Toxic Content and User Engagement on Social Media: Evidence from a Field Experiment*

George Beknazar-Yuzbashev[†] **Rafael Jiménez-Durán**[‡]
Jesse McCrosky[§] **Mateusz Stalinski**[¶]

This draft: January 19, 2025
First Draft: December 29, 2022

Abstract

Most social media users have encountered harassment online, but there is scarce evidence of how this type of toxic content impacts engagement. In a pre-registered browser extension field experiment, we randomly hid toxic content for six weeks on Facebook, Twitter, and YouTube. Lowering exposure to toxicity reduced advertising impressions, time spent, and other measures of engagement, and reduced the toxicity of user-generated content. A survey experiment provides evidence that toxicity triggers curiosity and that engagement and welfare are not necessarily aligned. Taken together, our results suggest that platforms face a trade-off between curbing toxicity and increasing engagement.

JEL Classification: C93, D12, D83, D90, I31, L82, L86, M37, Z13

Keywords: toxic content, moderation, social media, user engagement, browser experiment

*This research is funded by the Becker Friedman Institute at the University of Chicago and the Program for Economic Research at Columbia University. We are grateful to Luca Braghieri, Leonardo Bursztyn, Annalí Casanueva Artís, Ruben Durante, Sarah Eichmeyer, Ruben Enikolopov, Chiara Farronato, Matthew Gentzkow, Horacio Larreguy, Alex Imas, Ro'ee Levy, John List, Alexey Makarin, Alejandra Agustina Martínez, Nikita Melnikov, Suresh Naidu, Maria Petrova, Andrea Prat, Chris Roth, Carlo Schwarz, Andrey Simonov, David Strömberg, Egon Tripodi, Joshua Tucker, and numerous seminar and conference participants for helpful comments and suggestions. Furthermore, we are indebted to the Mozilla Foundation for their help in promoting the study and to Onar Alili for his help in developing the browser extension. The browser experiment was approved by the University of Chicago Institutional Review Board (IRB22-0073) and the Columbia University Institutional Review Board (AAAT9887). It was pre-registered in the American Economic Association Registry (AEARCTR-0009628 and AEARCTR-0010138). The survey experiment was approved by the Humanities and Social Sciences Research Ethics Committee at the University of Warwick (HSSREC 101/23-24). It was pre-registered in the American Economic Association Registry (AEARCTR-0014362).

[†]Columbia University gb2683@columbia.edu.

[‡]Bocconi University, IGIER, Chicago Booth Stigler Center, and CESifo rafael.jimenez@unibocconi.it.

[§]Mozilla Foundation mccrosky@gmail.com.

[¶]University of Warwick and CAGE mateusz.stalinski@warwick.ac.uk.

1 Introduction

More than seven in ten Americans are active on social media (Kemp, 2024), and a majority of users report experiencing some form of online harassment during their lifetimes (Anti-Defamation League, 2024). Due to the links between inflammatory content and violence (Bursztyn et al., 2019; Müller and Schwarz, 2020, 2023b), as well as the impact of social media on mental health (Allcott et al., 2020, 2022; Braghieri et al., 2022) and politics (Zhuravskaya et al., 2020), the incentives of platforms to curb hate speech, harassment, and related forms of content—hereafter referred to as "toxic"—have been under public scrutiny in recent years.

This scrutiny has centered on the connection between toxic content and user engagement. A prevalent hypothesis is that social media algorithms, which are trained to maximize various forms of engagement, may inadvertently amplify toxic material.[1] This concern stems from seminal work in social psychology (Baumeister et al., 2001) showing that negative events and emotions have a disproportionate impact on human behavior—likely making negativity particularly engaging. However, credible causal evidence on how toxic content impacts user engagement, the mechanisms driving this effect, and its welfare effects remains scarce. One important reason for this gap is the challenge of naturally varying the toxicity of content displayed to users.

We overcome this challenge through a field experiment targeting three major social media platforms: Facebook, Twitter (currently X), and YouTube. We recruited 742 social media users, mostly through Twitter ads, to install a custom-built desktop browser extension. This extension recorded their online activity—encompassing over 11 million pieces of content consumed and 30,000 hours of social media use—and hid toxic text content across all three platforms *in real time*. By doing so, the intervention induced exogenous variation in users' exposure to toxicity.

We randomized participants into two groups: a control group without any hiding and a treatment group in which the extension seamlessly hid toxic material on all three platforms during a six-week period. To classify text content as toxic, the extension relied on a machine-learning algorithm that predicts the fraction of human annotators likely to consider the content toxic, defined as "a rude, disrespectful, or unreasonable comment that is somewhat likely to make you leave a discussion or give up on sharing your perspective."[2] Following a two-week

[1] Frances Haugen, Facebook's whistleblower, voiced this concern in her 2021 disclosure: "As long as your goal is creating more engagement, optimizing for likes, reshares and comments, you're going to continue prioritizing polarizing, hateful content," see: https://www.wsj.com/articles/facebook-whistleblower-frances-haugen-says-she-wants-to-fix-the-company-not-harm-it-11633304122, accessed: 2024-12-23.

[2] We use Unitary's Detoxify library, an open-source algorithm trained on a large dataset of online comments. Its performance, measured by the Area Under the Receiver Operating Characteristic Curve, was high at 0.9864 out of 1, only 0.22% lower than the top performer in Kaggle's 2018 Toxic Comment Classification Challenge.

baseline period of passively collecting users' online activity, the extension—without notice—began hiding all content with a toxicity score above 0.3. This threshold is meant to capture content that three out of ten annotators would label as toxic or very toxic. While toxicity algorithms are inevitably prone to subjectivity and measurement error (Gröndahl et al., 2018), our intervention was designed to mimic platforms' actual content moderation practices of hiding, deprioritizing, or "shadowbanning" content using toxicity thresholds (Katsaros et al., 2022; Ribeiro et al., 2022).

As a result of the intervention, the average toxicity score of text content displayed to users in the treatment group was 73% (0.9 standard deviations, SD) lower than in the control group during the treatment period. As a benchmark, the magnitude of this reduction resembles the difference in average toxicity exposure between users in the 50th and 2nd percentiles during the baseline period. Overall, the intervention hid 7% of posts, comments, and replies displayed in the browser across the three platforms for users in the treatment arm. Given that we recruited heavy desktop users who self-reported spending 61% and 56% of their Twitter and Facebook time, respectively, on a desktop device, we conclude that our intervention led to a substantial reduction in exposure to toxicity on social media, even when accounting for mobile app usage. We also infer that the recommendation algorithms did not respond to our intervention since we observe an identical average toxicity between the posts shown to the control group and those intended to be shown to the treatment group *before* hiding.

We now proceed to report our main findings. As pre-registered, our empirical strategy employs a difference-in-differences specification, which exploits both between- and within-individual variation to increase statistical power (McKenzie, 2012). The first set of results focuses on various engagement measures. Our preferred engagement index aggregates different metrics similar to those used in practice by platforms: active time spent, content consumed, reactions, posts, reposts, browsing sessions, ad impressions, ad clicks, and post clicks.[3] The hiding treatment reduced this index by 0.054 SD across all three platforms, with overall stronger effects on Facebook, followed by Twitter and YouTube. Notably, we see an average decrease of 1.3 minutes per day on Facebook, or 9.2% relative to the mean. As a benchmark, Allcott et al. (2022) reduce social media time by 56 minutes per day by paying users $2.50 per hour reduced. Additionally, we find decreases of 2.3 and 0.5 ad impressions per day on Facebook and Twitter (where ad data were available), corresponding to declines of 27% and 6% relative to their respective means. A

As benchmark, this performance compares to state-of-the-art spam detection algorithms (Tusher et al., 2024).

[3]For example, Twitter ranks posts using a score given by a weighted average of various metrics: https://github.com/twitter/the-algorithm-ml/tree/main/projects/home/recap, accessed: 2025-01-01.

back-of-the-envelope calculation using ad prices suggests a drop in the revenue of both platforms of $440 across all users during our intervention.[4] Taken together, we find that reducing exposure to toxicity on social media decreases user engagement across a variety of metrics.

Second, we find that hiding toxic content led to a substitution effect in active time spent on 38 pre-registered related websites where the intervention did not take place. On average, users spent 1.8 more minutes per day on these non-treated platforms, representing a 22% increase relative to the mean. This increase was primarily driven by additional time spent on other social media websites such as Reddit.

Lastly, we report evidence in favor of toxicity being contagious. The treatment—which did not hide any content produced by users themselves—significantly reduced the average toxicity of posts and comments created by our users both on Facebook and Twitter, respectively, by 30% and 25% relative to the mean.

One potential concern with interpreting our results is that the intervention may have reduced engagement through mechanisms beyond the reduction in toxicity. For example, deviating from an engagement-optimized algorithm could have mechanically lowered engagement, or hiding toxic posts might have altered the composition of content (e.g., reducing exposure to political posts). However, it is not the case that deviating from an engagement-maximizing algorithm will mechanically decrease engagement. For instance, a prominent study by Nyhan et al. (2023) provides evidence against this concern by showing that decreasing exposure to like-minded content on Facebook had a null effect on time spent. Moreover, we conduct an LLM-based topic analysis and find that the composition of topics was mostly unaffected by our intervention.

Even though we targeted heavy browser users, a key internal validity concern in any browser-based study is the potential for users to substitute browser usage with phone usage, leading to measurement error. We assuage these concerns using Twitter API data, which allows us to observe the number of posts created by our users and their source (mobile app vs. browser). While this evidence is only available for Twitter, we do not find that mobile and browser usage are substitutes: users post fewer tweets from their mobile app, easing concerns that our results are driven by a substitution away from the browser.

Lastly, a common internal validity concern in browser studies is attrition. In our study, the attrition rate after baseline is low—10.6% over the six weeks of intervention—compared to the average attrition rate of 15% reported in a meta-analysis of experiments published in economics journals (Ghanem et al., 2023). Furthermore, there is no significant differential attrition between

[4]The cost per thousand impressions in 2022 was $6.72 for Facebook in the US and $1.74 on Twitter. See https://www.guptamedia.com/cpmentry, accessed 2025-01-11.

treatment arms, and baseline exposure to toxicity and time spent do not predict attrition. Our main results remain robust to a range of additional checks, such as accounting for the staggered nature of treatment—which spanned three weeks—using a "stacked" specification (Cengiz et al., 2019; Baker et al., 2022) or alternative clustering of standard errors.

The previous evidence notwithstanding, the exact mechanism through which exposure to toxic content drives social media engagement remains unsettled. Its welfare effects also remain unclear: because users engage less after our intervention, a commonly made revealed-preference argument would conclude that they are worse off. However, as we argue in a companion theory paper (Beknazar-Yuzbashev et al., 2024), changes in engagement are not a reliable proxy for changes in welfare; they can move in opposite directions. For example, consider a user who dislikes encountering toxic posts but, upon seeing them, cannot resist reading the comment section and possibly replying. In this case, a higher fraction of toxic posts in their feed would increase their engagement but decrease their welfare.

To investigate the mechanisms and to measure the welfare effects of encountering toxicity, we conduct a complementary survey experiment with 4,120 participants recruited via Prolific. The experiment varies the type of posts shown to respondents, with some users seeing more toxic posts than others. Besides varying the level of toxicity, we also varied their type: some users encountered a hateful (but not profane) post while others encountered a profane (but not hateful) post. We measure two main outcomes: whether respondents click to view the comment sections of the posts, as an engagement metric, and their willingness to accept (WTA) to participate in a future task requiring them to read similar posts, as a welfare metric.

We find that respondents who encounter more toxic posts are 6.1 percentage points (18% relative to the mean) more likely to click to view the comment sections of the posts. This effect is not driven by remuneration considerations: respondents know that viewing the comments does not give them additional payments—if anything, it carries an opportunity cost for their time. A question asking their recall of the posts at the end of the survey rules out that the effect is driven by differential attention. This evidence—based on a different experimental design and sample—aligns with the results of our field experiment, thereby reinforcing the external validity of our findings. These results also further ease concerns that our field evidence on engagement is mechanically driven by a mere deviation from users' engagement-optimized algorithms, and reinforce that exposure to toxicity directly affects user engagement.

In terms of welfare, the evidence is mixed. There is suggestive evidence that respondents require a higher WTA to read the hateful post compared to a similar neutral post, which,

based on a compensating differential argument, indicates a loss in welfare. In contrast, offering respondents a post with profanity versus a similar but less toxic post does not significantly affect their welfare—statistically or economically. A more qualitative approach confirms that respondents see these posts differently: they consider the hateful post as less entertaining than a similar non-toxic post, while the profanity post as more entertaining than a similar non-toxic post. These results suggest that toxic posts may trigger participants' *curiosity*, prompting them to click and uncover comments, but with differing welfare implications depending on the type of toxicity. Besides providing some evidence of mechanisms, these findings confirm that engagement and welfare are not necessarily aligned.

The findings in this paper suggest a trade-off for platforms. Taking our results at face value, a similar intervention would decrease exposure to toxic content on these websites—directly and indirectly due to the contagion effect on content creation. However, this decrease would come at the cost of a lower engagement, ad clicks, and impressions. If ad prices do not increase enough to compensate, platform revenue would decrease.[5] Additionally, our evidence on spillovers to engagement on other websites suggests that platforms might not fully internalize the benefits of curbing toxicity. This evidence, paired with our findings that changes in engagement might not correspond to changes in welfare, suggests that platforms' private incentives to curtail toxicity might not necessarily align with social incentives.

This paper contributes to three strands of the literature. First, a burgeoning literature studies the effects of social media on a variety of outcomes, including political expression (Artís Casanueva et al., 2024; Enikolopov et al., 2020; Fujiwara et al., 2024; Guriev et al., 2021; Petrova et al., 2021), polarization (Allcott and Gentzkow, 2017; Boxell et al., 2024; Melnikov, 2021), hate crimes (Müller and Schwarz, 2020, 2023b; Bursztyn et al., 2019; Jiménez-Durán et al., 2022), and mental health and well-being (Allcott et al., 2020, 2022; Braghieri et al., 2022; Bursztyn et al., 2023)—see Zhuravskaya et al. (2020) and Aridor et al. (2024) for recent reviews. We contribute to this work by shedding light on one of the potential explanations for the documented harmful effects of social media, namely, the exposure of users to toxic content.

