Social Media Toxicity and Mental Health: Experimental Evidence*

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

March 2026

Preliminary Draft — Please Do Not Circulate

Abstract

We study whether content-level moderation that reduces exposure to social media toxicity, holding platform access constant, affects mental health. We deploy a cross-platform browser extension that randomly hides posts and comments exceeding a prespecified toxicity threshold and follow N=1,398 desktop users for multiple weeks across two waves surrounding the 2024 U.S. elections. Contrary to our expectation, reducing toxicity exposure *worsens* mental health: the index (higher=better) combining GAD-7 and PHQ-8 declines by about 0.12 SD. Item-level decompositions show broad increases in anxiety symptoms and depressive symptoms concentrated in negative self-evaluation, sleep disturbance, and depressed mood. Heterogeneity analyses indicates larger effects for respondents with *good* baseline mental health. The treatment increases willingness to follow social media accounts producing toxic content, suggesting a possible mechanism whereby low-toxicity environments are less useful for media-based emotion regulation. We conclude that curbing toxic content may impose unintended psychological costs.

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, Rafael Jimenez-Duran, John List, Andrew Oswald, Matthew Ridley, Chris Roth, and numerous seminar participants for helpful comments and suggestions. We thank Cody Prosperini, Kevin Chen, and Esha Parekh 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.

[†]Univeristy of Chicago, The Normal Lab gb2683@columbia.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 literature provides causal evidence that access to (or deactivation from) these platforms affects well-being. Randomized deactivations of Facebook improve subjective well-being and reduce depression, anxiety, and related symptoms (Allcott et al., 2020); quasi-experimental evidence from Facebook's staggered rollout finds adverse effects on student mental health (Braghieri et al., 2022); recent work shows that self-control problems and social externalities, including fear of missing out, can sustain social media use despite negative welfare effects (Allcott et al., 2022; Bursztyn et al., 2025). Yet credible evidence remains limited on which specific components of the online environment harm mental health.

A leading candidate is exposure to toxic or hateful content—content that classifiers such as Perspective label as "rude, disrespectful, or unreasonable" or "likely to make you leave a discussion." Toxicity is both prevalent and behaviorally potent: experimentally reducing exposure lowers user engagement, revealing a trade-off for platforms between time spent and user welfare (Beknazar-Yuzbashev et al., 2025). Regulation that tightens moderation standards (e.g., Germany's NetzDG) has been shown to shift online discourse and reduce offline hate crimes, underscoring the broader downstream effects of curbing hateful content (Jiménez Durán et al., 2022). Consistent with this, social media penetration has also been linked to increased xenophobic incidents (Bursztyn et al., 2019a). Whether hiding toxic content from individuals' feeds improves mental health—holding overall social media access constant—remains an open question.

We provide the first randomized field evidence on the mental-health consequences of content-level moderation that selectively hides toxic material while otherwise leaving platform access intact. We build a cross-platform browser extension that intercepts posts and comments on major platforms and hides content exceeding a prespecified toxicity threshold, as computed using state-of-the-art classifiers (Perspective-style toxicity scores grounded in human annotations; Wulczyn et al., 2017; Muralikumar et al., 2023). We then follow participants for multiple weeks, measuring changes in clinically validated anxiety and depression scales—the GAD-7 and PHQ-8 (Spitzer et al., 2006; Kroenke et al., 2009).

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) to control conditions that do not hide any content (pure control) or hide a matched random fraction of content to equalize hiding rates (active control).[1] 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 and is 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 midline survey and continues for an additional six weeks to an endline; mental health outcomes (PHQ-8, GAD-7) are measured at baseline, midline, and endline, alongside perceptions of discrimination, social desirability, and beliefs about others' prosociality.

Using a difference-in-differences design that leverages both midline and endline observations, we compare treated and control participants relative to baseline (N = 1,398). The midline-baseline comparison is our primary pre-registered specification. Our main finding is that reducing exposure to toxic content *worsens* mental health. Using a composite index (higher means better) combining GAD-7 and PHQ-8, the effects are about −0.12 SD at midline. On the clinical scales, average effects are +0.85 (0.13 SD) points on the GAD-7 (0–21) and +0.81 (0.11 SD) points on the PHQ-8 (0–24), indicating small but precisely estimated increases in symptom burden. Item-level decompositions show broad worsening across GAD-7 items and more concentrated increases on PHQ-8 items capturing negative self-evaluation, sleep disturbance, and depressed mood. We also compute elasticities of toxic-exposure reductions with respect to mental health: over the ranges observed in our data, a 1% decrease in toxicity exposure is associated with roughly a 0.1% increase in anxiety and depression symptoms.

To benchmark magnitudes, we consider evidence from a large-scale RCT on online mindfulness meditation. 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 SD reduction in an index that aggregates validated measures of anxiety, depression, and stress after two weeks, and a 0.46 SD reduction after four weeks. Relative to the mindfulness-app benchmark, our midline composite effect is about one-fourth the magnitude of Headspace's four-week improvement (approximately $0.12/0.46 \approx 0.26$), but in the

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

opposite direction. We also convert our midline symptom increases of +0.85 points on the GAD-7 and +0.81 points on the PHQ-8 into Quality-Adjusted Life Years (QALYs). Using the published PHQ/GAD→EQ-5D-3L mapping in Mukuria et al. (2025), this corresponds to \$166–\$249 per treated participant at midline; if the effect persisted for 12 weeks, the implied loss would be \$332–\$498 per treated participant.

Heterogeneity analyses show that these adverse effects are stable across waves (with no detectable moderation between the pre- and post-election periods) and do not vary with engagement intensity. By contrast, users with good baseline mental health experience larger deterioration: approximately +0.18 on anxiety, +0.15 on depression, and −0.17 on the index (high mental health subgroup totals in SD units).

As further evidence, we find that treated users are more willing to follow social media accounts that post toxic content, as captured by a behavioral measure in our surveys. We interpret this toxicity-seeking behavior as a potential coping strategy for emotional regulation. Building on theories of self-medication and media-based emotion regulation—which predict that people sometimes select arousing or adversarial content to manage affect (Darden and Papageorge, 2024; Khantzian, 1985, 1997; Zillmann, 1988; Gross, 1998, 2002)—we develop a theoretical model that rationalizes our findings.

To investigate additional channels, we examine secondary outcomes capturing self-relevant evaluations and perceived discrimination. We find no evidence that treatment affects mental health through these channels, nor that it changes beliefs about altruism among other social media users. In addition, treated users do not report perceiving changes in their social media experience, corroborating the non-intrusiveness of the intervention.

By contrast, treatment is associated with higher reported loneliness—around 0.3 SD. One interpretation is that some of the removed content—although toxic—may be embedded in ongoing social exchanges (e.g., arguments, heated discussions, or critical comments) that nonetheless provide a sense of connection or social presence. When these interactions are filtered out, users' feeds may become more "sanitized" without feeling dramatically different or calmer, but their perceived social connection may weaken, leading them to report greater loneliness.

To strengthen the interpretation of our main results, we conduct a battery of additional checks. First, the treatment effects are robust to using specifications that include both midline and endline survey observations. Using this approach, we estimate the effect of -0.12 SD on the index of mental health, 0.13 SD on Anxiety, and 0.11 SD on Depression. These mag-

nitudes are very close to the midline estimates. Second, we find no evidence of differential attrition by treatment status. Survey completion rates and extension-uninstallation probabilities are nearly identical across conditions at both midline and endline. Moreover, even among treated users with the highest potential exposure to toxic content early in the study, neither survey dropout nor extension removal is measurably higher; estimated differences are close to zero across time windows. Third, our results are not driven by the choice of control group. Estimates comparing the random-hiding control with the pure control show negligible differences in anxiety, depression, and the mental-health index. Fourth, our results are insensitive to how participants enter the study—treatment effects for social-media recruits and CloudResearch panelists are almost identical. Fifth, the effects also do not vary with respect to the position of mental health questions in the survey.[2] Sixth, to address concerns about false positives, we apply the Benjamini–Hochberg false discovery rate adjustment. The anxiety and depression effects remain significant at the 5% level, and effect sizes change little relative to unadjusted estimates.

We also find no evidence that the intervention induces large changes in overall desktop social media use: estimated effects are typically below 0.5 pp. Moreover, although the intervention does not directly screen users' own posts, prior evidence using a comparable setup shows that reducing exposure to toxicity decreases users' own posting of toxic content (Beknazar-Yuzbashev et al., 2025). If anything, this channel would bias against finding deterioration in mental health, rather than in favor of them. Finally, we examine whether the intervention could affect well-being by altering the thematic mix of content rather than toxicity per se. Existing evidence suggests that such reweighting is limited: in Beknazar-Yuzbashev et al. (2025), only 2 of 26 categories move modestly when toxicity is filtered, and contentious or political topics remain largely unaffected. This makes topic substitution an unlikely explanation for the observed psychological effects.

We contribute to the literature along five dimensions. First, we isolate a content-level mechanism—exposure to toxic material—behind the mental-health costs of social media documented in deactivation and rollout designs (Allcott et al., 2020; Braghieri et al., 2022). Unlike designs that jointly alter time online, information access, and social comparison, our intervention manipulates toxicity exposure while holding platform access constant. Second, we link content moderation to individual welfare in the mental-health domain: whereas re-

[2]Mental health questions were asked as a part of a broader survey, which also included questions on political polarization.

cent field evidence shows that reducing toxicity lowers engagement and documents contagion and substitution patterns (Beknazar-Yuzbashev et al., 2025), we show that selectively filtering toxicity worsens clinically validated anxiety and depression outcomes—establishing that the engagement–welfare trade-off extends to mental health. This alignment between their engagement results and our health effects clarifies that toxicity can be simultaneously engaging and harmful.

Third, we complement evidence that toxic content can increase clicking and engagement even when welfare does not move in the same direction, drawing on large-scale feed experiments and related survey and field evidence on engagement versus welfare (Guess et al., 2023; Nyhan et al., 2023). By targeting clinically validated outcomes, our design speaks directly to the welfare margin left open in that work. Fourth, methodologically, we extend the growing toolkit of browser-extension experiments that directly manipulate on-platform content without platform collaboration—beyond prior uses that primarily measured exposure or nudged settings (Levy, 2021; Robertson et al., 2023; Aslett et al., 2022, 2024; Yu et al., 2024; Allcott et al., 2024; Farronato et al., 2024).

Lastly, our results have design and policy implications: together with platform-scale experiments and theory on moderation incentives, our estimates speak to the logic of "freedom of speech, not freedom of reach"—that is, algorithmic demotion of toxic content as a potentially welfare-improving lever—while recognizing that platforms' private incentives may diverge from social welfare (Jiménez Durán et al., 2022; Andres and Slivko, 2021; Ahmad et al., 2024; Kominers and Shapiro, 2024; Madio and Quinn, 2024). Our counterintuitive results add to this discussion by showing that curbing toxic content can also impose unintended psychological costs.

The paper is structured as follows. Section 1 discusses the experimental design. Section 3 presents the results. Section 4 rationalizes the finding using a theoretical framework. Finally, Section 5 concludes.

2 Experimental Design

2.1 Experiment Overview

We conducted a field experiment to test the impact of online toxicity on social media on users' mental health. We structured the field experiments in two waves of data collection to capture both the period before U.S. elections (campaign) and the period after the campaign

and the election concluded in the data time frame. Each wave lasted approximately 13 weeks. The first wave started on July 16, 2024 and ended on November 5, 2024, while the second wave started on May 8, 2025 and ended on September 6, 2025. Both waves presented the same structure. We recruited our respondents from two main channels: via social media ads (primarily on X and Meta) and on CloudResearch, which is a high quality research survey panel. We provide details on the recruitment procedure in Section 2.3.

The total sample of users who completed the baseline survey and installed the browser extension was N=1,398. The purpose of the browser extension was to alter the respondents' social media feeds to seamlessly hide toxic posts and comments using state-of-the-art classifiers (more details about the browsing extension are available in Section 2.2). The targeted social media platforms were: Facebook, X, and Reddit. Figure 1 provides a graphical overview of the field experiment.

