks and measurement error, so the high- and low-baseline groups revert toward the mean at different rates absent treatment (Chay et al., 2005). We therefore let the cohort-time effect vary by pre-treatment-mental-health stratum H_i:

$$Y_{it} = \beta_1 \,\text{Treated}_{it} + \beta_2 \,\text{Treated}_{it} \times H_i + \alpha_i + \gamma_{t \times s(i) \times H_i} + \varepsilon_{it}. \tag{3}$$

With $\gamma_{t \times s(i) \times H_i}$, β_2 is identified from treated-vs-control comparisons within stratum-by-cohort-by-period cells, so the mean-reversion trend—common to treated and control within a stratum—is differenced out. For exploratory quartile analyses we replace the binary indicator H_i with pre-treatment-symptom quartile bins B_i under the same saturated structure.

Exposure response. The ITT estimates the effect of *assignment*. To place it on a realized-exposure scale, we instrument the realized share of toxic content E_{it} with treatment assignment:

$$Y_{it} = \pi \,\widehat{E}_{it} + \alpha_i + \gamma_{t \times s(i)} + \varepsilon_{it}, \qquad E_{it} = \rho \,\text{Treated}_{it} + \alpha_i + \gamma_{t \times s(i)} + u_{it}. \tag{4}$$

Because the intervention begins immediately after enrollment, exposure is measured only post-treatment ($E_{i0} = 0$), so ρ is the post-period treatment-vs-control gap in the share of toxic content and π rescales the reduced form by it. We read π as an exposure-scaled measure of the ITT, under the assumption that assignment affects outcomes only through exposure to toxic content. The null effect of the random-hiding control group vs. the no-hiding control group (Section 6) is consistent with this exclusion, since hiding non-toxic content at the same rate does not move outcomes. Because E_{it} is continuous, π is a dose-response coefficient rather than a binary take-up estimand, and we treat it as an interpretation of the

ITT, not a separate headline estimand.

4 Main Results

4.1 Main Effects on Mental Health

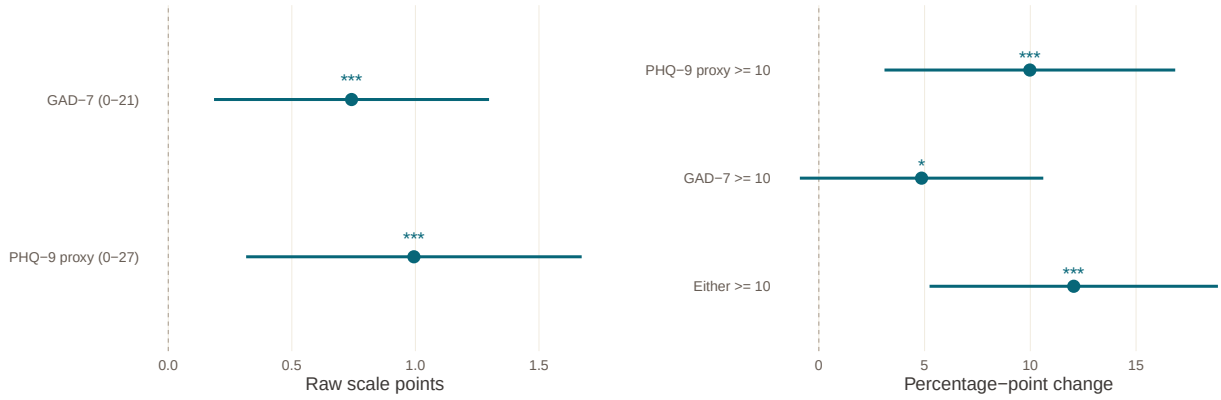
Clinical scales. Toxicity filtering worsens mental health. The standardized mental-health-symptoms index rises by 0.15σ ($p = 0.0030$). The anxiety and depression subindices both increase, with coefficients of 0.13σ ($p = 0.0099$) and 0.16σ ($p = 0.0075$), respectively. The same pattern is visible on the symptom scales: GAD-7 rises by 0.7 points on the 21-point scale and the PHQ-9 proxy rises by 1.0 points on the 27-point scale (Figure 3, left panel).

Threshold crossings. Toxicity filtering also shifts respondents across conventional moderate-symptom screening cutoffs (Figure 3, right panel). The share of individuals scoring at or above the PHQ-9 proxy cutoff of 10 rises by 10.0 percentage points (pp) and the share scoring at or above the GAD-7 cutoff rises by 4.9 pp.[16] The share breaching either cutoff rises by 12.1 pp.

Item-level decomposition. Item-level results are presented in the Online Appendix. For the GAD-7, coefficients are positive across all seven items. The largest increases are on feeling afraid as if something awful might happen (GAD7), feeling nervous or anxious (GAD1), trouble relaxing (GAD4), and becoming easily annoyed or irritable (GAD6). For the PHQ-8, the deterioration is more concentrated: the three items with the largest coefficients are trouble falling or staying asleep (PHQ3), feeling down, depressed, or hopeless (PHQ2), and feeling bad about oneself or like a failure (PHQ6). No single item accounts for the result; the composite effects reflect pervasive but modest increases across multiple symptoms.

Welfare costs. To translate the symptom movement into a welfare metric, we use a deliberately conservative calibration that relies on a continuous individual-level mapping and values the implied loss at respondents' own reported income. This procedure yields a welfare cost of approximately $158 per treated participant at their own individual income over the treatment window. The calculation uses quality-adjusted life-years (QALYs), a standard unit that weights time by health-related quality of life. The Online Appendix documents the aggregation and income-based dollar benchmark underlying this calibration.

[16] The depression threshold result holds at marginal significance even under the conservative, unrescaled PHQ-8 ≥ 10 cutoff (Online Appendix).



(A) Clinical scales (B) Moderate-symptom threshold

Figure 3: Main Effect

Notes: The figure reports ITT estimates from the specification in Equation (1) on anxiety (GAD-7) and depression (PHQ-9 proxy) scales, as well as the crossings of participants into the scores that are classified as moderate depression or anxiety. Bars denote 95% confidence intervals based on standard errors clustered at the respondent level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

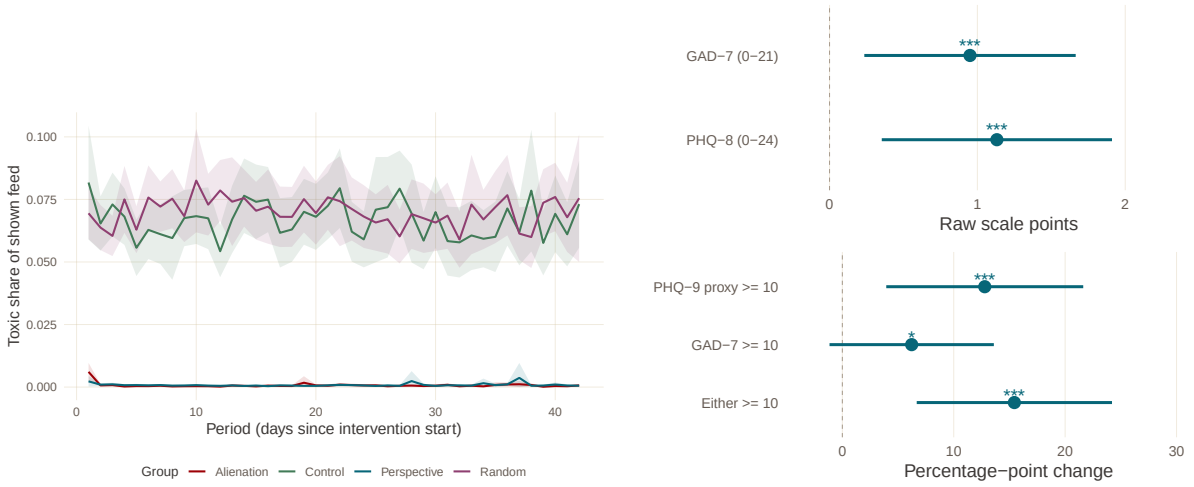
4.2 Treatment Intensity and Exposure-Scaled Effects

We next use the extension logs to characterize the treatment margin and then express the ITT estimates on a realized-exposure scale. Treatment assignment sharply reduces shown toxic content. Over the 42-day treatment window, the shown toxic share averages 7.75% in the control pool and 0.05% in the treatment pool; the corresponding first-stage estimate is a 7.81 pp decline.[17] The arm-level pattern in Figure 4 makes the same point: pure control and random hiding yield shown-toxic shares of 7.63% and 7.86%, while the Perspective and Perspective+Alienation treatment arms yield 0.07% and 0.03%. Random content removal does not reproduce the treatment margin, confirming that the first stage captures toxicity suppression rather than generic content reduction.

The extension logs do not reveal a material shift in platform-supplied toxic content. The offered toxic share is 7.8% in control and 8.6% in treatment, with the corresponding first-stage coefficient indistinguishable from zero. Companion evidence using the same browser extension finds no detectable platform-side adaptation over comparable horizons (Beknazar-Yuzbashev et al., 2025). The treatment therefore changed what respondents saw, not what the platform attempted to supply.

Scaling the reduced form by the shown-toxic-share first stage places the ITT on a realized-

[17] On Reddit, hiding a toxic comment also hides its nested replies, which the logs record as shown, introducing a small measurement error in the treatment-arm toxic shares. Because the first stage is large, this does not meaningfully affect our estimates. Moreover, since replies to toxic comments are likely more toxic than average, the error inflates the treatment share and makes the reported first stage a lower bound.



(A) Shown toxic share (B) Exposure-scaled coefficients

Figure 4: Exposure-Scaled Effect

Notes: This figure reports the realized reduction in toxic exposure induced by the intervention and the corresponding exposure-scaled treatment effects. Panel (a) reports the arm-level share of shown content that is toxic (shown toxic share) over the 42-day treatment window. Panel (b) reports coefficients rescaled to a 10 percentage-point reduction in the shown toxic share, with signs normalized so that positive values denote worse mental health under a reduction in toxic exposure. Within Panel (b), the upper sub-panel reports raw point changes in the GAD-7 and PHQ-8 scores, and the lower sub-panel reports percentage-point changes in the share of respondents scoring at or above the moderate-symptom cutoff on the PHQ-9 proxy, the GAD-7, or either scale. Bars denote 95% confidence intervals based on standard errors clustered at the respondent level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

exposure scale, as shown in Figure 4. A 10 percentage-point reduction in shown toxic share corresponds to an increase of 0.19σ in the mental-health-symptoms index. On the raw clinical scales, the same reduction corresponds to 1.0 GAD-7 points and 1.1 PHQ-8 points. These coefficients are not a separate preferred estimand; they are a scale conversion of the ITT under the exposure interpretation, useful for comparing the dose response across settings with different baseline toxic-share levels.

The complementary intensity question – whether the ITT varies with how much feed content a user consumes – has no clean answer in our design: unlike the compositional exposure measures, consumed volume is the margin treatment touches most directly, since hiding subtracts from what is shown and any behavioral response loads first on quantity, while the pre-treatment alternative, self-reported use, fails validation against measured behavior. The Online Appendix examines the best available margin, early shown feed-post volume, with the scope for treatment to enter the measure limited as far as the design allows. The interaction is positive and stable across constructions, suggesting that the harm concentrates among high-throughput feed consumers, but it remains a descriptive heterogeneity association and we do not build on it.

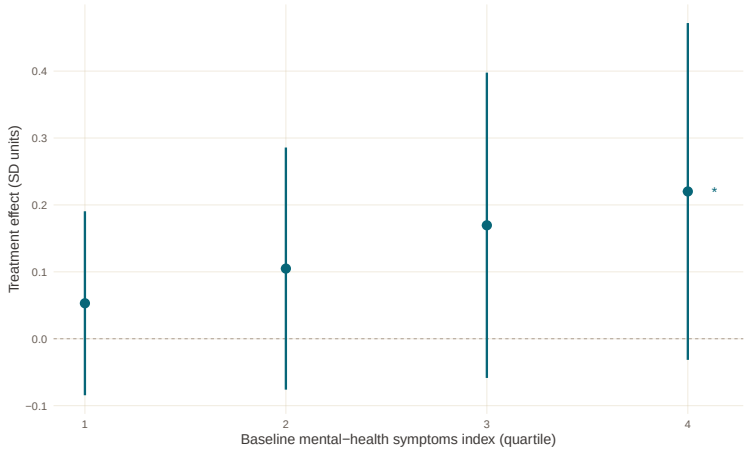


FIGURE 5: HETEROGENEITY BY PRE-TREATMENT MENTAL-HEALTH

Notes: This figure reports estimates of the effect of the intervention on the standardized mental-health-symptoms index, separately by quartile of pre-treatment mental health. Points denote quartile-specific intention-to-treat (ITT) coefficients, where quartiles are defined using pre-treatment symptoms. Bars denote 95% confidence intervals based on standard errors clustered at the respondent level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

4.3 Heterogeneous Treatment Effects

Exploratory subgroup estimates suggest that the adverse effect of toxicity filtering may be larger among respondents who entered the study with worse pre-treatment mental health, but the evidence is imprecise. Figure 5 plots quartile-specific ITT coefficients on the standardized mental-health-symptoms index. The point estimate rises monotonically from 0.05σ in the lowest pre-treatment-symptom quartile to 0.22σ in the highest. The high-minus-low linear combination is 0.17σ, so the ordering is visible in the point estimates but not statistically distinguishable from zero.

The preregistered median split gives the same ordering. The ITT coefficient is 0.09σ ($p = 0.12$) below the median and 0.21σ ($p = 0.02$) above it. The coefficient on treatment-by-high-pre-treatment symptoms is 0.12σ ($p = 0.25$). The difference between subgroups is therefore not statistically distinguishable from zero. We note the pattern as directionally consistent with a mechanism in which users with pre-existing mental health difficulties face greater costs from disruption to their content environment, without treating it as a reliable subgroup finding. The main result – toxicity filtering worsens mental health on average – does not depend on this interpretation.

Other demographic subgroup splits provide little evidence of systematic treatment effect heterogeneity. For sex, age, education, ethnicity, and income, the treatment-by-subgroup interaction coefficients are individually small and jointly insignificant for the mental-health-symptoms index. Usage-intensity splits – defined by above-median active days, above-median active time, and self-reported social-media use level – are similarly uninformative, with

interaction coefficients close to zero and no consistent pattern across definitions. The main finding is not driven by any identifiable demographic or usage subgroup.

Detailed median-split, clinical-scale quartile, and demographic subgroup results are presented in the Online Appendix.

4.4 Benchmarking Magnitudes

To place the magnitude of our headline ITT in context, we compare it with intended mental-health benefits reported in four experimental studies: a mindfulness app, an AI-powered cognitive behavioral therapy app, behavioral-activation psychotherapy, and a one-time psychedelic treatment.

We begin with the digital intervention closest to our six-week treatment window: a large-scale RCT of the Headspace mindfulness app. In Shreekumar and Vautrey (2024), participants randomized to receive access to the Headspace app experience sizeable improvements in mental health: the authors report a $0.38\,\sigma$ reduction in an index that aggregates validated measures of anxiety, depression, and stress after two weeks, and a $0.46\,\sigma$ reduction after four weeks. Relative to the mindfulness-app benchmark, our preregistered six-week ITT estimate of $0.15\,\sigma$ is about 33 percent of the magnitude of Headspace's four-week improvement, but in the opposite direction: toxicity filtering *worsens* mental health by roughly a third as much as a mindfulness intervention of comparable length *improves* it.

Angelucci et al. (2026) conduct a large-scale RCT among 1,964 Mexican women with mild to severe psychological distress, randomizing free access to an AI-powered, CBT-based mental health app for up to six months. Using PHQ-8 and GAD-7 as primary scales – the same instruments as our study – they document a $0.29\,\sigma$ improvement in a composite mental health index at the six-month endpoint, a magnitude the authors note is comparable to psychotherapy and pharmacotherapy. Relative to this benchmark, our headline ITT of $+0.15\,\sigma$ (worsening) is approximately 52 percent of the $0.29\,\sigma$ benefit in Angelucci et al. (2026), despite our treatment window being more than four times shorter (six weeks vs. six months).

Bhat et al. (2022) show that a brief, inexpensive course of behavioral activation durably reduces PHQ-9 depression by 0.23σ five years on for their most effective arm, with a 13 percentage-point fall in moderate-symptom prevalence and that it does so by reshaping beliefs about the self: less overconfidence, and in particular a lower propensity to feel like a failure. Our intervention reverses each of these margins. It raises the comparable symptom index by 0.15σ, lifts the share crossing the moderate-symptom threshold by 12.1 percentage points, and concentrates its symptom burden on the very item, negative self-evaluation, that

therapy attenuates, while lowering relative positive self-view.

A further benchmark comes from outside the digital domain. Francois et al. (2025) run the largest randomized controlled trial of psychedelics to date, partnering with an ayahuasca center in Brazil to randomize 429 first-time users to a one-time ayahuasca treatment or a ritualized placebo. Relative to placebo, ayahuasca raises happiness and lowers psychological distress by roughly 0.4σ six months later, with the gains concentrated among participants who were distressed at baseline. Our 0.15σ adverse effect is thus on the order of a third of this benefit in magnitude – again in the opposite direction – and notable because it arises from a passive change to the everyday content environment rather than from a deliberate, intense therapeutic experience.

Expressed per week of exposure, the unintended adverse effect of content moderation is therefore of the same order of magnitude as the intended benefit from a dedicated mental health care intervention. The parallel is striking: an algorithmic intervention whose purpose is to improve the online environment generates psychological harm at rates comparable to those at which a purposefully designed therapeutic tool generates improvement.

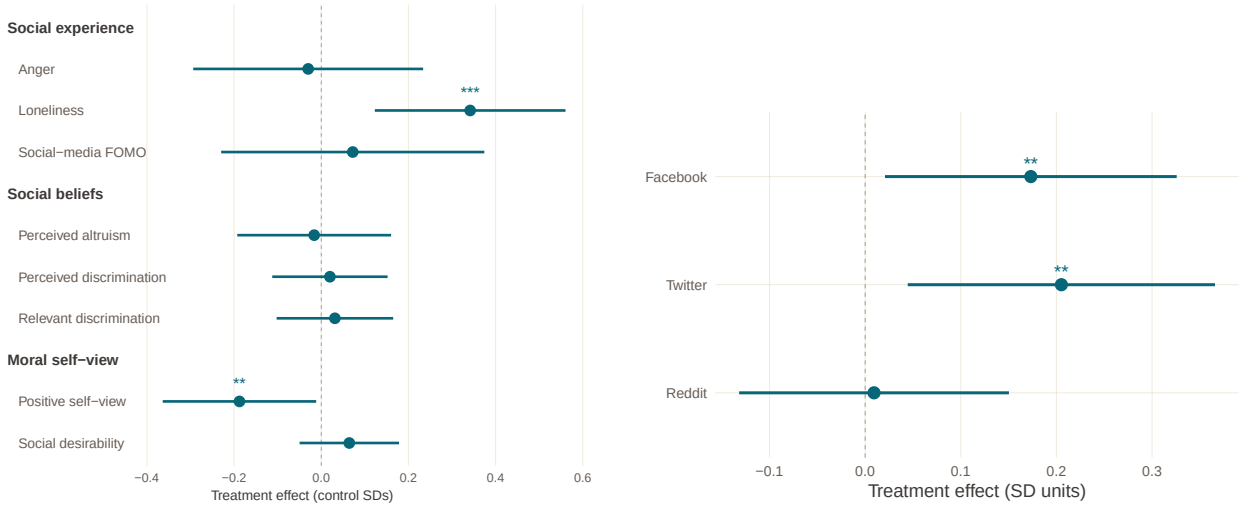
5 Mechanisms

The main results show that assignment to toxicity filtering worsens mental health over the six-week treatment window. The natural question is why. A simple content-quality interpretation would predict the opposite sign: removing hostile posts and comments should improve the social-media bundle that users consume. The mechanism evidence points instead to a different interpretation. Despite its noxious character, toxic content on social media is often embedded in social relationships, shared attention, and politically salient exchanges. Filtering it can therefore remove information about what proximate or identity-relevant users are saying, fighting over, and treating as important – including others fighting for causes the user cares about, so that its removal can leave the user's own point of view feeling more isolated. That proximity can be literal, through horizontal ties in the feed, or political, through attachment to conflicts and communities that make the content socially relevant.

This section organizes the evidence around that interpretation. We first show that the secondary outcomes point most directly to social connection. We then ask where the adverse ITT is largest. The pattern is strongest among users whose feeds contain more horizontal ties and among users for whom political content is more socially relevant. The final subsection puts these margins in the same specification and shows that they capture distinct dimensions of the same underlying idea: toxicity filtering is most costly when it removes content connected to "your people."

5.1 Social Connection

The secondary outcomes in Figure 6, panel A, point most directly to social connection. Toxicity filtering significantly raises the index of loneliness, by 0.34σ ($p = 0.003$),[18] while anger, FOMO, social desirability, perceived discrimination, and altruism move little. The pattern suggests that the adverse ITT is not primarily about affective arousal, fear of missing out, favorable self-presentation, or respondents revising their beliefs about the social environment; it is about feeling less connected to it. Details and Benjamini–Hochberg corrections across the secondary-outcome family are reported in the Online Appendix.