A subset of this literature has studied the effects of social media algorithms on user behavior, particularly on polarization (Levy, 2021; Nyhan et al., 2023). A recent study from the U.S. 2020

[5]While causal evidence on the elasticity of advertiser demand to toxicity is lacking, anecdotal evidence suggests that it might be small. Many advertisers reported brand safety concerns after Elon Musk's Twitter takeover in 2022. However, even if average ad prices (cost per thousand impressions) fell from $1.75 in 2022 to $0.68 in 2023, this decrease was historically small and ad prices reverted to $2.23 in 2024 (see Footnote 4). Moreover, Ahmad et al. (2024) find that advertisers are unaware of misinformation near their ads. This unawareness could also limit the response of ad prices.

Facebook and Instagram Election project shows that algorithms increase user engagement and exposure to uncivil content (Guess et al., 2023). Kalra (2024) shows similar evidence from a TikTok-like platform in India. To the best of our knowledge, our paper provides the first evidence of the causal effects of exposure to toxic content on user engagement and welfare.

This paper also belongs to a rapidly-growing economics literature that studies content moderation, including the empirical evaluation of policies countering misinformation (Henry et al., 2022; Guriev et al., 2023) and hate speech (Andres and Slivko, 2021; Müller and Schwarz, 2023a), as well as theoretical and structural work (Acemoglu et al., 2021; Liu, 2020; Liu et al., 2021; Madio and Quinn, 2024; Germano et al., 2022; Kominers and Shapiro, 2024). Previous empirical work could not isolate the effect of a reduced exposure on potential viewers, because most moderation interventions bundle the removal of material with sanctions to the content creators. For instance, when Twitter suspended Donald Trump on January 2021, users were less exposed to his posts but were also aware of his suspension. Isolating the effect of exposure is crucial because platforms may adopt strategies that balance exposing users to toxic content—thereby driving engagement—with imposing sanctions on the reported content, which can also enhance engagement (Jiménez Durán, 2022; Jiménez-Durán et al., 2022). Additionally, to the best of our understanding, ours is the first experimental evidence of substitution to other websites in response to a decrease in toxicity (see also Agarwal et al. (2022); Rizzi (2024) for more evidence on substitution in response to platform changes). Moreover, while Kim et al. (2021) provide evidence of the contagion of toxicity in a survey experiment, ours is the first causal evidence of contagion in the field.

Methodologically, this paper contributes to the literature by introducing a browser extension experimental design that directly manipulates content displayed to individuals. Browser extensions have been used in experiments (Aridor et al., 2024), primarily to record the content that individuals encounter (Levy, 2021; Beknazar-Yuzbashev and Stalinski, 2022; Farronato et al., 2023; Robertson et al., 2023; Aslett et al., 2024; Aridor, ming) but also to alter their social media settings and histories (Yu et al., 2024; Beknazar-Yuzbashev et al., 2024) and to provide information and nudges (Aslett et al., 2022; Yu et al., 2024; Zavolokina et al., 2024). Directly manipulating the content that users are exposed to is a particularly useful methodology in settings where platform collaboration is challenging. Similar applications include altering search results (Farronato et al., 2024), manipulating cookie consent interfaces (Farronato et al., 2024), and hiding ads (Allcott et al., 2024).[6] Lastly, we add to early applications of generative AI

[6]See Farronato et al. (2024) for an open-source browser extension tool for researchers.

to classification of media content (e.g., Djourelova et al., 2023) by using GPT to identify ads based on texts of social media posts and to categorize content by topic.

In what follows, Section 2 provides background information. Sections 3 and 4 outline the design and results of the browser and survey experiments, respectively. Section 5 concludes.

2 Background

2.1 Supported Platforms

Our hiding intervention encompasses three leading social media platforms: Facebook, YouTube, and Twitter (currently X). As of April 2024, the former two can boast of the top highest global number of users—3.1 billion (rank 1) and 2.5 billion (rank 2), respectively, with the latter's user base being 611 million.[7] The platforms are equally popular in the United States as they are worldwide. According to Pew Research, in 2024, 70% of US adults reported using Facebook. The proportion was equal to 85% for YouTube and 21% for Twitter/X.[8] With Facebook and YouTube selected for their sheer size and overall influence, we added Twitter to our analysis due to its special role as a modern digital agora, facilitating the dialogue between public figures and their followers, as well as politicians and the electorate.

An important aspect of our intervention is that we focus on hiding toxic *text* content. This feature makes Twitter and Facebook particularly suitable for our study due to their text-based discussion format. Specifically, Twitter encourages exchanges of brief statements, with a character limit of 280 symbols, while Facebook houses plenty of communities in the form of groups, supporting familial, professional, political, and other thematic discussions. YouTube differs from Facebook and Twitter in that the user's primary objective is watching videos, with the comment sections being an additional element. Beyond the three platforms, we measured user activity (time spent) on 38 additional sites (including Reddit, Quora, and Parler), where treatment did not take place.[9]

[7] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/, accessed: 2024-12-29.

[8] https://www.pewresearch.org/internet/fact-sheet/social-media/, accessed: 2024-12-29.

[9] Our platform choices warrant a question of why to stop at three. One reason is that the hiding intervention requires that the extension code be tailored to the DOM structure of each website on which it operates. Frequent alterations made by the websites' developers necessitate constant and careful maintenance of the add-on, which can only be extended to a limited number of platforms. Another factor that played a role in our decision was our interest in the spillovers from social media with the hiding intervention enabled to other related websites where the treatment did not apply.

2.2 Browser Extension

All participants installed our dedicated browser extension called *Social Media Research*, which enables the hiding intervention and records key outcomes. The extension was compatible with Chromium browsers such as Google Chrome, Edge, Opera, and Brave, and listed on Chrome Web Store. It was also available on Firefox via Firefox Browser Add-ons. Together, supported browsers account for 87% of the global market share for desktop browsers.[10] Extensions constitute a well-established element of modern browsing, with 56 million users of the iconic AdBlock.[11] Therefore, we expect that many prospective participants were familiar with the environment in which the study took place.

A major advantage of toxicity hiding implemented through an extension is that social media algorithms are unaware of the extension's actions, as it operates by changing the content of the website *after* it was loaded, without communicating anything to the host server. This minimizes the risk that any algorithm-induced adjustment in the content presented to the user could have occurred as a reaction to the intervention.

Lastly, conducting a social media experiment, involving broad data collection, via a browser extension developed and maintained by the research team is a major responsibility. Considering the privacy and safety of our participants as a priority, we ensured that the extension onboarding followed Firefox's best practices and was vetted by their add-on reviewer. Moreover, all data were encrypted when stored in our database, with the decryption key only known to the researchers. Details on installation, onboarding, and privacy policy are provided in the online appendix. Throughout the study, users could report issues and send questions to the research team via a feedback form placed on our Twitter page, which was followed by many participants. Technical problems were infrequent, and those that occurred were addressed expeditiously.

2.3 Toxicity Detection

2.3.1 Algorithms

Effective automated real-time content moderation is a necessity for social media platforms operating at a large scale. With the ever-growing volume of online conversations and financial as well as ethical considerations placing constraints on human moderation, the algorithms must play a central role in toxicity detection efforts. With that in mind, we evaluated the impact of hiding toxic content on social media as detected by state-of-the-art tools available.

[10] https://gs.statcounter.com/browser-market-share/desktop/worldwide, accessed: 2024-12-29.
[11] https://getadblock.com/en/, accessed: 2024-12-29.

One of the original solutions, published in 2017, is Perspective API, a machine learning technology identifying toxicity in text conversations. The API is widely used by commercial clients, including major publishers like *The New York Times*, *Le Monde*, or *The Financial Times*.[12] The need for constant improvement of the algorithms' precision led to the creation of Jigsaw challenges, hosted by Kaggle, a Google-affiliated machine learning company. These were toxicity detection competitions for machine learning solutions. The contestants could rely on two newly published data sets "containing over one million toxic and non-toxic comments from Wikipedia," marked by human raters. For example, Detoxify library ("original" model) provided by Unitary, a contestant, was trained to serve as a "multi-headed model that's capable of detecting different types of of toxicity like threats, obscenity, insults, and identity-based hate." Its performance in the first Jigsaw challenge was admirable, with a score of 98.64 (the highest score was 98.86). In addition, Unitary supplied a successful "multilingual" model.[13] Owing to the high quality performance combined with the prospect of working with a fast and easy-to-use library, we decided to adopt Detoxify as our main toxicity detection tool. Additionally, we chose Perspective API as our fallback option, which was helpful due to its support for a wide array of languages.

2.3.2 Toxicity Scores

According to the providers of the algorithms employed in our project, their models generate toxicity scores corresponding to the probability that a text is considered toxic. Specifically, they suggest considering 0.3 as the threshold where statements become "suspect," where the algorithm is uncertain (for reference, they suggest using 0.7 for research on harassment).[14] In order to better understand the meaning of this uncertainty, we need to scrutinize how the toxicity detection solutions were trained. For example, in the case of Wikipedia comments, several human reviewers classified each comment as "Very Toxic," "Toxic," "Not Toxic," or chose "I'm not sure." If 3 out of 10 people categorized a statement as toxic, the algorithms were trained to assign a score of 0.3. This interpretation holds for all algorithms prepared to compete in the Jigsaw challenges (such as Unitary's Detoxify). Specifically, the target levels of toxicity in the training and evaluation samples was described as "fractional values which represent the fraction of human raters who believed the attribute applied to the given comment". Lastly, it is important to consider the meaning of the words "toxic" and "very toxic" as presented to

[12] https://perspectiveapi.com/case-studies/, accessed 2024-12-29.
[13] https://github.com/unitaryai/detoxify, accessed 2022-09-07.
[14] https://developers.perspectiveapi.com/s/about-the-api-score, accessed 2022-09-03.

human raters whose input was used to train the algorithms. In this context, the term "toxic" is understood as "a rude, disrespectful, or unreasonable comment that is somewhat likely to make you leave a discussion or give up on sharing your perspective", whereas "very toxic" refers to "a very hateful, aggressive, or disrespectful comment that is very likely to make you leave a discussion or give up on sharing your perspective."[15] While "leaving a discussion" and "giving up on sharing your perspective" constitute only a part of these industry-standard definitions, one might expect that these would bolster the likelihood that detoxification using algorithms trained this way will increase user engagement. In this context, our estimates showing the negative impact of exposure to toxicity on various forms of user engagement (see Section 3.2) are conservative.

2.3.3 Limitations

While the tools enabling our intervention are a sign of substantial progress in the field of automated toxicity detection, they are by no means perfect. Unitary itself acknowledges the deficiencies of their technology, pointing out issues with data sets that are very different from the training one. They also emphasize that the toxicity scores might be excessively affected by profanity words, which in certain contexts may not necessarily be harmful. This, however, does not imply that Detoxify cannot detect context-dependent toxicity. For example, a misogynistic statement "Women are not as smart as men", though devoid of traditional markers of abusive language, is correctly identified as toxic, with a toxicity score of 0.63, which would lead to its hiding by our intervention.

At this point, one might pose a question about the extent to which the imperfections of the toxicity detection technology affect the relevance of our results. Our experiment investigates the effects of applying *currently* available state-of-the-art tools, which can be used by social media platforms, online fora, news providers etc., for the purpose of real-time hiding of toxic content. This is directly relevant to stakeholders interested in automated toxicity detection. Furthermore, as a close proxy, the results can also provide valuable lessons to platforms considering hybrid systems, with human moderators partially overseeing the decisions made by the algorithm. Lastly, we hope to inform developers of future toxicity detection technologies about the social implications of the existing solutions.

[15] https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification, accessed 2022-09-03.

3 Browser Experiment

3.1 Experimental Design

3.1.1 Overview

Figure 1 summarizes the study flow. All individuals who installed the browser extension and agreed to data collection were randomly assigned either a treatment or control condition. Each participant went through a 14-day baseline period, during which we collected data on users' social media activity, with no hiding of toxic content regardless of the group. Subsequently, for users in the treatment group, we enabled the intervention, hiding toxic text content on Twitter, Facebook, and YouTube, for six weeks. After the last recruited person completed the intervention period, we invited all participants to an endline survey, where we collected additional outcomes.

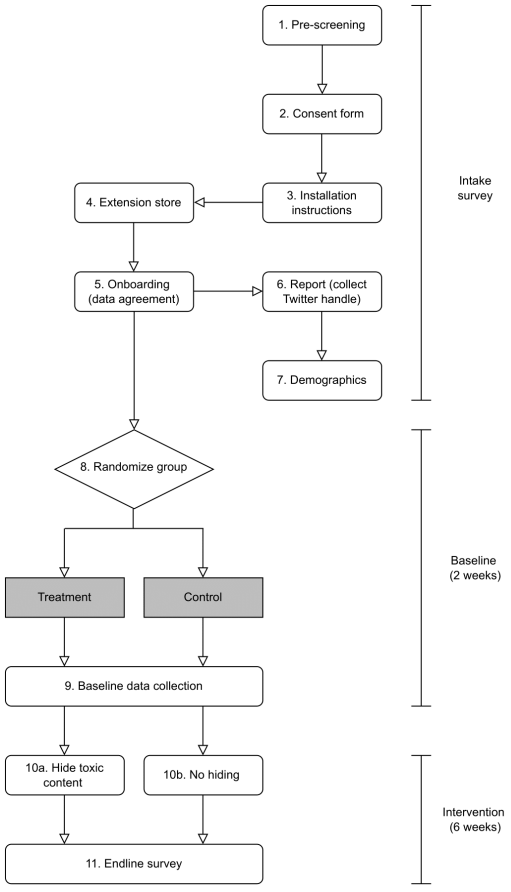


Figure 1: The Study Flow

3.1.2 Recruitment

Recruitment Ads. The recruitment process began on July 6th, 2022 and concluded on July 29th, 2022. We encouraged participation in the study using Twitter ads targeted at US-based English-speaking adults on desktop devices. Our decision to recruit on Twitter was motivated by the smaller size of its user base in comparison to Facebook and YouTube. We anticipated that if we enrolled participants via Twitter ads, there would be a relatively larger chance of them also using the other two social media sites.