In the baseline survey, we collected respondents' demographics, mental health status, and social media habits. After completing the survey, respondents were randomly assigned to experiment conditions. Participants were assigned to one of two main groups—*Control* or *Treatment*—and further split into two subconditions within each. The Control group comprised (i) a *passive* condition, in which social media feeds were not altered, and (ii) an *active* condition, in which random content was hidden. The Treatment group comprised two conditions corresponding to two toxicity-hiding algorithms, each applied to the social media feed. Random hiding in the active control group was calibrated so that, on average, the same proportion of content was hidden in this group and the treatment groups. We detail the rationale of the treatment manipulations in Section 2.4.

For the subsequent six weeks, respondents continued using their social media accounts and then completed a midline survey, which is our primary (pre-registered) outcome measurement. Respondents recruited via social media were recontacted by e-mail, while those recruited via CloudResearch were recontacted through the platform. After the midline survey, we intervened for additional six weeks before administering an endline survey using the same recontact channels. The endline survey is intended to assess the persistence of any effects observed at midline.

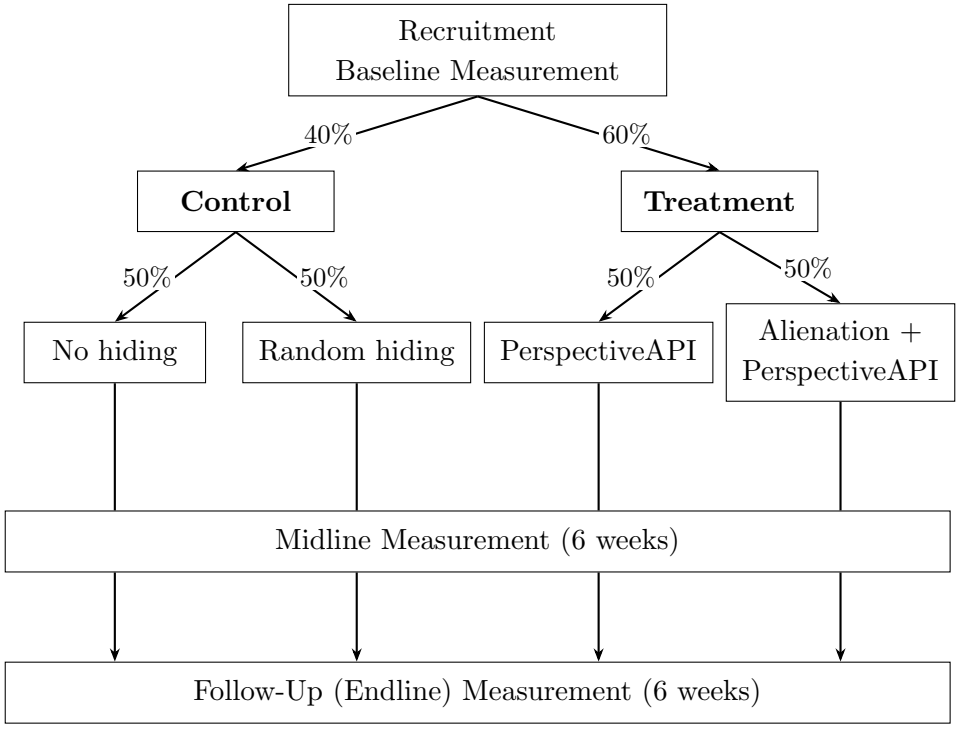


FIGURE 1: EXPERIMENTAL DESIGN OVERVIEW

2.2 Browser Extension

All participants installed our dedicated browser extension, Social Media Research, which implements the hiding intervention and records the key outcomes. The extension runs on Chromium-based desktop browsers (e.g., Chrome, Edge, Opera, Brave) and was distributed through the Chrome Web Store.[3] Browser extensions have become an increasingly common research tool for measuring online exposure and behavior in situ and for manipulating aspects of the browsing environment without platform collaboration (e.g., Levy, 2021; Beknazar-Yuzbashev et al., 2025; Beknazar-Yuzbashev and Stalinski, 2022; Aridor, 2022).

We extend this methodology by using the add-on to introduce exogenous variation in exposure to toxic content on three platforms—Facebook, X, and Reddit—while leaving platform access otherwise unchanged. The extension operates on the client side: after the platform serves and renders content in the browser, the extension modifies the page's Document Object Model (DOM) to hide targeted posts and comments. Because the intervention

[3]Chromium-based browsers cover the vast majority of desktop browsing in our study population, so focusing on this ecosystem allowed us to implement a uniform intervention while maintaining broad coverage. Consistent with the widespread adoption of browser extensions, we expected many participants to be familiar with installing and using add-ons.

does not communicate any signals back to the host server, platform ranking algorithms are unlikely to adjust the content supplied to users in response to the treatment over the study horizon. This feature is central to our identification strategy, as it reduces the scope for endogenous platform responses that could confound treatment effects.

The extension primarily relies on toxicity screening using Perspective API. Operationally, it maps text to a continuous score interpretable as the probability that the content would be labeled "toxic" by human annotators. This interpretation follows directly from how widely used toxicity models are trained: labels are aggregated across multiple raters, and the target is a fractional outcome equal to the share of annotators applying the "toxic" (or related) label to a given snippet (e.g., a score of 0.3 corresponds to roughly 3 out of 10 annotators classifying the text as toxic).[4]

2.3 Recruitment

We recruited our respondents using two different strategies. One approach was leveraging the use of social media ads to advertise the academic study on Facebook and X, while the other consisted of recruiting respondents on the online survey panel CloudResearch. Our ads were presenting our study as an opportunity to improve one's social media experience and earn a monetary reward upon the installation of the browsing extension. Upon clicking on the ads or clicking on the study link on CloudResearch, the respondents were re-directed to the Chrome Web Store on the page of our study where they could learn about the study and how to contact us for any question related to the study. Crucially, our recruitment strategy targeted specifically desktop users of the treated social media platforms.

We paid extra care to never mention the real functionalities of the browsing extension to reduce the concerns of the experimenter demand effect, focusing instead on the vague fact that the study was meant to improve their social media experience.

After completing the installation, the respondents filled in the baseline survey of our study where we informed them about the possibility of winning an additional $10 by answering two subsequent surveys within 6 and 13 weeks from that moment. For the respondents who were allocated to the treatment group, we then switched on the hiding algorithm.

[4]In the annotation protocols used to train industry-standard toxicity classifiers, "toxic" is commonly defined as content that is rude, disrespectful, or unreasonable and likely to make someone disengage from a discussion, while "very toxic" refers to highly hateful or aggressive content very likely to trigger disengagement. This framing is useful for interpreting why removing such content can mechanically affect engagement, as discussed in Beknazar-Yuzbashev et al. (2025).

2.4 Randomization and Treatment Groups

After completing the baseline survey, respondents were randomized in two stages. First, 60% were allocated to the *Treatment* arm and 40% to the *Control* arm. Second, respondents were split 50/50 within each arm, yielding four experimental conditions of equal size within arm: 30% in each Treatment condition and 20% in each Control condition.

In the passive control condition (20%), the extension did not hide any posts or comments. The only function that the extension performed was to record posts and comments the participants were exposed to and compute the toxicity scores. In the Treatment conditions, 30% of respondents were assigned to an extension that used Perspective API to flag and hide toxic posts and comments—those with scores above 0.33 (at least 1/3rd of people agree that a statement is toxic). The other 30% were assigned to toxicity hiding using a combination of Perspective (the same threshold) and Alienation, a new classifier developed by Jigsaw. Alienation is a proprietary, experimental attribute of the Google Jigsaw Perspective API. They define alienation content as language which, "Portrays someone as inferior, implies a lack of belonging, or frames the statement in an us vs. them context."

The remaining 20% were assigned to an active control condition, in which the extension hides content at the same rate as in the Treatment conditions but without targeting toxicity. This condition enables us to identify the effect of hiding toxic content *holding constant the overall amount of hidden content*, by contrasting the hiding of toxic versus non-toxic posts. To calibrate the random-removal rate used in the active control, we ran a pilot version of the experiment to estimate the average hiding rate in the Treatment conditions.

Table 1 shows balance for treatment vs. control groups on demographic variables (age, sex, education, race, political affiliation) and baseline characteristics (recruitment mode, baseline anxiety and depression). Our analysis do not indicate any detectable imbalances.

We take great care to ensure that our treatments vary exposure to toxicity alone, and do not introduce confounding effects by affecting user experience in other ways. To that end, we programmed the extension to never hide content in viewport (visible on the page at the time of being identified by the extension). This ensures that all toxic content that we hide is always hidden *off-screen*. This minimizes the chances that hiding content is ever visible to participants. The vast majority of content (more than 85%) is outside of viewport, implying that this restriction does not meaningfully reduce the first stage of our intervention. Secondly, we introduced the random hiding control group to experimentally show that our

results are driven solely by changes to toxicity exposure and not by the act of hiding a particular share of content.

As pre-registered, in our main analysis we pool respondents across the two Control conditions and across the two Treatment conditions. Our results support this decision—baseline-to-midline changes in toxicity exposure do not differ across the two Treatment algorithms.

Table 1: Balance Table by Treatment

	Control (N=540)		Treatment (N=854)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
Age	43.73	14.12	42.45	13.55	−1.27	0.76
Bachelor	0.20	0.40	0.14	0.35	−0.06	0.02
Democrat	0.45	0.50	0.45	0.50	0.00	0.03
Male	0.43	0.50	0.47	0.50	0.04	0.03
White	0.66	0.48	0.69	0.46	0.03	0.03
SampleCR	0.42	0.49	0.43	0.50	0.01	0.03
BaselinePHQ	6.35	5.81	6.19	5.66	−0.16	0.32
BaselineGAD	5.81	5.83	5.50	5.51	−0.31	0.31

Notes: This table reports summary statistics and balance tests by treatment assignment (Control vs. Treatment). Age is measured in years. Bachelor indicates completion of a bachelor's degree. Democrat indicates Democratic party affiliation. SampleCR indicates recruitment via CloudResearch (versus social media ads). BaselinePHQ and BaselineGAD are baseline mental health scores. Robust standard errors are reported.

2.5 Outcome Variables

Primary Outcomes The primary outcomes of our experiments are two psychometric scales commonly used in the medical profession to diagnosed anxiety and depression disorders: GAD-7 and PHQ8. The two scales describe two different sets of symptoms. The respondents can indicate the frequency of those symptoms in the previous two weeks. For each scale, we sum the points for each question. Moreover, we create a composite index which is the average between the PHQ8 score and the GAD7 score, which is then z-scored. Primary outcomes are elicited in all three phases of the experiment.

Mechanisms We explore different potential mechanisms that might lead to changes in people's mental health after an exogenous variation in the toxicity exposed to them on social media. Specifically, we explore two potential mechanisms: perceived discrimination and moral superiority.

On the one hand, we hypothesize that a reduction in online toxicity on social media might reduce the perceived level of discrimination in many contexts (i.e. gender discrimination, sexual harassment, racial and sex orientation discrimination, age discrimination). Therefore, the respondents might infer that the society they live in is a better place which leads to an improvement in their mental health. To measure these beliefs on perceived discrimination, we display 6 scenarios of discrimination to the respondents and elicit the prevalence of these events happening.

On the other hand, the respondents might derive some form of utility from feeling morally superior to the average social media user given the widespread toxicity on social media. Thus, we conjecture that a reduction in online toxicity might not result in an improvement in mental health, but in a potential reduction. To test this, we ask the respondents to answer 5 questions from the Marlow-Crowne index which is a scale commonly used to detect the propensity to give socially desirable answers. This index collects statements that resemble sanctity and extreme moral superiority. A high score on the index might suggest moral superiority as well as willingness to give socially desirable answers.

To rule out the latter, we ask the respondents to state their beliefs about the fraction of bonus payment that social media users donated to Save the Children USA in an unrelated experiment. The mechanisms are elicited in both midline and endline.

Behavioral Measures Beknazar-Yuzbashev et al. (2025) point out that social media users reduce their online engagement when toxicity is hidden. For this reason, we want to investigate whether people are more interested in seeking toxic content when it is hidden. If the evidence is supportive of this pattern, then it will shed light on an addictive component of toxic content.