(A) Secondary outcomes (B) Platform heterogeneity

Figure 6: Secondary outcomes and platform heterogeneity

Notes: This figure reports estimates of the effect of the intervention on secondary outcomes (Panel (a)) and the implied per-platform effects on mental health (Panel (b)). Panel (a) reports intention-to-treat (ITT) coefficients on the secondary survey outcomes, standardized in control-group standard-deviation units. Panel (b) reports implied per-platform ITT coefficients on the standardized mental-health-symptoms index, obtained from the early-platform-mix specification, with Facebook as the omitted platform. Bars denote 95% confidence intervals based on standard errors clustered at the respondent level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

This evidence suggests that toxic content may be part of the social presence that users experience on these platforms. Even when a post or comment is hostile, it can signal that other people are participating in the same arena and reacting to the same events. Removing that content may therefore reduce exposure to hostility while also thinning the user's sense

[18] This is the six-week post-treatment estimate, identified from the preregistered within-person comparison between the pre-treatment and six-week surveys (342 observations). These items were introduced by the October 2024 pre-registration amendment, so only the second recruitment wave answered them at the six-week survey; the first wave answered them only at the twelve-week follow-up and therefore contributes a cross-section rather than a within-person comparison. Pooling that cross-section with the six-week estimate by inverse-variance weighting gives a smaller but still significant 0.21σ ($p = 0.005$), statistically indistinguishable from the six-week estimate (equality test $p = 0.121$) in the Online Appendix.

of shared participation. The point is not that toxic content is desirable in itself, but that it is bundled with social information.

The moral-self-view outcomes point in a related but secondary direction. Positive self-view falls under treatment, consistent with the idea that filtering toxic content removes a comparison point against which respondents can see themselves favorably; however, this coefficient does not survive Benjamini–Hochberg correction across the secondary-outcome family (Online Appendix). We therefore do not treat this as the central channel. Rather, it supports the broader interpretation that toxic content supplies social reference points, not just negative information.

Platform heterogeneity points in the same direction. The subgroup-specific coefficients are positive for Facebook and X, and close to zero for Reddit (Panel B of Figure 6). This split is useful because the platforms organize consumption differently. Facebook and X more often place toxic content inside ongoing social exchange: personal feeds, replies, quote posts, comment threads, and public conversations involving socially salient accounts. Reddit is more often organized around topical communities and information search, with weaker or less persistent author-level ties. The platform pattern is therefore consistent with the social-connection interpretation: filtering is more costly where toxic content is embedded in social-network consumption than where it is closer to informational consumption.

5.2 Horizontal Ties

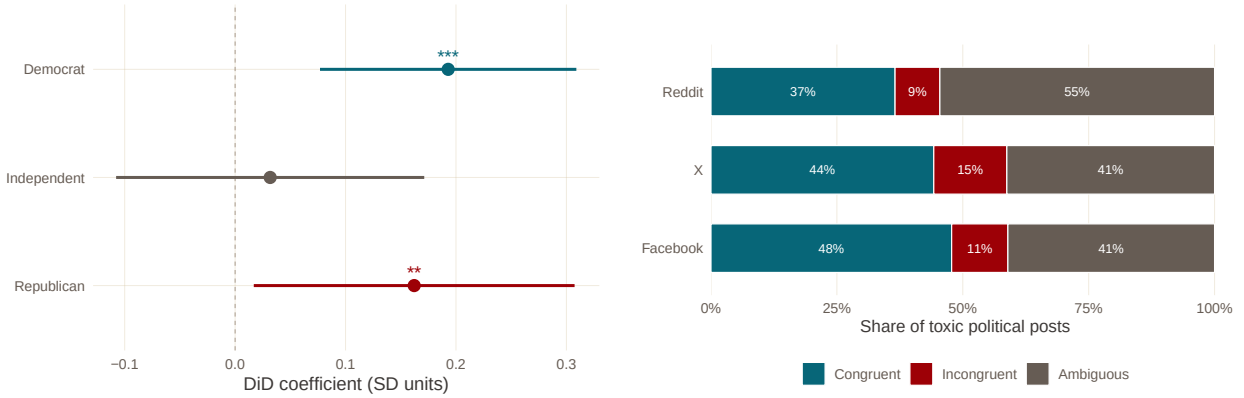
The share of impressions coming from horizontal authors differs across platforms. The next piece of evidence therefore asks whether horizontal ties matter conditional on platform mix. If the cost of filtering toxicity came only from changing the quality of consumed content, there would be little reason for the effect to depend on whether the hidden content came through horizontal ties. Instead, the first column of Table 2 shows that, controlling for early platform shares, the adverse ITT is larger for users with a higher peer share – a higher share of feed impressions from horizontal authors – with a coefficient of 0.11σ ($p = 0.013$). Section 3.4 discusses how we distinguish feed impressions from horizontal and vertical authors.

This heterogeneity is informative because horizontal ties are the part of the feed most closely tied to the network feature of social media. A toxic post from a friend, acquaintance, or socially proximate account carries different information than a toxic post from an impersonal source: it conveys what people in the user's own social environment are discussing, and how. The larger coefficient among users with more horizontal ties therefore suggests the mental-health cost is concentrated where toxic content is embedded in social relationships rather than broadcast.

5.3 Political Attachment

Political attachment provides a second margin of proximity. The horizontal-tie evidence shows that author-level social proximity matters, but it does not exhaust the relevant social environment. Content can also be connected to the user through political communities, conflicts, and identities, even when the author is not personally close.

The subgroup estimates by party affiliation in Figure 7, panel A, point in that direction. The ITT coefficients are positive for Democrats and Republicans, while the coefficient for independents is close to zero. Users with clearer partisan attachment therefore bear larger costs from toxicity filtering than users without such an attachment.



(A) Party-specific ITT coefficients (B) Toxic political content by platform

Figure 7: Political attachment and toxic political content

Notes: This figure reports estimates of the effect of the intervention on mental health by political affiliation (Panel (a)) and the corresponding composition of toxic political content in respondents' feeds (Panel (b)). Panel (a) reports party-specific intention-to-treat (ITT) coefficients on the standardized mental-health-symptoms index. Panel (b) reports post-weighted shares of toxic political posts over treatment days 1–42, separately by platform, among respondents identifying with or leaning toward the Democratic or Republican party. A post is classified as congruent when its classified partisan side matches the respondent's party and as incongruent when it matches the opposing party; the ambiguous category comprises political posts with mixed, non-U.S., unclear, or otherwise non-clean partisan-side labels. Bars in Panel (a) denote 95% confidence intervals based on standard errors clustered at the respondent level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

The content evidence in panel B points in the same direction. Among toxic political posts with a discernible affiliation, the overwhelming majority are party-congruent with the user's own affiliation. The toxic content removed by the extension is therefore often tied to the user's political side, not only to hostile out-party speech. This helps explain why politically charged toxic content can be socially relevant rather than merely aversive.

Baseline political indices sharpen this interpretation because the significant HTE is concentrated in MDI rather than spread across the political battery. Respondents with higher pre-treatment MDI show a larger adverse response to toxicity filtering (Table 2, second column); the corresponding coefficients for dehumanization, political polarization, support

for political violence, and spiral of silence are smaller and not statistically distinguishable from zero (reported in the Online Appendix). This pattern is informative: the political margin is not generic partisan intensity, willingness to speak, or broad hostility. It is the morally charged relation to the political out-party captured by MDI: viewing the out-party as threatening, evil, or lacking humanity. The party split therefore appears to reflect more than partisan identity as a demographic marker. It reflects whether political conflict is central to the user's feed and social world.

5.4 Social and Political Proximity

Table 2: Peer-share and MDI heterogeneity combined

	(1)	(2)	(3)	(4)	(5)
Treated × peer share (per +1 SD)	0.110**		0.140***	0.141***	0.141***
	(0.044)		(0.045)	(0.046)	(0.046)
Treated × MDI (per +1 SD)		0.116***	0.129***	0.128***	0.128***
		(0.042)	(0.042)	(0.042)	(0.042)
Treated × peer share × MDI				0.023	0.023
				(0.042)	(0.042)
Observations	862	862	862	862	862
Baseline symptoms controls		✓	✓	✓	✓
Platform share controls	✓		✓	✓	✓
Toxicity-dose control					✓

Notes: This table reports estimates from six-week post-treatment difference-in-differences regressions for the standardized mental-health-symptoms index. All columns are fit on a single common sample of 431 users, held fixed across the table: users with a measured peer share, an early platform mix (coverage gate of at least three observed active days), a pre-treatment moral disengagement index (MDI), pre-treatment symptoms, and an early toxicity dose clearing its own three-day coverage gate. The dose is required throughout because the final column controls for it, so that every column uses the identical observations. The peer share is the horizontal-author share of post impressions; the peer share and the MDI are standardized on this common sample, so every interaction coefficient is expressed per one-standard-deviation (+1 SD) increase. Three control blocks enter where indicated in the footer: platform-share controls interact the treatment dummy with early Reddit and Twitter usage shares (Facebook is the omitted platform, and the corresponding rows are not displayed); baseline-symptoms controls interact the treatment dummy with a high-pre-treatment-symptoms indicator and stratify the cohort-by-time fixed effect on that indicator; and, in the final column, a toxicity-dose control interacts the treatment dummy with the early share of platform-supplied posts that are toxic (defined in Section 3.4; construction and balance in the Online Appendix), entered only as a control so that its coefficient is not displayed. The treated main effect is the omitted-cell intercept under these interactions and is not displayed; the headline no-controls intention-to-treat (ITT) estimate on this common sample is $0.18\,\sigma$ (standard error 0.06). Because the peer share and the MDI are treatment-invariant covariates rather than randomly assigned treatment cells, the interaction coefficients should be interpreted as descriptive heterogeneity associations, not as identified differential ITTs. Standard errors are clustered at the respondent level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

The final step is to put the two margins together. Table 2 enters each margin alone – each column carrying only the controls its confound requires, as flagged in the footer – and then both jointly, all on one common sample. The peer share and baseline MDI both remain positive and statistically significant when entered in the same specification, and the final column shows they also survive netting out the amount of toxicity removed (the early toxic-post dose, entered as a control). The interaction between the two is null.

The evidence therefore does not require the cost of filtering toxicity to be concentrated only among users who are both socially proximate to their feed and politically attached. Instead, social proximity and political self-relevance appear to be separable margins.

A remaining concern is that the peer-share and MDI margins proxy not for where hidden content sits in the user's social world but simply for how much toxicity was removed. If higher-peer-share or higher-MDI users faced a larger toxicity dose, and the adverse effect scaled with that dose, the heterogeneity would carry no social-proximity interpretation. The final column of Table 2 already nets out the early toxic-post dose. Further analysis in the Appendix corroborates this alternative more fully on the same common sample. Three facts weigh against it. First, neither heterogeneity variable carries much dose information: the peer share is, if anything, negatively correlated with the early toxic-share dose, and the MDI correlation is near zero. Second, the outcome does not scale with the dose, i.e. the treatment-by-dose interaction is itself indistinguishable from zero. Third, the peer-share and MDI coefficients are essentially unchanged when the treatment-by-dose interaction is added and when each variable is residualized on the dose and re-standardized; the early supplied toxic share is itself approximately balanced across arms once the start-date cohort is netted out. The heterogeneity therefore does not appear to be an artifact of the amount of toxicity removed.

This synthesis gives the mechanism evidence a single interpretation. Toxicity filtering has the largest private cost when hidden content is connected to the user's relevant social environment. That connection can be literal, through horizontal ties in the feed. It can also be political, through communities, conflicts, and identities that make the content socially meaningful even when the author is not personally close.

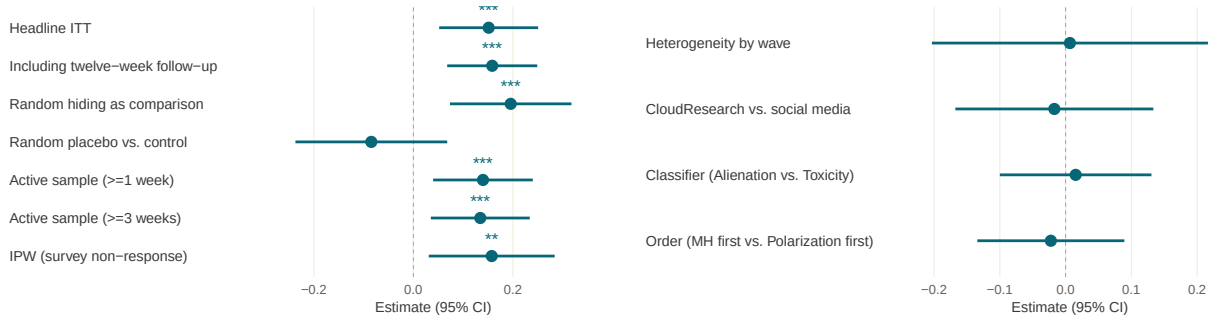
Social media may be becoming more media-like, but the network component still matters for welfare.

6 Robustness and Alternative Interpretations

The headline ITT is that assignment to toxic-content filtering raises the mental-health-symptoms index by 0.15σ. This section asks two questions about that estimate. First, is the coefficient fragile to natural alternatives in horizon, sample, weighting, and implementation? Second, does the four-arm design support the interpretation that the adverse ITT comes from reducing toxic-content exposure rather than from generic feed disruption or the desktop-only scope of the intervention?

6.1 Robustness of the Estimate

We first ask whether the mental-health-symptoms coefficient depends on a narrow analysis choice. Figure 8 reports the same ITT under the main alternatives available in the design: adding the twelve-week follow-up survey to the estimation sample, reweighting for survey non-response, estimating the coefficient among respondents who remained extension-active for longer windows, and allowing the coefficient to differ by recruitment wave, recruitment source, questionnaire order, and classifier.



(A) Treated (B) Treated × subgroup

Figure 8: Robustness panel

Notes: This figure reports estimates from difference-in-differences regressions for the standardized mental-health-symptoms index, where each row is a separate estimate and higher values indicate worse mental health. Panel (a) reports the treated coefficient for specifications without a subgroup interaction, together with the combined active-stratum intention-to-treat (ITT) effect for the ≥1-week and ≥3-week active-sample cuts; for these combined effects, the standard error is the linear-combination standard error computed from the individual-clustered covariance matrix. Panel (b) reports the treatment-by-subgroup coefficient (`Treated` × subgroup) for specifications that include a subgroup interaction. Bars denote 95% confidence intervals based on standard errors clustered at the respondent level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

The estimate is stable across these checks. Adding the twelve-week follow-up survey leaves the pooled post-period ITT close to the post-treatment-survey coefficient, so the result is not an artifact of ending the analysis at the preregistered endpoint. Reweighting post-treatment-survey respondents by the inverse of their predicted completion probability produces a similar coefficient, which limits the role of selection on observed pre-treatment covariates into survey response; this complements the differential-attrition tests in the Online Appendix. Restricting attention to respondents who remained active on the extension for at least one or three weeks also leaves positive coefficients. These activity windows are realized after assignment, so we read them as descriptive compliance-window checks rather than as alternative causal estimands; their purpose is to show that the adverse ITT is not driven by users who installed the extension and immediately stopped using it.

Nor is the coefficient concentrated in one implementation stratum. The treatment-by-wave, treatment-by-recruitment-source, and treatment-by-survey-order coefficients are all

close to zero relative to the main ITT. The recruitment-source split compares respondents recruited through the CloudResearch panel with those recruited through social-media ads; the classifier split compares the toxicity-only and alienation-augmented arms. Both deliver interaction coefficients indistinguishable from zero. Taken together, the robustness panel leaves the main estimate tied to the randomized assignment rather than to a particular horizon, weighting rule, compliance window, recruitment channel, questionnaire order, or classifier.

Lastly, we see no evidence of differential attrition: treatment assignment does not meaningfully predict either post-treatment survey completion or extension activity through the post-treatment survey, and this conclusion is unchanged after adding pre-treatment symptom and demographic controls and their treatment interactions. Appendix explores this in more detail. The inverse-propensity-weighted estimate in Figure 8 reaches the same conclusion from the selection-on-observables direction: reweighting post-treatment-survey completers by predicted completion probabilities leaves the mental-health ITT close to the unweighted estimate.

6.2 Validity of the Interpretation

Assignment to filtering sharply reduces realized toxic exposure, but it also removes content and changes the feed. The first interpretive concern is therefore generic disruption: perhaps any content removal, regardless of toxicity, worsens mental health. The random-hiding arm addresses that concern directly by removing a matched fraction of content without targeting toxicity. If generic disruption were the operative margin, random hiding should also worsen mental health. It does not. Relative to pure control, the random-hiding coefficient on the mental-health-symptoms index is -0.08σ (s.e. $= 0.08$), which is small, statistically indistinguishable from zero, and opposite-signed to the treatment effect. Conversely, when the random-hiding arm is used as the comparison group, the toxicity-filtering ITT remains positive (Figure 8). The four-arm design therefore points away from generic hiding and toward the toxicity of the removed content as the relevant treatment margin.

A second alternative is transitory adjustment: treated users may initially react to the sudden removal of familiar content and then converge back as they adapt. Figure 9 estimates a dynamic ITT with separate coefficients at the post-treatment and follow-up surveys. The symptom index rises by 0.17σ at the six-week post-treatment survey and remains elevated by 0.13σ at the twelve-week follow-up survey. The follow-up coefficient is less precise, as expected given attrition, but the adverse ITT does not vanish. That persistence is difficult to reconcile with a purely transitory adjustment response; it is more consistent with a durable

change in the experienced social environment.

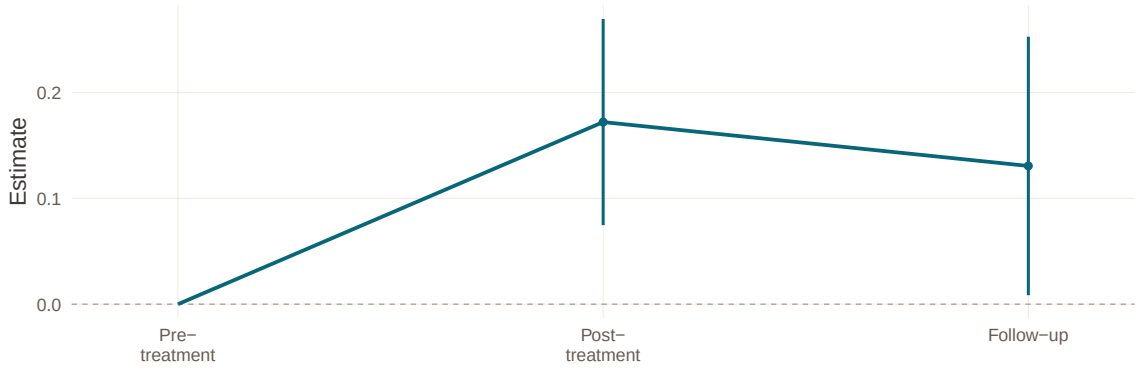


FIGURE 9: MENTAL-HEALTH-SYMPTOMS ITT DYNAMICS AT THE POST-TREATMENT AND FOLLOW-UP SURVEYS

Notes: This figure reports intention-to-treat (ITT) estimates of the effect of the intervention on the mental-health-symptoms index from a dynamic difference-in-differences specification, estimated separately for the six-week post-treatment survey and the twelve-week follow-up survey. The pre-treatment period is normalized to zero. The outcome is standardized in control-group standard-deviation units, with higher values indicating worse mental health. Points denote ITT coefficients, and vertical lines denote 95% confidence intervals based on standard errors clustered at the respondent level.

A broader concern is substitution outside the treated desktop feeds. If users replace hidden toxic content with equally or more toxic content elsewhere, the intervention may not reduce their overall toxicity exposure and could, in the extreme, raise it. We examine this concern in two ways. First, within the same platforms, treatment does not shift reported use from desktop to unfiltered mobile devices, so the extension appears to reduce toxicity on the platforms it targets rather than merely moving the same consumption to phones. Second, while prior evidence using the same extension shows some substitution toward other social-media sites, that substitution does not flow toward sites classified as more toxic. The available behavioral checks also do not show active demand for toxic accounts. The full substitution evidence is reported in the Online Appendix.

Finally, we rule out the case that the effect is driven by the respondents in treatment group realizing the exact purpose and functionality of the extension. We explicitly asked the respondents whether they had experienced any change in their social media experience during the weeks of the experiment, and we find no evidence that treated respondents reported more change: on a five-point scale, the treated–control gap is a precise null (-0.03, s.e. 0.10, $p = 0.73$).[19]

[19]The item was introduced partway through data collection and has no pre-treatment counterpart, so we pool all post-treatment responses – both the six-week and twelve-week surveys and both recruitment waves – to maximize power, comparing treated and control respondents within start-date-cohort-by-survey cells with standard errors clustered at the respondent level.