To attract a broad subject pool, we relied on a variety of ad designs, including video ads with social media themed animations (Figure A1a in the appendix), static ads drawing attention to our gift card raffle (Figure A1b in the appendix), and ads offering a report on user's social media stats (Figure A1c in the appendix). Individuals who clicked on the link in the ads were directed to a Qualtrics environment for the intake survey. During the intake survey, we provided everyone with a link to the appropriate extension store, based on the browser detected by Qualtrics, and offered an animated GIF explaining the installation process (see Figure A2 in the appendix). In addition to our main method of recruitment, we benefited from promotion of our study by the Mozilla Foundation. The foundation's official Twitter account (@mozilla) retweeted a recruitment post (Figure A1d in the appendix) tailored to their followers (278.3 thousand as of August 2022). The prospective recruits who clicked on a link in the post were directed to a landing page, which was a simplified version of the intake survey.

Targeting Desktop Users. A consequential choice that we made when planning ad campaigns was to target users on desktop devices. In this case, we faced a trade-off. Individuals viewing Twitter on desktop during recruitment were more likely to regularly access social media platforms this way, thus allowing the browser extension (which does not operate on mobile devices) to capture a higher proportion of their activity and moderate a greater share of the content they are exposed to. Ultimately, this consideration prevailed over the concern about the impact on external validity—desktop users could be a special segment of the population. The alternative, allowing recruitment on mobile devices, carried a significant risk of hiding very little toxicity. Our decision led to recruiting a sample with a high share of social media consumption on desktop devices (detoxified and recorded by the extension)—users' reported desktop shares of Twitter and Facebook consumption are 60.5% and 56.2% respectively. Thanks to that, our intervention amounted to hiding a considerable proportion of users' overall social media diet, even when taking into account mobile app activity.

Obfuscation of Extension Purpose. After clicking the link to the extension store provided in the intake survey, participants could explore the extension listing, followed by installation and onboarding. Prospective users could read that the extension "can improve [their] user experience on Twitter, YouTube, and Facebook," and that it "may optimize [their] Twitter, YouTube, and Facebook pages by changing page content." In order to obfuscate the exact purpose of the study, we chose to describe the functionality in general high-level terms that, among other things, could include hiding toxic content. The wording of the store listing and onboarding is provided in the online appendix. Notably, neither the recruitment ads nor the onboarding materials ever mention toxicity or hate speech directly. The way in which we describe the extension is consistent with many possible ways of improving user experience. Thus, we minimize the risk that participants select into our study on the basis of their preferences regarding toxic content.

Endline Recruitment. Recruitment to the endline survey started on September 28, 2022, soon after the last participant's six-week intervention period concluded. The link to the survey was included in a browser notification (opening a new tab) sent to all users through the extension. We supplemented this process by sending emails to participants who provided a valid email address during the intake survey. As promised during enrollment, everyone who kept the extension enabled until the end of the study was entered into a raffle with three available prizes: $50, $150, and $300 gift cards.

3.1.3 Sample

User ID. Individuals who installed the browser extension were assigned a unique user ID on their first visit to one of the supported social media platforms. All data recorded by the extension was stored in the database under the user ID. Since ID assignment was performed at the browser level, in the intake survey we instructed participants to install the extension for only one browser—their main one—to minimize the risk that the user could experience different treatments. Furthermore, the user ID was placed in the extension storage, which should all but eliminate the possibility that the same person could be represented by two different user IDs. Even if someone accidentally uninstalled or disabled the extension, they should still be assigned the same ID on re-entry. Hence, we are confident that user IDs provide a reliable system of identifying participants.

Main Sample. We detected 755 user IDs pertaining to individuals who were active on the last day of the baseline period or later. Before the intervention period, the user experience in

the treatment group and the control group was identical. After a minimal cleaning procedure, which involved discarding user IDs that were associated with more than one Twitter or Facebook handle, we arrived at the main sample of 742 users.[16]

Covariates. We used Twitter handles collected by the extension to match participants to the Twitter API dataset to obtain covariates related to their previous Twitter activity—such as the number of years on the platform, the number of likes, friends, and followers. The extension retrieved at least one Twitter handle for 86.3% of users, based on whether the handle was available on the page while the participant was browsing. We expect handle availability to be independent of treatment assignment. In particular, if the handle was obtainable given the user's interface, the extension would have picked it up during the baseline period, where there was no difference in user experience across groups. In addition to obtaining Twitter API data, we relied on Twitter handles to match participants to their intake survey, where we elicited their demographics and data on their social media usage on desktop. We were able to match 522 individuals (70.4%) to their responses. We collected Twitter handles before treatment assignment, therefore, the matched individuals should constitute an as-if random subset of the main sample.

Sample Balance. Data from the Twitter API and the intake survey allowed us to create a rich balance table, depicted in Table A2 in the appendix. The main sample is well-balanced: only one out of sixteen covariates indicates significant differences by treatment assignment at the 5% and the 10% levels.

Endline Survey Sample. Based on matching the endline survey to the extension data by Twitter handles, we identified 384 participants—51.8% of users in the main sample. This includes 51.4% and 52.1% of users in the treatment and control group, respectively.

3.1.4 Treatment

Participants who installed the browser extension and agreed to data collection were randomly assigned either a treatment or a control condition. Of the 742 users in the main sample, 391 (52.7%) were assigned the treatment condition and 351 (47.3%) were in the control condition.

[16]The handful of discarded IDs pertain to cases, where despite our efforts, the same person experienced multiple treatments or re-entered the study at a late stage with a different user ID. We can identify these problematic cases thanks to the extension collecting Facebook and Twitter handles.

During the intervention period, our browser extension hid toxic text content on Twitter, Facebook, and YouTube for all individuals in the treatment group. The extension identified and analyzed each post, comment, and reply every time a user accessed a page on any of the three sites. Based on the text of each element, a toxicity score between 0 and 1 was assigned. The extension hid all content with the score exceeding a fixed threshold.

Analyzing Text Content. The extension sent text content to be evaluated to our server. There, we detected the language of the text. If the language was English, we relied on the "original" model provided by Unitary's Detoxify library (see Section 2.3). Otherwise, we applied one of the multilingual models, which together support 16 additional languages.[17] Given that our recruitment ads targeted US-based English-speaking adults, we anticipated that the overwhelming majority of content will be covered by the "original" model. Nevertheless, we chose to add fallback options for elements in other languages to increase the strength of the intervention, and welcome participants of various ethnicities.

Hiding Threshold. For all users in the treatment group, we adopted a hiding threshold of 0.3. This rule means that posts and comments with a toxicity score greater than 0.3 were hidden by the extension. To interpret the intervention in light of this threshold, we need to recall the meaning of toxicity scores, introduced in Section 2.3.2. In particular, the score of 0.3 reflects that 3 out of 10 human raters would label a text as toxic. We considered this threshold a meaningful candidate for hiding—one that could be reasonably implemented by a platform. Ultimately, the optimal threshold depends on the application. For example, if we intended to remove a piece of content from a website entirely or block the author, a more stringent criterion would be appropriate. Our choice of the threshold also reflects our ex-ante hope to examine whether substantial detoxification can improve user engagement and reduce the toxicity of content generated by users, or perhaps reveal a trade-off between these two objectives. We consider our efforts a starting point in this type of analysis with the intention of offering a benchmark for future, perhaps less intensive, interventions.

Our data suggests that our choice of a fairly low threshold level was less consequential than anticipated. In particular, Figure A3 in the appendix demonstrates that the distribution of toxicity of above-the-threshold content is skewed to the right.

[17] If the language was French, Italian, Russian, Portuguese, Spanish, or Turkish, we used the "multilingual" model by Unitary. In all other cases, we applied Perspective API—an alternative toxicity detection technology—which additionally supports multiple other languages: Arabic, Chinese ("zh"), Czech, Dutch, German, Hindi ("hi", "hi-Latn"), Indonesian, Japanese, Korean, and Polish.

Speed of Hiding. The immediate hiding of toxic content was of critical importance for the project. To ensure it, the app on our server—processing all statements and assigning them toxicity scores—was served by 8 machines, providing 4 GB RAM and 2 vCPUs each, with requests efficiently distributed among them during peak times. To evaluate our efforts, we collected data on the hiding speed on Twitter, measured as the difference between the toxic element being loaded on the page and being hidden. The median hiding speed was equal to 407 milliseconds. The histogram, presented in Figure A4 in the appendix, confirms that most hiding occurred in a fraction of a second, ensuring an uninterrupted experience from the perspective of the user. Moreover, content on social media is pre-loaded in batches ahead of where the user is on the page (i.e., content is loaded by the browser and can be assessed by our server before it appears on the screen). This further enables us to achieve our objective of hiding toxic content before users can see it.

Style of Hiding. In addition, we minimized the traces left on the page by our hiding intervention. Figure A5 in the appendix demonstrates user experience in the Facebook feed (and on group pages)—with hidden posts seamlessly replaced by the content below. Figure A6 in the appendix offers an example of a comment section under a post in the original state and with the intervention provided by our extension. In general, the hiding of posts both in the feed and in the comment sections for all platforms should not have been easily noticeable to a casual user. On Twitter and Facebook, posts were hidden together with their visible comments, and comments with their visible replies. If a toxic comment/reply on Twitter was a part of a thread, all subsequent replies were hidden too, as they would not make sense without the toxic element.[18] On YouTube, we were hiding toxic comments together with "Show replies" button (if unwrapped) and nested replies (if visible). Figure A7 in the appendix depicts an example of the hiding intervention on YouTube.

Twitter-specific Functionality. In order to induce greater exogenous variation in exposure to toxic content between the treatment and the control group on Twitter, the extension seamlessly unwrapped "Show more replies" sections at the bottom of comments under a post, where the platform places more toxic elements (see Figure A8 in the appendix). The functionality was

[18]This introduced a challenge that we could not fully address with the extension: for toxic replies on Twitter when they were marked with a vertical line connecting elements of a thread—we could not entirely remove the line from an element preceding the hidden one. Crucially, our data suggest that the hiding intervention did not negatively impact the user experience. This is indicated by the lack of differential attrition in our experiment—the overall attrition was low and the survival rate was actually higher in the treatment group, albeit insignificantly.

enabled in both the treatment and the control groups during the baseline and the intervention period, so that the addition of hiding was the only thing that we experimentally varied across the conditions at the beginning of the intervention.

3.1.5 Outcomes

Exposure to Toxicity. Before we begin investigating the main questions, we report the proportion of content on Twitter, Facebook, and YouTube that the extension hid during the study, as well as the average toxicity scores of content displayed to users on the three platforms. These measures allow us to understand the strength of the intervention and can be interpreted as a "first stage." For the treatment group, we simultaneously present the toxicity of content offered by a platform (what they intended to display before hiding applied) and toxicity of content shown. Comparing the former measure to the toxicity of content in the control group (over time) is helpful in discerning any potential learning by social media algorithms.

Time on Social Media. We report the total amount of active time that users spend consuming feeds and comment sections on Twitter, Facebook, and YouTube on a given day. To compute this, we collect two types of information. First, we collect the timestamps of all new elements (e.g., posts and comments) loaded by the browser on each platform—this indicates active interaction with the website, such as scrolling (this allows us to avoid attributing to consumption the time that the user spends away from the computer with the platform tab open on the screen). To refine this measure, we ping the browser, approximately every minute, to check what site is open in the current tab, which allows us to identify when the user switches away from the platform.

We measure durations of each browsing session on a particular platform, defined as the time elapsed between the first record that a user started using it (a new element displayed or a ping related to the platform) and the user either (1) switching to another platform (indicated by the extension recording a new element from another platform or a ping related to another platform) or (2) becoming inactive, i.e., there have been no new elements loaded for at least 3 minutes. We aggregate session durations at platform-day level to obtain the final measure of the time spent. Separately, we report the number of sessions. Given that in browser studies there is no straightforward way of defining active time spent (Aridor et al., 2024), we demonstrate robustness of our active time results to different inactivity thresholds (1, 2, 4, 5, and 10 minutes) in the online appendix.

Content Consumption. As a basic measure, we record the quantity of content displayed to users on each platform as a proxy for content consumption—we call this measure content *shown*. However, when evaluating the effect of exposure to toxicity on content consumption, we mostly rely on the quantity of content *offered* by the platform—inclusive of the hidden elements. By including the hidden elements in the count, we are certain that any negative effects are caused by a genuine reduction in user engagement and not simply a mechanical consequence of the hiding process. We also distinguish between content in the feed and in the comment sections (for Facebook and Twitter). The former category is more relevant to advertising due to the positioning of ad slots—ads are typically placed in user feeds in between posts from followed accounts or friends.

Ad Impressions. In order to further illuminate the effects of the intervention on predictors of advertising revenue, we report ad impressions on Twitter and Facebook. Following the pre-registration, we do not measure ads on YouTube, as they appear in video form. On Twitter, the browser extension can directly identify ads displayed to the user, and we compute their total number per day. Measuring ad impressions on Facebook proved more challenging, as the extension could not identify them in real time. However, the extension recorded the text of each Facebook post shown to the user. We fine-tuned gpt-3.5-turbo to identify ads based on the text of a post. We used 199 posts for this exercise—70% for fine-tuning and 30% for evaluation. We achieved a precision rate of 100% and a recall rate of 89.5%. Any negative treatment effect on the number of ads displayed to users is unlikely to be driven by the mechanical consequences of hiding—only 0.8% of ads during the intervention period were toxic in the treatment condition and therefore hidden. Despite that, as our main measure of ad impressions, we use ads *offered*, defined in a way analogous to content offered (see the previous paragraph). This conservative measure precludes the possibility of any mechanical effects.

Ad Clicks. In addition to ad impressions, a pre-registered outcome, we also report ad clicks on Twitter and Facebook as a part of exploratory analysis. We rely on information from periodically pinging the browser to check the current tab's website. We record an ad click if shortly after being shown a social media ad we detect that a user left the platform for a website which is neither one of the three treated platforms nor one of the additional 38 related websites that we track.[19] This is a proxy for leaving the social media site for an advertiser's webpage,

[19]The list of the 38 related websites is provided in the online appendix. We define "shortly after" as before another 12 new elements on the page load *and* within 2 minutes. These criteria reflect the fact that elements

although measurement error is a limitation, because a move to an unrelated website after seeing an ad would also be counted as an ad click. Separately, we report the click rate—the proportion of ads displayed to the user with an associated ad click.