To study the demand for toxic content, we present the respondents with the opportunity to follow different X accounts. Crucially, we cross-randomize the respondents to two conditions. In one condition, the respondents see 3 accounts that are tweeting neutral contents spanning across different neutral topics. In the other condition, the respondents see the same 3 accounts tweeting the same content as the alternative condition, but they are exposed to an additional account that is tweeting toxic content on a political topic.

We measure in both conditions how many accounts they would like to follow in a procedure similar to Item Count Method in order to reduce social desirability bias (Bursztyn et al., 2019b). If the respondents report a propensity to follow more accounts in condition

with the toxic account relative to the condition without the toxic account, then it will be evidence in favor of the toxicity seeking hypothesis.

2.6 Empirical Strategy

Treatment and outcomes. Let i index individuals and $t \in \{\text{baseline}, \text{midline}, \text{endline}\}$ index survey waves. Define D_i^{tox}, $D_i^{\text{tox+alien}}$, and D_i^{rand} as mutually exclusive assignment dummies (with *No hiding* as the omitted category). Note that D_i^{tox} corresponds to toxicity hiding based on a Perspective scores, $D_i^{\text{tox+alien}}$ corresponds to toxicity hiding based on Perspective and Alienation scores, and D_i^{rand} corresponds to random hiding a matched fraction of content irrespective of toxicity. For pooled "any toxicity removal" analyses, we set $D_i \equiv \mathbb{1}\{D_i^{\text{tox}} = 1 \text{ or } D_i^{\text{tox+alien}} = 1\}$. Let Post_t indicate periods after baseline. Outcomes include (i) a Mental-Health Index (MHI), constructed as a z-score where higher values reflect worse symptoms; (ii) standardized Anxiety and Depression indices (coded so that higher values indicate better outcomes); and (iii) raw clinical scales—the GAD-7 and PHQ-8—which we invert so that higher values denote fewer symptoms.[5]

Estimands and baseline specification. Our primary estimand is the ITT effect of toxicity hiding on mental-health outcomes. We estimate

$$y_{it} = \beta\, D_i \times \text{Post}_t \;+\; \gamma\, D_i \times \text{Post}_t \times \text{WaveTwo}_i \;+\; \alpha_i \;+\; \tau_t \;+\; \varepsilon_{it}, \tag{1}$$

where y_{it} is an outcome for individual i at time t, α_i are individual fixed effects, τ_t are time (survey-wave) fixed effects, and WaveTwo_i indicates assignment to Wave 2. Coefficient β captures the pooled ITT during Wave 1, while $\beta + \gamma$ captures the Wave 2 ITT; we report both, as well as the pooled effect with γ set to zero. Standard errors are clustered at the individual level. For transparency across outcomes, we standardize Anxiety and Depression indices to the baseline-control mean and variance and invert GAD-7/PHQ-8 so that signs are aligned (higher = better mental health), while the MHI is constructed so that improvements appear as negative coefficients.

[5] GAD-7 (Spitzer et al., 2006) and PHQ-8 (Kroenke et al., 2009) are validated measures of anxiety and depression. Toxicity scores are derived from human-annotated training data using the Perspective-style classifiers described in Wulczyn et al. (2017); see also Muralikumar et al. (2023) for external evaluation and limitations.

Multi-arm specification. To separate the two active filtering rules and the placebo control, we also estimate

$$y_{it} = \sum_{k \in \{\text{tox},\, \text{tox+alien},\, \text{rand}\}} \beta_k \, D_i^k \times \text{Post}_t \;+\; \alpha_i \;+\; \tau_t \;+\; \varepsilon_{it}, \tag{2}$$

with *No hiding* omitted. Here, β_{tox} isolates the effect of toxicity filtering at the 0.33 threshold, $\beta_{\text{tox+alien}}$ adds the Alienation screen, and β_{rand} is a placebo that absorbs generic UI/attention changes from hiding a similar share of content.

Dynamic effects. To study the timing of impacts, we estimate an event-study with periods relative to treatment onset:

$$y_{it} = \sum_{\ell \neq -1} \delta_\ell \left(D_i \times \mathbb{1}\{t - t_0 = \ell\} \right) \;+\; \alpha_i \;+\; \tau_t \;+\; \varepsilon_{it}, \tag{3}$$

where t_0 is the baseline and the $\ell = -1$ bin is omitted. Parallel pre-trends correspond to $\delta_\ell \approx 0$ for $\ell < 0$; treatment effects materialize in $\ell \geq 0$.

Identification. Under individual random assignment, the ITT estimators in (1)–(2) identify causal effects provided (i) the extension only alters visibility of targeted content; (ii) there is no differential survey measurement error across arms; and (iii) attrition is conditionally random given baseline covariates and fixed effects. The *Random hiding* arm helps rule out confounds from reduced content volume or UI salience that are unrelated to toxicity. We assess balance at baseline and explore robustness to differential attrition using the pre-specified covariate set; throughout, inference is based on individual-clustered standard errors.

Heterogeneity and secondary analyses. We pre-specify heterogeneity by baseline mental health, minority status, sex, and political views, via interactions of $D_i \times \text{Post}_t$ with the subgroup indicator (and the corresponding main effects). We further examine non-mental-health outcomes (perceived discrimination, social desirability, beliefs about others' prosociality) to probe channels distinct from symptom change, as well as willingness to follow toxic accounts on social media. All heterogeneity and secondary outcomes are reported with the same FE structure and clustering as in (1).

3 Results

3.1 First Stage

Figure 2 plots, over time, the share of toxic content shown in the feed for treated and control users, averaged at the user level. The treated series (content shown) drops to approximately zero immediately after installation and remains there, indicating that exposure suppression is strong and persistent. The share of toxic content does not differ between the active and pure control groups, consistent with the fact that the former is based on *random* hiding of posts and comments.

Table B.1 indicates that the intervention substantially reduced the likelihood that toxic content was shown to users without affecting the proportion of toxic content that the platform intended to display to users prior to filtering (content offered). In the specification without wave interactions, treatment lowers the probability that content is shown by about 7.0 percentage points, relative to a baseline of roughly 7.1% (a 98.6% reduction). Allowing for wave heterogeneity yields an almost identical treatment effect in Wave One (-0.071) and a statistically indistinguishable effect in Wave Two (sum: $-0.071 + 0.002 = -0.069$). Thus, the reduction in content shown is large in relative terms and stable across recruitment waves.

3.2 Effects of Hiding Toxic Content on Mental Health

Figure 3 report DiD estimates at midline with individual and time fixed effects. Outcomes are standardized; higher values of Anxiety and Depression indicate worse mental health, whereas the composite index is coded so that higher values indicate better mental health. The treatment increased Anxiety by 0.13 SD and Depression by 0.11 SD; the composite declines by 0.12 SD

The treatment—which reduced toxicity exposure—*worsened* mental health on both scales that we employed: GAD-7 and PHQ-8 (Table B.2) . At midline, the estimated effects are +0.85 points on the GAD-7 and +0.81 points on the PHQ-8. For scale benchmarks, the GAD-7 ranges from 0 to 21 and the PHQ-8 from 0 to 24. Thus, the point estimates correspond to roughly 4.0% of the full GAD-7 range ($0.85/21 \approx 4.0\%$) and 3.4% of the PHQ-8 range ($0.81/24 = 3.4\%$).

On GAD–7, conventional cut points are 5 (mild), 10 (moderate), and 15 (severe) (e.g., Spitzer et al., 2006). On PHQ–8, commonly used category thresholds mirror the PHQ–9:

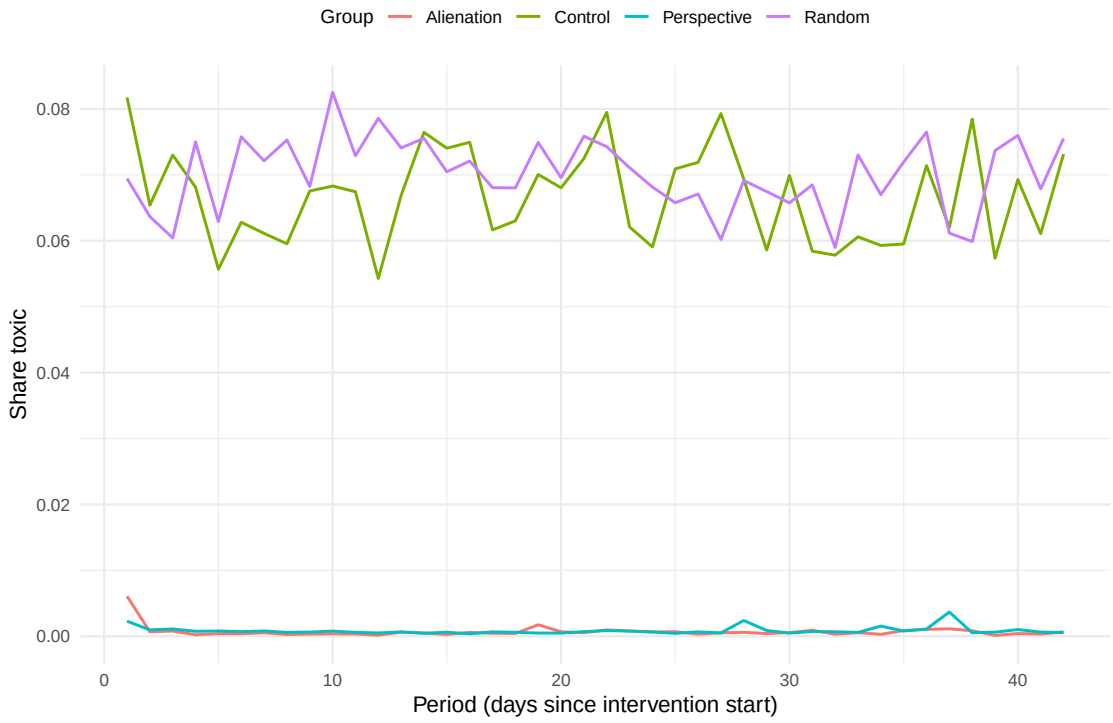


FIGURE 2: FIRST STAGE

Notes: This figure plots the share of toxic content shown to users over time (days since intervention start), separately for different treatment and control groups, averaged at the user level. We classify a post or comment as toxic if its Perspective score exceeds 0.33.

0–4 (none/minimal), 5–9 (mild), 10–14 (moderate), 15–19 (moderately severe), and 20–24 (severe) (e.g., Dhingra et al., 2011; Kroenke et al., 2009). While our average effects are modest fractions of the total scale ranges, they shift scores upward, potentially nudging some respondents near category thresholds (e.g., around 5 or 10) into a higher-severity band. For additional context, typical minimal clinically important differences are about 4 points on GAD–7 and roughly 3–4 points on PHQ–9 (often used as a guide for PHQ–8), so the average effects here are smaller than those individual-level MCIDs even though they are statistically precise in our estimates (Toussaint et al., 2020; Bauer-Staeb et al., 2021).

In the first stage, treatment substantially reduced toxicity exposure ($\Delta Tox_{\text{mid}} < 0$, $\Delta Tox_{\text{end}} < 0$). At the same time, the ITT estimates for Anxiety and Depression are positive (symptoms increase). Consequently, we compute the elasticities of reducing toxic exposure on mental health: a 1% decrease in toxicity exposure is associated with roughly a 0.1% increase in anxiety/depression symptoms over the ranges observed in our data.