7 Conclusion

This paper provides novel randomized field evidence on the mental-health consequences of content-level moderation that hides toxic content from social-media feeds while leaving platform access intact. Across two recruitment waves spanning the 2024 U.S. election campaign and a later non-campaign period, hiding posts and comments that exceed pre-specified toxicity thresholds (Perspective-style scores grounded in human annotation) *worsens* mental health. In difference-in-differences specifications with respondent and cohort-time fixed effects, treated participants exhibit increases in symptom burden on both the GAD-7 and PHQ-8 and a rise in a composite mental-health-symptoms index. The estimated impact is 0.15σ – broad-based across clinically validated outcomes. On the raw clinical scales, GAD-7 rises by 0.7 points and the PHQ-9 proxy rises by 1.0 points. The deterioration is not confined to mean scores: the share of respondents crossing the moderate-symptom threshold rises by 12.1 percentage points on the combined PHQ-9 proxy and GAD-7 cutoff. Using a deliberately conservative QALY-based calibration, we compute an implied welfare loss of about $158 per treated participant over the six-week treatment window.

These findings speak to a central open question in the literature: which components of the online environment drive the mental-health costs of social media? Prior work documents that deactivating platforms can improve well-being and mental health (Allcott et al., 2020, 2025), that the diffusion of social media during Facebook's rollout worsened student mental health (Braghieri et al., 2022), and that persistent usage can reflect self-control problems (Allcott et al., 2022) and network externalities that trap users on the platform (Bursztyn et al., 2025). We move from platform access to a sharply defined content lever. By manipulating exposure to toxic content while holding overall access fixed and by incorporating both a pure control and an active (random-hiding) control, our design isolates viewer-side exposure to toxicity from other channels. For this component, the answer is negative: viewer-side exposure to toxic content is not a source of the mental-health costs of social media – removing it does not alleviate them, and instead makes them worse. Together with evidence that reducing toxicity also lowers engagement (Beknazar-Yuzbashev et al., 2025), our results imply that removing toxic content can be costly on both the engagement and the mental-health margin.

Why should removing hostile content make users worse off? The mechanism evidence points away from content quality and toward social connection. The largest and most robust effect among the secondary outcomes is on loneliness. The heterogeneity points the same way. The cost is concentrated where toxic content reaches the user through peers rather than impersonal accounts, and on the platforms organized around personal ties rather than topical communities. It is also larger among users with a morally charged attachment to

political conflict, for whom hostile political content with a discernible lean — most of it congenial to their own side — is socially meaningful, not merely unpleasant. Toxic content, it appears, comes bundled with a sense of social presence, and filtering removes both.

Three limitations bound the interpretation. First, our estimand is viewer-side: we measure what filtering does to the people whose feeds are cleaned, not to the targets of toxic speech or to the offline harms that motivate moderation in the first place. Our welfare figure is therefore not a number to net against those benefits. Second, the extension operates on desktop browsers, so we recruit users who consume social media there and filter only that share of their consumption. Extending this evidence to users who consume social media predominantly through the mobile apps is a natural next step. Third, our horizon is weeks. The adverse effect is still present at the twelve-week follow-up, but longer-run adaptation — in habits, in whom users follow, and in what platforms supply — lies beyond our design.

Our counterintuitive findings raise a natural question: does this mean content moderation is net harmful and should be abandoned? The answer is clearly no. There are compelling reasons – including offline violence, discrimination, and the well-documented aggregate harms of hateful content – to continue moderating the content on large platforms. The point is rather that the *form* of moderation matters. Our intervention suppressed toxic content above a fixed toxicity threshold, removing user agency in the process. A less disruptive alternative would be for platforms to flag posts exceeding a toxicity threshold – displaying a warning or a "view anyway" prompt – rather than hiding them outright. This approach preserves the user's ability to engage with the content when they choose to do so, while still introducing friction that may reduce unreflective consumption and limit exposure among users who neither want nor benefit from the material. Platforms could additionally allow users to set their own toxicity thresholds. Finally, because the cost concentrates where toxic content arrives through peers, platforms could moderate peer-authored content more lightly than content from public figures and institutions — either automatically, or by letting users opt out of moderation of peer content themselves.

References

Ahmad, Wasiq, Ananya Sen, Charles Eesley, and Erik Brynjolfsson, "Companies inadvertently fund online misinformation despite consumer backlash," *Nature*, 2024, *630* (8015), 123–131.

Allcott, Hunt, Jose Castillo, Matthew Gentzkow, Lukas Musolff, and Timo Salz, "Sources of Market Power in Web Search: Evidence from a Field Experiment," 2024. mimeo.

_ , Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow, "The Welfare Effects of Social Media," *American Economic Review*, 2020, *110* (3), 629–676.

_ , Matthew Gentzkow, and Lena Song, "Digital Addiction," *American Economic Review*, 2022, *112* (7), 2424–2463.

_ , _ , Benjamin Wittenbrink, Juan Carlos Cisneros, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Sandra González-Bailón, Andrew M Guess, Young Mie Kim et al., "The Effect of deactivating Facebook and Instagram on users' emotional state," Technical Report, National Bureau of Economic Research 2025.

_ , _ , Ro'ee Levy, Adriana Crespo-Tenorio, Natasha Dumas, Winter Mason, Devra Moehler, Pablo Barbera, Taylor Brown, Juan Carlos Cisneros et al., "The effects of political advertising on Facebook and Instagram before the 2020 US election," *Nature Human Behaviour*, 2026, *10* (5), 884–895.

Andres, Rafael and Olga Slivko, "Combating Online Hate Speech: The Impact of Legislation on Twitter," Discussion Paper, ZEW 2021.

Angelucci, Manuela, Raissa Fábregas, and Antonia Vazquez, "The Well-Being Effects of Digital Mental Health Care," Technical Report, ROCKWOOL Foundation Berlin (RFBerlin) 2026.

Aslett, Katherine, Andrew M. Guess, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker, "News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions," *Science Advances*, 2022, *8* (18), eabl3844.

Beknazar-Yuzbashev, George, Rafael Jiménez-Durán, and Mateusz Stalinski, "A model of harmful yet engaging content on social media," in "AEA Papers and Proceedings," Vol. 114 American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203 2024, pp. 678–683.

_ , _ , Jesse McCrosky, and Mateusz Stalinski, "Toxic Content and User Engagement on Social Media: Evidence from a Field Experiment," CESifo Working Paper 11644, CESifo 2025.

Benjamini, Yoav and Yosef Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, 1995, *57* (1), 289–300.

Bhat, Bhargav, Jonathan de Quidt, Johannes Haushofer, Vikram Patel, Gautam Rao, Frank Schilbach, and Pierre-Luc Vautrey, "The Long-Run Effects of Psychotherapy on Depression, Beliefs, and Economic Outcomes," April 2022. Working paper. Pre-registered as AEARCTR-0003823.

Biasi, Barbara, Michael S. Dahl, and Petra Moser, "Career Effects of Mental Health," Working Paper 29031, National Bureau of Economic Research July 2021.

Brady, William J., Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel, "Emotion Shapes the Diffusion of Moralized Content in Social Networks," *Proceedings of the National Academy of Sciences*, 2017, *114* (28), 7313–7318.

Braghieri, Luca, Ro'ee Levy, and Alexey Makarin, "Social media and mental health," *American Economic Review*, 2022, *112* (11), 3660–3693.

Burke, Moira and Robert E. Kraut, "The Relationship between Facebook Use and Well-Being Depends on Communication Type and Tie Strength," *Journal of Computer-Mediated Communication*, 2016, *21* (4), 265–281.

Bursztyn, Leonardo, Benjamin Handel, Rafael Jiménez-Durán, and Christopher Roth, "When product markets become collective traps: The case of social media," *American Economic Review*, 2025, *115* (12), 4105–4136.

_ , Georgy Egorov, Ruben Enikolopov, and Maria Petrova, "Social media and xenophobia: evidence from Russia," Technical Report, National Bureau of Economic Research 2019.

Chay, Kenneth Y, Patrick J McEwan, and Miguel Urquiola, "The central role of noise in evaluating interventions that use test scores to rank schools," *American Economic Review*, 2005, *95* (4), 1237–1258.

Council on Foreign Relations, "Hate Speech on Social Media: Global Comparisons," 2019.

Currie, Janet and Mark Stabile, "Child mental health and human capital accumulation: The case of ADHD," *Journal of Health Economics*, November 2006, *25* (6), 1094–1118.

_ and _ , "Mental Health in Childhood and Human Capital," in Jonathan Gruber, ed., *An Economic Perspective on the Problems of Disadvantaged Youth*, Chicago: University of Chicago Press for NBER, 2009, pp. 115–148.

de Quidt, Jonathan and Johannes Haushofer, "Depression for Economists," Working Paper 22973, National Bureau of Economic Research December 2016.

Durán, Rafael Jiménez, Karsten Müller, and Carlo Schwarz, "The Online and Offline Effects of Content Moderation: Evidence from Germany's NetzDG," *Available at SSRN 4230296*, 2025.

Eisenberg, Daniel, Ezra Golberstein, and Justin B. Hunt, "Mental Health and Academic Success in College," *The B.E. Journal of Economic Analysis & Policy*, September 2009, *9* (1), Article 40.

Enikolopov, Ruben, Maria Petrova, Gianluca Russo, and David Yanagizawa-Drott, "Socializing Alone: How Online Homophily Has Undermined Social Cohesion in the US," 2024.

Farronato, Chiara, Andrey Fradkin, and Colin Karr, "Webmunk: A New Tool for Studying Online Behavior and Digital Platforms," 2024. Working paper/tool description.

Franco, Catalina, Akshay Moorthy, and Sara Abrahamsson, "Gender Differences in Recognizing Depression and Seeking Help," 2025.

Francois, Patrick, Matt Lowe, and Ieda Matavelli, "Psychedelics and Well-Being: An Experiment in Brazil," December 2025. Working paper. Pre-registered as AEARCTR-0010400.

Friedrich, Mary Jane, "Depression Is the Leading Cause of Disability Around the World," *JAMA*, April 2017, *317* (15), 1517.

Gauthier, Germain, Roland Hodler, Philine Widmer, and Ekaterina Zhuravskaya, "The political effects of X's feed algorithm," *Nature*, 2026, *652* (8109), 416–423.

Goldman, Nina, Melek Alemdar, Herlind Megges, Naka Matsumoto, Eric Schoenmakers, Pauline van den Berg, Mathias Lasgaard, Julie Christiansen, Niina Junttila, Andreas Goldman, Debora Draxl, Austen El-Osta, and Pamela Qualter, "National policy responses to address loneliness: A global scoping review of 194 WHO member states," *Health Policy*, 2026, *165*, 105553.

Graham, Jesse, Jonathan Haidt, and Brian A. Nosek, "Liberals and Conservatives Rely on Different Sets of Moral Foundations," *Journal of Personality and Social Psychology*, 2009, *96* (5), 1029–1046.

Guess, Andrew M., Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taka Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow et al., "How do social media feed algorithms affect attitudes and behavior in an election campaign?," *Science*, 2023, *381* (6656), 398–404.

Guess, Andrew M, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow et al., "Reshares on social media amplify political news but do not detectably affect beliefs or opinions," *Science*, 2023, *381* (6656), 404–408.

Hughes, Mary Elizabeth, Linda J Waite, Louise C Hawkley, and John T Cacioppo, "A short scale for measuring loneliness in large surveys: Results from two population-based studies," *Research on aging*, 2004, *26* (6), 655–672.

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes, "Affect, Not Ideology: A Social Identity Perspective on Polarization," *Public Opinion Quarterly*, 2012, *76* (3), 405–431.

Jiménez-Durán, Rafael, "The economics of content moderation: Theory and experimental evidence from hate speech on Twitter," *George J. Stigler Center for the Study of the Economy & the State Working Paper*, 2023, (324).

Jin, Shan, Tiancheng Liu, Zepeng Sun, and Xiaomeng Zhang, "Social isolation induced negative emotions affect economic behavior: A natural experiment study," *Journal of Development Economics*, 2026, *179*, 103677.

Kalmoe, Nathan P. and Lilliana Mason, *Radical American Partisanship: Mapping Violent Hostility, Its Causes, and the Consequences for Democracy*, Chicago: University of Chicago Press, 2022.

Kim, Jae Woo, Andrew Guess, Brendan Nyhan, and Jason Reifler, "The Distorting Prism of Social Media: How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity," *Journal of Communication*, 2021, *71* (6), 922–946.

Kominers, Scott Duke and Jesse M. Shapiro, "Content Moderation with Opaque Policies," Technical Report, National Bureau of Economic Research 2024.

Kroenke, Kurt, Tara W. Strine, Robert L. Spitzer, Janet B. W. Williams, Joyce T. Berry, and Ali H. Mokdad, "The PHQ-8 as a Measure of Current Depression in the General Population," *Journal of Affective Disorders*, 2009, *114* (1-3), 163–173.

Kteily, Nour, Emile Bruneau, Adam Waytz, and Sarah Cotterill, "The Ascent of Man: Theoretical and Empirical Evidence for Blatant Dehumanization," *Journal of Personality and Social Psychology*, 2015, *109* (5), 901–931.

Levy, Ro'ee, "Social media, news consumption, and polarization: Evidence from a field experiment," *American economic review*, 2021, *111* (3), 831–870.

Madio, Leonardo and Martin Quinn, "Content Moderation and Advertising in Social Media Platforms," *Journal of Economics & Management Strategy*, 2024.

Mandile, Simona, "The dark side of social media: Recommender algorithms and mental health," Technical Report, CESifo Working Paper 2025.

Milli, Smitha, 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*, 2025, *4* (3), pgaf062.

Motte, Elliot, "Insult Politics in the Age of Social Media," 2026.

Müller, Karsten and Carlo Schwarz, "Fanning the flames of hate: Social media and hate crime," *Journal of the European Economic Association*, 2021, *19* (4), 2131–2167.

Muralikumar, Meena Devii, Yun Shan Yang, and David W. McDonald, "A Human-centered Evaluation of a Toxicity Detection API: Testing Transferability and Unpacking Latent Attributes," *ACM Transactions on Social Computing*, 2023, *6* (1–2), 1–38.

Musacchio Schafer, Katherine, Jacob Franklin, Peter J. Embí, and Colin G. Walsh, "Loneliness, Anxiety Symptoms, Depressive Symptoms, and Suicidal Ideation in the All of Us Dataset," *JAMA Network Open*, 2026, *9* (3), e260596.

Noelle-Neumann, Elisabeth, "The Spiral of Silence: A Theory of Public Opinion," *Journal of Communication*, 1974, *24* (2), 43–51.

Nyhan, Brendan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y. Chen, Hunt Allcott, Taka Brown, Adriana Crespo-Tenorio, Drew Dimmery et al., "Like-minded sources on Facebook are prevalent but not polarizing," *Nature*, 2023, *620* (7972), 137–144.

OECD, *Health at a Glance 2019: OECD Indicators*, Paris: OECD Publishing, 2019.

_ , *A New Benchmark for Mental Health Systems: Tackling the Social and Economic Costs of Mental Ill-Health* OECD Health Policy Studies, Paris: OECD Publishing, 2021.

Pascoe, Elizabeth A. and Laura Smart Richman, "Perceived Discrimination and Health: A Meta-Analytic Review," *Psychological Bulletin*, 2009, *135* (4), 531–554.

Piccardi, Tiziano, Martin Saveski, Chenyan Jia, Jeffrey Hancock, Jeanne L Tsai, and Michael S Bernstein, "Reranking partisan animosity in algorithmic social media feeds alters affective polarization," *Science*, 2025, *390* (6776), eadu5584.

Rathje, Steve, Jay J. Van Bavel, and Sander Van der Linden, "Out-Group Animosity Drives Engagement on Social Media," *Proceedings of the National Academy of Sciences*, 2021, *118* (26), e2024292118.

Schmer-Galunder, Sonja, Ruta Wheelock, Zaria Jalan, Alyssa Chvasta, Scott Friedman, and Emily Saltz, "Annotator in the Loop: A Case Study of In-Depth Rater Engagement to Create a Prosocial Benchmark Dataset," in "Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)," Vol. 7 2024, pp. 1319–1328.

Shreekumar, Advik and Pierre-Luc Vautrey, "Managing Emotions: The Effects of Online Mindfulness Meditation on Mental Health and Economic Behavior," June 2024. Manuscript, Massachusetts Institute of Technology. Version: June 14, 2024.

Spitzer, Robert L., Kurt Kroenke, Janet B. W. Williams, and Bernd Löwe, "A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7," *Archives of Internal Medicine*, 2006, *166* (10), 1092–1097.

Steele, Claude M., "The Psychology of Self-Affirmation: Sustaining the Integrity of the Self," in Leonard Berkowitz, ed., *Advances in Experimental Social Psychology*, Vol. 21, New York: Academic Press, 1988, pp. 261–302.

Stray, Jonathan, Ian Baker, George Beknazar-Yuzbashev, Ceren Budak, Julia Kamin, Kylan Rutherford, Mateusz Stalinski, Tin Acosta, Chris Bail, Michael Bernstein et al., "The prosocial ranking challenge: reducing polarization on social media without sacrificing engagement," *arXiv preprint arXiv:2603.19626*, 2026.

Tappin, Ben M. and Ryan T. McKay, "The Illusion of Moral Superiority," *Social Psychological and Personality Science*, 2017, *8* (6), 623–631.

Taylor, Shelley E. and Jonathon D. Brown, "Illusion and Well-Being: A Social Psychological Perspective on Mental Health," *Psychological Bulletin*, 1988, *103* (2), 193–210.

Valkenburg, Patti M. and Jochen Peter, "Social Consequences of the Internet for Adolescents: A Decade of Research," *Current Directions in Psychological Science*, 2009, *18* (1), 1–5.

Wills, Thomas Ashby, "Downward Comparison Principles in Social Psychology," *Psychological Bulletin*, 1981, *90* (2), 245–271.

Wulczyn, Ellery, Nithum Thain, and Lucas Dixon, "Ex Machina: Personal Attacks Seen at Scale," in "Proceedings of the 26th International Conference on World Wide Web (WWW '17)" ACM 2017, pp. 1391–1399.

Yu, Xiao, Muhammad Haroon, Eszter Hargittai Menchen-Trevino, and Magdalena Wojcieszak, "Nudging recommendation algorithms increases news consumption and diversity on YouTube," *PNAS Nexus*, 2024, *3* (12), pgae518.

Online Appendix

Social Media Toxicity and Mental Health

The most recent version is available here.

Contents

A Data, Outcomes, and Estimands **1**
A.1 Sample Composition and Pre-Treatment Mental Health 1

B Main Results **1**
B.1 Main Effects on Mental Health 1
B.2 Treatment Intensity and Exposure-Scaled Effects 4
B.3 Heterogeneous Treatment Effects 4

C Mechanism Evidence **8**
C.1 Secondary Outcomes 8
C.2 Exposure-Scaled Effects on Secondary Outcomes 8
C.3 Political Indices and Moral Disengagement 8

D Robustness and Alternative Interpretations **10**
D.1 Substitution 10
D.2 Attrition 11

E Feed-Based Exposure Measures: Construction, Balance, and Robustness **13**
E.1 Author Classification and the Peer Share 13
E.1.1 Author-composition stability 14
E.2 Arm Balance of the Measures and Their Coverage Gates 16
E.3 Robustness of the Peer-Share Construction 18
E.4 Choice of the Early-Measure Criteria 19
E.5 Robustness to the Alternative Criteria 20

F Usage Intensity: Shown Feed-Post Volume **21**
F.1 Construction and Choice of Criteria 21
F.2 Results 22

G Welfare Translation: Mapping, Aggregation, and Sensitivity **25**
G.1 External mapping 25
G.2 Aggregation 26
G.3 Uncertainty and sensitivity 26
G.4 Dollar benchmarks 27

H Content Classification **28**
H.1 Political content among toxic posts 28
H.2 Comparing the toxicity and alienation classifiers 29

I Preregistration Clarifications and Deviations **31**

J Instructions **35**

A Data, Outcomes, and Estimands

A.1 Sample Composition and Pre-Treatment Mental Health

This subsection supplements the main paper's "Data, Outcomes, and Estimands" section. Table A.1 reports sample composition and pre-treatment covariate balance across the four randomized treatment arms among all pre-treatment-survey respondents. None of the fourteen covariates reported in the table indicates imbalance, which is consistent with a successful randomization of participants into treatment groups.

Figure A.1 describes pre-treatment mental-health severity in the analysis sample used for the six-week ITT estimates.