Reactions, Posts, and Post Clicks. We capture alternative forms of user engagement, including those that require a visible action. We report the number of user's reactions (such as likes) and posts (that include, for Twitter, retweets). Moreover, as part of exploratory analysis, we create a measure of post clicks on Twitter and Facebook. Specifically, we compute the number of times a user accesses a comment section, which is what happens after a post click.[20]

Index of Engagement. We report two indices of overall user engagement, combining various metrics. First, we provide the *original* index, which is defined in our pre-registration. We compute the equally weighted average of the z-scores of its three components: time spent on social media, quantity of content displayed to users, and alternative forms of engagement (reactions, posts, and retweets). We report the index for each treated platform separately (Twitter, Facebook, YouTube), as well as for all platforms combined. Second, we provide the *preferred* index. This index includes the three components of the original index as well as the number of browsing sessions, the number of ads displayed to the user, the number of ads clicked, and the number of posts clicked. We report these for each platform as well as for all platforms combined. The preferred index is a more comprehensive measure of user engagement, and reflects outcomes that we did not expect to be able to compute before the experiment. For example, at the time, using generative AI to identify ads on Facebook was not feasible.

Contagiousness. To investigate the contagiousness of toxicity, we calculate the average daily toxicity scores of the posts, comments, and replies published by each participant.[21]

Substitution. We also consider potential substitution effects to platforms where the intervention did not take place. The extension measured the time spent by users on 38 pre-registered social-networking websites (the list is in the online appendix). We also record time spent on

are loaded in batches—the ad is likely to be recorded with the same or a very similar timestamp as posts next to it. In the online appendix, we show robustness to alternative criteria.

[20] On Twitter, we directly observe when a user views the comment section of a post. On Facebook, we rely on the fact that the platform does not show more than two comments related to a post in the feed. Hence, if we record at least three comments in a row, we can interpret it as the user viewing comments following a click.

[21] Please note that, unlike what we stated in the pre-registration, we cannot include likes and retweets in our analysis of contagiousness. This is because any effect on these endpoints would be explained mechanically: the content with toxicity exceeding 0.3 that users could have shared or reacted to would be hidden in the treatment.

any of the other websites combined, as well as the time that the browser window is inactive. In order to measure time we ping the browser, approximately every minute, to check what site is open in the current tab. As our main endpoint, we report the total number of minutes spent on all of the 38 sites (as pre-registered). Separately, we classify the websites into three categories: social media, messaging, and other, to check if a particular type of website drives any substitution patterns. Lastly, we quantify spillovers to mobile devices using Twitter API data. For most users (86.3%), we are able to obtain data on the number of posts and reposts made using the Twitter app on Android/iOS, although due to technical limitations these data exist only for a subset of days.

Heterogeneity. As indicated in the pre-registration, we explore two angles of heterogeneity. First, we split the sample into two parts according to the toxicity of the content consumed during the baseline period. To that end, we ranked individuals by the average toxicity score and categorized them relative to the median person. Considering the above-the-median individuals gives us insight into the effects on users who might exhibit higher tolerance for toxic content, or perhaps even a degree of preference for it. This interpretation stems from the possibility that platforms may optimize what they display to users at the individual level, and thus the heterogeneity in toxicity scores likely reflects what platforms know about each participant. The second angle of heterogeneity is by platform. Due to the fundamental differences between Facebook, Twitter, and YouTube, we focus primarily on platform-specific investigation, reporting our results for each website separately. However, we do provide treatment effects on indices of user engagement for all platforms combined (as specified above).

Endline Survey. We collected three additional outcomes in the endline survey. First, we measured the willingness to pay/accept for using the extension for one extra month. Second, we elicited toxicity ratings for seven statements representing different types of toxicity. This allows us to test whether any potential contagiousness of toxicity may be driven by its normalization over the intervention period. Lastly, we rely on measures proposed by Allcott et al. (2020) to elicit users' subjective well-being. Details about the survey outcomes are in the online appendix.

3.1.6 Descriptive Statistics

Panels A and B of Table A1 in the appendix display descriptive statistics for users and Twitter accounts in our main sample, and compare them to representative samples. The representative sample of Twitter users comes from the American Trends Panel (ATP) of September 2020, which

is a nationally representative panel of U.S. adults provided by the Pew Research Center. The representative sample of Twitter accounts originates from English Tweets collected in August 2020 from the 1% random sample of Twitter's API. Our sample of users is comparable to a representative sample of U.S. Twitter users in terms of age and sex, but it oversamples Democrats and undersamples Independents. Additionally, Twitter accounts in our sample tend to be older and have fewer followers, with an approximately similar number of accounts followed relative to accounts from the random sample of Tweets.

Panel C of Table A1 in the appendix reports summary statistics for a subset of our outcomes based on the 14-day baseline period. On average, users spend roughly 57 minutes per day on the three platforms—17 minutes on Facebook, 29.5 minutes on Twitter, and 10.5 minutes on YouTube. They consume 2.6 times more content on Twitter than on Facebook, and content consumption on YouTube is half of that on Facebook. Elements in comment sections constitute 35% of Facebook content displayed to users and 30% in the case of Twitter. The average toxicity score per unit of content (both consumed and produced) is almost double on Twitter in comparison to Facebook.

3.1.7 Empirical Strategy

At the core of our identification strategy is the use of the baseline period to establish the benchmark levels of activity, such as time on social media or content consumption, for each individual. This baseline should allow us to estimate the effects of the intervention with more power (McKenzie, 2012). In our pre-registration, we indicated our intention to evaluate the outcomes using a difference-in-differences approach, where we rely on the two-week baseline and the six-week intervention periods. Given that we randomly assigned treatment to each participant, the parallel trends assumption is satisfied by design. Furthermore, the median person was actively using their browser on 14 out of the 14 days of the baseline, with the median total time equal to 1812 minutes (2.16 hours per day). The first quartile values were 12 days and 879 minutes (1.05 hours per day), respectively. The high level of activity during the baseline, even for the left tail of the distribution, indicates that it was a reliable measure of users' typical activity.

We adopt the two-way fixed effects model (TWFE) as our main specification. First, for each participant, we define time periods t as days in the study relative to their individual start time. Second, we generate a treatment dummy D_{it}, indicating whether the hiding intervention was on for individual i in period t. Lastly, we regress the outcome variable Y_{it} on the treatment

dummy D_{it} with individual fixed effects α_i and period fixed effects δ_t:

$$Y_{it} = \alpha_i + \delta_t + \beta^{TWFE} D_{it} + \epsilon_{it}. \tag{1}$$

We use Driscoll and Kraay standard errors to account for serial and cross-sectional dependence, as we have a relatively long panel of individuals (Cameron and Miller, 2015), but we also discuss robustness to standard errors clustered at the individual level.[22]

Although our recruitment period was very short (about 3 weeks), one may be concerned that our participants enrolled in the study on different days and, therefore, treatment started for them at different times. According to the newest difference-in-differences literature (see Baker et al., 2022; Chabé-Ferret, 2021, for a review), the staggered treatment could lead to bias in the TWFE estimator. As a robustness check, we report the stacked difference-in-difference regressions (Cengiz et al., 2019; Baker et al., 2022), which address this problem. This involves extending specification (1) by including start date × period fixed effects.

Lastly, it is important to note that our extension does not record uninstallation events by users, which necessitates inferring attrition from user activity. All regression specifications presented in the main text of the paper rely on panels involving participants who were active on the last day of the study (day 56) or later. We tracked users' browser activity for some time after day 56, which means that this should be a precise measure of survival. In Section 3.3.3, we discuss robustness to an alternative definition of attrition—one in which survival is implied by being active on day 42 or later, i.e., at least two weeks before the end of the intervention.

3.2 Results

In this section, we present the findings of the browser experiment.

3.2.1 Exposure to Toxicity

During the intervention period, the extension automatically hid toxic text content for each treated user on the three supported platforms. The hidden content corresponds to 6.6% of total content that the platforms intended to display to users in their browser—7.2 % on Twitter, 4.9 % on Facebook, and 6.3 % on YouTube. Given that our participants reported spending, on average, 60.5% of their Twitter time and 56.2% of their Facebook time on a desktop device, a back-of-the-

[22] Driscoll and Kraay (1998) provide a nonparametric estimator that is robust to heteroscedasticity and very general forms of spatial and temporal dependence. This method requires a large number of time periods, which is a plausible assumption in our setting with 56 time periods per individual. See Alvarez and Argente (2022) for another example of a use case where this assumption is plausible.

envelope calculation suggests that the extension hid 4.4% of their entire Twitter diet—taking into account mobile usage—and 2.8% for Facebook. We conclude that the intervention, despite being introduced solely on desktop devices, considerably varied exposure to social media toxicity.

The hiding intervention resulted—by design—in a decrease in the average toxicity of users' desktop feeds and comment sections. Figure 2 depicts the average toxicity score of elements on the three supported platforms over the course of the study, split by treatment condition. For treated individuals, the figure provides both the level of toxicity of the content that platforms *intended* to display to the user, i.e., before the extension hid toxic content (dashed line), and the toxicity of content actually *displayed* to the user (solid line). For participants in the control, the figure plots the toxicity of displayed content. The graph demonstrates a sharp drop in the average exposure to toxic content in the treatment group. At the same time, it is clear that the average toxicity of elements that would have been displayed during the intervention period does not differ by treatment arm—it is 0.063 in the control and 0.064 in the treatment. This similarity suggests that the algorithm did not learn or respond by adjusting the toxicity offered to the treatment group. These levels contrast with the mean toxicity of 0.017 that was shown to users in the treatment group after the conclusion of the baseline period. Overall, the hiding intervention reduced the toxicity of content the participants were exposed to by 73% across the three platforms.

Table A3 in the appendix demonstrates the relevant difference-in-differences results. We find that the treatment lowered the average toxicity score of content displayed to users by about 2 pp on Facebook, 4.8 pp on Twitter, and 3.4 pp on YouTube. All of these results are significant at the 1% level. We also report significant reductions in toxic content (toxicity score exceeding 0.3) as a share of total content displayed to users—by 3.3 pp on Facebook, 6.9 pp on Twitter, and 5.3 pp on YouTube.

3.2.2 Time on Social Media

We present the results on the active time users spend on social media, summarized in Table 1. The intervention reduces the active time spent on Facebook by 1.3 minutes per day, or 9.2% relative to the mean. We report a similar effect on YouTube, with magnitude of 0.6 minutes per day, a drop of 6.8%. Both of these results are significant at the 5% level. The intervention did not significantly affect the active time spent on Twitter, although the point estimate is negative. We also report that the number of separate browsing sessions on Twitter and YouTube fell as a result of the intervention. Specifically, users accessed Twitter 1.4 fewer times a day and

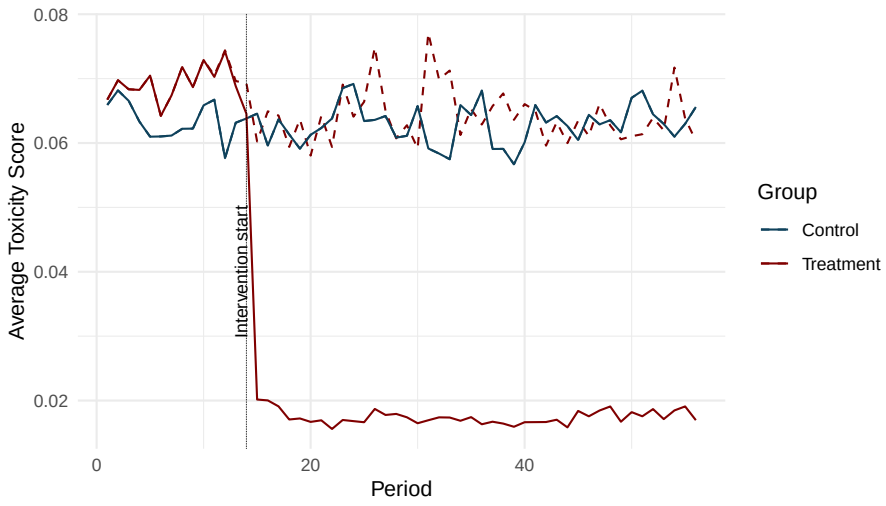


FIGURE 2: AVERAGE TOXICITY OF CONTENT SHOWN TO USERS DURING THE STUDY

Note: The figure depicts the average toxicity of posts, comments, and replies shown to users on each day of the study (relative to when a given participant started), separately for the control group and the treatment group. The dashed line for the treatment group demonstrates the average toxicity of elements that the platforms intended to show to the user before any hiding was applied by the extension. The data presented here encompasses the three supported platforms (Twitter, Facebook, and YouTube). The dashed vertical line ("Intervention start") indicates day 15—the first day of the intervention period.

YouTube 0.9 fewer times a day. Taking all of these results together, we conclude that toxicity *improves* user engagement across platforms, either by increasing time spent (intensive margin) or increasing the number of browsing sessions (extensive margin).

TABLE 1: EFFECT OF INTERVENTION ON ACTIVE TIME SPENT ON SOCIAL MEDIA

	Facebook		Twitter		YouTube	
	Time	Sessions	Time	Sessions	Time	Sessions
Treated	−1.314***	−0.467	−0.351	−1.377***	−0.611**	−0.886**
	(0.402)	(0.315)	(1.001)	(0.463)	(0.269)	(0.381)
Mean	14.31	3.91	23.76	5.63	9.04	4.07
SD	35.53	16.45	48.53	22.59	25.32	19.62
N	32 312	32 312	37 184	37 184	36 456	36 456

Note: This table reports estimates from Equation (1) for our main experimental sample. The dependent variables are the total active time in minutes spent on social media platforms (Time) and the number of separate browsing sessions (Sessions). Precise definitions of these outcomes are provided in Section 3.1.5. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active on the last day of the study (day 56) or later. Driscoll-Kraay standard errors are parenthesized. *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

3.2.3 Content Consumption

Table 2 summarizes the main results on another key measure of engagement: the quantity of posts and comments that users consume. The intervention significantly reduced content consumption on Facebook, including content that the platform *offered*, i.e., including the hidden

elements. Thus, the negative effect on this measure of consumption cannot be explained by the mechanical effect of hiding, and indicates a genuine reduction in this form of user engagement. Specifically, we observed that the hiding intervention decreased content consumption by at least 16.5 elements a day, a result significant at the 1% level. This magnitude represents a 23% decrease relative to the mean quantity of content throughout the study, or 0.08 SD. We do not detect significant effects on content offered on Twitter and YouTube.