Figure 4 decomposes the average effects into item responses for the GAD–7 (Panel 4a)

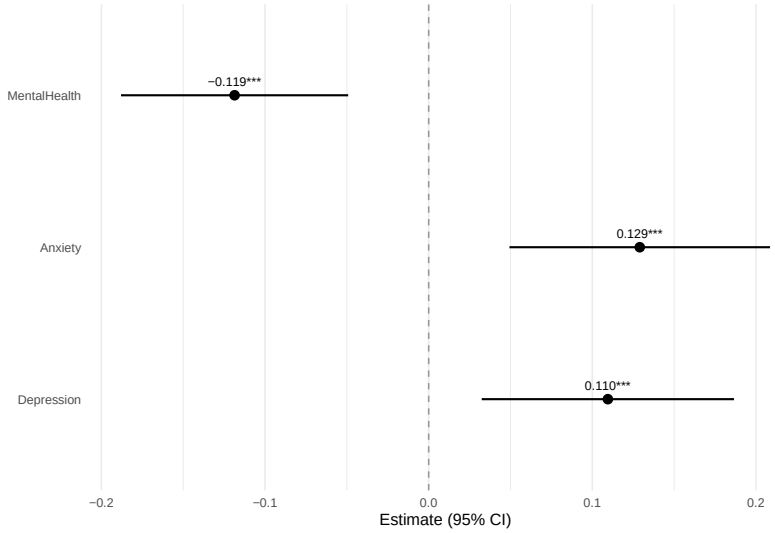


FIGURE 3: TREATMENT EFFECTS ON MENTAL HEALTH

Notes: This figure displays point estimates and 95% confidence intervals from difference-in-differences regressions (Equation 1) of standardized indices of Anxiety (GAD-7) and Depression (PHQ-8) on treatment status at midline. A higher value of Anxiety and Depression variables indicate worse mental health. Mental Health variable is a composite index based on both variables and coded so that higher values indicate better mental health. All values are in units of standard deviation. Standard errors are clustered at the individual level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

and PHQ–8 (Panel 4b) scales. Consistent with the main results, the treatment—a reduction in toxicity exposure—*worsens* mental health: positive coefficients indicate higher symptom frequency on the 0–3 item scales.

The overall deterioration in the PHQ8 is driven by three items:

- **PHQ6** (*"Feeling bad about yourself—or that you are a failure or have let yourself or your family down"*): the largest PHQ item effect.
- **PHQ3** (*"Trouble falling or staying asleep, or sleeping too much"*).
- **PHQ2** (*"Feeling down, depressed, or hopeless"*).

Other items are small and statistically indistinguishable from zero.[6] In terms of scale units, effects of 0.18–0.21 correspond to roughly one-fifth of a response category on the 0–3 item scale.

The deterioration in GAD7 is broad-based, with the largest increase on:

- **GAD7** (*"Feeling afraid as if something awful might happen"*),

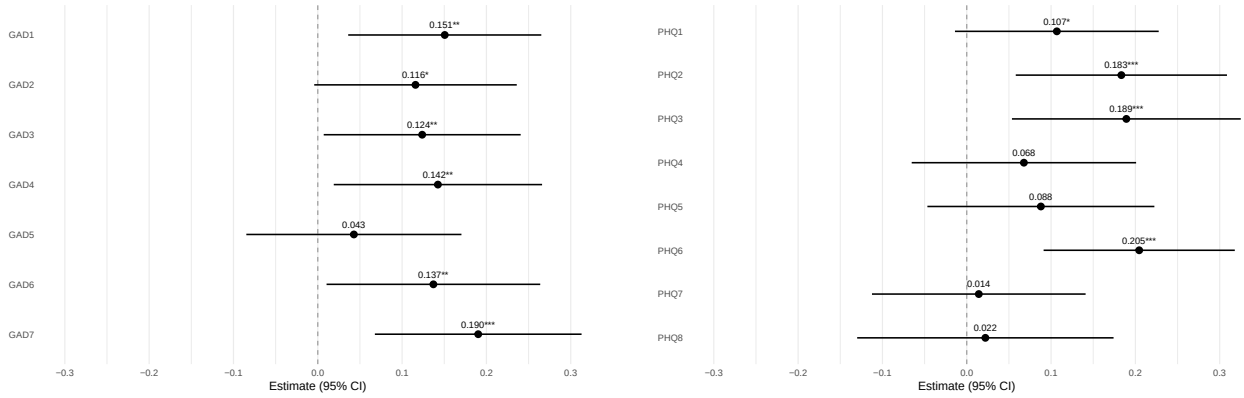
and additional significant increases on:

[6](PHQ1 interest/pleasure, PHQ4 fatigue, PHQ5 appetite, PHQ7 concentration) or slightly negative (PHQ8 psychomotor change).

- **GAD1** (*"Feeling nervous, anxious, or on edge"*),
- **GAD4** (*"Trouble relaxing"*),
- **GAD6** (*"Becoming easily annoyed or irritable"*),
- **GAD3** (*"Worrying too much about different things"*).

The restlessness item (GAD5) is small and not significant. Magnitudes of 0.12–0.19 are about one-eight to one-fifth of a response category on the 0–3 item scale.

Together, these patterns indicate that the treatment most strongly increased negative self-evaluation (PHQ6), sleep disturbance (PHQ3), and depressed mood (PHQ2), on the depressive dimension, and generalized worry/anxious arousal (GAD1–GAD4, GAD7) along with irritability (GAD6) on the anxiety dimension. While the effects are fractions of a single response category per item, they are pervasive across multiple core symptoms, which aggregates to the significant scale-level worsening documented above.



(A) GAD COEFFICIENT PLOT (B) PHQ COEFFICIENT PLOT

FIGURE 4: TREATMENT EFFECTS ON ITEMS OF GAD AND PHQ

Notes: This figure displays point estimates and 95% confidence intervals from difference-in-differences regressions (Equation 1) of individual scale items on treatment status at midline. Each item is scored 0–3 based on symptom frequency over the previous two weeks. Panel (a) shows GAD-7 items: GAD1 (feeling nervous/anxious), GAD2 (unable to stop worrying), GAD3 (worrying too much), GAD4 (trouble relaxing), GAD5 (restlessness), GAD6 (easily annoyed/irritable), GAD7 (feeling afraid). Panel (b) shows PHQ-8 items: PHQ1 (little interest/pleasure), PHQ2 (feeling down/depressed/hopeless), PHQ3 (sleep problems), PHQ4 (fatigue), PHQ5 (appetite problems), PHQ6 (feeling bad about oneself), PHQ7 (concentration problems), PHQ8 (psychomotor changes). Positive coefficients indicate higher symptom frequency. Standard errors are clustered at the individual level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Monetization of the Effect We proceed in three steps commonly used in health-economics applications. First, we map changes in the GAD-7 and PHQ-8 scores to changes in a generic

health-utility index (EQ-5D-3L) using published mapping evidence that links PHQ-9 and GAD-7 to EQ-5D-3L utility as a function of age (Mukuria et al., 2025). Second, we convert the implied utility decrement into QALYs by multiplying by the relevant exposure window; our primary (pre-registered) endpoint is the midline at six weeks, and we also report valuations over the full 12-week horizon to reflect persistence through endline. Third, we monetize the implied QALY losses using standard U.S. willingness-to-pay benchmarks. Following common practice, we report values using 100,000–150,000 per QALY as the main range, and we provide sensitivity to wider benchmarks (Neumann et al., 2014; Institute for Clinical and Economic Review , ICER). Applying this approach to our midline estimates (a +0.85 point increase in GAD-7 and a +0.81 point increase in PHQ-8), we obtain an implied per-participant welfare loss on the order of a few thousandths of a QALY over six weeks, corresponding to roughly a few hundred dollars per treated participant under conventional valuation ranges.

Heterogeneity by Baseline Mental Health. Figure 5 indicates that the treatment's adverse effects are concentrated among respondents with *high baseline mental health*. Recall that higher scores on *Anxiety* and *Depression* denote worse outcomes, and that more negative values on the *MentalHealth* index indicate deterioration.

Among those with high baseline mental health, treatment increases Anxiety by 0.18 SD ($p < 0.01$) and Depression by 0.15 SD ($p < 0.01$), corresponding to a 0.17 SD decline in the composite Mental Health index ($p < 0.01$). Effects for respondents with low baseline mental health point in the same adverse direction but are smaller in magnitude and statistically insignificant across all three outcomes (-0.07, 0.07, and 0.07 SD, respectively). The difference between subgroups is itself statistically significant for all three measures: 0.10 SD for Mental Health ($p < 0.05$), 0.11 SD for Anxiety ($p < 0.05$), and 0.08 SD for Depression ($p < 0.10$). That is, while the treatment worsens mental health for both subgroups, the deterioration is substantially larger for those who entered the study in better psychological health.

The self-medication hypothesis holds that individuals select stimuli to modulate aversive affective states in targeted ways (Khantzian, 1985, 1997). In the media domain, mood management and selective-exposure theories similarly predict that people choose content to optimize momentary mood or regulate arousal (e.g., distraction, reappraisal, catharsis) (Zillmann, 1988; Knobloch-Westerwick, 2014). Modern emotion-regulation frameworks (Gross, 1998; Gross and Souleles, 2002) help explain how specific media can serve antecedent- or

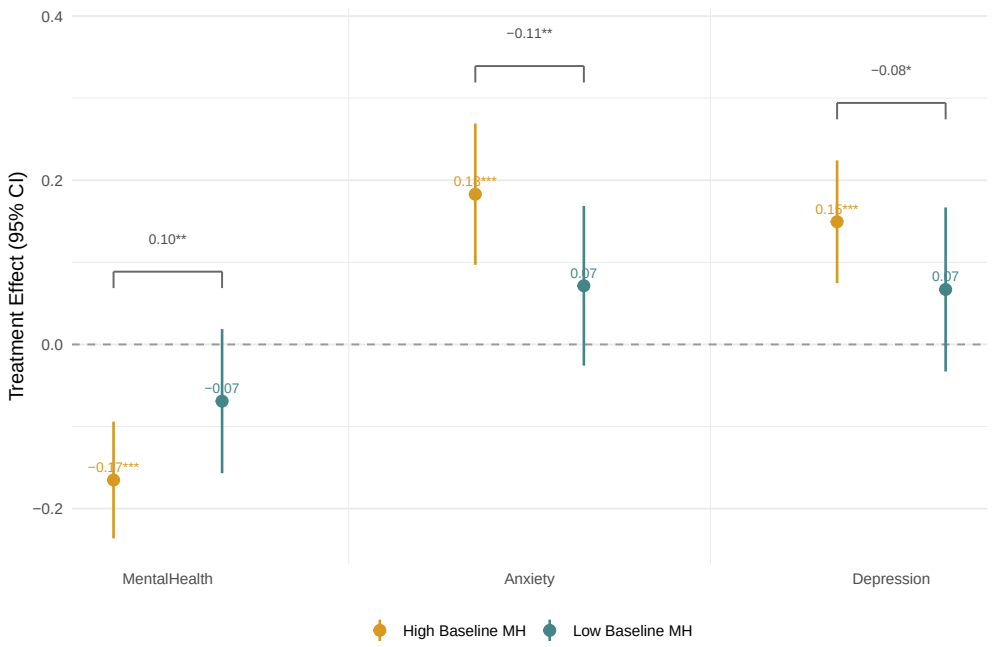


FIGURE 5: HETEROGENEITY BY BASELINE MENTAL HEALTH

Notes: This figure displays coefficient estimates and 95% confidence intervals for the interaction between treatment status and high baseline mental health. High baseline mental health is defined as scoring above the median on the baseline mental health index. Higher values of Anxiety and Depression indicate worse mental health; lower values of the Mental Health index indicate worse mental health. Standard errors are clustered at the individual level. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

response-focused strategies (e.g., attentional deployment via engrossing content, cognitive change via reframing narratives). Consistent with this view, reviews document that people *use* digital media for coping and emotion regulation, especially under stress (Wolfers and Schneider, 2021; Wolfers and Utz, 2022; Gioia et al., 2021).

If some users strategically consume provocative content to regulate affect (e.g., to vent, distract, or cognitively reframe), then algorithmically reducing such content can remove a (perceived) coping tool. Our estimates show larger harm among those with *good* baseline mental health (High MH), which is consistent with a *rational self-medication* account in which initially non-vulnerable users employed this content functionally: removing it disrupted established regulation routines, yielding worse anxiety, depression, and a lower overall mental-health index. This interpretation does not imply that such coping is universally beneficial—prolonged reliance on media for regulation can carry costs (e.g., displacement, rumination, or problematic use) even if it provides short-term relief—but it clarifies why restricting provocative content can backfire for some subgroups.