Table A.1: Sample Composition and Balance by Treatment Arm

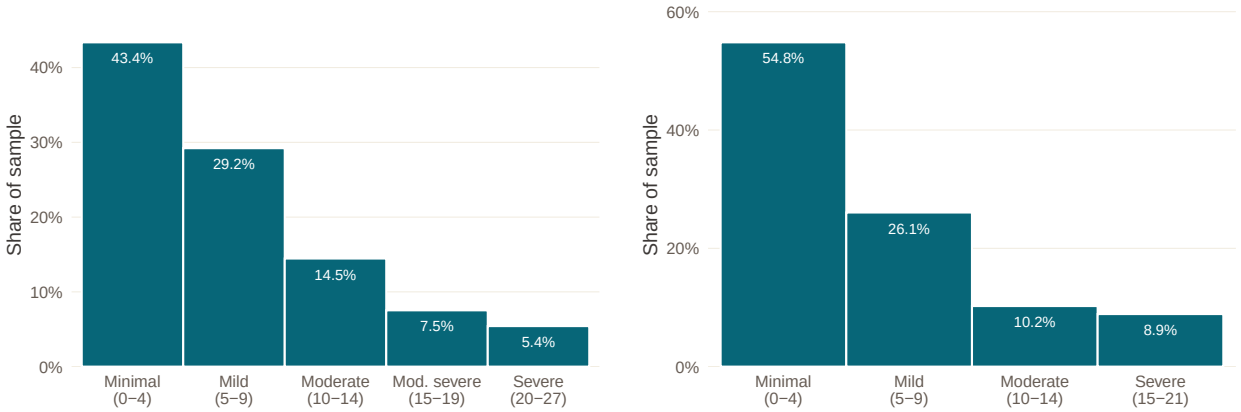
	Control (N=257)		Random (N=283)		Toxicity (N=390)		Alienation (N=464)		
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	F-test p
Age (years)	43.99	13.32	43.49	14.83	42.13	13.16	42.72	13.88	0.336
Male	0.44	0.50	0.42	0.50	0.47	0.50	0.48	0.50	0.452
White	0.67	0.47	0.65	0.48	0.68	0.47	0.69	0.46	0.692
Bachelor's degree or higher	0.51	0.50	0.52	0.50	0.47	0.50	0.48	0.50	0.561
Democrat	0.44	0.50	0.46	0.50	0.48	0.50	0.42	0.49	0.295
Republican	0.23	0.42	0.21	0.41	0.20	0.40	0.20	0.40	0.698
Independent	0.26	0.44	0.24	0.43	0.24	0.43	0.27	0.44	0.781
Recruited via Cloud Research	0.43	0.50	0.41	0.49	0.45	0.50	0.42	0.49	0.840
Daily social-media use (1–7 scale)	4.04	1.90	4.13	1.87	4.01	1.80	4.08	1.93	0.829
Desktop share, Facebook	0.59	0.28	0.59	0.28	0.61	0.27	0.62	0.27	0.643
Desktop share, Reddit	0.65	0.29	0.61	0.30	0.64	0.29	0.66	0.28	0.326
Desktop share, Twitter/X	0.58	0.32	0.58	0.32	0.54	0.31	0.59	0.30	0.180
PHQ-8 depression score	6.03	5.76	6.64	5.85	6.38	5.73	6.03	5.60	0.463
GAD-7 anxiety score	5.53	5.69	6.06	5.96	5.50	5.41	5.51	5.59	0.546

Notes: Sample is all randomized users who completed the pre-treatment survey (N = 1,394), before conditioning on post-treatment-survey completion or extension activity. Assignment was performed client-side by the extension at enrollment via simple (independent) randomization – each participant's arm is an independent draw rather than a balanced or quota-based allocation – so realized arm sizes are themselves random and are not constrained to be equal across arms. Identification requires only that assignment be orthogonal to participants' potential outcomes, which the covariate balance reported here supports; the unequal realized counts across arms carry no implication for it. The two active-classifier arms (Toxicity and Alienation) are pooled into a single treatment indicator in all headline specifications, so the split between them does not enter the intention-to-treat estimand. Cells report arm-specific means and standard deviations. For indicator rows, means are shares. The PHQ-8 depression score is the observed depression sum (range 0–24); the GAD-7 anxiety score is the observed anxiety sum (range 0–21). Daily social-media use is a self-reported ordinal scale (1 = under 30 min/day to 7 = over 4 h/day). Desktop platform shares are self-reported sliders on the 0–1 scale. The final column reports the p-value of an F-test of joint equality of the four arm means (one-way regression of each covariate on the arm indicators); no covariate is imbalanced across arms. The table is included as a pre-treatment randomization check; the identifying variation is random assignment within the experimental design.

B Main Results

B.1 Main Effects on Mental Health

This subsection reports the supporting estimates behind the main paper's "Main Effects on Mental Health" subsection. Table A.2 reports the full main-ITT coefficient table for



(A) PHQ-9 PROXY (B) GAD-7

FIGURE A.1: PRE-TREATMENT MENTAL HEALTH IN THE ANALYSIS SAMPLE

Notes: Left panel: PHQ-9 severity buckets applied to the rescaled PHQ-9 proxy, computed as PHQ-8 $\times 9/8$ (range 0–27). Right panel: GAD-7 severity buckets on the pre-treatment GAD-7 sum. Bars show the share of the pooled analysis sample in each bucket.

the standardized mental-health-symptoms index, the anxiety and depression subindices, and the raw GAD-7 and PHQ-8 scales. Table A.3 reports the threshold-crossing specifications, including the raw PHQ-8 ≥ 10 rule and the PHQ-9-proxy cutoff used in the main text. Figure A.2 decomposes the GAD-7 and PHQ-8 estimates into item-level coefficients.

TABLE A.2: MAIN DID COEFFICIENTS ON MENTAL-HEALTH OUTCOMES

	MentalSymptoms	Anxiety	Depression	GAD	PHQ
Treated	0.151***	0.130***	0.156***	0.742***	0.884***
	(0.051)	(0.050)	(0.058)	(0.284)	(0.308)
Control mean	0.000	0.000	0.000	5.648	6.111
Control SD	1.000	1.000	1.000	5.656	5.354
Observations	1312	1312	1312	1312	1312

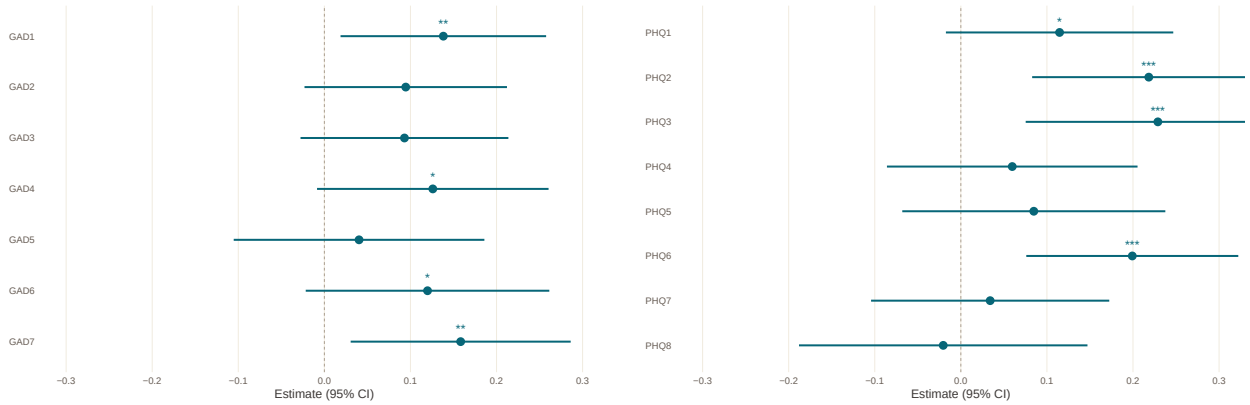
Notes: This table reports difference-in-differences estimates of the effect of the intervention on mental-health outcomes, obtained from the main intention-to-treat (ITT) specification described in the main paper. Higher values of all outcomes indicate worse mental health. The first three outcomes are expressed in control-group standard-deviation units, while the GAD-7 and PHQ-8 outcomes are reported on their raw symptom scales. All specifications include respondent fixed effects and cohort-by-time fixed effects. The estimation sample is 656 of the 664 analysis-sample participants; the eight participants alone in their start-date cohort are absorbed by the cohort-by-time fixed effects, leaving the coefficients unchanged. Standard errors are clustered at the respondent level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Our experimental infrastructure served two separately preregistered studies drawing on the same user pool and survey instrument, which raises the question of how outcomes should be partitioned into testing families. The companion study of political attitudes (AEARCTR-0013997) specifies 4 primary outcomes, which are outside this paper's estimand. Even if they are grouped with the mental-health-symptoms endpoint, the main ITT inference is unchanged: with 0.0030 as the unadjusted p-value, the largest possible Benjamini–Hochberg q-value in the resulting 5-outcome primary family is 0.0149.

Table A.3: Treatment Effects on Moderate-Symptom Threshold Crossings

	PHQ-8 ≥ 10	PHQ-9 proxy ≥ 10	GAD-7 ≥ 10	Either ≥ 10 (PHQ-8)	Either ≥ 10 (PHQ-9 proxy)
Treated	0.068*	0.100***	0.049*	0.093***	0.121***
	(0.035)	(0.035)	(0.029)	(0.034)	(0.035)
Control mean	0.237	0.281	0.231	0.298	0.320
Control SD	0.426	0.450	0.422	0.458	0.467
Observations	1312	1312	1312	1312	1312

Notes: This table reports linear-probability difference-in-differences estimates of the intention-to-treat (ITT) effect of the intervention on the probability of scoring at or above the moderate-symptom cutoff at the six-week post-treatment survey. Estimates and standard errors are reported on the 0–1 probability scale; multiply by 100 to interpret them as percentage points. The PHQ-9 proxy rescales the observed PHQ-8 sum by 9/8; for integer PHQ-8 scores, the condition $(9/8) \cdot \text{PHQ-8} \geq 10$ reduces to PHQ-8 ≥ 9. The specification uses the same fixed effects and clustering as Table A.2, and, as there, on 656 of the 664 participants (eight alone in their start-date cohort are absorbed by the cohort-by-time fixed effects). *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.



(A) GAD-7 items (B) PHQ-8 items

Figure A.2: DiD Coefficients on the GAD-7 and PHQ-8 Components

Notes: This figure reports item-level intention-to-treat (ITT) estimates of the effect of the intervention on the individual GAD-7 (Panel (a)) and PHQ-8 (Panel (b)) component items, obtained from the same difference-in-differences specification as Table A.2. Points denote item-level ITT coefficients, and bars denote 95% confidence intervals based on standard errors clustered at the respondent level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

An intentionally broader calculation pools the mental-health-symptoms endpoint, political-attitude outcomes from the common survey instrument, and the nine secondary outcomes in Table A.11, giving 16 hypotheses. In that expanded family, loneliness and mental-health symptoms have Benjamini–Hochberg q-values of 0.0238. The worst case for these two estimates after adding further outcomes is that all added outcomes have larger p-values, so their ranks are unchanged while the family size grows. Under that worst case, adding 17 outcomes leaves the q-value at 0.0491; adding 18 raises it to 0.0506. Thus both estimates survive BH correction at the 5% level for any expanded family that adds at most 17 outcomes to this already broad family, regardless of the added outcomes' realized p-values.

B.2 Treatment Intensity and Exposure-Scaled Effects

This subsection supplements the main paper's "Treatment Intensity and Exposure-Scaled Effects" subsection. Table A.4 reports the arm-level exposure measures over the 42-day treatment window. Table A.5 reports the corresponding first-stage coefficients for shown toxic share (exposure to toxic content experienced by the user after any filtering) and toxic share supplied (exposure to toxic content that the platform intended for the user to have prior to any filtering). Table A.6 reports the IV exposure-response coefficients before the main-text rescaling to a 10 percentage-point reduction in shown toxic share.

Table A.4: Treatment Intensity by Treatment Arm

	Pooled comparison		Arm-level breakdown			
	Treatment	Control	Alienation	Toxicity	Pure control	Random
Shown toxic share	0.05	7.75	0.03	0.07	7.63	7.86
Hidden toxic share	98.91	3.81	99.34	98.44	0.00	7.18
Overall content hidden	8.31	3.18	8.99	7.58	0.00	6.04

Notes: This table reports user-level means of the treatment-intensity measures over the 42-day treatment window, separately by treatment arm. The shown toxic share is shown toxic content divided by shown content; the hidden toxic share is hidden toxic content divided by total toxic content; and overall content hidden is hidden content divided by total content, averaged over a user's active days. All shares are reported in percent.

Table A.5: First-Stage on Realized and Supplied Toxic Exposure

	Shown toxic share	Supplied toxic share
Treated	−0.078***	0.003
	(0.003)	(0.004)
Observations	1312	1312

Notes: This table reports first-stage estimates of the effect of treatment assignment on realized and supplied toxic exposure. The columns report the coefficient on the treatment indicator for the shown toxic share and the supplied toxic share, respectively. The shown-share column measures realized toxic exposure after hiding, while the supplied-share column measures the toxic share of content supplied before hiding. Standard errors are clustered at the respondent level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

B.3 Heterogeneous Treatment Effects

This subsection supplements the main paper's "Heterogeneous Treatment Effects" subsection. Table A.7 and Figure A.3 report the preregistered median split by pre-treatment mental-health symptoms. Figure A.4 shows the corresponding exploratory quartile estimates using the raw GAD-7 and PHQ-8 pre-treatment scales. Figure A.5 reports additional subgroup splits. Table A.8 reports heterogeneity by the early platform-served toxic share, entered continuously across observed-day windows.

Table A.6: IV Exposure-Response Coefficients Using Shown Toxic Share

	MentalSymptoms	Anxiety	Depression	GAD	PHQ
Shown toxic share	−1.939***	−1.669**	−1.993***	−9.501***	−11.319***
	(0.654)	(0.647)	(0.748)	(3.647)	(3.973)
Control mean	0.000	0.000	0.000	5.648	6.111
Control SD	1.000	1.000	1.000	5.656	5.354
Observations	1312	1312	1312	1312	1312

Notes: The endogenous regressor is shown toxic share. MentalSymptoms, Anxiety, and Depression are in control-SD units; GAD-7 and PHQ-8 are raw symptom scales. Higher values of all outcomes indicate worse mental health. The table keeps the endogenous regressor on the share scale; the main text reports the same object rescaled to a 10 percentage-point reduction in shown toxic share. The estimation sample is 656 of the 664 analysis-sample participants; the eight participants alone in their start-date cohort are absorbed by the cohort-by-time fixed effects. Standard errors are clustered at the respondent level.

Table A.7: Heterogeneity by Pre-Treatment Mental-Health Symptoms

	MentalSymptoms	Anxiety	Depression
Treated	0.089	0.089	0.079
	(0.056)	(0.061)	(0.058)
Treated × BaselineSymptoms high	0.116	0.062	0.158
	(0.102)	(0.100)	(0.117)
Control mean	0.000	0.000	0.000
Control SD	1.000	1.000	1.000
Observations	1296	1296	1296

Notes: This table reports estimates from a difference-in-differences specification that allows the intention-to-treat (ITT) effect to vary by pre-treatment mental-health symptoms. The *Treated* row reports the ITT coefficient for respondents below the median of pre-treatment mental-health symptoms, and the *Treated* × *BaselineSymptoms high* row reports the coefficient on the treatment indicator interacted with being above that median, so the above-median ITT is the sum of the two rows. The mental-health-symptoms (MentalSymptoms), anxiety (Anxiety), and depression (Depression) indices are each expressed in control-group standard-deviation units; higher values indicate worse mental health. The cohort-by-time fixed effects are stratified by pre-treatment symptoms, so the estimation sample is 648 of the 664 analysis-sample participants; the sixteen participants alone in their cohort-by-time-by-stratum cell are absorbed by these fixed effects. Standard errors are clustered at the respondent level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

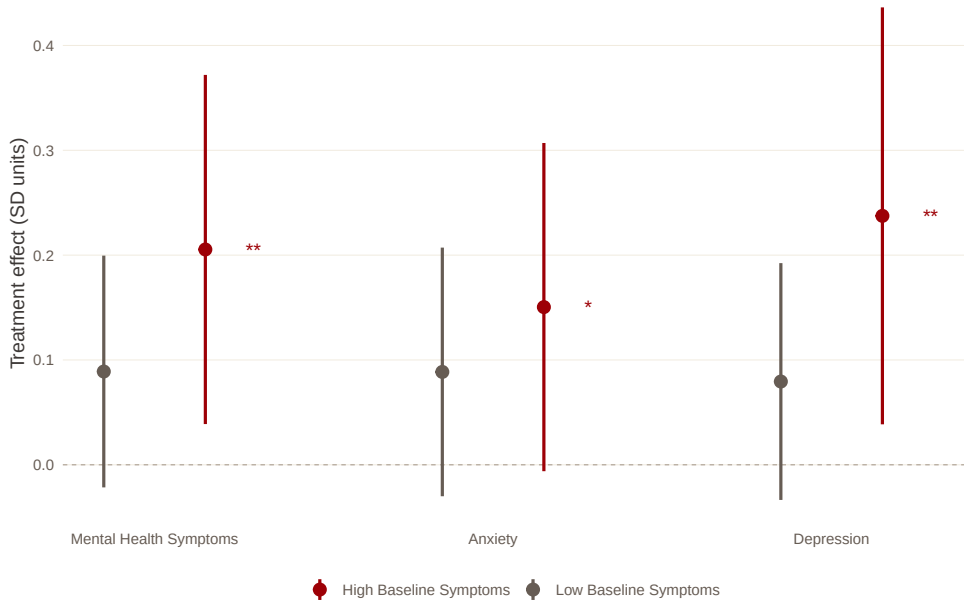
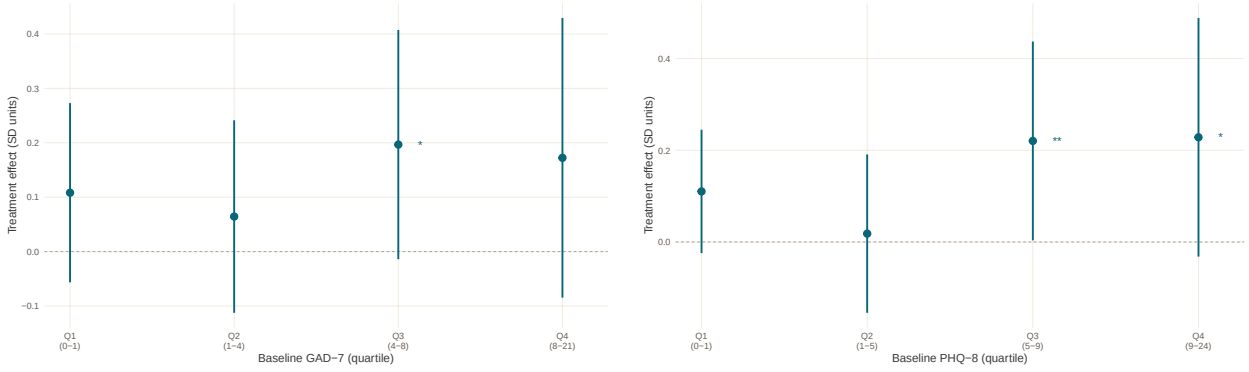


FIGURE A.3: HETEROGENEITY BY PRE-TREATMENT MENTAL-HEALTH SYMPTOMS

Notes: This figure reports estimates from the difference-in-differences specification of Table A.7, which allows the intention-to-treat (ITT) effect to vary by pre-treatment mental-health symptoms. Points denote subgroup-specific ITT coefficients: the low-symptoms point is the *Treated* coefficient (the ITT for respondents below the median of pre-treatment mental-health symptoms), and the high-symptoms point is the sum of the *Treated* and *Treated* × *BaselineSymptoms high* coefficients, with its standard error computed for that linear combination from the respondent-clustered covariance matrix. Outcomes are expressed in control-group standard-deviation units. Standard errors are clustered at the respondent level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.



(A) PRE-TREATMENT GAD-7 QUARTILE (B) PRE-TREATMENT PHQ-8 QUARTILE

FIGURE A.4: BIN-SPECIFIC ITT COEFFICIENTS ACROSS PRE-TREATMENT GAD-7 AND PHQ-8 QUARTILES

Notes: This figure reports quartile-specific intention-to-treat (ITT) estimates of the effect of the intervention on the mental-health-symptoms index. Points denote quartile-specific ITT coefficients, where quartiles are defined separately using pre-treatment GAD-7 (Panel (a)) and pre-treatment PHQ-8 (Panel (b)). Bars denote 95% confidence intervals based on standard errors clustered at the respondent level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

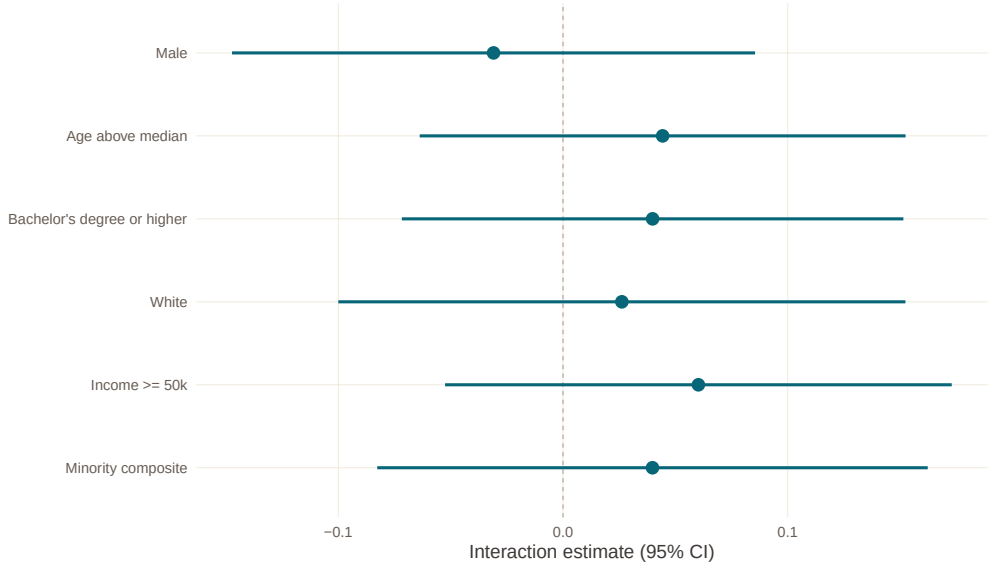


FIGURE A.5: DEMOGRAPHIC HETEROGENEITY

Notes: This figure reports subgroup-specific ITT estimates of the effect of the intervention on the mental-health-symptoms index across demographic and media-use subgroups. Each subgroup is estimated from a separate difference-in-differences specification with respondent and cohort-time fixed effects in which the subgroup indicator – sex, age above the sample median, education (bachelor's degree or higher), ethnicity (white), income (at least $50k), and minority status – enters only through its interaction with treatment; the plotted point is the *Treated* × *subgroup* coefficient, that is, the differential ITT for members of the subgroup. Bars denote 95% confidence intervals based on standard errors clustered at the respondent level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

TABLE A.8: HETEROGENEITY BY EARLY PLATFORM-SERVED TOXIC SHARE

	5 observed days	3 observed days	1 observed day
Treated	0.163***	0.165***	0.162***
	(0.056)	(0.056)	(0.056)
Treated × share toxic	−0.009	−0.018	−0.020
	(0.029)	(0.029)	(0.037)
Observations	1106	1106	1106

Notes: This table reports heterogeneous treatment effects on the mental-health-symptoms index by the share of early platform-served posts that are toxic, entered as a continuous interaction rather than a median split. The dose is the share of platform-served posts (excluding comments and ads) that exceed the toxicity threshold, measured from the first five elements of each browsing session and averaged over the respondent's first five, three, and one observed active days (columns, left to right), and standardized in the control sample, so the *Treated* × *share toxic* coefficient is the change in the ITT per one-standard-deviation increase in the early toxic share. Because the dose is time-invariant, its level is absorbed by the respondent fixed effect, and only the interaction is identified. Each column reports a separate difference-in-differences specification with respondent and cohort-time fixed effects. The mental-health-symptoms index is expressed in control-group standard-deviation units. Standard errors are clustered at the respondent level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

C Mechanism Evidence

C.1 Secondary Outcomes

Table A.9 reports ITT estimates for the secondary survey outcomes. Table A.10 gives recruitment-wave estimates for loneliness and their inverse-variance-weighted average. Tables A.11 and A.12 report multiple-testing adjustments and demographic subgroup splits.