Table 2: Effect of Intervention on Content Consumed on Social Media

	Facebook		Twitter		YouTube	
	Shown	Offered	Shown	Offered	Shown	Offered
Treated	−20.081***	−16.479***	−12.563	0.177	−4.295**	−1.940
	(2.528)	(2.634)	(7.990)	(8.143)	(1.670)	(1.681)
Mean	69.81	71.25	182.04	187.26	37.39	38.35
SD	205.30	210.87	384.20	396.91	124.53	127.54
N	32 312	32 312	37 184	37 184	36 456	36 456

Note: This table reports estimates from Equation (1) for our main experimental sample. The dependent variables are the number of posts and comments shown to users (Shown) and the number of posts and comments offered to users; those displayed on their feeds and comment sections plus the content mechanically hidden by the extension (Offered). The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active on the last day of the study (day 56) or later. Driscoll-Kraay standard errors are parenthesized. *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Our conservative measure of content consumption—content offered—provides a lower bound (in absolute value) on any negative treatment effect. While we use it as our main consumption outcome, this does not imply that the effect on content *shown* is necessarily uninformative. In the feed and in long comment sections, hidden elements are instantly replaced by the content below—they are pulled up. Furthermore, even in the event that there is no available replacement, or if we consider the implications of elements being loaded in batches, a lower quantity of content shown during a browsing session indicates that users decided not to scroll further or seek more content in place of what was hidden—a meaningful decision. Table 2 indicates that the intervention led to a significant reduction in the quantity of content displayed to users on both Facebook (29%) and YouTube (11%).

To further shed light on our main estimates, we split our data by whether a piece of content in question appears in the feed or in a comment section—Table 3 presents the results. It is notable that the intervention reduced the conservative measure of both consumption of feed content and comment section content on Facebook, the former of which is particularly consequential for ad impressions—as the platform places ads in between posts. Consumption of feed content fell by at least 11.5 per day (a 25% change). The reduction for comment sections content was at least 5 per day (a 20% change). We also observe a negative effect on consumption of content in

comment sections on Twitter, using the conservative metric of comments offered—at least 7.1 fewer elements per day (a 13% change).

Table 3: Effect of Intervention on Content Offered by Conversation Type

	Facebook		Twitter		YouTube
	Feed	Comments	Feed	Comments	Comments
Treated	−11.532***	−4.950***	7.323	−7.146***	−1.940
	(2.024)	(1.792)	(6.240)	(2.600)	(1.681)
Mean	46.20	25.05	132.10	55.16	38.35
SD	137.05	101.80	264.87	178.38	127.54
N	32 312	32 312	37 184	37 184	36 456

Note: This table reports estimates from Equation (1) for our main experimental sample. The dependent variables are the number of elements in the feed offered to users (Feed) and the number of elements in the comment sections offered to users (Comments). In both cases, we include both content displayed to users plus the content mechanically hidden by the extension (i.e., content offered). The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active on the last day of the study (day 56) or later. Driscoll-Kraay standard errors are parenthesized. *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Taking all of the content consumption results together, we demonstrate strong negative effects of the intervention on Facebook, and weaker evidence of disengagement on Twitter (only for comments). Results for YouTube should be treated with caution, as the significant decrease of content shown may be attributable, in large part, to the fact that the YouTube comment section does not have infinite scroll (thus making the effect mechanical). These results complement the earlier evidence on time spent on social media (Section 3.2.2) and corroborate that toxicity is a driver of user engagement across platforms.

3.2.4 Ad Impressions and Clicks

The intervention reduced ad consumption on Facebook and Twitter. All of the ad impression results reported in Table 4 rely on the conservative measure of content offered applied to ads, thus they cannot be driven by any mechanical effects of hiding. We find that the average number of ads displayed to users on Facebook fell by at least 2.3 per day, a drop of 27% relative to the mean. The intervention also reduced ad consumption on Twitter by at least 0.6 per day, a decrease of 5.7%. The former result is significant at the 1% level whereas the latter is significant at the 10% level. Furthermore, we report that the intervention reduced the number of ad clicks on Facebook and Twitter—for both platforms the effect is significant at the 5% level. Lastly, we document that the effect on clicks is not driven by a change in the click rate (i.e., more clicks per impression), but more likely by a lower number of impressions. Overall, we demonstrate that exposure to toxicity improves metrics predictive of advertising revenue, such as ad impressions and ad clicks.

Table 4: Effect of Intervention on Ad Consumption and Ad Clicks

	Facebook			Twitter		
	Offered	Clicked	Click Rate	Offered	Clicked	Click Rate
Treated	−2.297***	−0.038**	0.003	−0.559*	−0.056***	0.001
	(0.427)	(0.016)	(0.005)	(0.322)	(0.018)	(0.003)
Mean	8.59	0.20	0.0399	9.81	0.287	0.0404
SD	31.81	0.921	0.12	19.12	1.08	0.114
N	32 312	32 312	10 671	37 184	37 184	17 483

Note: This table reports estimates from Equation (1) for our main experimental sample. The first dependent variable is the number of ads displayed to users plus any ads mechanically hidden by the extension (Offered). The second dependent variable is the number of ad clicks (Clicked). Lastly, we report the the proportion of ads shown to the user that they clicked (Click Rate). Precise definitions of these outcomes are provided in Section 3.1.5. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active on the last day of the study (day 56) or later. Driscoll-Kraay standard errors are parenthesized. *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

3.2.5 Reactions and Post Clicks

We report the effects of the intervention on alternative forms of user engagement. First, we do not find strong evidence that hiding toxicity affects the total number of reactions (such as likes), posts, and retweets (on Twitter). As specified in Table 5, the intervention lowers the outcome on Facebook by 0.55 action per day, which corresponds to a 10.2% change with respect to the study mean. This result is significant at the 10% level. However, we find a null effect on Twitter, and a marginally significant effect in the opposite direction on YouTube. Liking or retweeting a post is a visible action that may be interpreted as an endorsement. Hence, even if users are more willing to spend time on or consume toxic content, that may not translate into whether they want to react to, repost, or post about it.

Second, we consider a private action in the form of clicking to uncover a comment section associated with a post. The intervention reduces the number of post clicks by 0.6 per day on Facebook (a 24% change relative to the mean) and 0.6 per day on Twitter (a 13% change). Both results are significant at the 1% level. This result is consistent with the evidence reported earlier, which indicates that toxicity increases user engagement that is not publicly observed.

3.2.6 Index of Engagement

Following Section 3.1.5, we report the effects on two indices of engagement: original and preferred. The original index is narrower and is a subset of the preferred one. Table 6 summarizes the results. Combining all the treated platforms, the intervention reduces the original and preferred indices of user engagement. Both effects are significant at the 5% level. Splitting the results by platform, we find evidence that hiding toxicity reduces engagement on Facebook

Table 5: Effect of Intervention on Reactions and Post Clicks

	Facebook		Twitter		YouTube
	Reactions and Posts	Post Clicks	Reactions and Posts	Post Clicks	Reactions and Posts
Treated	−0.546*	−0.581***	0.496	−0.625***	0.160*
	(0.298)	(0.120)	(0.569)	(0.154)	(0.083)
Mean	5.34	2.38	11.51	4.68	0.45
SD	20.84	7.26	37.03	12.38	4.07
N	32 312	32 312	37 184	37 184	36 456

Note: This table reports estimates from Equation (1) for our main experimental sample. The dependent variables are the total number of reactions (such as likes), posts, and retweets (Reactions and Posts) and the number of post clicks (Post Clicks). Precise definitions of these outcomes are provided in Section 3.1.5. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active on the last day of the study (day 56) or later. Driscoll-Kraay standard errors are parenthesized. *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

(according to both indices), Twitter (according to the preferred index), and YouTube (according to the preferred index). We conclude that exposure to toxic content on social media increases user engagement across a variety of metrics. Furthermore, while the effects on Facebook are of a higher magnitude, each platform contributes to the overall negative effect.

Table 6: Effect of Intervention on Index of Engagement

	Original Index				Preferred Index			
	All Platforms	Facebook	Twitter	YouTube	All Platforms	Facebook	Twitter	YouTube
Treated	−0.032**	−0.064***	−0.009	−0.010	−0.054***	−0.063***	−0.030**	−0.015**
	(0.013)	(0.012)	(0.017)	(0.010)	(0.010)	(0.011)	(0.013)	(0.006)
Mean	0.0298	0.0583	0.007 01	0.0166	0.0232	0.0432	0.003 99	0.0188
SD	0.913	1.07	0.864	0.824	0.76	0.875	0.72	0.54
N	37 184	32 312	37 184	36 456	37 184	32 312	37 184	36 456

Note: This table reports estimates from Equation (1) for our main experimental sample. The dependent variables are the original index of user engagement (Original Index) and the preferred index of user engagement (Preferred Index). The indices are defined in Section 3.1.5. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active on the last day of the study (day 56) or later. Driscoll-Kraay standard errors are parenthesized. *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

We also present (Figure A9 in the appendix) estimates from the event study specification for all platforms combined that evaluate the intervention's impact on the preferred index of engagement on a week-by-week basis. This allows us to analyze the dynamics of the treatment effects. The negative effect on user engagement is stable across the intervention period, and there is no evidence that the overall result is driven by period-outliers.

3.2.7 Contagiousness

We now report results regarding contagiousness of toxicity, i.e., whether higher exposure to toxicity increases production of toxic content. Table 7 displays the estimates on the toxicity

of posts and comments written by users. The intervention had a significantly negative impact on the average toxicity scores of content they publish on Facebook and Twitter—conditional on posting. The effects are significant at the 5% level and quantitatively similar on both platforms: −0.011 on Facebook and −0.019 on Twitter. These can be interpreted as a 30% and a 25% reduction in the content toxicity relative to the mean, or a decrease of 0.13 SD and 0.13 SD, respectively. We do not find a significant effect on YouTube. This is understandable, as the primary motivation for visiting YouTube is to watch videos, which may interrupt the link between consumption and the production of toxic comments. Taken together, we find broad evidence consistent with the hypothesis that toxicity on social media is contagious.

Table 7: Effect of Intervention on Toxicity of Content Produced

	Facebook	Twitter	YouTube
Treated	−0.011**	−0.019***	−0.037
	(0.006)	(0.006)	(0.035)
Mean	0.0371	0.0748	0.0805
SD	0.0848	0.145	0.198
N	6411	8962	1341

Note: This table reports estimates from Equation (1) for our main experimental sample. The dependent variable is the average toxicity of the published content, conditional on posting. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active on the last day of the study (day 56) or later. Driscoll-Kraay standard errors are parenthesized. ∗,∗∗ , and ∗∗∗ denote significance at the 10%, 5%, and 1% levels, respectively.

3.2.8 Substitution

We report that reducing exposure to toxicity on treated platforms led to positive spillover effects on the combined total time spent on 38 pre-registered social-networking websites (a list comprised of both social media platforms and messengers like WhatsApp), where the intervention did *not* take place. Table 8 presents the results. In particular, the hiding intervention led to users spending, on average, 1.8 more minutes per day on the non-treated platforms, a result significant at the 5% level. This masks important heterogeneity, as we uncover that this effect is driven by spending more time on other *social media platforms* (1.9 more minutes per day, significant at the 1% level) as opposed to spending more time on *messaging* apps. We observe no substitution towards other websites.

Table 8 also presents the effects of our intervention on posts and reposts on Twitter made on mobile devices (i.e., using Twitter app on Android or iOS). Hiding toxicity significantly reduces both outcomes, suggesting that there was no substitution towards mobile devices as a result of filtering out toxicity on desktop. If anything, there, we provide evidence of *negative* spillover

effects to mobile apps of platforms on which we hide toxicity on desktop.

TABLE 8: EFFECT OF INTERVENTION ON SUBSTITUTION TO RELATED SITES AND MOBILES

	Desktop					Mobile	
	All	Social Media	Messengers	Other	Inactive	Posts	Reposts
Treated	1.821**	1.959***	−0.138	−1.690	−7.684*	−0.859***	−1.997**
	(0.807)	(0.710)	(0.220)	(3.120)	(4.371)	(0.272)	(0.838)
Mean	8.30	7.34	0.952	88.39	175.27	2.80	3.28
SD	37.92	35.72	11.84	158.13	272.01	7.35	20.07
N	37 184	37 184	37 184	37 184	37 184	8666	8666

Note: This table reports estimates from Equation (1) for our main experimental sample. The first dependent variable is the total time (in minutes) spent on 38 pre-registered platforms where the intervention did not take place (All). The following three dependent variables pertain to the time spent on subsets of the 38 websites by category: social media websites (Social Media), messaging apps (Messengers), and other sites (Other). The fifth dependent variable pertains to the total time when the browser window was open but inactive on the operating system of the user (Inactive). The final two dependent variables are the number of posts (Posts) and reposts (Reposts) made from mobile devices on Twitter. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active on the last day of the study (day 56) or later. Driscoll-Kraay standard errors are parenthesized. *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

3.2.9 Heterogeneity

Table A4 in the appendix presents the results of heterogeneity analysis by baseline levels of exposure to toxic content. We find that both the intervention's negative effects on user engagement (summarized by the preferred index) and contagiousness are significantly stronger for users with above-median exposure to toxic content in the baseline period. The estimates are consistent across platforms and when aggregating them together. Following our interpretation of baseline exposure to toxicity as a proxy for individual-level tolerance for toxicity, we conclude that the analysis offers suggestive evidence that people with some degree of tolerance/preference for toxic content disengage more when such content is hidden.

3.2.10 Endline Survey

Lastly, we briefly report the results of the endline survey. We do not find significant effects of the intervention on the willingness to pay/accept for using the extension for an extra month, measures of subjective well-being, and toxicity ratings of seven social media posts containing toxic content of various types and intensities. For these outcomes, we performed between-subjects analysis with a reduced sample size (the take-up rate was slightly above 50%). Both factors contributed to lower statistical power. Hence, the null results on survey outcomes should not be interpreted as definitive. Instead, we recommend a better powered analysis of toxicity's impact on these measures as an idea for future research work. More details on the endline survey results are provided in the online appendix.

3.3 Robustness and Potential Concerns

3.3.1 Attrition

Overall, attrition in our study is low for this type of field experiment, especially taking into account that the intervention lasted for six weeks (56 days). Specifically, we report that at least 89.4% of users who started the intervention period survived to the end of the study, i.e., completed all 56 days. This includes 91% of users in the treatment group and 87.5% of users in the control group. Table A5 in the appendix shows a regression of the survival dummy on the treatment dummy. The coefficient interaction coefficient is insignificant at the 10% level, indicating no evidence of differential attrition. Ex-ante it was natural to worry that the hiding intervention could result in higher attrition in the treatment group if it is not sufficiently seamless. Our data dispel this concern. If anything, attrition is lower in the treatment group, albeit insignificantly.