3.3 Effect of Hiding Content on Toxicity Seeking Behavior

We explore what is the effect of hiding toxic content on toxicity seeking behavior on social media in Table B.3. In the pooled ("Midline + Endline") specification, adding the toxic account does not change follow demand in the Control group (-0.045, SE $= 0.102$), but it raises demand among respondents whose feeds were previously polished by the intervention: the interaction term *Treated* $\times$ *Toxic* is 0.287 (SE $= 0.132$), implying a within-Treated increase of about 0.242 accounts (from 1.033 to 1.275), i.e., roughly 21 % to 23 % of baseline. This pattern is consistent with a *rational self-medication* interpretation grounded in mood-management and selective-exposure theory: once background exposure to toxicity is reduced, some users selectively seek higher-arousal negative content when it is available.

3.4 Effect of Hiding Toxic Content on Secondary Outcomes

For each respondent and wave, we constructed three secondary outcomes. *MoralSuperiority* is the standardized (mean 0, SD 1) index built from items capturing the sense that one's moral views or behavior are superior to others'. *Discrimination* is the mean of standardized ratings for six scenarios.[7] *DiscriminationRelevant* is the mean of standardized ratings for only those scenarios that apply to the respondent's own group membership (women; non–white; women in hiring; straight white men in hiring; everyone for supervisor mistreatment; age ≥ 65 for ageism).

Table B.4 shows no reliable effects on the secondary indices. Point estimates for *Discrimination* and *DiscriminationRelevant* are small relative to their standard errors and indistinguishable from zero in both panels. For *MoralSuperiority*, negative point estimates appear at some specifications, but they are small and not robust; we therefore treat them as null.

Given these null results, we find no evidence that changes in perceived discrimination mediate treatment effects. Prior work documents a common *illusion of moral superiority* (Tappin and McKay, 2017) and links self–enhancing beliefs to affect and well–being (Taylor and Brown, 1988). Two processes could, in principle, connect exposure reductions to affect—*downward social comparison* (Wills, 1981) and *self–affirmation* (Steele, 1988)—but

[7] 1) Women facing sexual harassment (ranging from inappropriate jokes to unwanted advances) in the workplace. 2) Members of ethnic minorities and LGBTQ+ individuals experiencing hate crime (ranging from abusive language to assault). 3) Despite having equal qualifications, women are not getting hired for a managerial role. 4) Despite having equal qualifications, straight white men are not getting hired because they do not add to the diversity of the workplace. 5) Supervisors in a workplace being excessively critical and demeaning towards employees. 6) Older individuals are denied equal opportunities because of their age.

our estimates provide at most weak evidence of any shift in *MoralSuperiority*, and no change in discrimination perceptions.[8]

Table B.5 likewise shows no significant effects on perceived altruism of other users. If anything, treated respondents rate others as slightly less altruistic and do not report feeling morally superior in relative terms, but these differences are not statistically distinguishable from zero.

The pattern in Table B.6 suggests that hiding toxic content from users' feeds leaves their reported social media experience largely unchanged on most dimensions, but is associated with a sizeable and statistically significant increase in loneliness.

Across both the midline and the pooled specifications, the treatment effect on FOMO (Fear of missing out), perceived changes in the feed, and anger is small in magnitude and imprecisely estimated: coefficients are close to zero and not statistically significant. This is consistent with the idea that participants do not experience the toxicity-removal intervention as a salient or disruptive change in how they use or feel about the platform in general. In other words, cleaning up the feed does not seem to make people feel notably less angry, nor does it make them feel that "something is different" or that they are missing out more or less.

By contrast, the treatment is consistently associated with higher reported loneliness: around 0.30–0.32 SD at midline and in the pooled specification, significant at conventional levels. The fact that the loneliness effect is stable in size and precision across both sets of models (and survives individual and time fixed effects) suggests a robust unintended consequence of the intervention. One interpretation is that some of the content being removed—although toxic—might also be embedded in ongoing social exchanges (e.g., arguments, heated discussions, or critical comments) that nonetheless provide a sense of connection or social presence. When those interactions are filtered out, users' feeds may become more "sanitized" without feeling dramatically different or calmer, but their underlying sense of social connection appears to weaken, leading them to report feeling lonelier.

3.5 Robustness Checks

Figure 6 summarizes the results of our robustness checks.

[8] Perceived discrimination is generally associated with poorer mental health (e.g., Pascoe and Richman, 2009), but our null effects suggest this is not the operative pathway here.

Including Endline Observations. Using a specification that leverages both midline and endline observations, we estimate treatment effects almost identical in magnitude to those estimated using the main specification (that compares midline to baseline). In particular, we report an effect of -0.12 SD on the index of mental health, 0.13 SD on Anxiety, and 0.11 SD on Depression.

Heterogeneity by Waves. Figure 6 show little evidence that the treatment's mental–health effects differ by recruitment wave. In the pooled model, treatment worsens outcomes on all three measures. The wave–interaction terms are small and statistically indistinguishable from zero.

Translating these to wave–specific totals (main + interaction), the treatment effects in *Wave Two* (2025), relative to *Wave One* (pre–2024 U.S. elections), are similar. Because the interaction estimates are close to zero and imprecise, we do not detect meaningful variation in treatment harm across the pre-election (Wave One) and post-election (Wave Two) periods. Substantively, this suggests that the adverse mental–health consequences of reducing provocative content are not confined to a specific political cycle or the heightened salience of the 2024 election period; the pattern generalizes to the subsequent year.

Active Users. Figure 6 shows that the adverse mental-health effects of treatment do not meaningfully vary with engagement intensity. Across five engagement definitions—active after one week, active after three weeks, above-median active days, above-median active time, and high baseline media usage—the interaction terms are small and statistically insignificant. The main treatment effects remain stable and significant throughout. We find no evidence that greater engagement amplifies or attenuates treatment harm.

CloudResearch vs. Social Media. The treatment effect does not differ by recruitment source. The main effects for the social-media subsample are significant across all outcomes (*MentalHealth* -0.148, *Anxiety* $+0.168$, *Depression* $+0.129$), while the *Treated* $\times$ *CloudResearch* interactions are small and insignificant ($+0.045$ [SE 0.055], -0.061 [SE 0.068], -0.031 [SE 0.057]). The results are not driven by a specific recruitment subsample.

Order Effects. Re-estimating the main models with an interaction for whether mental-health questions appeared before or after the polarization block yields no evidence of priming

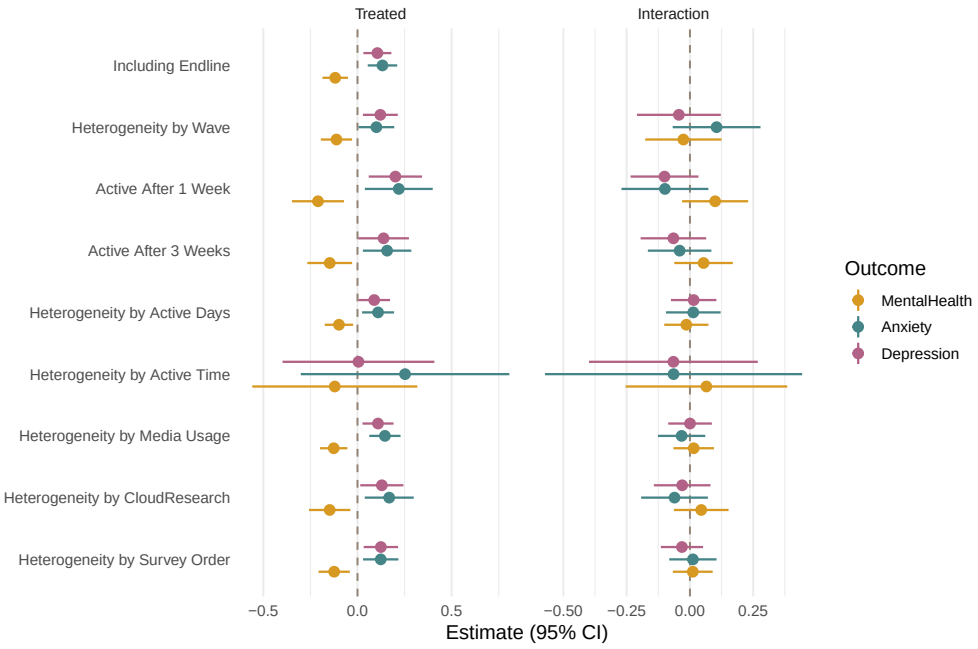


FIGURE 6: ADDITIONAL HETEROGENEITY ANALYSES AND ROBUSTNESS CHECKS

Notes: This figure displays coefficient estimates and 95% confidence intervals comparing the treatment group with the control group under different specifications. The panel shows midline estimates. Standard errors are clustered at the individual level.

or fatigue confounds. The main treatment effects are stable (*MentalHealth* −0.124, *Anxiety* +0.123, *Depression* +0.124), and the *Treated* × *MentalHealthFirst* interactions are negligible (+0.011 [SE 0.040], +0.012 [SE 0.048], −0.031 [SE 0.043]).

Random Hiding Only. Figure C.2 contrast the treatment group with the random-hiding control at midline. Across Anxiety, Depression, and the composite index, differences are perfectly in line, if anything stronger, with those found in our main results, supporting our pooling of the two controls in the main analysis.

Attrition. We find no systematic differential attrition by treatment. Figure C.1a and Figure C.1b show that survival rates over time, based on users' activity, are virtually identical across experimental conditions. Pooling waves, treatment assignment does not predict midline survey completion: the treatment–control difference is −0.007 percentage points (SE 0.028) against a baseline completion rate of approximately 50% (Table B.7, Column 1). This null holds when we interact treatment with high baseline mental health (Column 2), control for predicted survival probability (Column 3), and estimate attrition separately within the control and treatment subsamples (Columns 4–5).

4 Toxic Content as Self-Medication

4.1 Environment

Time is discrete $t = 0, 1, 2, \ldots$. Each individual i has a stock of mental health $M_{it} \in \mathbb{R}_+$. Baseline symptoms are $s^0_{it} = s(M_{it}, \varepsilon_{it})$ with $s_M < 0$ and ε_{it} an i.i.d. affective shock (stressful day, bad work news, rumination, etc.).

There are two media goods:

- **Background toxic content** $b_{it} \geq 0$: posts/comments that the platform would display to the user for free, absent the extension. Think of this as "naturally supplied toxicity."
- **Searched toxic content** $x_{it} \geq 0$: additional toxic content the user actively looks for (follows a toxic account, clicks through, switches platform, disables the extension on mobile, etc.). This is the counterpart of x_t in Darden and Papageorge (2024).

Toxic content—whether background or searched—is an *affect-regulation tool*. Consuming $q_{it} \equiv b_{it} + x_{it}$ units delivers instantaneous relief:

$$s_{it} = s^0_{it} - \phi(q_{it}), \qquad \phi'(\cdot) > 0, \ \phi''(\cdot) < 0. \tag{4}$$

The idea is perfectly consistent with media-based emotion-regulation and mood-management theories: people use media to distract, to vent, to affiliate, or even to *wallow* when wallowing is what restores perceived fit with their current mood (Zillmann, 1988; Knobloch-Westerwick, 2015; Gross, 1998, 2002).

The cost is that toxic content aggravates future mental health:

$$M_{i,t+1} = (1-\delta)M_{it} + \underbrace{I_{it}}_{\text{other investments}} - \psi(q_{it}) + \eta_{i,t+1}, \qquad \psi'(\cdot) > 0. \tag{5}$$

Hostile exchanges, doom-scrolling, and morally outraged content are physiologically and cognitively depleting; they also keep people inside online dynamics that amplify outrage and negativity (Brady et al., 2021; Van Bavel et al., 2024). Appendix A provides additional background to our theoretical results.