Table A.9: ITT coefficients on secondary outcomes

	MechLoneliness	MechAnger	MechFomo	Discrimination	DiscriminationRelevant	MoralSuperiority	RelativeMoralSuperiority	PerceptionOfAltruism	FollowToxic
Estimated effect	0.342***	−0.030	0.072	0.020	0.031	0.064	−0.188**	−0.016	0.213
	(0.112)	(0.135)	(0.154)	(0.068)	(0.068)	(0.058)	(0.090)	(0.090)	(0.155)
Observations	342	342	342	1312	1312	1312	1312	1312	699

Notes: This table reports the average ITT coefficients for the secondary outcomes shown in the main text, standardized in control-group standard-deviation units. The final column (*FollowToxic*) reports the toxic-list follow-choice item-count experiment: the reported coefficient is the *Treated* × toxic-offer interaction, measured on the raw account-count scale, and is discussed together with the substitution evidence in Section D.1. The loneliness, anger, and FOMO items were added by the October 2024 pre-registration amendment and fielded at the six-week survey only in the second recruitment wave, so they are estimated on fewer observations (per-column N). Standard errors are clustered at the respondent level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.10: Pooled loneliness coefficient across recruitment waves

	Full-panel DiD	Wave 1 Cross-Section	IVW Pooled
Treated	0.310***	0.072	0.212***
	(0.098)	(0.118)	(0.076)
Observations	485	313	798

Notes: The first two columns report ITT coefficients for standardized loneliness in the two recruitment waves, estimated on the sample of follow-up sample (respondents with a pre-treatment survey and either the post-treatment survey or the follow-up). Wave 2 uses the full-panel difference-in-differences specification over the pre-treatment, six-week, and twelve-week surveys; Wave 1 uses the twelve-week follow-up cross-section with start-date-cohort fixed effects. The Wave 2 column's larger observation count than the six-week estimate in Table A.9 reflects both the added twelve-week survey and this broader sample. The final column gives their inverse-variance-weighted average. A test of equality of the two wave-specific ITT coefficients yields $p = 0.121$. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

C.2 Exposure-Scaled Effects on Secondary Outcomes

Table A.13 expresses the secondary-outcome reduced forms per 10 percentage-point reduction in shown toxic share. Treatment assignment instruments shown toxic share, as in Appendix B.2.

C.3 Political Indices and Moral Disengagement

Table A.14 compares heterogeneity by five pre-treatment political indices. Only the coefficient on treatment assignment interacted with the MDI is statistically distinguishable from

Table A.11: Benjamini–Hochberg q-values across secondary outcomes

Outcome	Estimate	p-value	BH q-value
Loneliness	0.342	0.003	0.023
Anger	-0.030	0.824	0.857
Social-media FOMO	0.072	0.640	0.857
Perceived discrimination	0.020	0.769	0.857
Relevant discrimination	0.031	0.647	0.857
Social desirability	0.064	0.269	0.606
Positive self-view	-0.188	0.037	0.167
Perceived altruism	-0.016	0.857	0.857
Toxic follow choice	0.213	0.171	0.512

Notes: Benjamini–Hochberg q-values for the secondary-outcome family. The survey coefficients match Table A.9, and the follow-choice coefficient is treatment assignment interacted with a toxic offer in the item-count experiment.

Table A.12: Demographic subgroup splits for secondary outcomes

Subgroup split (Treated × …)	Loneliness	Anger	Social-media FOMO	Perceived discrimination	Relevant discrimination	Perceived altruism	Social desirability	Positive self-view
Sex (Male)	-0.236* (0.130)	0.168 (0.136)	0.006 (0.190)	-0.050 (0.082)	-0.007 (0.084)	-0.008 (0.105)	0.019 (0.067)	-0.061 (0.099)
Age (above median)	-0.054 (0.109)	-0.157 (0.190)	0.071 (0.224)	0.084 (0.090)	0.114 (0.089)	0.093 (0.116)	-0.002 (0.070)	-0.016 (0.107)
Education (BA+)	-0.041 (0.125)	0.063 (0.147)	0.130 (0.180)	-0.018 (0.078)	0.035 (0.082)	0.062 (0.105)	0.090 (0.065)	0.025 (0.099)
Ethnicity (White)	-0.027 (0.141)	-0.006 (0.135)	-0.149 (0.212)	-0.002 (0.087)	-0.022 (0.089)	0.036 (0.110)	-0.104 (0.078)	0.006 (0.112)
Income (≥ 50k)	-0.315** (0.132)	0.322** (0.144)	-0.083 (0.184)	-0.030 (0.081)	-0.070 (0.085)	0.162 (0.105)	-0.066 (0.068)	0.076 (0.107)
Baseline SM use (high)	0.160 (0.125)	0.233* (0.139)	-0.241 (0.178)	0.000 (0.078)	0.067 (0.081)	-0.156 (0.104)	0.097 (0.065)	0.026 (0.101)
Minority (!Male — !White)	0.240* (0.135)	0.002 (0.140)	0.099 (0.206)	0.140* (0.083)	0.118 (0.089)	0.060 (0.112)	0.012 (0.067)	0.020 (0.108)

Notes: This table reports estimates of the effect of the intervention interacted with demographic subgroups. Each cell reports the coefficient on the treatment indicator interacted with the indicated pre-treatment subgroup split. The outcomes are the secondary outcomes reported in the main text, excluding follow choice. Standard errors are clustered at the respondent level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.13: Effects of a 10 percentage-point reduction in shown toxic share on secondary outcomes

	MechLoneliness	MechAnger	MechFomo	Discrimination	DiscriminationRelevant	MoralSuperiority	RelativeMoralSuperiority	PerceptionOfAltruism
10 pp reduction in shown toxic share	0.472***	−0.041	0.100	0.025	0.040	0.082	−0.240**	−0.021
	(0.166)	(0.187)	(0.213)	(0.087)	(0.088)	(0.075)	(0.116)	(0.115)
Control mean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Control SD	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Observations	342	342	342	1312	1312	1312	1312	1312

Notes: This table reports instrumental-variables estimates of the exposure-response relationship for the secondary survey outcomes, where shown toxic share is instrumented with treatment assignment; the first-stage specification is the one reported in Appendix B.2. Coefficients are scaled and sign-normalized to report the implied change from a 10 percentage-point reduction in shown toxic share. Outcomes are standardized in control-group standard-deviation units. The estimates rescale the intention-to-treat coefficients in Table A.9. Standard errors are clustered at the respondent level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

zero. Table A.15 decomposes the incremental explanatory power of the MDI-item interactions, suggesting that the effect is not driven by a single item.

Table A.14: Heterogeneity by pre-treatment political indices

	MDI	Dehumanization	Polarization	Violence support	Spiral of silence
Interaction	0.090***	0.048	0.028	−0.012	−0.018
	(0.035)	(0.037)	(0.032)	(0.038)	(0.031)
Control mean	0.000	0.000	0.000	0.000	0.004
Control SD	1.000	1.000	1.000	1.000	1.000

Notes: Each column reports the coefficient on treatment assignment interacted with one pre-treatment political index, standardized in the control sample, so every coefficient is expressed per one-standard-deviation (+1 SD) increase. All specifications include baseline-symptom controls: the treatment indicator interacted with a high-pre-treatment-symptoms indicator, with the cohort-by-time fixed effect stratified on that indicator. The outcome is expressed in control-group standard-deviation units. Standard errors are clustered at the respondent level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.15: Shapley decomposition of MDI-item incremental partial R^2

Item	Share	Shapley partial R2
Out-party are a threat (MDI1)	39.3%	0.0055
Out-party are evil (MDI2)	39.1%	0.0055
Out-party lack humanity (MDI3)	21.6%	0.0030

Notes: Shapley shares allocate the joint incremental partial R^2 of the MDI-item interactions across the three component items. Their joint incremental partial R^2 is 0.0141.

D Robustness and Alternative Interpretations

D.1 Substitution

A natural alternative interpretation of our results is that toxicity filtering does not reduce users' exposure to toxic social-media environments, but instead displaces that exposure elsewhere. The concern has three versions. First, because the extension operates on desktop browsers, treated users could move consumption of the same platforms to unfiltered mobile devices. Second, if toxicity is hidden on the treated platforms, users could compensate by spending more time on more-toxic websites outside the intervention. Third, treated users could actively seek toxic accounts when given an opportunity to do so. The evidence does not support these substitution channels.

Table A.16 tests the mobile version directly. The outcome is the self-reported desktop share of platform use, measured before and after treatment for Facebook, Reddit, Twitter, and the user's average across platforms. A negative coefficient would indicate substitution

from desktop to mobile use. The estimates are close to zero across all four columns, including the pooled platform measure, so treated users do not appear to move their use to unfiltered mobile devices.

TABLE A.16: DESKTOP SHARE OF PLATFORM USE BY TREATMENT ASSIGNMENT

	Facebook	Reddit	Twitter	All platforms
Treated	0.001	−0.015	0.031	0.000
	(0.025)	(0.025)	(0.033)	(0.019)
Control mean	0.567	0.622	0.602	0.602
Control SD	0.306	0.309	0.314	0.257
Observations	1136	1028	866	1300

Notes: This table reports estimates of the effect of the intervention on the self-reported desktop share of platform use. The specifications are intention-to-treat (ITT) difference-in-differences regressions with respondent and cohort-by-time fixed effects; standard errors are clustered at the respondent level. The columns cover the pre-treatment and post-treatment surveys. The "All platforms" column averages the per-platform sliders across the platforms each respondent reports using. A negative coefficient would indicate substitution toward mobile use. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.17 addresses a broader "whack-a-mole" concern using off-platform browsing data from the disjoint sample in Beknazar-Yuzbashev et al. (2025). That study tracked desktop browsing outside the treated platforms, allowing external social-media destinations to be grouped into more-toxic and less-toxic buckets. If hiding toxicity merely pushed users toward toxic corners of the open web, the treatment effect should be larger for the more-toxic bucket. The pattern goes the other way: the increase in external social-media time is concentrated in the less-toxic bucket, while time on more-toxic sites is a precise null. This evidence comes from a separate experiment, so we use it as supporting evidence on a substitution mechanism rather than as part of the present ITT analysis.

Finally, the follow-choice experiment asks whether treated respondents actively seek toxic social-media accounts when offered a toxic option. Respondents were cross-randomized between a neutral list and a list that included a politically toxic account, and the coefficient of interest is the *Treated* × toxic-offer interaction, reported in A.9. Results do not show evidence that treated users systematically seek more-toxic content elsewhere.

Taken together, the device, external-site, and follow-choice checks point against the interpretation that the mental-health effect is driven by compensatory movement toward more-toxic exposure outside the treated desktop feeds.

D.2 Attrition

There are two relevant margins to consider with respect to potential differential attrition. The first is survey non-response: the post-treatment-survey outcome is observed only for respondents who complete that survey. The second is extension drop-off: the headline sample requires at least some recorded extension activity by the post-treatment survey. The concern

Table A.17: Substitution into more- versus less-toxic external sites

	Less toxic	More toxic
Treated	2.407***	−0.385
	(0.746)	(0.303)
Control mean	8.112	2.066
Control SD	40.534	10.330
Observations	33 880	18 088

Notes: This table reports estimates using off-platform browsing data from Beknazar-Yuzbashev et al. (2025), a separate experiment with a disjoint sample. The dependent variable is daily active minutes on a bucket's external social-media destinations over the balanced eight-week panel, with treatment turning on after the two-week baseline on Facebook, Twitter, and YouTube. The two buckets partition the external social-media destinations tracked in that study, classified using GPT-5 model: the more-toxic bucket comprises 4chan, Gab, Parler, Gettr, CloutHub, MeWe, Rumble, Reddit (an external site for that study), VK, Tumblr, Steemit, and Weibo, and the less-toxic bucket comprises LinkedIn, Spotify, Pinterest, Medium, Substack, Blogger, LiveJournal, Vimeo, Meetup, Quora, Twitch, TikTok, Instagram, and Nextdoor; messaging apps and non-social destinations are excluded. Each column reports a separate difference-in-differences regression of these minutes on a treatment indicator, with user and cohort-by-period fixed effects, restricted to users who ever visit that bucket, so the two columns are estimated on different samples. Standard errors are Driscoll–Kraay. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

is therefore differential selection by assignment, especially selection related to pre-treatment symptoms.

Table A.18 studies both margins in the attrition-analysis sample ($N = 1270$). The left block uses post-treatment survey completion as the dependent variable; the right block uses an indicator for remaining active on the extension through post-treatment day 42.

Table A.18: Differential attrition

	Survey completion					Extension to post-treatment survey				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Constant	0.517***	0.551***	0.521***	0.432***	0.399***	0.785***	0.797***	0.814***	0.624***	0.540***
	(0.023)	(0.027)	(0.032)	(0.071)	(0.113)	(0.019)	(0.021)	(0.025)	(0.069)	(0.107)
Treatment	0.009	0.008	0.058	0.044	0.106	−0.026	−0.026	−0.054	−0.043	0.088
	(0.029)	(0.029)	(0.041)	(0.036)	(0.140)	(0.024)	(0.024)	(0.033)	(0.033)	(0.135)
N	1,270	1,270	1,270	1,262	1,262	1,270	1,270	1,270	1,262	1,262
Baseline symptoms		✓	✓	✓	✓		✓	✓	✓	✓
Baseline symptoms × Treated			✓	✓	✓			✓	✓	✓
Demographics				✓	✓				✓	✓
Demographics × Treated					✓					✓
Joint F-stat (Treatment + Treatment×X = 0)			1.521	1.233	0.926			1.337	0.863	0.665
Joint p-value			0.219	0.292	0.508			0.263	0.422	0.757

Notes: This table reports OLS linear-probability estimates of differential attrition, estimated among enrollees with any recorded extension activity after assignment, with heteroskedasticity-robust (HC1) standard errors. The dependent variable is post-treatment survey completion in the left block and an indicator for extension activity through post-treatment day 42 in the right block. Only the constant and the pooled treatment coefficient are displayed. Within each outcome, the columns add controls cumulatively: treatment only; pre-treatment-symptom controls; the pre-treatment-symptom × treatment interaction; demographic controls; and demographic × treatment interactions. Checkmarks indicate the included control blocks. The joint Wald rows test the null hypothesis that the treatment coefficient and all treatment-by-covariate interactions are jointly zero. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

The table shows little evidence of differential attrition. Treatment assignment does not meaningfully predict post-treatment survey completion or continued extension activity. This holds as the specifications progressively add pre-treatment symptoms, allow the treatment

coefficient to vary with symptom severity, and then add demographic controls and treatment-by-demographic interactions. The joint Wald tests in the interaction columns likewise do not reject the differential-attrition null across either outcome.

E Feed-Based Exposure Measures: Construction, Balance, and Robustness

The mechanism analysis conditions on feed-based measures with no pre-treatment analogue: the early platform mix, the early supplied toxic-post share, and the full-window peer share. The two early measures share one scaffold – average the first five qualifying days falling within two weeks of treatment start, requiring at least three, where a qualifying day is a day with any logged platform activity for the platform mix and a day with any supplied main-feed posts for the toxicity measure – and the peer share averages its daily shares over the full six-week logging window, combining each day's Facebook and X shares with weights frozen at the user's early platform-time allocation. Ranking qualifying days rather than calendar days equalizes the amount of behavior sampled per user, so lighter users do not mechanically carry noisier measures. Because the measures and their coverage requirements all condition on post-randomization behavior, the constructions carry testable implications, and this appendix reports them in four parts: the construction of the author classifier and the peer share, together with an external diagnostic showing that toxicity hiding does not move feed-author composition (Appendix E.1); treated–control balance of the measures, of pre-treatment covariates on the estimation sample and of the coverage gates (Appendix E.2); robustness of the main heterogeneity table to each ingredient of the peer-share construction (Appendix E.3); and the outcome-blind diagnostics disciplining the five-day, three-day-gate, two-week criteria, together with robustness to moving them (Appendices E.4 and E.5).

E.1 Author Classification and the Peer Share

The feed-composition measures start from the deidentified post-level substrate used by the analysis pipeline. It contains Facebook and X post impressions only: comments are excluded, Reddit has no author-position input, and the post must have a non-missing author key before deidentification. While the author key is still available, the pipeline computes three author-level quantities within platform: the author's reach across study participants (the number of respondents in whose logged feeds the author appears), median likes from parsed like-count strings, and ever-verified status. It then copies those author-level quantities onto each post-impression row for that author, so later steps can classify impressions without retaining the author key. All author identifiers, URLs, handles, raw like strings, and element IDs are dropped before the project-local substrate is written.

Each post impression then receives an author-position label: horizontal (H, peer), vertical (V, broadcast), ambiguous (A), or missing. The strict classifier is our preferred author-level

measure. It is an ordered rule: label the impression V if the author has reach at least 5, or median likes at least 100, or is ever verified and has reach at least 2; label it H if the author has reach one, is not ever verified, and has missing median likes or median likes no larger than 20; label all remaining cases A. Reach alone would not support the horizontal label: being seen by a single study respondent is not the same thing as a peer tie, since unique entity accounts, niche pages, and public figures can also appear in only one respondent's logs. The engagement and verification screens move such single-respondent but entity-like authors out of the peer bucket.

Because the horizontal rule is deliberately conservative, we also carry a relaxed classifier that differs in exactly one constant: the median-likes cap an author may attract while still counting as horizontal rises from 20 to 50. The relaxed classifier therefore reassigns single-respondent, never-verified authors with median likes between the two caps from the ambiguous bucket to the peer bucket, leaving all vertical labels unchanged. It serves as the alternative classifier in the peer-share robustness exhibit (Table A.21).

For each classifier, we aggregate the post-level labels to the respondent-level peer share described in the main text: within each day, the share of classifiable post impressions labeled horizontal, computed separately on Facebook and X and blended with weights frozen at the user's early Facebook-versus-X platform-time allocation – the early platform mix with Reddit dropped and renormalized – with the daily shares averaged over the full logging window. Freezing the weights closes the one channel through which the long measurement window could import treatment response: platform allocation is the margin treatment reaches most directly, and peer shares differ sharply by platform – Facebook feeds are dominated by horizontal authors, X feeds by vertical ones – so an average with realized weights would move if treated users reallocated attention between the platforms, even with both platforms' compositions unchanged. Beyond precision, the full window also serves the measure's purpose: a short window would capture the mix of platforms a user happened to open in a particular week rather than the standing composition of their feed. Missing labels are excluded from the denominator, while ambiguous labels remain in the denominator but count as neither peer nor broadcast. Figure A.6 reports the post-weighted distribution of strict horizontal, ambiguous, and vertical labels separately for Facebook and X. The net peer share, defined as the peer share minus the vertical share, enters the robustness exhibit (Table A.21) as the alternative treatment of vertical authors.