Furthermore, we consider two likely channels that could have led to differential attrition regarding types of individuals leaving. Given the character of our hiding intervention, people with preference for toxic content or those with high levels of social media activity might be more likely to drop out of the treatment group. First, we extend the regression analysis by including the average toxicity score of content displayed to the user during the baseline (a proxy for tolerance for toxicity) and the average time spent on social media during the baseline. Neither covariate is a significant predictor of overall attrition. We further extend the regression by including the interactions of both measures with the treatment dummy. The interaction coefficients are insignificant. Thus, neither baseline toxicity nor activity explain attrition by treatment condition, reducing concerns of differential attrition regarding types of individuals.

3.3.2 Topic Analysis

One concern about the study is that, by hiding toxicity on social media, the intervention indirectly changed the topical composition of content shown to users. To address this problem, we used gpt-4o to classify users' posts and comments *shown* into 26 topic categories. We selected these topics to match the "interest" categories that advertisers can use to target users on Twitter. To economize resources and time, we asked ChatGPT to compute the percent of elements which talk about each topic, truncated each element to a maximum of 280 characters, and considered at most 250 elements per day. More details, including the prompt, are available in the online appendix. Data is aggregated at the individual-day level, so that we can conduct estimation using specification (1).

Figure 3 provides treatment effects on the proportion of content displayed to users by category. We report significant effects at the 5% level only for 2 of the 26 categories—the intervention marginally increases the proportion of content on personal finance and movies and television. Importantly, we find null effects for contentious categories such as "society" and "law, government and politics." For comparison purposes, we report the point estimate on the proportion of toxic content shown to users. It showcases that any effects on topical composition of content displayed to users are very small in comparison to the change in exposure to toxicity.

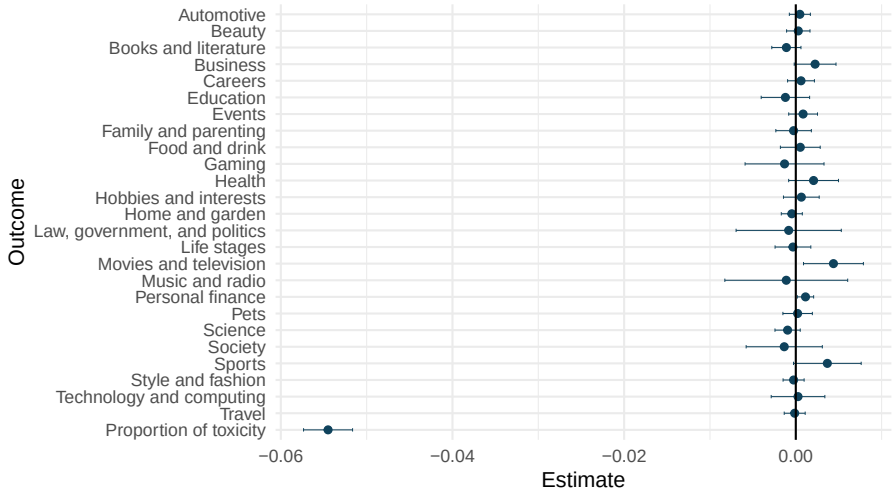


FIGURE 3: EFFECT OF INTERVENTION ON DISTRIBUTION OF TOPICS SHOWN TO USERS

Note: This figure presents coefficients from regressions estimated according to Equation (1). The dependent variable is the proportion of content displayed to the user that pertains to a specific topic category. Separately, we provide a regression where the dependent variable is the proportion of toxic content shown to the user. Details of the topic classification procedure are outlined in Section 3.3.2. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active on the last day of the study (day 56) or later. We report 95% confidence intervals based on Driscoll-Kraay standard errors.

3.3.3 Alternative Specifications

Alternative Attrition Cutoff. Due to our inability to observe unistallation events, we need to infer attrition from user activity. The regression specifications presented in the main text of the paper are based on panels involving participants who were active on the last day of the study (day 56) or later. Table A6 in the appendix presents robustness of our user engagement and contagiousness results to an alternative attrition cutoff. There, we rely on regressions involving participants who were active on day 42 or later, thus requiring survival up until at least two weeks before the end of the intervention. Using this specification, the intervention has a significant effect (at the 5% level) on the preferred index of user engagement aggregated across all platforms, as well as specifically on Facebook and Twitter. Similarly, the effects on

contagiousness are significant for all platforms combined, as well as specifically on Facebook and Twitter.

Stacked Regression. We now discuss robustness of our results to the stacked regression specification, which is meant to address the potential bias caused by the staggered nature of treatment (Cengiz et al., 2019; Baker et al., 2022). We emphasize that this concern is unlikely to apply in our setting, as the staggering occurred over a short period—users were recruited and assigned to treatment within a short span of three weeks. Table A7 in the appendix presents the estimates. User engagement results, summarized by the preferred index, are robust to the alternative specification for all platforms combined (at the 1% level) as well as for Facebook (at the 1% level), Twitter (at the 1% level), and YouTube (at the 10% level) separately. Contagiousness results are robust to the stacked regression approach for all platforms combined, on Facebook, and on Twitter (at the 5% level).

Alternative Standard Errors. We also discuss robustness of our user engagement and contagiousness results to clustering standard errors at the individual level rather than relying on Driscoll and Kraay (1998) standard errors. For user engagement outcomes, we have a long 56-period panel of observations that justifies the "large T" assumption (Cameron and Miller, 2015) required to use Driscoll and Kraay standard errors (see Section 3.1.7). For the contagiousness results, since the outcome measure is conditional on posting, we do not have a full 56-period panel for many individuals. Table A8 in the appendix provides the estimates. User engagement results, summarized by the preferred index, are robust to the alternative standard errors for all platforms combined (at the 10% level) and on Facebook (at the 5% level). Contagiousness results are also robust to the alternative standard errors for all platforms combined, on Facebook, and on Twitter (all at the 5% level).

Taking the above exercises together, the overall conclusions of the paper, that toxicity increases user engagement and that it is contagious, are generally robust to alternative specifications. In particular, when considering user engagement as well as contagiousness on all treated platforms combined, we find significant negative effects of the intervention using an alternative attrition cutoff, the stacked regression specification, and alternative standard errors.

4 Survey Experiment

This section describes the design and results of an additional survey experiment that we conducted to supplement the findings of the browser experiment described in Section 3.

4.1 Experiment Design

Figure 4 summarizes the experiment design. We explore it in more detail below.

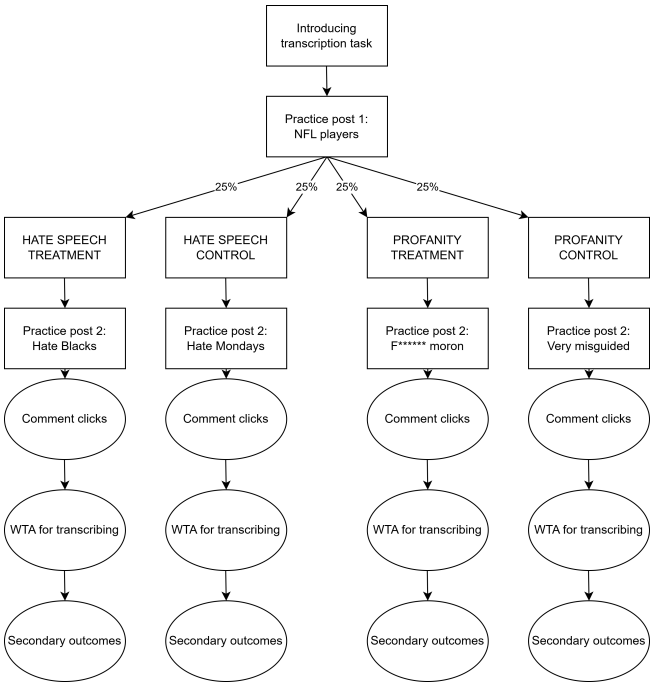


FIGURE 4: THE SURVEY EXPERIMENT FLOW

Note: The figure provides an overview of the flow of the survey experiment. It shows the order of information blocks (rectangles) and outcome measures (ovals). Participants are randomly assigned to one of four experimental conditions: hate speech treatment, hate speech control, profanity treatment, and profanity control.

4.1.1 Study Flow

Recruitment. We recruited N=4,120 US adults on Prolific, a survey platform commonly used for conducting online experiments, for a study called "Social Media Posts and Training Algorithms."[23] We report that 4,048 individuals (98.3%) completed the main section of the study, including collection of all primary outcomes. To maximize statistical power, and accord-

[23]Prolific has been consistently used in high-quality research in Economics. See Peer et al. (2022) for a discussion of Prolific's data quality in comparison to similar platforms.

ing to our pre-registration, this sample includes 411 individuals from a pilot study. We discuss robustness to dropping these individuals.

Introduction. After collecting basic demographic variables, we inform participants that we are recruiting for a follow-up task of transcribing 100 social media posts. The task consists of transcribing the text of a post from its image. We enumerate possible uses of data collected during the transcription task, such as training an algorithm that helps people avoid mistakes when posting on social media. This helps us divert attention from the intervention, which focuses on varying exposure to toxic content, and thus reduces potential experimenter demand effects.

Transcription Task. Afterwards, we explain the process that will be used to select the participant recruited for the task of transcribing 100 social media posts. We use a second-price auction: We select a person with the lowest willingness to accept and pay them the second lowest bid as compensation for completing the task. We also describe how to transcribe posts (where to type up the text and that proceeding requires entering the minimum number of characters based on the length of the post). We test comprehension of our instructions using four questions, two of which pertain to the second-price auction. It is important to emphasize that the transcription exercise ensures that participants read the posts carefully without revealing anything about the purpose of the study.

First Practice Post. Subsequently, participants complete two practice transcriptions. The images of posts look like Facebook posts, with the author name, profile picture, and the date of posting all censored. Figure A10a in the appendix shows an example with the first practice post, concerning retirements of NFL players. This post was shown to all participants in an identical way, regardless of the treatment group. It is non-toxic according to the Perspective API classifier (0.02 toxicity score).

Intervention. The second post is where the intervention occurs. Our intervention follows a 2×2 design. The first layer of the intervention varies the *level of toxicity* of the post: whether the participants see the original toxic post (treatment) or a slightly modified version of the post with reduced toxicity (control). The second layer varies the *type of toxicity* of the post: whether the original post contains identity hate or profanity. Overall, participants are randomly assigned into one out of four experimental conditions: hate speech treatment, hate speech

control, profanity treatment, or profanity control.

In the *hate speech treatment* condition, the post involves discrimination of a particular group of people: "What is so difficult to understand? I hate Blacks." In the associated *hate speech control*, the post is "What is so difficult to understand? I hate Mondays." We vary only one word to keep the structure and complexity of the post the same, with the length differing by only one character. At the same time, we introduce a significant variation in toxicity. The former post has a toxicity score of 0.8 according to Perspective API, whereas the latter one has a toxicity score of 0.29. This difference implies that the former post would be hidden in the treatment group of the browser experiment (Section 3), whereas the latter would not.

In the *profanity treatment* condition, participants were given the following post: "You are a f***ing moron." Please note that in the actual experiment the profane word was not censored. In the *profanity control*, they were shown a mellow version of this post: "You are very misguided." Both posts have 23-24 characters. As before, we aim to keep the structure and complexity of the post the same, with the length differing by only one character. Moreover, we preserve the direct meaning (both statements imply that the person is wrong or doing something incorrectly), while varying the toxicity of the language used to express the sentiment.[24] According to Perspective API, the former post has a toxicity score of 0.98, whereas the latter has a score of 0.28.

Comments. Regardless of the experimental condition, the comment section for the second practice post is always comprised of three comments. The comments are identical across different levels of toxicity within each type of toxicity, i.e., the same comments for the hate speech treatment and control, and the same comments for the profanity treatment and control.

4.1.2 Outcomes

Following our pre-registration, we collect two primary outcomes: user engagement and willingness to accept (WTA) to transcribe 100 social media posts.

User Engagement. We measure user engagement by observing whether the participant clicks "View 3 comments" at the bottom of the treated post to uncover the comment section. The link is formatted and situated within the post to closely resemble how such links appear on Facebook (see Figure A10a in the appendix for an example). Upon clicking, an image with comments following Facebook's formatting is displayed (Figure A10b in the appendix). We stress that the

[24] In the hate speech conditions, reducing toxicity inherently changes the meaning of the sentence. In the profanity conditions, however, it is possible to preserve much of the meaning without resorting to profanity.

participants were informed that uncovering the comments of a post had no impact on their pay or performance on the task. Since they had to transcribe only the text of the post, if anything, uncovering the comments carried only an opportunity cost in terms of time spent.

WTA to Transcribe. We elicit the WTA to transcribe 100 social media posts after the second practice post is completed. Participants use a slider that can move from \$2 to \$30 to choose the WTA with precision of \$0.1. Participants are reminded that the person with the lowest compensation will be invited to transcribe 100 social media posts and paid the second smallest compensation. If there are multiple people with the smallest compensation, one of them will be chosen at random. The lower bound is selected to reflect that the minimum compensation per hour is \$8 and that the task will take at least 15 minutes.

Secondary Outcomes. After the WTA elicitation, we collect several secondary outcomes. In particular, we measure the impact on recall of toxicity, recall of hate speech, and recall of treatment posts. To minimize fatigue and keep the survey short, we randomly assign participants one of these outcomes (1/3rd chance for each outcome) and ask them the relevant question.[25] For the recall of toxicity, we ask whether or not any of the posts displayed before were toxic and we gave the participants the same definition of toxicity that we used in the field experiment (Section 2.3.2). We measure the recall of hate speech in an analogous way.[26] For the recall of treatment posts, we ask participants which one of four similar-sounding options was one of the posts they were asked to transcribe. Lastly, we ask participants to rate on a scale from 0 to 100 how entertaining the training posts were.

Heterogeneity. We pre-registered the following three angles of heterogeneity analysis. First, we look at heterogeneous treatment effects by gender. Second, we consider the importance of being a member of a minority group. We classify participants as such if they (1) select an answer other than "white" in the ethnicity question, (2) identify as a Hispanic, Latino, or a person of Spanish origin, or (3) report a different sexual orientation than heterosexual. Third, as religiosity shapes a lot of norms of behavior, we look at heterogeneity with respect to being religious (i.e., choosing a different answer than "no religion" when asked about religion or belief).