4.2 Platform intervention as a price shock

Your browser extension does *not* change what the platforms *offer*; it only prevents a subset from being *shown*. In the notation above this means: the offer b^o_{it} stays the same in control

and treatment, but the *effective* free background supply becomes

$$b_{it} = \begin{cases} b^o_{it} & \text{if } i \text{ in control,} \\ (1-\theta)b^o_{it} & \text{if } i \text{ in treatment,} \end{cases}$$

with $\theta \in (0,1)$ the hiding rate empirically close to 1.[9] Importantly, users can still *search* for toxicity, but searched units now cost $c(\cdot)$ in time/attention/frictions. The cost of x_{it} can be written $p_t x_{it}$ with p_t the marginal search cost; the extension raises p_t by blocking the cheapest units of toxicity.

4.3 Individual problem

In each period the individual chooses x_{it} and I_{it} to maximize

$$U = \sum_{t=0}^{\infty} \beta^t \left[u\big(-s(M_{it}, \varepsilon_{it}) + \phi(b_{it} + x_{it}) \big) - p_t x_{it} - C(I_{it}) \right], \tag{6}$$

subject to (5), $x_{it} \geq 0$, $I_{it} \geq 0$. Take $u'(\cdot) > 0$, $u''(\cdot) < 0$, $C'(\cdot) > 0$. We can think of u as turning "lower symptoms" into flow utility.

Let λ_{it} be the multiplier on (5). The FOC for x_{it} is

$$u'\big(\,\cdot\,\big)\,\phi'(q_{it}) = p_t + \lambda_{it}\psi'(q_{it}). \tag{7}$$

The left-hand side is the *marginal benefit of extra toxic content*: it is higher when current symptoms are bad (so u' is high) and when toxicity is affectively potent (high ϕ'). The right-hand side is the *full marginal cost*: out-of-pocket search cost p_t plus the shadow value of the damage to mental-health capital, $\lambda_{it}\psi'(\cdot)$.

We observe:

- In control, $p_t = p$ and $b_{it} = b^o_{it}$.
- In treatment, $p_t = \bar{p} > p$ (search is harder) and $b_{it} = (1-\theta)b^o_{it} < b^o_{it}$, so q_{it} falls even if x_{it} stayed the same.

[9] Our first-stage figure shows shown-toxicity essentially at 0 while offered-toxicity tracks control; this is exactly $b_{it} = (1-\theta)b^o_{it}$ with $\theta \approx 1$.

4.4 Static mental-health effect

Evaluate (4) at the post-intervention optimum. Let q^C and q^T denote control and treated total toxicity. Then

$$s^T - s^C = \big(s^0 - \phi(q^T)\big) - \big(s^0 - \phi(q^C)\big) = \phi(q^C) - \phi(q^T) > 0$$

whenever $q^T < q^C$, because $\phi' > 0$. The extension *must* make q fall for at least some users, because both the free component b and the price of x move in the wrong direction. Hence the model necessarily delivers worse symptoms on average.

Proposition 1 (Content removal worsens mental health) *Suppose u and ϕ are strictly increasing and concave and ψ is weakly convex. For any individual with an interior demand for toxic content in the pre-treatment regime, a policy that (i) lowers the free background level b and (ii) weakly raises the user price p strictly lowers optimal total toxicity q^* and strictly raises contemporaneous symptoms s. Formally,*

$$(b', p') \prec (b, p) \quad \Rightarrow \quad q^*(b', p') < q^*(b, p) \quad \Rightarrow \quad s(b', p') > s(b, p).$$

4.5 Toxicity seeking

Differentiate the FOC (7) totally with respect to b:

$$u''\phi' \frac{\partial s}{\partial b} + u'\phi'' \left(1 + \frac{\partial x}{\partial b}\right) = \lambda\psi'' \left(1 + \frac{\partial x}{\partial b}\right).$$

Because $u'' < 0$, $\phi'' < 0$, and $\psi'' \geq 0$, the term multiplying $(1 + \partial x/\partial b)$ is *negative*. Solving for $\partial x/\partial b$ yields

$$\frac{\partial x}{\partial b} = -1 + \frac{u''\phi' \, \partial s/\partial b}{\lambda\psi'' - u'\phi''}.$$

The first term is -1: if you give one more unit of free toxicity, the person would like one less unit of searched toxicity. The second term is small in magnitude (it is essentially a curvature correction). Hence generically

$$\frac{\partial x}{\partial b} \approx -1,$$

i.e. if you remove one free unit ($\mathrm{d}b = -1$), the optimal searched quantity rises by almost exactly one unit. This is *compensatory seeking* and is a direct analogue of the self-medication substitution toward alcohol in Darden and Papageorge (2024). In our experiment the exten-

sion made $\mathrm{d}b \ll -1$ (it basically made $b \approx 0$), so the model says x should jump up—precisely what we find in the follow-account task.

Proposition 2 (Compensatory search) *Let $x(b,p)$ be the user's optimal searched toxicity. If $\phi'' \leq 0$ and $\psi'' \geq 0$, then $x_b(b,p) \leq -1$ locally around an interior solution. Hence, a one-unit reduction in free toxicity triggers at least a one-unit increase in searched toxicity, up to the corner $x = 0$.*

4.6 Why the healthy get hurt

The key object in (7) is λ_{it}, the marginal value of preserving one unit of mental-health stock. In a standard health-capital problem this multiplier is *larger* when the person is sicker, because the marginal value of health is higher. Here, however, the extension acts like a *temporary* price increase on self-medication. People who start from *very poor* mental health (M_{i0} low) are typically already at the corner $x_{it} = 0$ even in control: the future penalty $\lambda_{it}\psi'(0)$ is too high relative to their current benefit. For them, reducing b_{it} barely moves total q_{it}, so symptoms hardly change.

By contrast, people who start from *good* mental health (M_{i0} high) have small λ_{it}: losing a bit of stock tomorrow is not very costly. They therefore choose an *interior* level of q_{it} in control. For an interior solution we can write the Hicksian demand for x_{it} as

$$x_{it} = x\big(p_t, b_{it}; M_{it}, \varepsilon_{it}\big), \quad x_p < 0,\ x_b < 0.$$

If we now execute the platform intervention, both p_t and $(1-\theta)$ move unfavorably, so x_{it} must fall unless the person compensates again by conscious search outside the extension. But even if they do search, they cannot quite restore q_{it} to the pre-intervention level because each extra followed account or off-platform visit is costly. Hence q_{it} declines *most* for those who had an interior x to begin with—the good-mental-health group. Since s_{it} is decreasing in q_{it}, this group displays the largest *increase* in anxiety/depression measures. This reproduces your heterogeneity result.

Formally:

Proposition 3 (Heterogeneous treatment harm) *Suppose (i) there exists $\bar{M}$ such that for $M_{it} < \bar{M}$ the optimal choice is $x_{it} = 0$ when $b_{it} = b^o$ and $p_t = p$; and (ii) for $M_{it} > \bar{M}$ the solution is interior. Then for any policy $(b^o, p) \to ((1-\theta)b^o, \bar{p})$ with $\theta \in (0,1)$ and $\bar{p} > p$*

we have

$$s_{it}^T - s_{it}^C = 0 \quad \textit{for } M_{it} < \bar{M},$$

and

$$s_{it}^T - s_{it}^C > 0 \quad \textit{for } M_{it} > \bar{M}.$$

Intuition. Low-M types are already at $x = 0$ and marginal utility from toxicity is low relative to the shadow value of mental health. Cutting b therefore just removes a good they were barely using. High-M types equate marginal benefit to marginal cost; cutting b or raising p upsets that equality, so they must move left along a concave $\phi(\cdot)$. With concavity, a small leftward move raises symptoms more than proportionally. ■

4.7 Discussion of the Assumptions

Three established strands in communication and media-psychology make our assumptions precise:

1. **Selective exposure and mood repair.** Decades of work show that media choice is goal-driven: people in a given affective state pick content that helps them *reach the state they want*, not necessarily the most pleasant content. In an environment of media abundance, users will even choose *negatively valenced* or *indignant* material if it fits their regulatory goal (e.g. to maintain arousal, to affirm values, to signal virtue). Knobloch-Westerwick (2015) summarizes this for modern, on-demand media. This gives us $\phi'(\cdot) > 0$ even when the content is toxic.
2. **Media-induced recovery.** Experimental work collected in Reinecke and Eden (2016) demonstrates that even tense, dramatic, or sad clips can restore subjective energy and psychological resources *when they are absorbing*. Recovery does not require hedonic positivity; it requires attention capture and temporary detachment. That is exactly what an outraged thread or a snarky, toxic post does. This supports writing the contemporaneous benefit as $\phi(q)$ instead of assuming that toxic content is always costly.
3. **Depressive Internet memes and online venting.** Clinical and cyberpsychology studies show that people with emotion-regulation difficulties often rate "depressive" or self-deprecating memes as *helpful*, because they enable expression, affiliation, and humor (Akram et al., 2020). That is psychologically very close to using negative, even hostile political content to vent. This strand says that a user can rationally select *negative* content for a *positive* regulatory function, which is the key step in our

reinterpretation of self-medication.

Taken together, these strands mean that the crucial assumptions—(i) toxic content delivers short-run relief, (ii) it may be reached through effortful search, and (iii) removing it can hurt people who relied on it—are all anchored in existing literature.

5 Conclusions

This paper provides the first randomized field evidence on the mental-health consequences of content-level moderation that hides toxic material 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 individual and time fixed effects, treated participants exhibit increases in symptom burden on both the GAD-7 and PHQ-8 and a decline in a composite mental-health index. The estimated impacts are modest in absolute magnitude (0.12 standard deviations) but statistically precise and broad-based across clinically validated outcomes.

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), 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 and social externalities (Allcott et al., 2022). 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 generic changes in feed volume or interface salience. Together with recent large-scale evidence that reducing toxicity lowers engagement (Beknazar-Yuzbashev et al., 2025), our results imply that the engagement–welfare trade-off extends to mental health: content that is both prevalent and engaging can also function as an input into users' affective states in ways that are not uniformly improved by moderation.

Our heterogeneity results and behavioral evidence point toward an emotion-regulation interpretation. The adverse effects are concentrated among users with *good baseline mental health*: for this subgroup, toxicity hiding produces substantially larger deteriorations. Con-

sistent with theories of self-medication and media-based emotion regulation, which predict that individuals sometimes select arousing or adversarial content to manage affect, treated users also display greater willingness to follow accounts posting toxic content when given the opportunity, suggesting compensatory demand for such material. At the same time, we find little evidence that the mental-health effects operate through perceived discrimination, moral-superiority beliefs, or changes in perceived altruism of others. Instead, the most robust secondary outcome is an increase in reported loneliness, consistent with the possibility that some toxic material is embedded in social interaction that nonetheless provides a sense of connection or social presence; filtering it can make feeds feel more sanitized without alleviating negative affect, while weakening perceived social connectedness.

Several limitations motivate future work. First, classifier-based interventions necessarily inherit the imperfections of automated toxicity detection, including context dependence and potential differential error across groups; continued validation, auditing, and transparency around thresholds remain important. Second, external validity beyond desktop users and the three platforms covered by our extension is an open question, particularly in environments where substitution across outlets is easier or where moderation policies induce equilibrium responses by platforms or users. Third, our primary horizon is weeks; longer-run adaptation, habit formation, and dynamic responses in content selection are natural next steps. Finally, while we focus on anxiety and depression, other welfare margins—including loneliness and sleep, stress biomarkers, and broader measures of life satisfaction—may reveal additional trade-offs and help clarify when, and for whom, curbing toxic content improves welfare.

References

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

Akram, Umair, Jennifer Drabble, Glhenda Cau, Fraser Hershaw, Ashileen Rajenthran, Mollie Lowe, Carissa Trommelen, and Jason G. Ellis, "Exploratory study on the role of emotion regulation in perceived valence, humour and beneficial use of depressive internet memes in depression," *Scientific Reports*, 2020, *10* (899).

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.

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

Aridor, Guy, "Drivers of digital attention: Evidence from a social media experiment," Technical Report, CESifo Working Paper 2022.

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.

_ , Ziv Sanderson, William Godel, Nathaniel Persily, Jonathan Nagler, and Joshua A. Tucker, "Online searches to evaluate misinformation can increase its perceived veracity," *Nature*, 2024, *625* (7995), 548–556.