E.1.1 Author-composition stability

Because the peer share is measured over post-treatment impressions, a natural concern is that it could itself be affected by treatment assignment: toxicity hiding might change which authors the platform serves, so the peer share would partly be an outcome rather than a stable feed-composition measure. This concern is weaker for post impressions than for comments. The extension records platform-served posts, including posts subsequently hidden, so hiding does not mechanically remove an impression or its author metadata from the peer-share

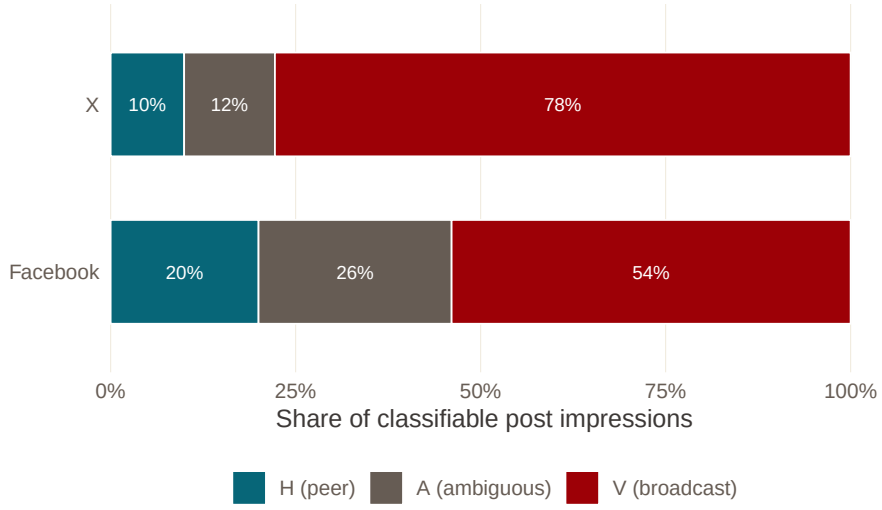


FIGURE A.6: AUTHOR-POSITION COMPOSITION BY PLATFORM

Notes: This figure reports the post-weighted distribution of author-position labels among classifiable Facebook and X post impressions in the deidentified horizontal/vertical substrate. Bars report post-weighted shares, where labels follow the strict classifier and comprise horizontal (H, peer), ambiguous (A), and vertical (V, broadcast). Missing labels are excluded from the denominator, while ambiguous labels remain in the denominator. Reddit is omitted because the logs do not contain author-position inputs for Reddit.

denominator. The residual threat is an algorithmic or browsing channel: by changing what users open, view, or engage with, hiding could lead the platform to serve a different author distribution. Consistent with the platform-supplied interpretation, treatment assignment does not detectably change the toxic share of platform-supplied content; the diagnostic below asks whether the same stability holds for author composition. The present experiment does not contain a pre-treatment browsing window with author identities, but the earlier toxicity-hiding field experiment in Beknazar-Yuzbashev et al. (2025) does. That experiment used the same toxicity-hiding technology and includes a 14-day baseline before hiding begins. We therefore use that study's data as an external diagnostic for whether toxicity hiding moves feed-author composition.

The diagnostic uses Facebook and Twitter posts in that study's main sample. We identify authors using the display name for Facebook and the profile URL for Twitter, the most stable keys for the two platforms. We compute all metrics separately for each respondent-platform pair and then pool Facebook and Twitter using the primary platform keys. The first three rows of Table A.19 are baseline-to-post similarity metrics: the coefficient is the randomized ATE of treatment assignment on the similarity between a respondent's post-treatment author distribution and that respondent's baseline author distribution. The last two rows are within-window composition metrics estimated in a two-period DiD with respondent-platform fixed effects and platform-specific post-window fixed effects.

The five metrics target the peer-share concern directly. The share of post-treatment impressions from baseline authors measures how much of the post-treatment feed comes

from accounts already seen before hiding begins. Author-share overlap and cosine similarity compare the full impression-weighted author distribution before and after treatment. Unique authors per 100 impressions tests whether the feed becomes broader after scaling out changes in usage volume. The author HHI tests whether treatment changes concentration in a small number of accounts.

TABLE A.19: AUTHOR-COMPOSITION STABILITY IN THE EARLIER TOXICITY-HIDING EXPERIMENT

Metric	Estimand	Offered coefficient	Shown coefficient	CI bound / SD
Post impressions from baseline authors	ATE	−0.008 [−0.037, 0.020]	−0.009 [−0.037, 0.020]	0.16
Author-share overlap	ATE	−0.009 [−0.032, 0.013]	−0.010 [−0.032, 0.012]	0.17
Cosine similarity	ATE	0.003 [−0.034, 0.040]	0.003 [−0.034, 0.040]	0.13
Unique authors per 100 impressions	DiD	−0.287 [−1.917, 1.343]	0.121 [−1.510, 1.752]	0.11
Author HHI	DiD	−0.001 [−0.012, 0.009]	−0.001 [−0.012, 0.009]	0.10

Notes: This table reports diagnostics for author-composition stability using post impressions from the disjoint toxicity-hiding experiment of Beknazar-Yuzbashev et al. (2025). The baseline period is Periods 1–14 and the post period is Periods 15–56. Offered impressions are all platform-served posts, including posts hidden by the extension, while shown impressions exclude hidden posts. Each coefficient cell reports the point estimate with the 95% confidence interval beneath it. The ATE rows regress the user-platform baseline-to-post similarity metric on treatment assignment and a platform indicator; the DiD rows include user-platform and platform-by-post fixed effects. Standard errors are clustered by respondent. The final column reports, for offered impressions, the larger absolute endpoint of the 95% confidence interval divided by the control-sample standard deviation.

The estimates are close to zero throughout. For offered impressions, the 95% intervals rule out shifts larger than 0.10–0.17 control-SD units for the full-distribution and concentration metrics, and the shown-impression estimates are nearly identical. Thus, in a comparable experiment where author composition can be measured before treatment, toxicity hiding changes neither the recurring author base nor the breadth or concentration of the author distribution. This supports our use of the post-treatment peer share in the present study as a feed-composition measure: the share is computed after treatment because no pre-treatment author window is available, but the evidence from that study indicates that the author base underlying such a share is largely treatment-invariant.

E.2 Arm Balance of the Measures and Their Coverage Gates

Reading a treatment-by-measure interaction as heterogeneity requires that treatment not move the measure, and because each measure is post-randomization, that requirement has a

testable implication: treated–control balance of the measure itself. Selection into measurement is testable on the same grounds: each coverage gate conditions on post-randomization activity, so if treatment changed early activity, gating would select different kinds of users into the two arms' estimation samples. Table A.20 reports both sets of tests; each row is a user-level regression of a measure or coverage margin on treatment assignment with start-date-cohort fixed effects and heteroskedasticity-robust standard errors. The exposure-measure rows are computed on the 431 users for whom every measure is jointly available, exactly the sample the mechanism estimates use – so they test balance on the estimation sample itself. The coverage rows instead use all 664 main-sample users, coding users with no qualifying activity in the two-week window as zero days.

TABLE A.20: ARM BALANCE OF THE FINALIZED EXPOSURE MEASURES AND THEIR COVERAGE GATES

	Treated − control	Std. error	p	Control mean	N
Exposure measures and baseline covariates (Table 1 common sample)					
Supplied toxic-post share	+0.0046	(0.0069)	0.51	0.0743	431
Platform share: Facebook	-0.0368	(0.0400)	0.36	0.5107	431
Platform share: Reddit	-0.0001	(0.0319)	1.00	0.2395	431
Platform share: X	+0.0368	(0.0360)	0.31	0.2497	431
Peer share (full window)	-0.0166	(0.0166)	0.32	0.2314	431
Baseline demographics/media use (joint)			0.14		429
Coverage (all users)					
Early active days (platform logs)	+0.06	(0.39)	0.88	8.72	664
Early qualifying days (feed impressions)	+0.01	(0.44)	0.98	7.13	664
Passes coverage gate ($g \geq 3$, platform)	-0.0311	(0.0315)	0.33	0.8617	664
Passes coverage gate ($g \geq 3$, dose)	-0.0210	(0.0392)	0.59	0.7194	664
Passes coverage gate ($g \geq 5$, platform)	+0.0166	(0.0370)	0.65	0.7470	664
Passes coverage gate ($g \geq 5$, dose)	-0.0089	(0.0418)	0.83	0.6206	664

Notes: This table reports cohort-netted treated-vs-control differences: each cell is the coefficient on treatment assignment from a user-level regression with start-date-cohort fixed effects and heteroskedasticity-robust standard errors. Exposure-measure rows are computed on the common sample of 431 users, on which every measure is jointly available; coverage rows are computed on all 664 main-sample users, with users without qualifying activity coded as zero days. Coverage rows count distinct qualifying days within the first two weeks – in the platform time logs and in the main-feed impression stream, respectively – and the gate rows are indicators for clearing the corresponding requirement. The baseline demographics/media-use row reports the p-value of a heteroskedasticity-robust joint Wald test of age, sex, race, education, party affiliation, recruitment source, and self-reported social-media use in a treatment-assignment regression net of start-date-cohort fixed effects; no single coefficient is displayed because the row summarizes a joint test. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Nothing is imbalanced. On the estimation sample, the daily-average toxic-post share is balanced (treated–control difference 0.0046 on a control mean of 0.0743, $p = 0.506$), as are the three platform shares and the full-window peer share. The omnibus pre-treatment covariate check on the same sample – a joint Wald test of demographics, party affiliation, recruitment source, and self-reported social-media use in a treatment-assignment regression – is also non-significant ($p = 0.143$), so conditioning the mechanism sample on measurement coverage does not undo randomization on observables. The coverage margins carry the selection check: neither the count of early qualifying days ($p = 0.882$ in the platform logs,

$p = 0.979$ in the feed-impression stream) nor passage of the three-day gate ($p = 0.325$ platform, $p = 0.593$ dose) differs by arm. The strict five-day gate – the sample-composition robustness line of Appendix E.5 – is likewise balanced ($p = 0.655$ platform, $p = 0.832$ dose).

E.3 Robustness of the Peer-Share Construction

The peer share bundles four construction choices: the author classifier, the rule combining each day's Facebook and X shares, the measurement window, and how vertical authors enter. Each row of Table A.21 changes exactly one of these ingredients and re-fits the peer-alone specification of the main table – treatment interacted with the peer share, early platform-mix controls, cohort-by-time fixed effects – for the standardized mental-health-symptoms index and its anxiety and depression subindices. Every row is estimated on its own common sample, derived exactly as for the main table, with the peer share re-standardized on that sample, so each cell is the change in the six-week post-treatment ITT per one-standard-deviation increase in the peer share.

Table A.21: Robustness of the peer-share heterogeneity to the ingredients of the peer-share construction

Peer-share construction	Mental Health Symptoms	Anxiety	Depression
Main specification (reference)	0.110** (0.044)	0.083* (0.049)	0.125*** (0.042)
Alternative classifier	0.117*** (0.043)	0.098** (0.047)	0.123*** (0.043)
No platform weighting	0.107** (0.044)	0.081* (0.049)	0.122*** (0.042)
Early window	0.099** (0.043)	0.046 (0.048)	0.140*** (0.042)
Net peer share (peer minus vertical)	0.121*** (0.043)	0.100** (0.045)	0.129*** (0.044)

Notes: Each cell is the coefficient on treatment interacted with the peer share (+1 SD) from the peer-alone specification of the main table – a six-week post-treatment difference-in-differences regression with individual and cohort-by-time fixed effects and early platform-mix controls – with standard errors clustered at the respondent level in parentheses. Outcomes are control-standardized indices; anxiety and depression are the GAD and PHQ subindices of the mental-health-symptoms index. Each row swaps exactly one ingredient of the finalized peer share (strict author classifier; Facebook and X daily shares blended by early platform-time weights; full-window average; gross horizontal share) and is estimated on its own common sample, derived exactly as for the main table, with the peer share re-standardized on that sample. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

No single ingredient carries the result. Replacing the frozen early platform weights with realized within-day pooling of the Facebook and X posts leaves the symptoms coefficient at 0.11σ, and replacing the gross horizontal share with the net horizontal-minus-vertical difference gives 0.12σ, both on the same sample as the finalized measure. Swapping the strict author classifier for the relaxed variant – which raises the median-likes cap defining a horizontal author from 20 to 50 (Appendix E.1) – gives 0.12σ ($p = 0.007$), marginally larger and more precisely estimated than under the strict classifier. Measuring the peer share on the same early five-day, two-week scaffold as the other measures, rather than over the full window, restricts the sample to the 412 users with early classifiable posts and gives 0.10σ ($p = 0.023$) – the direct check that the full-window choice, made for precision, is not what generates the heterogeneity. Across all five constructions the heterogeneity loads on the

depression subindex – significant at the 1% level in every row – with anxiety consistently the weaker subindex.

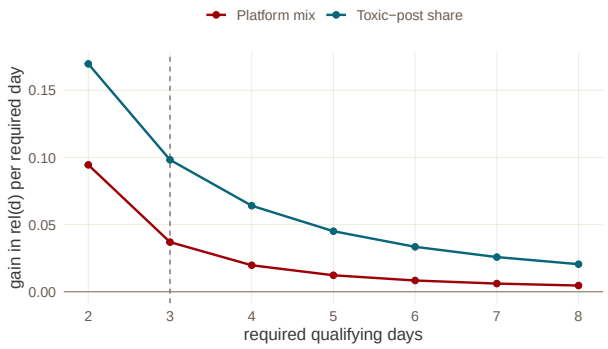
E.4 Choice of the Early-Measure Criteria

The early-measure criteria involve three constants, each with a distinct measurement trade-off. A stricter required-day gate ($g \geq 3$) improves within-user averaging but excludes users with sparse activity. A longer calendar window ($w = 14$) improves coverage but lengthens the period over which treatment could in principle affect the measure. A larger day count ($K = 5$) improves averaging but moves the measure farther into the treated period. The diagnostics below evaluate the marginal gain from increasing each constant along the corresponding margin. For both the platform mix and the toxic-post share, the gains are concentrated at low values and flatten near the chosen criteria. Figure A.7 reports these marginal gains. The diagnostics are outcome-blind – they use control-arm behavior and coverage only.

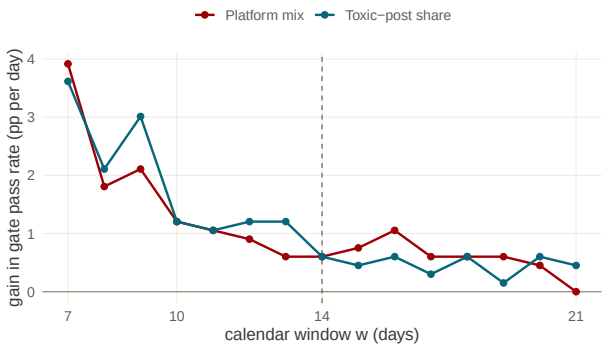
Panel (a) evaluates the gate. A user measured on exactly d days carries a daily mean whose implied reliability is $\text{rel}(d) = \sigma_u^2/(\sigma_u^2 + \sigma_e^2/d)$, computed from control-arm variance components (σ_u^2 between users, σ_e^2 across days within user). The cost side is steady rather than diminishing – each additional required day excludes about 5 percentage points of users (69% of main-sample users clear $g \geq 3$ on the toxicity margin, 60% clear $g \geq 5$) – so the gate sits at the bend of the noisier measure's curve: past three days, reliability gains shrink while the exclusion cost does not.

Panel (b) evaluates the window. At a fixed day count, extending the calendar window does not change a dense user's measure – their first five qualifying days are already inside it; the window's only benefit is coverage, letting sparse users accumulate the three days the gate requires, and its cost is scope, since every calendar day is another day over which treatment could in principle move the measure. The marginal coverage curve empties fast: by the start of the plotted range at $w = 7$ the bulk of the mass is already in, and past $w = 14$ an extra calendar day adds about half a percentage point of coverage on either margin (69% of users clear the toxicity gate within two weeks, 72% within three). Two weeks also spans two complete weekly usage cycles, so weekend-only users are sampled twice; the window stops where the curve has emptied.

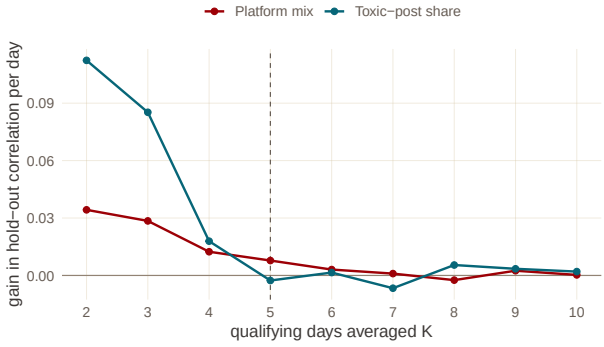
Panel (c) evaluates the day count against a hold-out target: the control-arm correlation between the first-K-days mean and the same user's daily mean over days 22–42, disjoint from every early window, on the gated sample held fixed along the curve. Both curves climb through the fourth day and are flat by the fifth – so past five days an additional averaged day buys essentially no reliability while pushing the average one day deeper into the treated period.



(A) Gate: marginal reliability per required day



(B) Window: marginal coverage per calendar day



(C) Day count: marginal hold-out reliability per averaged day

Figure A.7: What one more day buys: marginal returns behind the early-measure criteria

Notes: Each panel plots, for the early platform mix and the early toxic-post share, the improvement bought by one additional day along one constant of the five-day, three-day-gate, two-week scaffold; dashed lines mark the chosen constants. Panel (a): the gain in implied reliability $\text{rel}(d) = \sigma_u^2/(\sigma_u^2 + \sigma_e^2/d)$ from each additional required qualifying day, computed from control-arm variance components (σ_u^2 between users, σ_e^2 across days within user; for the platform mix, averaged over the three platform-share columns). Panel (b): the gain in the share of users whose third qualifying day falls within the calendar window, in percentage points per day, computed on all main-sample users. Panel (c): the gain in the control-arm correlation between the first-K-days mean and the same user's daily mean over days 22–42, disjoint from every early window, on the fixed $w = 14$, $g \geq 3$ gated sample.

E.5 Robustness to the Alternative Criteria

The five-day, three-day-gate, two-week constants reach the main table through three channels only: the platform-mix controls, the toxicity-share control in the final column, and the common-sample gate. Each row of Table A.22 rebuilds both early measures under the indicated criteria, applies that variant's own coverage gate, re-derives the common sample exactly as the main table does, and re-fits the joint specification. The results hold across all specifications.

Table A.22: Robustness of the mechanism estimates to the early-measure criteria

Criteria (K-g-w)	N users	Treated × peer (+1 SD)	Treated × MDI (+1 SD)
Main criteria (5-3-14)	431	0.140*** (0.045)	0.129*** (0.042)
Three-day average (3-3-14)	431	0.137*** (0.045)	0.130*** (0.042)
Strict five-day gate (5-5-14)	373	0.113** (0.046)	0.147*** (0.049)
Relaxed criteria (3-2-10)	441	0.086** (0.038)	0.130*** (0.043)
Ten-day window (5-3-10)	405	0.128*** (0.046)	0.132*** (0.045)
Three-week window (5-3-21)	451	0.134*** (0.043)	0.127*** (0.040)

Notes: Each cell is a coefficient from the joint specification of the main table (peer share and MDI entered together, with platform-mix controls): a six-week post-treatment difference-in-differences regression with individual and cohort-by-time-by-baseline-symptoms fixed effects and standard errors clustered at the respondent level. The outcome is the standardized mental-health-symptoms index. Each row rebuilds the two early measures under the indicated day-count–gate–window criteria, applies that variant's own coverage gate, and re-derives the common sample exactly as the main table does; the peer share is held at its finalized full-window construction, including its blend weights, and the peer share, the MDI, and the toxicity share are re-standardized on each row's own user frame.*, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

F Usage Intensity: Shown Feed-Post Volume

The experiment randomizes the filtering regime, not the amount of use, so whether the six-week ITT varies with usage intensity can only be asked of a measured activity margin. Like the feed-based exposure measures of Appendix E, any such margin is measured after randomization; unlike them, it is the margin treatment touches most directly. The composition measures behind the main heterogeneity results – the platform mix, the supplied toxic-post share, the peer share – are shares of the feed and its supply rather than quantities, trait-like objects that treatment has little room to move, which is what lets that appendix section build the case for reading them as proxies for pre-treatment variables. Consumed volume has no such insulation: hiding subtracts from the shown count mechanically, and if treated users respond behaviorally to a changed feed, the response loads first on how much they consume. However well the measure tests, it cannot be promoted to a pre-treatment proxy on those terms. This is why we rely on it less than on the composition measures: no main-text result builds on the analysis below, and every interaction in this section is a descriptive heterogeneity association, not a dose-response. We report the exercise nonetheless, because shown feed-post volume is the only usable measure of usage intensity available to us. The natural pre-treatment alternative – self-reported social-media minutes – carries almost no signal about measured behavior: its correlation with extension-measured minutes over the 42-day window, adjusted by each user's self-reported desktop share, is 0.12 ($N = 251$ control-pool users), too little to rank users by intensity.