[25] In the pilot study, all participants were asked only about the toxicity recall. The other secondary outcomes were not measured in the pilot study, so the overall sample size for them is slightly smaller.

[26] Meta's definition of hate speech is provided: "Hate speech is a direct attack against people on the basis of protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease."

Lastly, as part of the exploratory analysis, we discuss heterogeneity by age (younger vs. older than 35).

4.1.3 Descriptive Statistics

Table A9 in the appendix provides descriptive statistics about our sample and the primary outcome variables. The sample is balanced on gender, with 49.4% of participants declaring as male, and exhibits considerable diversity—45.7% of participants belong to an ethnic or sexual minority. The sample is not skewed toward young participants, with the average age of 41.6. We also capture a lot of individuals with high household incomes—48.8% declare income in excess of $70,000. Our participants are highly active on social media. On average, they log into social media platforms on 6.3 out 7 days a week, with 74% of people doing it every day. Thus, both this sample and the sample of users in the browser extension experiment (Section 3) consist overwhelmingly of social media users.

We also report statistics about the outcome variables. The proportion of people who uncover the comment section below the treated post (practice post 2) equals 34.3%. This is a considerable proportion given that participants have no material incentive to check out the comments. This is especially clear as they have already completed practice post 1 (41.7% uncovered the comments for that post). Regarding the other primary outcome, we report that the average WTA for transcribing 100 social media posts is $13.38, slightly above Prolific's recommended pay for an hour of work ($12).

4.2 Results

In this section, we discus the results of the survey experiment.

4.2.1 User Engagement

Table 9 presents the results for user engagement. Following the pre-registration, as our main specification, we conduct a pooled test of both treatments (hate speech and profanity) against both controls. We find that toxicity increases user engagement, measured as the likelihood of clicking to uncover the comment section, by 18% relative to the mean (column 1). Specifically, the difference in the likelihood of clicking between toxic posts and their non-toxic versions is 6.1 pp ($p<0.001$), corresponding to an effect size of 0.13 SD. The effect is stronger after restricting the sample to people who passed all comprehension checks (column 2).

Columns 3-6 show the results for specific types of toxicity. The effect on user engagement

Table 9: User Engagement in the Survey Experiment

	Pooled		Hate		Profanity	
	All	Comp.	All	Comp.	All	Comp.
Toxic	0.061***	0.086***	0.094***	0.104***	0.028	0.069**
	(0.015)	(0.021)	(0.021)	(0.030)	(0.021)	(0.030)
Mean	0.34	0.42	0.35	0.43	0.34	0.41
SD	0.47	0.49	0.48	0.49	0.47	0.49
N	4049	2144	2021	1072	2028	1072

Note: The table is based on a sample of all people who completed the second practice post (N=4,049). Column 1 shows the results of a regression of a dummy variable equal to one if a participant clicked to uncover the comment section below the second practice post and zero otherwise on a dummy variable equal to one if they were assigned one of the toxic treatments (either the hate speech treatment or the profanity treatment) and zero otherwise. Column 2 provides the same regression but with a sample restricted to individuals who passed all comprehension checks. Columns 3-4 pertain to specifications analogous to those in Columns 1-2 but with the sample restricted to individuals who were in the hate speech treatment or the hate speech control. Columns 5-6 pertain to specifications analogous to those in Columns 1-2 but with the sample restricted to individuals who were in the profanity treatment or the profanity control. Robust standard errors are in parentheses. *significant at 10%; ** significant at 5%; *** significant at 1%.

is particularly strong for hate speech: a 9.4 pp increase in the likelihood of clicking (p<0.001), corresponding to 0.2 SD. The effect for profanity is weaker, at 2.8 pp (p=0.179), though still detectable for the sample of users who passed all comprehension checks (6.9 pp, p=0.022). In the online appendix, we demonstrate that all results reported in Table 9 are robust to excluding pilot observations.

These results offer an additional piece of evidence supporting the conclusion that social media toxicity increases user engagement. The result naturally complements the browser experiment (Section 3), which focuses on the policy of hiding content above a specified toxicity threshold and relies on observed online behavior over the period of 8 weeks.

4.2.2 WTA for Transcribing Posts

Table 10 demonstrates treatment effects on the willingness to accept (WTA) for transcribing 100 social media posts. First, the hate speech treatment increases the WTA by $0.62 (p=0.038) in comparison to the hate speech control, an effect of 0.09 SD (column 1). Second, the profanity treatment has no effect on the WTA, with the coefficient close to zero and the p-value of 0.912 (column 4).

We interpret these findings as follows. The goal of the WTA outcome is to test the connection between increases in engagement and welfare. If increases in engagement were a good proxy for increases in welfare, we should observe that the WTA is *lower* for toxic posts than their detoxified versions—individuals require a lower compensation to transcribe posts that increase their engagement. Our results indicate that this is not the case. Despite being ex-ante powered to detect effects as small as 0.12 SD, we find no evidence of toxicity lowering the WTA for

Table 10: WTA for Transcribing 100 Social Media Posts

	Hate			Profanity		
	All	WTA Comp.	Comp.	All	WTA Comp.	Comp.
Toxic	0.624**	0.671*	0.493	0.033	-0.108	-0.074
	(0.301)	(0.380)	(0.412)	(0.297)	(0.369)	(0.400)
Mean	13.61	13.40	13.26	13.15	12.78	12.76
SD	6.77	6.75	6.75	6.70	6.53	6.55
N	2020	1264	1072	2028	1253	1072

Note: The table is based on a sample of all people who provided the willingness to accept for transcribing 100 social media posts (N=4,048). Column 1 shows the results of a regression of the willingness to accept (WTA) for transcribing 100 social media posts on a dummy variable equal to one if the participant was assigned the hate speech treatment and zero if they were assigned the hate speech control. Column 2 provides the same regression but with a sample restricted to individuals who passed the comprehension checks about the WTA elicitation. Column 3 provides the same regression as Column 1 but with a sample restricted to individuals who passed all comprehension checks. Column 4 shows the results of a regression of the WTA for transcribing 100 social media posts on a dummy variable equal to one if the participant was assigned the profanity treatment and zero if they were assigned the profanity control. Column 5 provides the same regression but with a sample restricted to individuals who passed the comprehension checks about the WTA elicitation. Column 6 provides the same regression as Column 4 but with a sample restricted to individuals who passed all comprehension checks. Robust standard errors are in parentheses. *significant at 10%; ** significant at 5%; *** significant at 1%.

neither hate speech nor profanity.

On the contrary, if anything, we find suggestive evidence of the opposite effect in the case of hate speech. The WTA is significantly *higher* in the hate speech treatment than in the associated control. This result remains significant at the 10% level after restricting the sample to individuals who passed the comprehension checks about the second price auction (column 2). However, it is not significant (although the point estimate is positive) in the sample of those who passed all comprehension checks (column 3) or after excluding pilot observations (online appendix).

These findings are consistent with theoretical arguments that highlight how user engagement might not be a good proxy for welfare. In a companion paper (Beknazar-Yuzbashev et al., 2024), we propose a simple model to argue that content can be simultaneously harmful—in the sense that it lowers user utility—but engaging. This happens when there is enough complementarity between harmful content and time spent on social media (or other forms of engagement). For example, users may dislike reading toxic posts, but conditional on seeing them, they may choose to increase the time spent on the platform in order to respond to the posts or participate in the related discussion.

4.2.3 Additional Results

Table 11 presents means and standard deviations of the secondary outcome variables. First, we report that in the hate speech treatment, 94.3% of participants recall seeing a toxic practice

post. The corresponding proportion in the hate speech control group is 7.3%. Similarly, 93.2% of people in the profanity treatment recalls toxicity, whereas 24.3% do in the profanity control. These patterns indicate that we successfully induced variation in exposure to toxicity, as perceived by the participants. Second, 94% of people in the hate speech treatment recalls seeing posts containing hate speech. The analogous proportion in the profanity treatment is 40.4%. Together, these recall results confirm that our distinction between the two types of toxicity (profanity and hate) is shared by the respondents: they indeed perceive both treatment groups as more toxic and the hate speech group specifically as more likely to contain hate speech.

Table 11: Secondary Outcomes in the Survey Experiment

	Hate Control	Hate Treatment	Profanity Control	Profanity Treatment	Test
Recalls Toxicity	0.073 (0.261)	0.943 (0.231)	0.243 (0.429)	0.932 (0.251)	<0.001
Recalls Hate	0.024 (0.152)	0.940 (0.237)	0.037 (0.189)	0.404 (0.491)	<0.001
Recalls Treatment	0.993 (0.081)	0.987 (0.115)	0.987 (0.113)	1.000 (0.000)	0.233
Entertainment	35.237 (26.581)	23.923 (26.353)	33.780 (26.022)	38.881 (28.495)	<0.001

Note: The table presents means and standard deviations (in parentheses) of four secondary outcome variables by experimental condition. The outcomes are as follows with the corresponding sample sizes provided in brackets: (1) whether an individual considers any of the practice posts toxic (N=1,623), (2) whether they consider any of the practice posts to contain hate speech (N=1,211), (3) whether they correctly recall the text content of the second practice post from a menu of similar-sounding options (N=1,213), (4) entertainment rating (from 0 to 100) of the practice posts (N=3,636). The last column shows a p-value for a joint test of equality of the outcomes across the four experimental conditions.

Table 11 also provides evidence that our results are unlikely to be driven by differential attention; that is, that toxicity might be more engaging because it draws more attention. The ability of participants to recall the text of the second practice post from a menu of similar-sounding options is unaffected by experimental condition (p=0.233 in a joint test).

The fourth outcome reported in Table 11 is the entertainment rating of the practice posts, which might help reconcile why the WTA point estimates tend to go in opposite directions between the hate speech content and the profanity content. In particular, participants find the post displayed in the profanity treatment more entertaining (38.9 out of 100) than in the profanity control (33.8). However, the result for the hate speech conditions is in the opposite direction, with the more toxic version of the post being less entertaining (23.9 vs. 35.2). These results are consistent with a story in which both profane and hateful posts trigger participants' curiosity (and hence lead them to uncover the comments), but for potentially different reasons, with different welfare implications. Profane treatment (toxic) posts offer participants more entertainment, while hate treatment posts offer participants less entertainment (perhaps even outrage).

Lastly, we briefly summarize the results on heterogeneity of the treatment effects. In the online appendix, we report that for neither primary outcome gender, minority status, or reli-

giosity is a significant moderator at the 5% level. However, we find one interesting exploratory result. We previously reported that both hate speech and profanity increase user engagement, but the former has a stronger effect (Section 4.2.1). Heterogeneity analysis by age indicates that the weaker effect for profanity is driven by young people. For individuals aged 35 or older, the effect size is 6 pp, which is 10 pp (p=0.027) higher than for individuals under the age of 35.

5 Conclusion

This paper studies how the toxicity of users' feeds and comment sections can impact their consumption and production of social media content. In principle, toxic content could reduce user engagement—indeed, the definition of toxicity utilized by the leading toxicity detection algorithms includes its propensity to make people leave a discussion. Yet, we find evidence that lower exposure to toxicity can actually reduce different measures of engagement. We also find evidence supporting the concerns that toxic online behavior can be contagious and we document some substitution patterns towards other social media websites. The lower engagement we document does not mean that users are worse off in an environment with low toxicity. In fact, we provide evidence that revealed-preference arguments based on engagement do not apply in this setting—welfare and engagement do not necessarily move in the same direction.

We hope that our findings inform policy; both, platforms and regulators are likely to favor less severe moderation tools over outright removing content or banning users. One option is reducing prominence of toxic content on the platform (reducing its visibility)—often referred to as "freedom of speech, not freedom of reach," akin to our hiding intervention. At the same time, our findings warn that platforms might not have enough incentives to mitigate the reach of toxic content. Nevertheless, these findings should be complemented by future work. In particular, we stress the importance of conducting platform-side experiments unbeknownst to users to increase the external validity of the results.

References

Acemoglu, D., A. Ozdaglar, and J. Siderius (2021). Misinformation: Strategic sharing, homophily, and endogenous echo chambers. Technical report, National Bureau of Economic Research.

Agarwal, S., U. M. Ananthakrishnan, and C. E. Tucker (2022). Content moderation at the infrastructure layer: Evidence from parler. *Available at SSRN 4232871*.

Ahmad, W., A. Sen, C. Eesley, and E. Brynjolfsson (2024). Companies inadvertently fund online misinformation despite consumer backlash. *Nature 630*(8015), 123–131.

Allcott, H., L. Braghieri, S. Eichmeyer, and M. Gentzkow (2020). The welfare effects of social media. *American Economic Review 110*(3), 629–76.

Allcott, H., J. Castillo, M. Gentzkow, L. Musolff, and T. Salz (2024). Sources of market power in web search: Evidence from a field experiment. Technical report, mimeo.

Allcott, H. and M. Gentzkow (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives 31*(2), 211–36.

Allcott, H., M. Gentzkow, and L. Song (2022). Digital addiction. *American Economic Review 112*(7), 2424–2463.

Alvarez, F. and D. Argente (2022). On the effects of the availability of means of payments: The case of uber. *The Quarterly Journal of Economics 137*(3), 1737–1789.

Andres, R. and O. Slivko (2021). Combating online hate speech: The impact of legislation on twitter. Technical report, ZEW Discussion Papers.

Anti-Defamation League (2024). Online hate and harassment: The american experience 2024. Available at https://www.adl.org/resources/report/online-hate-and-harassment-american-experience-2024.

Aridor, G. (Forthcoming). Measuring substitution patterns in the attention economy: An experimental approach. *RAND Journal of Economics*.

Aridor, G., R. Jiménez-Durán, R. Levy, and L. Song (2024). The economics of social media. *Journal of Economic Literature*.

Aridor, G., R. Jiménez Durán, R. Levy, and L. Song (2024). Experiments on social media.

Artís Casanueva, Annalí, V. A., S. Sardoschau, and K. Saxena (2024). Late adoption and collective action: Social media expansion and the diffusion of black lives matter.

Aslett, K., A. M. Guess, R. Bonneau, J. Nagler, and J. A. Tucker (2022). News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Science advances 8*(18), eabl3844.

Aslett, K., Z. Sanderson, W. Godel, N. Persily, J. Nagler, and J. A. Tucker (2024). Online searches to evaluate misinformation can increase its perceived veracity. *Nature 625*(7995), 548–556.