Bauer-Staeb, Clarissa, Daphne-Zacharenia Kounali, Nicky J. Welton, Emma Griffith, Nicola J. Wiles, Glyn Lewis, Julian J. Faraway, and Katherine S. Button, "Effective Dose 50 Method as the Minimal Clinically Important Difference: Evidence from Depression Trials," *Journal of Clinical Epidemiology*, 2021, *137*, 200–208.

Bavel, Jay J. Van, Colin E. Robertson, Karen del Rosario, Jacob Rasmussen, and Steve Rathje, "Social media and morality," *Annual Review of Psychology*, 2024, *75*, 311–340.

Beknazar-Yuzbashev, George and Mateusz Stalinski, "Do social media ads matter for political behavior? A field experiment," *Journal of Public Economics*, 2022, *214*, 104735.

_ , 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.

Brady, William J., Killian McLoughlin, Tuan N. Doan, and Molly J. Crockett, "How social learning amplifies moral outrage expression in online social networks," *Science Advances*, 2021, *7* (33), eabe5641.

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

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," NBER Working Paper 26567, National Bureau of Economic Research 2019.

_ , _ , _ , and _ , "Social media and xenophobia: evidence from Russia," Technical Report, National Bureau of Economic Research 2019.

Darden, Michael E and Nicholas W Papageorge, "Rational self-medication," *Economics & Human Biology*, 2024, *53*, 101350.

Dhingra, Satvinder S., Kurt Kroenke, Matthew M. Zack, Tara W. Strine, and Lina S. Balluz, "PHQ-8 Days: A Measurement Option for DSM-5 Major Depressive Disorder (MDD) Severity," *Population Health Metrics*, 2011, *9*, 11.

Durán, Rafael Jiménez, Karsten Müller, and Carlo Schwarz, "The Effect of Content Moderation on Online and Offline Hate: Evidence from Germany's NetzDG," CEPR Discussion Paper 17554, CEPR 2022.

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

Gioia, Francesca, Valeria Rega, and Valentina Boursier, "Problematic Internet Use and Emotional Dysregulation Among Young People: A Literature Review," *Clinical Neuropsychiatry*, 2021, *18* (1), 41–54.

Gross, David B. and Nicholas S. Souleles, "Do Liquidity Constraints and Interest Rates Matter for Consumer Behavior? Evidence from Credit Card Data," *Quarterly Journal of Economics*, 2002, *117* (1), 149–185.

Gross, James J., "The Emerging Field of Emotion Regulation: An Integrative Review," *Review of General Psychology*, 1998, *2* (3), 271–299.

_ , "Emotion Regulation: Affective, Cognitive, and Social Consequences," *Psychophysiology*, 2002, *39* (3), 281–291.

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.

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.

Khantzian, Edward J., "The Self-Medication Hypothesis of Addictive Disorders: Focus on Heroin and Cocaine Dependence," *American Journal of Psychiatry*, 1985, *142* (11), 1259–1264.

_ , "The Self-Medication Hypothesis of Substance Use Disorders: A Reconsideration and Recent Applications," *Harvard Review of Psychiatry*, 1997, *4* (5), 231–244.

Knobloch-Westerwick, Silvia, *Choice and Preference in Media Use: Advances in Selective Exposure Theory and Research*, New York: Routledge, 2014.

_ , *Choice and Preference in Media Use: Advances in Selective Exposure Theory and Research*, New York: Routledge, 2015.

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.

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

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

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.

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.

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.

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.

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

Reinecke, Leonard and Allison L. Eden, "Media use and recreation: Media-induced recovery as a link between media exposure and well-being," in Leonard Reinecke and Mary Beth Oliver, eds., *The Routledge Handbook of Media Use and Well-Being*, New York: Routledge, 2016, pp. 106–117.

Robertson, Rebecca E., Jonathan Green, Daniel J. Ruck, Katherine Ognyanova, Christo Wilson, and David Lazer, "Users choose to engage with more partisan news than they are exposed to on Google Search," *Nature*, 2023, *618* (7964), 342–348.

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.

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.

Toussaint, Anne, Paul Hüsing, Antje Gumz, Katja Wingenfeld, Martin Härter, Elisabeth Schramm, and Bernd Löwe, "Sensitivity to Change and Minimal Clinically Important Difference of the 7-Item Generalized Anxiety Disorder Questionnaire (GAD-7)," *Journal of Affective Disorders*, 2020, *265*, 395–401.

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

Wolfers, Lara N. and Frank M. Schneider, "Using Media for Coping: A Scoping Review," *Communication Research*, 2021, *48* (8), 1070–1094.

_ **and Sonja Utz**, "Social Media Use, Stress, and Coping," *Current Opinion in Psychology*, 2022, *45*, 101305.

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.

Zillmann, Dolf, "Mood Management Through Communication Choices," *American Behavioral Scientist*, 1988, *31* (3), 327–340.

For online publication only:

Social Media Toxicity, Emotions, and Mental Health: Experimental Evidence

George Beknazar-Yuzbashev Francesco Capozza
Kylan Rutherford Mateusz Stalinski

A Proofs for Section 4

This appendix supplies existence and the proofs of Propositions 1–3.

A.1 Preliminaries

Let time be $t = 0, 1, \ldots$ and define

$$q_t \equiv b_t + x_t, \qquad s_t \equiv s(M_t, \varepsilon_t) - \phi(q_t),$$

with $s_M < 0$, $\phi' > 0$, $\phi'' \leq 0$. Mental health evolves as

$$M_{t+1} = (1 - \delta)M_t + I_t - \psi(q_t) + \eta_{t+1}, \tag{8}$$

with $\delta \in (0, 1)$, $\psi' > 0$, $\psi'' \geq 0$. Period utility is

$$u_t = u(-s_t) - p_t x_t - C(I_t),$$

where $u(\cdot)$ is increasing and concave, $C(\cdot)$ is increasing, convex, differentiable, and $\lim_{I\to\infty} C'(I) = \infty$. The agent discounts at $\beta \in (0, 1)$.

We make two standard assumptions:

(A1) u, ϕ, ψ, C are continuous and twice continuously differentiable on $\mathbb{R}_+$.

(A2) b_t and p_t are bounded and nonnegative, and (ε_t, η_t) are bounded supports.

Under (A1)–(A2) the Bellman operator associated with this problem is a contraction on the space of bounded continuous functions and a value function V exists and is concave in (M_t, b_t); see, e.g., *Stokey, Lucas and Prescott (1989), ch. 4*. We therefore solve the problem period by period.

A.2 One-period problem and FOC

Because our experiment is short and because x_t only shows up in $q_t = b_t + x_t$, it is enough to study the static problem for a fixed (M, b, p) and fixed continuation value $V_M \equiv \beta\mathbb{E}[V_M(M_{t+1})]$:

$$\max_{x\geq 0} \; F(x; b, p, M) \equiv u\big(-s(M, \varepsilon) + \phi(b + x)\big) - px - V_M\psi(b + x). \tag{9}$$

Write $q \equiv b + x$. Then

$$F_x = u'(\cdot)\,\phi'(q) - p - V_M\psi'(q).$$

An interior optimum $x^* > 0$ must satisfy

$$u'(\cdot)\,\phi'(q^*) = p + V_M\psi'(q^*). \tag{10}$$

Note that $F_{xx} = u''(\cdot)(\phi'(q^*))^2 + u'(\cdot)\phi''(q^*) - V_M\psi''(q^*) \le 0$ because $u'' \le 0$, $\phi'' \le 0$, $\psi'' \ge 0$. Hence the interior solution, when it exists, is unique.

A.3 Proof of Proposition 1 (content removal worsens symptoms)

Proof. Take two regimes:

$$\mathcal{C} : (b, p) = (b^o, p) \quad \text{(control)}, \qquad \mathcal{T} : (b, p) = ((1-\theta)b^o, p+\Delta p),\ \theta \in (0,1),\ \Delta p \ge 0 \quad \text{(treatment)}.$$

Let x^C and x^T be the respective interior optima. By Proposition 2,

$$x^T = x^C - \theta b^o + \underbrace{\frac{\partial x}{\partial p}}_{<0}\Delta p + o(\theta b^o, \Delta p).$$

Hence total toxicity in treatment is

$$q^T = (1-\theta)b^o + x^T = (1-\theta)b^o + x^C - \theta b^o + x_p\Delta p + o(\cdot) = b^o + x^C + x_p\Delta p + o(\cdot) = q^C + x_p\Delta p + o(\cdot).$$

Because $x_p < 0$ (higher price reduces optimal search) and $\Delta p \ge 0$, we have $q^T \le q^C$ with strict inequality whenever $\Delta p > 0$. Since current symptoms are

$$s^R = s(M, \varepsilon) - \phi(q^R), \qquad R \in \{\mathcal{C}, \mathcal{T}\},$$

and $\phi' > 0$, a weak fall in q implies a weak rise in s. Therefore $s^{\mathcal{T}} \ge s^{\mathcal{C}}$, with $>$ if $\Delta p > 0$. This is exactly the experimental contrast: the extension both hid content and raised the implicit user cost of seeing it again, so treatment users must exhibit (weakly) worse contemporaneous symptoms. ■

A.4 Proof of Proposition 2 (compensatory search)

Proof. Consider (9) at an interior optimum. Because $q = b + x$ enters only through the sum, we have

$$\frac{\partial F_x}{\partial b} = u''(\cdot)(\phi'(q))^2 + u'(\cdot)\phi''(q) - V_M\psi''(q) = \frac{\partial F_x}{\partial x} = F_{xx}.$$

By the implicit-function theorem,

$$\frac{\partial x^*}{\partial b} = -\frac{F_{xb}}{F_{xx}} = -\frac{F_{xx}}{F_{xx}} = -1.$$

So, locally around an interior point, a one-unit fall in free background toxicity b causes a one-unit rise in searched toxicity x^*. This is an *exact* result and it rests only on the additivity $q = b + x$ and on twice differentiability. At the corner $x^* = 0$, the derivative is $\partial x^*/\partial b \in [-1, 0]$ by standard Kuhn–Tucker envelope conditions: the agent cannot reduce x below zero, so the most x can rise when b falls is 1. This proves the proposition. ■

A.5 Proof of Proposition 3 (heterogeneous treatment effects)

Proof. Let $V_M(M)$ be the marginal value of mental-health capital in the pre-treatment regime. From the envelope condition of the dynamic problem,

$$V_M(M) = u'(\cdot)\, s_M(M, \varepsilon) + \beta(1 - \delta)\, \mathbb{E}[V_M(M') \mid M].$$

Because $s_M < 0$ and $u' > 0$, the first term is negative; the second term is positive. Standard arguments for concave health-capital models imply that $V_M(M)$ is *decreasing* in M: when you are healthier, an extra unit of health is worth less. (Formally, $V_{MM} < 0$ because the flow marginal utility is decreasing in M and the continuation value is concave in M.)

Return to the static FOC (10). For a given (b, p), define

$$G(q; M) \equiv u'(\cdot)\, \phi'(q) - V_M(M)\, \psi'(q).$$

Our FOC is $G(q; M) = p$. The derivative of G with respect to M is

$$\frac{\partial G}{\partial M} = -V'_M(M)\psi'(q) > 0,$$

because $V'_M(M) < 0$ and $\psi'(q) > 0$. Hence high-M individuals have a *higher* $G(\cdot)$ schedule

and therefore a larger interior solution $q^*(M)$ for any given (b, p). Intuitively, healthier people internalize the future penalty less and so consume more toxic content.