F.1 Construction and Choice of Criteria

The preferred measure is the daily mean of main-feed posts shown – supplied posts minus hidden ones, with comments and ads excluded – over the first five observed active days

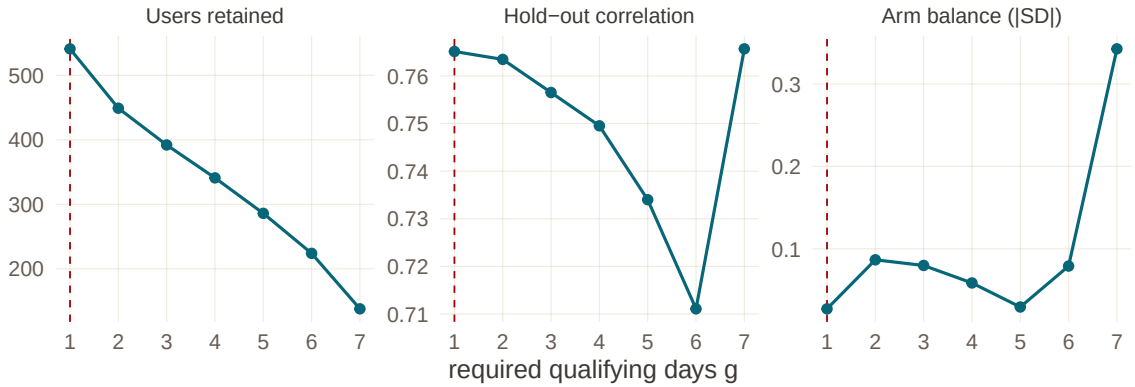
($K = 5$) within one week of treatment start ($w = 7$), requiring at least one qualifying day ($g \geq 1$). It is standardized on control-pool moments, winsorizing at the control-pool 99th percentile first so that a handful of extreme feeds do not set the scale of the regressor; the cap matters for the unit, not the inference (Panel A of Table A.23).

Restricting to main-feed posts rather than to all shown content minimizes the role of user choice in what gets counted. Feed delivery is comparatively passive – given the time a user spends scrolling, the platform decides which posts enter the viewport – so shown feed posts track feed-only behavior. Comment impressions instead record which posts the user clicks into, a choice that is endogenous twice over: it responds to the content encountered, and treatment changes what there is to click on. Empirically the broader measure behaves accordingly, diluting the interaction toward zero (Panel A of Table A.23).

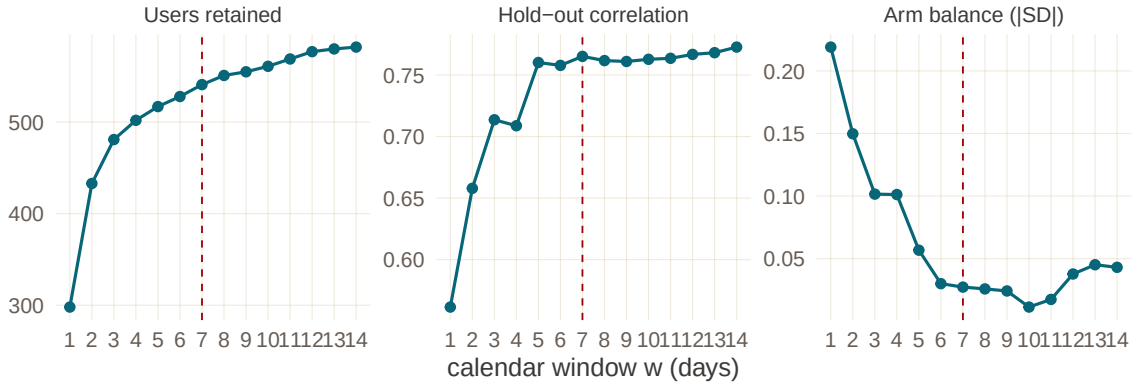
The criteria follow the same principles that set the exposure measures' five-day, three-day-gate, two-week scaffold (Appendix E.4); applied to a volume regressor, they land on different values. Figure A.8 reports outcome-blind diagnostics along each criterion – retained sample, hold-out reliability (the control-pool correlation between the early measure and the same user's days-22–42 volume, disjoint from every early window), and arm balance. Panel (a) shows why the gate is minimal: requiring repeated early activity conditions the sample on the very margin being measured, and buys nothing in exchange – raising the gate from one to three qualifying days drops the sample from 541 to 392 users while hold-out reliability is flat (0.77 vs. 0.76) and balance does not improve. Panel (b) shows why the window stops at one week: reliability climbs through the first seven days ($\rho = 0.71$ at $w = 3$, 0.77 at $w = 7$) and is flat thereafter (0.77 at $w = 14$), the noisy point imbalances of the shortest windows – never significant – settle near zero by the first full week, and every added calendar day admits later, more treatment-exposed days into the measure; one week is the shortest window that spans a complete weekly usage cycle. Panel (c) shows why five days are averaged: the hold-out correlation rises steeply through the third averaged day (0.74 at $K = 3$) and gains little past the fifth, while each additional averaged day pushes the measure one active day deeper into the treated period. The stability behind these curves is itself worth noting: one week of feed throughput predicts weeks four through six at $\rho = 0.77$, so consumed volume is a persistent user trait, which is what makes an early, less treatment-exposed measurement viable at all. Panel B of Table A.23 reports select alternatives. The least treatment-exposed construction available – the volume of the first observed active day alone – carries the interaction by itself (0.059σ, $p = 0.046$); the gated construction mirroring the exposure measures' scaffold – three qualifying days required, the first three averaged – gives a marginal 0.055σ ($p = 0.099$) on the smaller gated sample; and the three-day-window and two-week constructions move the coefficient little.

F.2 Results

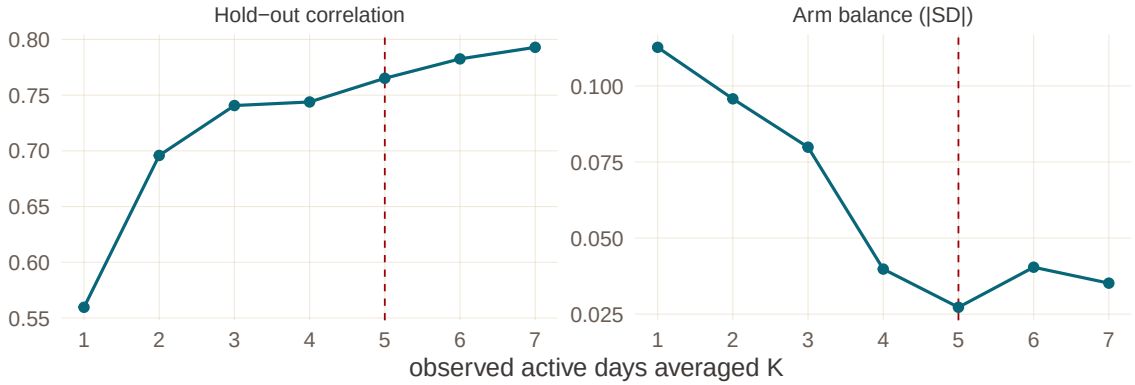
Panel A of Table A.23 reports the preferred measure. The interaction is 0.061σ (se 0.031, $p = 0.048$) per control-SD of early feed-post volume: the deterioration is larger among



(A) GATE: REQUIRED QUALIFYING DAYS g ($K = 5$, $w = 7$)



(B) WINDOW: CALENDAR DAYS w ($K = 5$, $g \geq 1$)



(C) DAY COUNT: OBSERVED ACTIVE DAYS AVERAGED K ($g \geq 1$, $w = 7$)

FIGURE A.8: OUTCOME-BLIND DIAGNOSTICS BEHIND THE SHOWN FEED-POST CRITERIA

Notes: Each panel varies one constant of the preferred five-day, one-day-gate, one-week construction of shown feed-post volume, holding the other two fixed, and reports the retained sample, the hold-out reliability, and the absolute arm imbalance of the resulting measure; the day-count panel omits the retained-sample series, which does not vary with K at a fixed gate and window. Dashed lines mark the chosen constants. Hold-out reliability is the control-pool correlation between the early measure and the same user's mean daily feed-post volume over days 22–42, disjoint from every early window. Arm balance is the absolute cohort-netted treated-vs-control difference in the standardized measure, in control-SD units, from a user-level regression with start-date-cohort fixed effects and heteroskedasticity-robust standard errors. The diagnostics are outcome-blind: they use behavior and coverage only, never the mental-health outcomes.

treated users with heavier feeds. The measure is balanced by arm (0.03σ, $p = 0.778$) – the testable implication of reading the interaction as heterogeneity rather than treatment response. Winsorization is presentational, not load-bearing: the uncapped measure gives 0.044σ ($p = 0.032$), if anything more sharply estimated, with the smaller per-SD coefficient reflecting the tail-inflated standard deviation of the raw measure. The all-content analogue is 0.030σ ($p = 0.352$), the dilution anticipated above. Panel C disciplines the interpretation against the paper's other margins. Heavy consumers skew toward particular platforms, so the volume interaction could be platform composition in disguise; adding the Treated × platform-share interactions instead sharpens it slightly (0.068σ, $p = 0.036$). Jointly with the peer-share, MDI, and pre-treatment-symptom margins, the coefficient is 0.079σ ($p = 0.015$, $N = 517$), while the peer-share (0.085σ, $p = 0.019$) and MDI (0.118σ, $p = 0.005$) margins retain their own coefficients: the volume margin is additional to, not a repackaging of, the composition margins.

Magnitude matters for the reading. If harm scaled proportionally with consumed volume, the ITT at the mean (0.162σ in this sample) and the dispersion of volume (CV = 1.45) would imply an interaction of roughly 0.24σ per control SD, before any attenuation for early-measure error; the estimate is about a quarter of that. The data therefore do not support harm proportional to throughput; they suggest, more modestly, that the ITT is concentrated among high-volume feed consumers.

Table A.23: Shown feed-post volume and ITT heterogeneity

Panel A. Preferred first-week construction						
Measure	Users	Control mean	CV	Treatment × measure	Arm balance	Hold-out ρ
Main-feed posts shown	541	134	1.45	0.061** (0.031)	0.03 (p=0.778)	0.77
Main-feed posts shown, unwinsorized	541	134	1.48	0.044** (0.020)	0.07 (p=0.495)	0.76
All shown content	615	321	1.35	0.030 (0.032)	0.02 (p=0.793)	0.76
Panel B. Criterion checks for main-feed posts shown						
Construction	Users	Control mean	CV	Treatment × measure	Arm balance	Hold-out ρ
First active day only (K=1, g≥1, w=7)	541	89	1.58	0.059** (0.029)	0.11 (p=0.252)	0.56
Three-day average, three-day gate (K=3, g≥3, w=7)	392	147	1.25	0.055* (0.033)	0.15 (p=0.205)	0.73
Three-day window (K=3, g≥1, w=3)	481	130	1.37	0.071** (0.030)	0.10 (p=0.330)	0.71
Preferred first-week measure (K=5, g≥1, w=7)	541	134	1.45	0.061** (0.031)	0.03 (p=0.778)	0.77
Two-week window, all days (K=all, g≥1, w=14)	582	136	1.47	0.061** (0.030)	0.02 (p=0.840)	0.87
Panel C. Horse race for the preferred feed-post measure						
Specification	Term	Users		Estimate (SE)	p-value	
Feed posts only	Treatment × feed posts shown	541		0.061** (0.031)	0.048	
+ platform shares	Treatment × feed posts shown	541		0.068** (0.032)	0.036	
+ platform shares, peer share, MDI, pre-treatment symptoms	Treatment × feed posts shown	517		0.079** (0.032)	0.015	
+ platform shares, peer share, MDI, pre-treatment symptoms	Treatment × peer share	517		0.085** (0.036)	0.019	
+ platform shares, peer share, MDI, pre-treatment symptoms	Treatment × MDI	517		0.118*** (0.042)	0.005	

Notes: Shown feed-post volume counts main-feed posts delivered to the screen after filtering (supplied minus hidden; comments and ads excluded), averaged over the first five observed active days within one week of treatment start with at least one qualifying day, winsorized at the control-pool 99th percentile, and standardized on control-pool moments. Panel A also reports its unwinsorized version and the all-shown-content analogue from the full impression panel. Panel B reports alternative day-count–gate–window criteria for the feed-post measure; each row re-derives its own sample, winsorization cap, and control moments. Panel C reports the preferred measure with progressively richer controls. Interaction estimates come from six-week post-treatment difference-in-differences regressions of the mental-health-symptoms index (control-SD units) with individual and cohort-by-time fixed effects and standard errors clustered at the respondent level (in parentheses); Panel C's final specification adds the peer-share, MDI, and pre-treatment-symptom margins and stratifies the cohort-by-time fixed effects by pre-treatment symptoms. Arm balance is the cohort-netted treated-vs-control difference in the standardized measure, with the heteroskedasticity-robust p-value in parentheses. Hold-out ρ is the control-pool correlation between the early measure and the same user's average daily volume over days 22–42. Because shown volume is measured after randomization, all interaction coefficients are descriptive heterogeneity associations. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

G Welfare Translation: Mapping, Aggregation, and Sensitivity

This appendix documents the welfare-calibration exercise summarized in the welfare-translation subsection of the paper

G.1 External mapping

We map symptom scores to EQ-5D-3L health-state utility using the published mapping of Mukuria et al. (2025), whose recommended specification is a four-component adjusted limited-dependent-variable mixture model (ALDVMM) estimated on pooled clinical-trial data. The model takes PHQ-9, GAD-7, and age as inputs and returns a predicted EQ-5D-3L utility on the UK value set. We access the authors' published lookup tables directly – the supplementary workbook `MOESM2_ESM.xlsx`, archived with a regeneration recipe under `data/external/mukuria2025/` – and do not re-estimate the mapping.

The exercise uses four specifications, all drawn from the same recommended four-component ALDVMM:

1. **Joint PHQ + GAD + age** – the full three-input lookup (`aldvmm_phq_gad_age.csv`); our headline specification.
2. **Joint, PHQ-8 × 9/8** – identical inputs, but PHQ-8 is first rescaled to the PHQ-9 range (see below).
3. **PHQ + age (GAD ignored)** – the model's PHQ-marginal lookup (`aldvmm_phq_age.csv`).
4. **GAD + age (PHQ ignored)** – the model's GAD-marginal lookup (`aldvmm_gad_age.csv`).

PHQ-8-as-PHQ-9 convention. Our instrument fields the eight-item PHQ-8 (range 0–24), whereas the mapping is developed on the nine-item PHQ-9 (range 0–27). The default (`phq8_as_phq9`) passes the PHQ-8 score into the lookup unchanged, implicitly treating the omitted ninth (self-harm) item as contributing zero; the sensitivity (`phq8_rescaled`) multiplies by 9/8 before lookup. GAD-7 enters on its original scale.

Interpolation. Because the tables are tabulated on the integer score grid, `utility_lookup()` evaluates non-integer points (e.g. a pre-treatment value shifted by a fractional ITT) by multilinear interpolation: each input is bracketed between its neighbouring grid nodes and utility is the weighted average over the 2^d surrounding corners, where d is the number of inputs. The interpolator returns `NA` rather than extrapolating outside the published support (PHQ-9 0–27, GAD-7 0–21, age 18–99), so an out-of-range participant drops from the average rather than receiving an imputed value.

Table A.24: Effect of aggregation choice on the welfare translation

Aggregation	Δu
Plug-in at sample means	−0.0203
Individual-level, homogeneous effect	−0.0195
Individual-level, quartile-heterogeneous effect	−0.0176

Notes: This table reports the per-participant utility change Δu under three aggregation choices. All three rows use the recommended four-component ALDVMM with the PHQ-8 treated as PHQ-9. The plug-in row evaluates the nonlinear mapping at the sample-mean pre-treatment point; the homogeneous row applies the sample-average ITT to each participant's own pre-treatment values and averages the resulting utility change; and the heterogeneous row uses pre-treatment-PHQ-quartile-specific ITTs, so the highest-symptom quartile—which lies on a steeper part of the utility surface—contributes its share of the aggregate.

G.2 Aggregation

Let $u(\cdot)$ denote the mapping above and let $(\widehat{\Delta\text{PHQ}}, \widehat{\Delta\text{GAD}})$ be the estimated six-week ITT. We report three ways of turning per-symptom effects into a per-participant utility change Δu, compared in Table A.24:

- **Plug-in at sample means** evaluates the mapping once at the sample-mean pre-treatment point $(\overline{\text{PHQ}}, \overline{\text{GAD}}, \overline{\text{age}})$ and again at that point shifted by the ITT, then differences. It is the simplest route but, because $u(\cdot)$ is nonlinear, it is biased for the population-average utility change.
- **Individual-level, homogeneous** (our headline) applies the common ITT shift to *each participant's own* pre-treatment values $(\text{PHQ}_i, \text{GAD}_i, \text{age}_i)$, forms $\Delta u_i = u(\text{post}_i) - u(\text{pre}_i)$, and averages. This is the conceptually correct mean utility change: it averages the map of the shift rather than mapping the shift of the average.
- **Quartile-heterogeneous** replaces the common shift with pre-treatment-PHQ-quartile-specific ITTs – estimated from `i(BaselinePHQQuartile, Treated)` with the cohort-time fixed effects interacted with the quartile – so participants in the highest-symptom quartile, who sit on a steeper part of the utility surface, contribute their own slope to the aggregate.

The gap between the plug-in and individual-level figures is the Jensen term induced by curvature of $u(\cdot)$; it is small here because the ITT is modest relative to the grid spacing, but we treat the individual-level number as primary on principle.

G.3 Uncertainty and sensitivity

We separate *sampling* uncertainty from *specification* uncertainty.

Cluster bootstrap. We resample participants (by `ID`) with replacement, re-estimate the fixed-effects ITT on each resample (clustering on the resampled-participant index), recompute the individual-level Δu, and read percentile confidence intervals off the bootstrap distribution. To fold mapping uncertainty into the *same* interval, every outer draw also makes

Table A.25: Cluster bootstrap over individuals and mapping specifications

Metric	Estimate	95% CI	
		Lower	Upper
Δu	−0.0195	−0.0297	−0.0048
Cost at \$50,000/QALY	\$113	\$28	\$171
Cost at \$100,000/QALY	\$225	\$55	\$342
Cost at \$150,000/QALY	\$338	\$83	\$513
Cost at own household income	\$158	\$38	\$243

Notes: This table reports full-sample point estimates and cluster-bootstrap 95% confidence intervals for the per-participant utility change Δu and the associated dollar metrics. The bootstrap resamples participants with replacement and, within each outer draw, uniformly selects one of the available mapping specifications (joint, PHQ-only, GAD-only) to capture across-specification uncertainty. The reported bounds are the 2.5th and 97.5th percentiles of the bootstrap distribution from $B = 1000$ draws under a fixed seed. The dollar rows report per-participant welfare costs over the six-week post-treatment window, where QALYs are computed as $\Delta u \times 6/52$.

an inner uniform draw over the available lookups – joint, PHQ-only, GAD-only – so the reported band reflects both who is in the sample and which mapping is used. We use $B = 1000$ draws under a fixed seed; intervals are the 2.5th and 97.5th percentiles. Table A.25 reports the bootstrap summary for Δu and the dollar metrics.

Specification range. As a second, deterministic layer, we recompute the headline translation under each of the four mapping specifications and report the span (reported in the main text). This is a range across modelling choices, not a confidence interval, and should be read as such.

Threshold-crossing cross-check. As a coarse, distribution-free check on the continuous estimate, we ask how many treated participants the intervention pushes across a clinical severity cutoff. We form each treated participant's untreated counterfactual at the six-week post-treatment survey by subtracting the ITT from their observed score, bin both the observed and counterfactual scores into the standard PHQ-8 (5/10/15/20) and GAD-7 (5/10/15) severity bands, and count net upward crossings. We then price only the crossers, valuing each move at the utility difference between the midpoints of the destination and origin bands while holding the other symptom at its sample mean. Because it discards within-band variation and prices at representative midpoints, this estimate is deliberately coarser than the continuous one; it serves to confirm the order of magnitude.

G.4 Dollar benchmarks

We monetize QALY losses two ways. First, at three willingness-to-pay benchmarks from the cost-effectiveness literature – \$50,000, \$100,000, and \$150,000 per QALY (Neumann et al., 2014; Institute for Clinical and Economic Review , ICER). Second, at a sample-anchored

value: each participant's self-reported household-income bucket is mapped to its midpoint (with the open-ended top bucket anchored conservatively at $175,000), and we report both an *own-income* figure (each participant valued at their own income) and a *population-mean income* figure (everyone valued at the sample-mean income, taken across treated and control). The income anchor is a descriptive benchmark, not a claim about the social value of a QALY.

All figures are evaluated over the six-week post-treatment horizon, so QALY $= \Delta u \times 6/52$; we report no annualization or decay scenario, consistent with treating the six-week post-treatment survey as the estimand. Dollar values are welfare costs.

H Content Classification

We classify the content of the toxic posts shown to participants—the margin the intervention manipulates—to characterize the material the hiding technology removes. The classification, performed with GPT-5, proceeds in two stages. First, a gate labels each toxic post impression (Perspective toxicity above 0.33) as political or not, where *political* denotes content concerning elections, parties, candidates, officials, government policy, courts, public agencies, civil rights as a political issue, foreign policy, or political identity. Second, for political posts, the classifier identifies the partisan target (who the post is about) and the stance toward that target; from these two labels we derive the partisan side the post is congenial to, with hostility toward one side coded as congenial to the other, and posts with mixed, non-U.S., or otherwise unclear partisan content labeled ambiguous. The gate is applied to toxic posts only, so the shares below describe the political composition *within* toxic content, not the political share of the entire feed.

H.1 Political content among toxic posts

Figure A.9 reports, separately for Democrats and Republicans, the average user-level share of toxic posts that are political on Facebook and X. For each participant we take the share of their toxic post impressions that the gate labels political over the first six weeks of the intervention, and the bars plot the equal-weighted mean of that share across participants, with 95% confidence intervals from the across-participant standard error. Party is the pre-treatment self-report, with independent leaners folded to the party they lean toward.

The political composition of toxic content is governed first by platform. On Facebook it is small and essentially identical across parties: 9.2% of the toxic posts Democrats see are political (264 users) against 9.3% for Republicans (112 users). On X the political share is about three times larger and tilts partisan: 25.1% for Democrats (230 users) versus 33.8% for Republicans (87 users), a gap whose confidence intervals are nearly disjoint. Because party is not randomized and the gate covers only toxic posts, these are descriptive feed-composition shares, not estimates of any treatment-induced change.

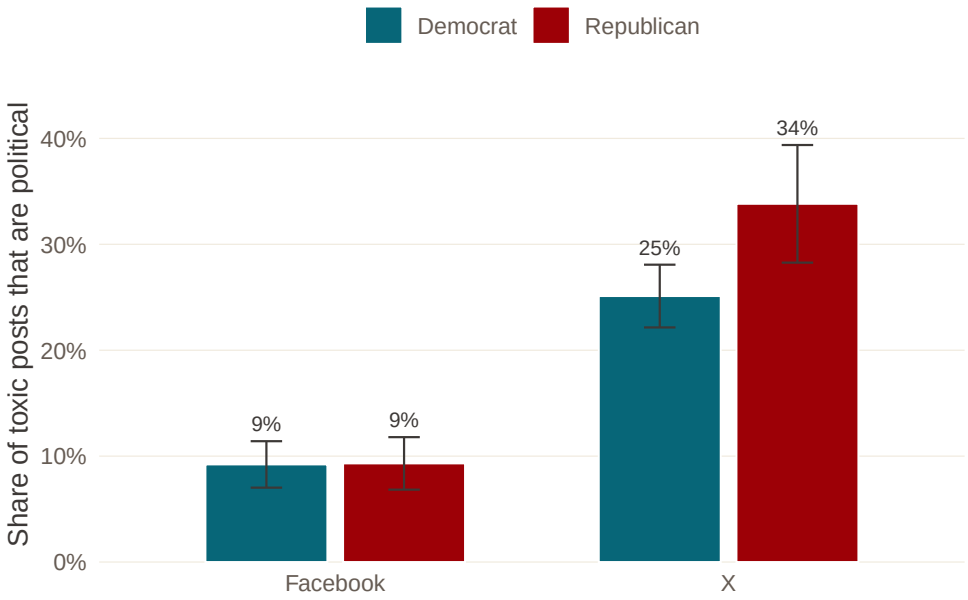


FIGURE A.9: POLITICAL CONTENT AMONG TOXIC POSTS, BY PARTY AND PLATFORM

Notes: This figure reports the equal-weighted across-participant mean of the per-participant share of toxic post impressions that the political gate labels political, computed over the first 42 days after treatment assignment, separately for Democrats and Republicans on Facebook and X. Toxic posts are impressions with Perspective toxicity above 0.33; the political gate is applied to toxic posts only, so the shares are within-toxic shares rather than shares of the full feed. Whiskers denote 95% confidence intervals based on the across-participant standard error. Party is the pre-treatment self-report, with independent leaners folded to their leaned party; Reddit is omitted. Because party is not randomized, the shares are descriptive.

H.2 Comparing the toxicity and alienation classifiers

The two treated arms differ only in the hiding rule. The Toxicity arm hides a post when its Perspective toxicity score exceeds 0.33; the Alienation arm hides the superset flagged either by Jigsaw's alienation classifier (score above 0.7) or by Perspective toxicity above 0.33. Averaging daily shares within each of the 411 users in the two treated arms (each user weighted equally), the Toxicity rule removes 7.6% of a user's feed and the combined rule 9.1%. The alienation classifier therefore adds only 1.5%—content that is alienating but not toxic, the material that passes through the Toxicity arm but is removed by the Alienation arm. The two arms' hide sets overlap almost entirely (Jaccard 0.83), which is why the design preregisters pooling them. The arms remove comparable amounts in practice: the extension hid 7.6% of the average Toxicity-arm user's feed and 9.0% of the average Alienation-arm user's feed. Table A.26 and Figure A.10 summarize the comparison.

Underlying the arms are two distinct classifiers. Across the same posts the Perspective toxicity and Jigsaw alienation scores are strongly but imperfectly related (Pearson 0.57, Spearman 0.49); Figure A.11 plots their joint distribution. At the operative thresholds the toxicity rule flags 10.1% of posts and the alienation rule 6.0%, agreeing on the 3.5% flagged by both, with a Jaccard of 0.28 and Cohen's κ of 0.39 (Table A.27).

Table A.26: What each hiding rule removes

Hiding rule	Share of feed hidden
Perspective (P > 0.33)	7.6%
Perspective + Alienation (A > 0.7 or P > 0.33)	9.1%
Added by Alienation (alienating, not toxic)	+1.5 pp
Hide-set overlap (Jaccard)	0.83

Notes: This table reports, over the 411 participants in the two treated arms (each user weighted equally), user-level averages of the daily share of feed each rule removes. The increment is content scored as alienating (A > 0.7) but not toxic (P ≤ 0.33). Jaccard is the overlap of the two (nested) hide sets.

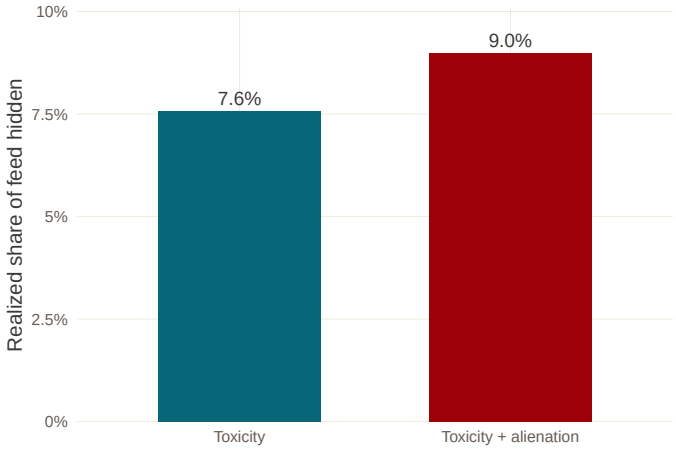


Figure A.10: Realized share of feed hidden, by arm (daily average)

Table A.27: Classifier agreement at the operative thresholds

	Alienating (A > 0.7)		
Toxic (P > 0.33)	Yes	No	Total
Yes	3.5%	6.6%	10.1%
No	2.5%	87.4%	89.9%
Total	6.0%	94.0%	100.0%

Notes: This table reports classifier agreement at the operative thresholds. Cells are shares of the 6 463 665 distinct posts. The off-diagonal "alienating, not toxic" cell (2.5% of distinct posts) is content that the alienation classifier flags but Perspective does not; the user-averaged feed analogue—the margin the Alienation arm removes from feeds—is reported in the arm comparison above.

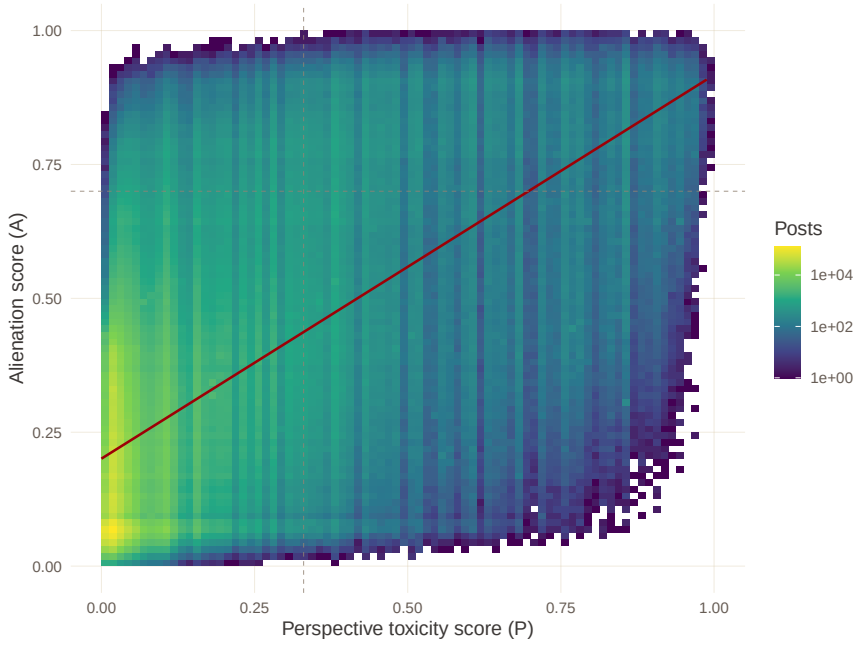


FIGURE A.11: JOINT DISTRIBUTION OF PERSPECTIVE TOXICITY AND ALIENATION SCORES

Notes: This figure reports the joint distribution of Perspective toxicity and alienation scores as a binned two-dimensional density on a logarithmic color scale over the 6 463 665 distinct posts; the red line is the OLS fit. The dashed lines mark the operative thresholds (Perspective 0.33, alienation 0.7).

I Preregistration Clarifications and Deviations

This experiment was registered in the AEA RCT Registry as AEARCTR-0013987 before participant enrollment. The treatment assignment, filtering thresholds, pooled treatment comparison, and six-week primary endpoint were implemented as preregistered. Below, we discuss deviations from the preregistration and, where necessary, present evidence that they do not materially affect the study's substantive conclusions. Table A.28 anchors the quantitative part of this discussion: starting from the paper's preferred specification in column (1), it unwinds the analysis deviations one at a time until, in column (4), the outcome, specification, and sample restrictions match the registered analysis.

The construction of the mental-health outcome used in the paper differs from the preregistered construction. The preregistration specifies an index formed by averaging the raw PHQ-8 and GAD-7 scale scores and then standardizing the resulting composite. Following the standardized-index construction of Kling et al. (2007), we instead standardize the fifteen component items before averaging them, preventing differences in item-level variances from implicitly assigning different weights to the items. We then re-standardize the resulting index to have a mean of zero and a standard deviation of one in the control group. We also code the items from 0 to 3, following the validated instruments, rather than from 1 to 4 as erroneously stated in the preregistration. Because this latter change is purely additive,

it does not affect the standardized treatment estimate. The construction of the index is immaterial for the headline result: re-estimating the preferred specification on the same estimation sample with the exact registered composite in place of the item-level index leaves the estimate and its precision essentially unchanged (columns (1) and (2) of Table A.28).

The preregistration specifies a baseline-to-six-week difference-in-differences analysis with recruitment-cohort fixed effects, but does not explicitly state whether these would be interacted with time. In implementing this analysis, we address the now-standard concern raised by staggered-rollout designs by including individual fixed effects and the interaction of cohort and time fixed effects (Baker et al., 2022; Beknazar-Yuzbashev et al., 2025). Because this interaction was not explicitly stated in the preregistration, we disclose it here for completeness. Column (3) of Table A.28 estimates the registered composite under the difference-in-differences specification exactly as written in the preregistration—the pooled assignment indicator interacted with a post-period indicator, with recruitment-wave fixed effects only. The adverse ITT remains statistically distinguishable from zero under this specification; the preferred fixed effects sharpen precision but are not the source of the result.

The analysis sample used in the paper requires respondents to have completed the six-week survey and to have recorded at least some extension activity after assignment. This restriction was intended to exclude participants who installed the extension to enroll in the study but immediately uninstalled it without meaningfully entering the intervention period. Because the restriction was not preregistered and conditions the estimation sample on post-assignment behaviour (although pre-treatment in the sense that these users have not actually experienced treatment), we also repeat the analysis among all respondents who completed the six-week survey, irrespective of subsequent extension activity. The registered specification estimated on that unrestricted sample of 699 respondents yields the same conclusion (column (4) of Table A.28), so the restriction does not generate the finding. Consistent with this, treatment assignment does not predict entry into any of the observed samples among all 1394 enrollees: six-week survey completion, recording any extension activity, and inclusion in the analysis sample are all balanced across assignment (Table A.30).

Recruitment fell short of the preregistered target of approximately 2,578 participants: enrollment reached 1394, of whom 699 completed the six-week survey. The shortfall reflects per-participant recruitment costs through social-media advertising that were substantially higher than anticipated.

The preregistration names PHQ-8, GAD-7, and their composite as primary outcomes, without specifying a hierarchy among them or a multiple-testing procedure. Treating the three as a single primary family under the registered specification, all three remain significant at the 5% level after a retrospective Benjamini–Hochberg adjustment (Table A.29). Section B.1 reports complementary, intentionally broader family calculations that add the primary outcomes of the companion political-attitudes registration and the secondary-outcome family.

The preregistration also specified subgroup analyses beyond the baseline-symptom median split reported in the paper. For partisanship, it specified a binary split pooling each

party with its leaning independents; we instead report three direct self-identification cells (Democrat, Independent, Republican). For minority status, it referenced female, LGBTQ+, and ethnic-minority respondents; because sexual orientation was not collected, we classify respondents as minority if they are female or non-White, as is common. Relatedly, the respondent-relevant discrimination index assigns the age-discrimination scenario using an above-sample-median age rule rather than the preregistered age-50 cutoff, to maximize the power of the split; treatment effects on the discrimination outcomes are statistically indistinguishable from zero in any case.

After observing the Wave 1 six-week results, which indicated that toxicity filtering worsened rather than improved mental health, we amended the preregistration on October 21, 2024, before collecting the outcomes introduced by the amendment. The amendment added measures of loneliness, social-media fear of missing out, and anger to investigate possible mechanisms behind the unexpected adverse effect. These outcomes should therefore be understood as outcome-informed extensions rather than as components of the original preregistered analysis.

The preregistration specified all three UCLA-3 items, but the item measuring feeling isolated from others was inadvertently omitted during survey programming. We therefore construct a two-item loneliness index and do not interpret it as the validated UCLA-3 scale. The omission may reduce the index's coverage of the broader loneliness construct and its comparability with studies using the full UCLA-3, but it does not compromise random assignment or the identification of the treatment effect on the two administered items. We also used a five-category response scale referring to the preceding four weeks.

Several further deviations concern study timing and intervention implementation. The later follow-up was conducted twelve rather than ten weeks after baseline, and the second recruitment wave ran from May to September 2025 rather than beginning in fall 2024. Relative to the registry's statement that all above-threshold posts and comments would be hidden, one implementation detail differs, as described in the main paper's design section: toxic Facebook comments are replaced with a "Deleted comment" placeholder rather than removed entirely.

Two preregistered components were not implemented. First, the willingness-to-accept exercise involving temporary account deactivation was omitted. Second, the preregistration specified cross-randomizing respondents in the two control cells, after the six-week survey, into an intervention instrumenting engagement (with feed latency as the example) to identify the engagement channel. We abandoned this component because it proved too difficult to implement reliably and because the results overtook its rationale: the channel it was designed to isolate—toxicity filtering reducing engagement (Beknazar-Yuzbashev et al., 2025), with lower engagement in turn improving mental health (Allcott et al., 2025)—operates toward mental-health improvements and therefore cannot account for the adverse effect that is this paper's central finding. In addition, the toxic-account follow-choice outcome was redesigned as an item-count experiment and therefore identifies a different estimand from the preregistered direct follow-or-not-follow decision.

Finally, several analyses in the paper extend beyond the preregistration and should be read as additional or exploratory: the clinical-threshold outcomes and the PHQ-9 proxy, the item-level symptom decompositions, the exposure-scaled instrumental-variables estimates, the welfare calibration, the feed-composition and platform heterogeneity, the moral-disengagement and political-hostility splits, the demographic splits by age, education, and income, and the cross-wave inverse-variance-weighted aggregation of the loneliness estimates.

Table A.28: From the preferred specification to the preregistered analysis

	(1)	(2)	(3)	(4)
Treated	0.151***	0.158***	0.121**	0.138***
	(0.051)	(0.051)	(0.047)	(0.045)
Outcome	Item-level index	Registered composite	Registered composite	Registered composite
Fixed effects	Respondent, cohort × time	Respondent, cohort × time	Recruitment wave	Recruitment wave
Extension-activity requirement	Yes	Yes	Yes	No
p-value	0.003	0.002	0.010	0.002
Participants	656	656	664	699
Observations	1,312	1,312	1,328	1,398

Notes: This table reconciles the paper's preferred specification with the preregistered analysis, unwinding one deviation per column. Column (1) reports the headline intention-to-treat (ITT) estimate: the item-level standardized mental-health-symptoms index, respondent and cohort-by-time fixed effects, on the analysis sample. Column (2) re-estimates column (1) on the same estimation sample, replacing the outcome with the preregistered composite, which averages the raw PHQ-8 and GAD-7 scale scores and standardizes the average. Columns (3) and (4) estimate the preregistered difference-in-differences specification—the pooled assignment indicator interacted with a post-period indicator, recruitment-wave fixed effects, and no respondent fixed effects—on the analysis sample and on all six-week survey completers irrespective of extension activity, respectively. The *Treated* row reports the assignment-by-post (DiD) coefficient; all outcomes are in control-group standard-deviation units. The preregistration does not specify the composite's standardization reference; we use the control pre-treatment distribution, and changing the reference rescales the coefficient and standard error together without affecting the t-statistic. Columns (1) and (2) estimate on 656 of the 664 participants—the eight alone in their start-date cohort are absorbed by the cohort-by-time fixed effects—while columns (3) and (4) omit those fixed effects and retain all 664 (respectively 699). Standard errors are clustered at the respondent level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.29: Preregistered primary outcomes under the registered specification

	Composite	GAD-7	PHQ-8
Treated	0.121**	0.651**	0.588**
	(0.047)	(0.261)	(0.284)
Unadjusted p-value	0.010	0.013	0.039
BH q-value	0.019	0.019	0.039
Participants	664	664	664
Observations	1,328	1,328	1,328

Notes: This table reports the three preregistered primary outcomes under the registered specification of column (3) of Table A.28: the registered composite construction, the difference-in-differences with recruitment-wave fixed effects only, and the analysis sample. The composite is in control-group standard-deviation units; GAD-7 (0–21) and PHQ-8 (0–24) are on their raw symptom scales. Because the preregistration specifies no multiple-testing procedure, the Benjamini–Hochberg adjustment across the three outcomes is retrospective. Significance stars reflect unadjusted p-values. Standard errors are clustered at the respondent level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A.30: Treatment assignment and selection into the observed samples

	Completed six-week survey	Any extension activity	Analysis sample
Constant	0.506***	0.906***	0.469***
	(0.022)	(0.013)	(0.021)
Difference	−0.007	0.009	0.013
	(0.028)	(0.016)	(0.027)
Difference p-value	0.807	0.571	0.643
Included participants	699	1,270	664
N	1,394	1,394	1,394

Notes: This table reports linear-probability regressions of sample-inclusion indicators on the pooled treatment-assignment indicator among all 1394 randomized enrollees who completed the pre-treatment survey and installed the extension. *Constant* is the control-group inclusion rate and *Difference* is the treated-minus-control difference in that rate. The inclusion margins are: completion of the six-week survey; any recorded extension activity after assignment (the risk set for the attrition analysis in Appendix D.2); and the paper's analysis sample (both requirements). Heteroskedasticity-robust standard errors in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

J Instructions

Instructions for the field experiment are available at this link: here.

References

Allcott, Hunt, Matthew Gentzkow, Benjamin Wittenbrink, Juan Carlos Cisneros, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Sandra González-Bailón, Andrew M Guess, Young Mie Kim et al., "The Effect of deactivating Facebook and Instagram on users' emotional state," Technical Report, National Bureau of Economic Research 2025.

Baker, Andrew C, David F Larcker, and Charles CY Wang, "How Much Should We Trust Staggered Difference-in-Differences Estimates?," *Journal of Financial Economics*, 2022, *144* (2), 370–395.

Beknazar-Yuzbashev, George, Rafael Jiménez-Durán, Jesse McCrosky, and Mateusz Stalinski, "Toxic Content and User Engagement on Social Media: Evidence from a Field Experiment," CESifo Working Paper 11644, CESifo 2025.

Institute for Clinical and Economic Review (ICER), "2020 Value Assessment Framework: Proposed Changes," Technical Report, Institute for Clinical and Economic Review August 2019. Accessed 2026-01-10.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz, "Experimental Analysis of Neighborhood Effects," *Econometrica*, 2007, *75* (1), 83–119.

Mukuria, Clara, Matthew Franklin, and Sebastian Hinde, "Mapping PHQ-9 and GAD-7 scores onto the EQ-5D-3L utility index using age in a general population cohort," *The European Journal of Health Economics*, 2025, *26*, 63–70.

Neumann, Peter J., Joshua T. Cohen, and Milton C. Weinstein, "Updating Cost-Effectiveness — The Curious Resilience of the $50,000-per-QALY Threshold," *The New England Journal of Medicine*, 2014, *371* (9), 796–797.