Now impose treatment: $(b, p) \to ((1-\theta)b, p+\Delta p)$. For individuals already at the corner $x = 0$ in control, we have $q^C = b$ and, after treatment, $q^T = (1-\theta)b < q^C$: these people suffer a treatment-induced drop in q of exactly θb. For individuals at the interior in control, apply the implicit-function theorem to $G(q; M) = p$:

$$\frac{\partial q^*}{\partial p} = \frac{1}{-G_q} < 0, \qquad \frac{\partial q^*}{\partial b} = 0 \quad \text{(because } q = b + x \text{ and } x_b = -1\text{)}.$$

So their total toxicity falls by $|q_p|\Delta p$. Because $G_q = u''(\cdot)(\phi')^2 + u'\phi'' - V_M\psi'' \leq 0$, and V_M is *smaller* for high-M types, we have

$$|q_p(M_{\text{high}})| = \frac{1}{-G_q(M_{\text{high}})} > \frac{1}{-G_q(M_{\text{low}})} = |q_p(M_{\text{low}})|,$$

i.e. the high-M types reduce total toxicity more when p rises. Since $s = s^0 - \phi(q)$ with $\phi' > 0$, this also means

$$s^{\mathcal{T}}(M_{\text{high}}) - s^{\mathcal{C}}(M_{\text{high}}) > s^{\mathcal{T}}(M_{\text{low}}) - s^{\mathcal{C}}(M_{\text{low}}) \geq 0,$$

which is the heterogeneity pattern observed in the data: those with better baseline mental health (large M) suffer the largest symptom increase when toxic content is removed. ■

A.6 Linking back to the experiment

We can now read the four treatment arms in model terms:

- **Pure control:** $(b, p) = (b^o, p)$.
- **Random hiding:** $(b, p) = ((1-\theta)b^o, p)$ but the content removed is random, so ϕ does *not* decline much because non-toxic high-arousal material still shows up. Hence mental health changes little.
- **Toxicity (Perspective) and Toxicity+Alienation:** $(b, p) = ((1-\theta)b^o, \bar{p})$ *and* the remaining content is, by design, less suited for venting, so ϕ *and* b fall. This is the configuration that produces the large rise in s and the compensatory search.

The browser extension matters because it shifts p_t *without* changing b^o_{it}—the platform never learns. In the random-hiding arm the platform sets $(b, p) = ((1-\theta)b, p)$ but the nature of the content—i.e. its suitability for venting—does not change. In the model that is

equivalent to a pure b-shock. By Proposition 2, interior users exactly offset a pure b-shock ($x_b = -1$), so q and hence s do not move. This lines up with your empirical finding that the random-hiding arm is effectively a placebo on mental-health outcomes.

B Online Tables

Table B.1: Effects on content shown and offered

	Content Shown	Content Shown	Content Supplied	Content Supplied
(Intercept)	0.071***	0.071***	0.071***	0.072***
	(0.002)	(0.003)	(0.002)	(0.003)
Treated	−0.070***	−0.071***	0.010***	0.009**
	(0.002)	(0.003)	(0.003)	(0.004)
WaveTwo		−0.002		−0.003
		(0.005)		(0.005)
Treated × WaveTwo		0.002		0.004
		(0.005)		(0.007)
Num.Obs.	1270	1270	1270	1270

Notes: This table reports first-stage estimates of treatment effects on the share of toxic content in users' feeds. "Content Shown" is the share of toxic content actually displayed to users. "Content Supplied" is the share of toxic content the platform would have shown absent the intervention. The unit of observation is the user-day. Columns 1 and 3 report pooled treatment effects; Columns 2 and 4 allow for wave heterogeneity. Robust standard errors in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table B.2: Main results: treatment effects on mental health outcomes

	MentalHealth	Anxiety	Depression	GAD	PHQ
Treated	−0.119***	0.129***	0.110***	0.854***	0.813***
	(0.036)	(0.041)	(0.040)	(0.267)	(0.281)
Num.Obs.	1762	1762	1762	1763	1763
FE: ID	X	X	X	X	X
FE: Time × Intervention Start	X	X	X	X	X

Notes: This table reports difference-in-differences estimates from Equation 1. The dependent variables are standardized indices of Anxiety (GAD-7), Depression (PHQ-8), and a composite Mental Health index. Higher values of Anxiety and Depression indicate worse mental health; higher values of Mental Health indicate better mental health. All columns use midline observations only (6 weeks post-baseline); Columns 4–5 report non-standardized GAD-7 (scale from 0 to 21) and PHQ-8 (scale from 0 to 24) scores. All specifications include individual and time fixed effects. Standard errors clustered at the individual level are in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table B.3: Treatment Effects on Toxicity Seeking Behavior

	Accounts Followed: Midline	Accounts Followed: Midline + Endline
(Intercept)	1.075***	1.161***
	(0.085)	(0.078)
Treated	−0.051	−0.123
	(0.110)	(0.100)
Follow	0.043	−0.047
	(0.126)	(0.102)
Treated × Follow	0.224	0.285**
	(0.163)	(0.132)
Num.Obs.	699	1076

Notes: This table reports treatment effects on willingness to follow social media accounts. The dependent variable is the number of accounts respondents indicate they would follow. Respondents were cross-randomized to see either 3 neutral accounts (Follow = 0) or 3 neutral accounts plus 1 toxic account (Follow = 1), following an item-count method design to reduce social desirability bias. The Treated × Follow interaction captures whether treated respondents show differential demand for the toxic account. Robust standard errors in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table B.4: Treatment Effects on Perceived Discrimination

	Midline lhs: Discrimination	Midline lhs: DiscriminationRelevant	Midline + Endline lhs: Discrimination	Midline + Endline lhs: DiscriminationRelevant
Treated	0.309	0.063	0.476	0.843
	(0.924)	(1.086)	(0.881)	(1.007)
Num.Obs.	1398	1398	1840	1840
FE: ID	X	X	X	X
FE: Time	X	X	X	X

Notes: This table reports treatment effects on perceived discrimination. Discrimination is the mean of standardized ratings for six discrimination scenarios. Discrimination Relevant restricts to scenarios relevant to the respondent's own group membership. Higher values indicate greater perceived prevalence of discrimination. Standard errors clustered at the individual level are in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table B.5: Treatment Effects on Altruism and Moral Superiority

	Midline lhs: Altruism	Midline lhs: Rel.Superiority	Midline lhs: Superiority	Midline + Endline lhs: Altruism	Midline + Endline lhs: Rel.Superiority	Midline + Endline lhs: Superiority
Treated	−0.078	−0.151*	0.038	−0.091	−0.096	0.055*
	(0.081)	(0.080)	(0.034)	(0.073)	(0.073)	(0.032)
Num.Obs.	1398	1398	1396	1840	1840	1838
FE: ID	X	X	X	X	X	X
FE: Time	X	X	X	X	X	X

Notes: This table reports treatment effects on beliefs about others' altruism and moral superiority. Altruism measures beliefs about the fraction of bonus payment other social media users donated to charity in an unrelated experiment. Standard errors clustered at the individual level are in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively

Table B.6: Treatment Effects on Fomo, Changes in Experience, Loneliness, Anger

	Midline lhs				Midline + Endline lhs			
	Fomo	Changes	Loneliness	Anger	Fomo	Changes	Loneliness	Anger
Treated	0.065	0.162	0.319***	0.055	0.080	0.198	0.298***	0.067
	(0.134)	(0.157)	(0.104)	(0.121)	(0.116)	(0.148)	(0.089)	(0.102)
Num. Obs.	378	378	378	378	449	449	449	449
FE: ID	X	X	X	X	X	X	X	X
FE: Time	X	X	X	X	X	X	X	X

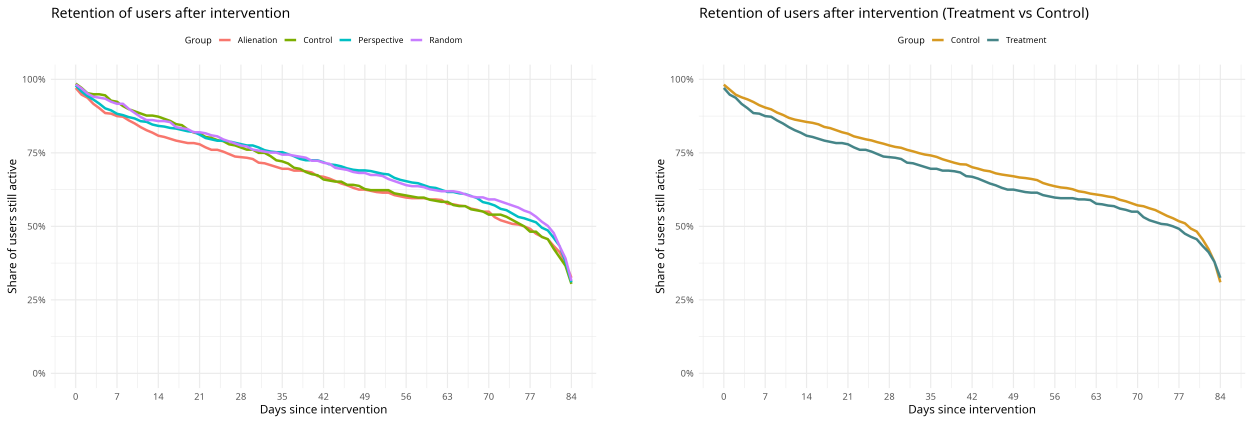
Notes: This table reports treatment effects on alternative outcomes measuring social media experience. Fomo measures fear of missing out. Changes captures whether respondents perceived changes in their social media feed. Loneliness and Anger are self-reported emotional states. Standard errors clustered at the individual level are in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table B.7: Midline Attrition

	(1)	(2)	(3)	(4)	(5)
(Intercept)	0.506***	0.496***	0.041	0.506***	0.481***
	(0.022)	(0.030)	(0.047)	(0.031)	(0.023)
GroupTreatment	−0.007	−0.043	−0.067		
	(0.028)	(0.039)	(0.058)		
HighMentalHealthTRUE		0.019			
		(0.043)			
GroupTreatment:HighMentalHealthTRUE		0.070			
		(0.055)			
PredSurvival			0.927***		
			(0.081)		
GroupTreatment:PredSurvival			0.119		
			(0.102)		
GroupFullRandom				−0.001	
				(0.043)	
GroupFullToxicity					0.040
					(0.034)
Num.Obs.	1394	1394	1394	540	854

Notes: This table reports OLS estimates of midline survey completion on treatment status. The dependent variable is an indicator equal to one if the respondent completed the midline survey. Column 1 estimates the overall treatment–control difference. Column 2 interacts treatment with an indicator for high baseline mental health (above-median composite index). Column 3 controls for predicted survival probability and its interaction with treatment. Columns 4 and 5 estimate attrition for active control vs. pure control and Alienation + Perspective vs. Perspective treatment subsamples, respectively. Standard errors are in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

C Online Figures



(A) Survival plot by Classifier (B) Survival plot by Treatment

Figure C.1: Survival Plot of Activity

Notes: This figure displays survival curves showing the share of users still active (using the browser extension) over time since intervention start. Panel (a) shows survival by classifier type (Alienation, Perspective, Control, Random); Panel (b) shows survival by treatment status (Treatment vs. Control).

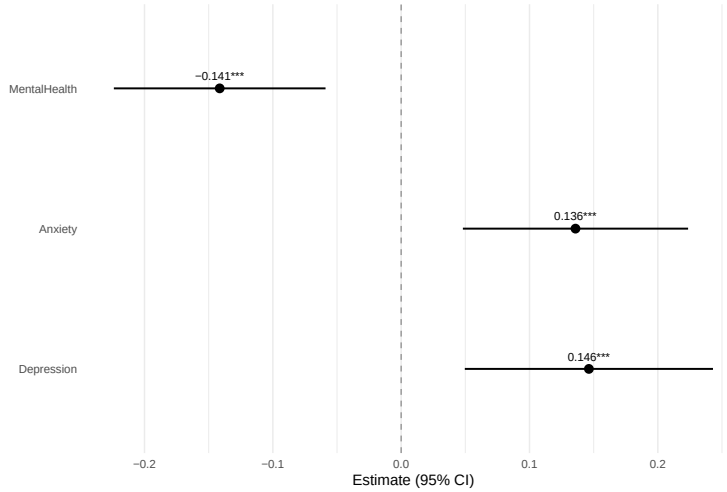


FIGURE C.2: RANDOM HIDING EFFECTS AGAINST PURE CONTROL

Notes: This figure displays coefficient estimates and 95% confidence intervals comparing the treatment group with the random-hiding control group only. The sample is restricted to treatment group and random-hiding control group respondents. The panel shows midline estimates. Standard errors are clustered at the individual level.

D Instructions

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