Baker, A. C., D. F. Larcker, and C. C. Wang (2022). How much should we trust staggered difference-in-differences estimates? *Journal of Financial Economics 144*(2), 370–395.

Baumeister, R. F., E. Bratslavsky, C. Finkenauer, and K. D. Vohs (2001). Bad is stronger than good. *Review of general psychology 5*(4), 323–370.

Beknazar-Yuzbashev, G., S. Ichiba, and M. Stalinski (2024). To the depths of the sunk cost: Experiments revisiting the elusive effect. *Available at SSRN*.

Beknazar-Yuzbashev, G., R. Jiménez-Durán, and M. Stalinski (2024). A model of harmful yet engaging content on social media. In *AEA Papers and Proceedings*, Volume 114, pp. 678–683. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Beknazar-Yuzbashev, G. and M. Stalinski (2022). Do social media ads matter for political behavior? a field experiment. *Journal of Public Economics 214*, 104735.

Boxell, L., M. Gentzkow, and J. M. Shapiro (2024). Cross-country trends in affective polarization. *Review of Economics and Statistics 106*(2), 557–565.

Braghieri, L., R. Levy, and A. Makarin (2022). Social media and mental health. *American Economic Review 112*(11), 3660–3693.

Bursztyn, L., G. Egorov, R. Enikolopov, and M. Petrova (2019). Social media and xenophobia: evidence from russia. Technical report, National Bureau of Economic Research.

Bursztyn, L., B. R. Handel, R. Jimenez, and C. Roth (2023). When product markets become collective traps: The case of social media. Technical report, National Bureau of Economic Research.

Cameron, A. C. and D. L. Miller (2015). A practitioner's guide to cluster-robust inference. *Journal of human resources 50*(2), 317–372.

Cengiz, D., A. Dube, A. Lindner, and B. Zipperer (2019). The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics 134*(3), 1405–1454.

Chabé-Ferret, S. (2021). *Statistical Tools for Causal Inference*. https://chabefer.github.io/STCI, accessed 2022-09-10.

Djourelova, M., R. Durante, E. Motte, and E. Patacchini (2023). Experience, narratives, and climate change beliefs. *Available at SSRN 4680430*.

Driscoll, J. C. and A. C. Kraay (1998). Consistent covariance matrix estimation with spatially dependent panel data. *Review of economics and statistics 80*(4), 549–560.

Enikolopov, R., A. Makarin, and M. Petrova (2020). Social media and protest participation: Evidence from russia. *Econometrica 88*(4), 1479–1514.

Farronato, C., A. Fradkin, and C. Karr (2024). Webmunk: A new tool for studying online behavior and digital platforms.

Farronato, C., A. Fradkin, and T. Lin (2024). Data sharing and website competition: The role of dark patterns. *Available at SSRN 4920040*.

Farronato, C., A. Fradkin, and A. MacKay (2023). Self-preferencing at amazon: evidence from search rankings. In *AEA Papers and Proceedings*, Volume 113, pp. 239–243. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Fujiwara, T., K. Müller, and C. Schwarz (2024). The effect of social media on elections: Evidence from the united states. *Journal of the European Economic Association 22*(3), 1495–1539.

Germano, F., V. Gómez, and F. Sobbrio (2022). Ranking for engagement: How social media algorithms fuel misinformation and polarization. *Available at SSRN 4238756*.

Ghanem, D., S. Hirshleifer, and K. Ortiz-Becerra (2023). Testing attrition bias in field experiments. *Journal of Human resources*.

Gröndahl, T., L. Pajola, M. Juuti, M. Conti, and N. Asokan (2018). All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*, pp. 2–12.

Guess, A. M., N. Malhotra, J. Pan, P. Barberá, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Dimmery, D. Freelon, M. Gentzkow, et al. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science 381*(6656), 398–404.

Guriev, S., E. Henry, T. Marquis, and E. Zhuravskaya (2023). Curtailing false news, amplifying truth. *Amplifying Truth (October 29, 2023)*.

Guriev, S., N. Melnikov, and E. Zhuravskaya (2021). 3g internet and confidence in government. *The Quarterly Journal of Economics 136*(4), 2533–2613.

Henry, E., E. Zhuravskaya, and S. Guriev (2022). Checking and sharing alt-facts. *American Economic Journal: Economic Policy 14*(3), 55–86.

Jiménez Durán, R. (2022). The economics of content moderation: Theory and experimental evidence from hate speech on Twitter. *Available at SSRN*.

Jiménez-Durán, R., K. Müller, and C. Schwarz (2022). The effect of content moderation on online and offline hate: Evidence from germany's netzdg. Technical report, CEPR Press Discussion Paper.

Kalra, A. (2024). Hate in the time of algorithms: Evidence from a large-scale experiment on online behavior. *Working Paper*.

Katsaros, M., K. Yang, and L. Fratamico (2022). Reconsidering tweets: Intervening during tweet creation decreases offensive content. In *Proceedings of the International AAAI Conference on Web and Social Media*, Volume 16, pp. 477–487.

Kemp, S. (2024). Digital 2024: Global Overview Report. Accessed December 31, 2023. Available at https://datareportal.com/reports/digital-2024-global-overview-report.

Kim, J. W., A. Guess, B. Nyhan, and J. Reifler (2021). The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication 71*(6), 922–946.

Kominers, S. D. and J. M. Shapiro (2024). Content moderation with opaque policies. Technical report, National Bureau of Economic Research.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review 111*(3), 831–70.

Liu, Y., P. Yildirim, and Z. J. Zhang (2021). Social media, content moderation, and technology. *arXiv preprint arXiv:2101.04618*.

Liu, Z. J. (2020). How allowing a little bit of dissent helps control social media: Impact of market structure on censorship compliance. *Johns Hopkins Carey Business School Research Paper* (20-13).

Madio, L. and M. Quinn (2024). Content moderation and advertising in social media platforms. *Journal of Economics & Management Strategy*.

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more t in experiments. *Journal of development Economics 99*(2), 210–221.

Melnikov, N. (2021). Mobile internet and political polarization. *Available at SSRN 3937760*.

Müller, K. and C. Schwarz (2020). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*.

Müller, K. and C. Schwarz (2023a). The effects of online content moderation: Evidence from president trump's account deletion. *Available at SSRN 4296306*.

Müller, K. and C. Schwarz (2023b). From hashtag to hate crime: Twitter and antiminority sentiment. *American Economic Journal: Applied Economics 15*(3), 270–312.

Nyhan, B., J. Settle, E. Thorson, M. Wojcieszak, P. Barberá, A. Y. Chen, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Dimmery, et al. (2023). Like-minded sources on facebook are prevalent but not polarizing. *Nature 620*(7972), 137–144.

Peer, E., D. Rothschild, A. Gordon, Z. Evernden, and E. Damer (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 1.

Petrova, M., A. Sen, and P. Yildirim (2021). Social media and political contributions: The impact of new technology on political competition. *Management Science 67*(5), 2997–3021.

Ribeiro, M. H., J. Cheng, and R. West (2022). Automated content moderation increases adherence to community guidelines. *arXiv preprint arXiv:2210.10454*.

Rizzi, M. (2024). Self-regulation of social media and the evolution of content: a cross-platform analysis. *Available at SSRN 5018309*.

Robertson, R. E., J. Green, D. J. Ruck, K. Ognyanova, C. Wilson, and D. Lazer (2023). Users choose to engage with more partisan news than they are exposed to on google search. *Nature 618*(7964), 342–348.

Tusher, E. H., M. A. Ismail, M. A. Rahman, A. H. Alenezi, and M. Uddin (2024). Email spam: A comprehensive review of optimize detection methods, challenges, and open research problems. *IEEE Access*.

Yu, X., M. Haroon, E. Menchen-Trevino, and M. Wojcieszak (2024). Nudging recommendation algorithms increases news consumption and diversity on youtube. *PNAS nexus 3*(12), pgae518.

Zavolokina, L., K. Sprenkamp, Z. Katashinskaya, D. G. Jones, and G. Schwabe (2024). Think fast, think slow, think critical: designing an automated propaganda detection tool. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–24.

Zhuravskaya, E., M. Petrova, and R. Enikolopov (2020). Political effects of the internet and social media. *Annual Review of Economics 12*, 415–438.

A ONLINE APPENDIX: Not for publication

The online appendix extends the analysis offered in the paper. Section A.1 provides additional figures and Section A.2 includes additional tables. Section A.3 discusses supplementary results and robustness checks for the browser experiment and Section A.4 does the same for the survey experiment. We also include materials accompanying both experiments, including the wording of outcome measures (Section A.5).

A.1 Additional Figures



(A) Standard Video Ad



(B) Standard Static Ad



(C) Learning Social Media Stats



(D) Retweet by Mozilla

Figure A1: Examples of Recruitment Ads on Twitter

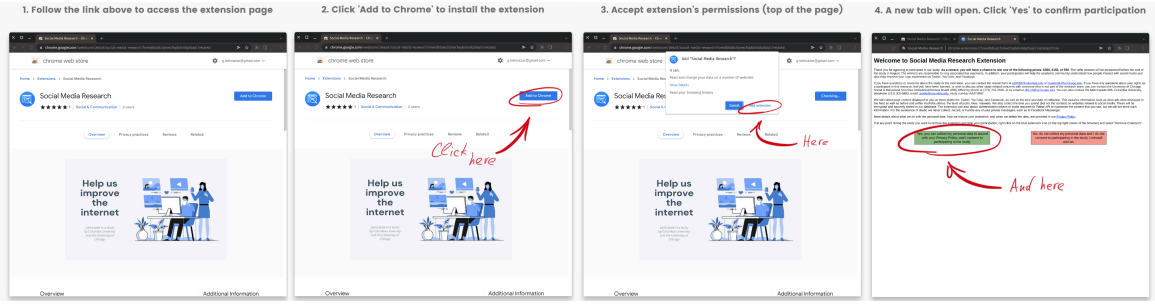


Figure A2: Screens from Installation Instructions GIF (Chrome Browser)

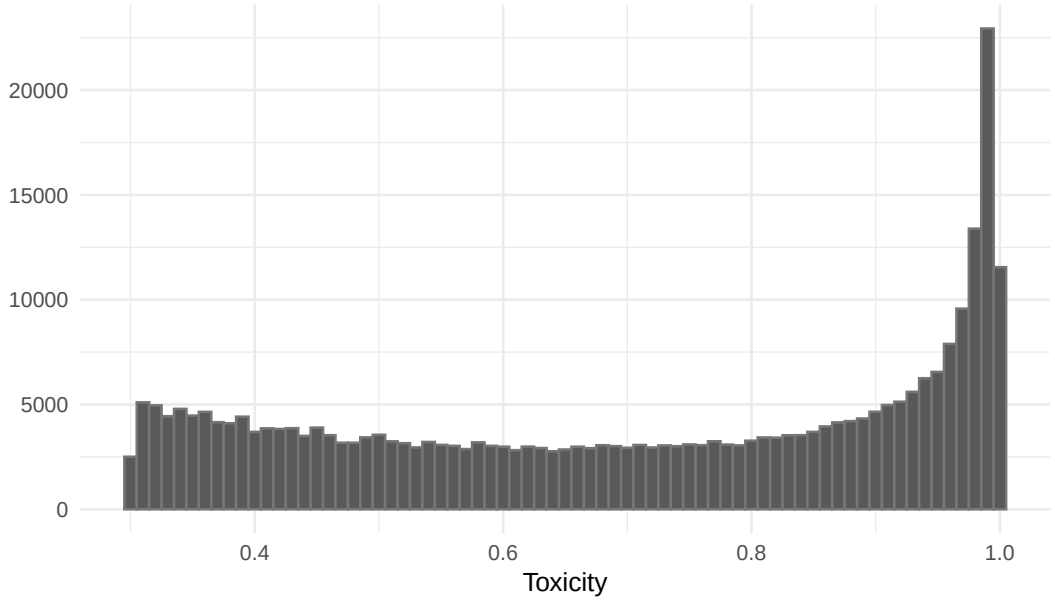


FIGURE A3: TOXICITY OF ELEMENTS ABOVE THE THRESHOLD

Note: The figure depicts the distribution of toxicity of posts, comments, and replies above the hiding threshold of 0.3. All elements with toxicity above the threshold were hidden for users in the treatment group during the intervention period. The data presented here encompasses the three platforms (Twitter, Facebook, and YouTube) and includes both the baseline and the intervention period.

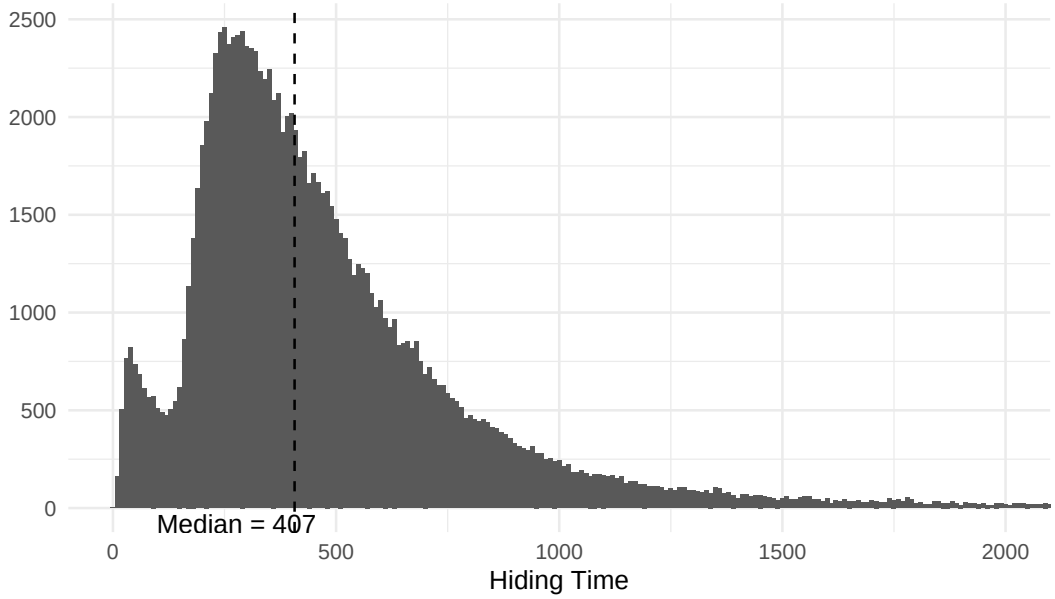
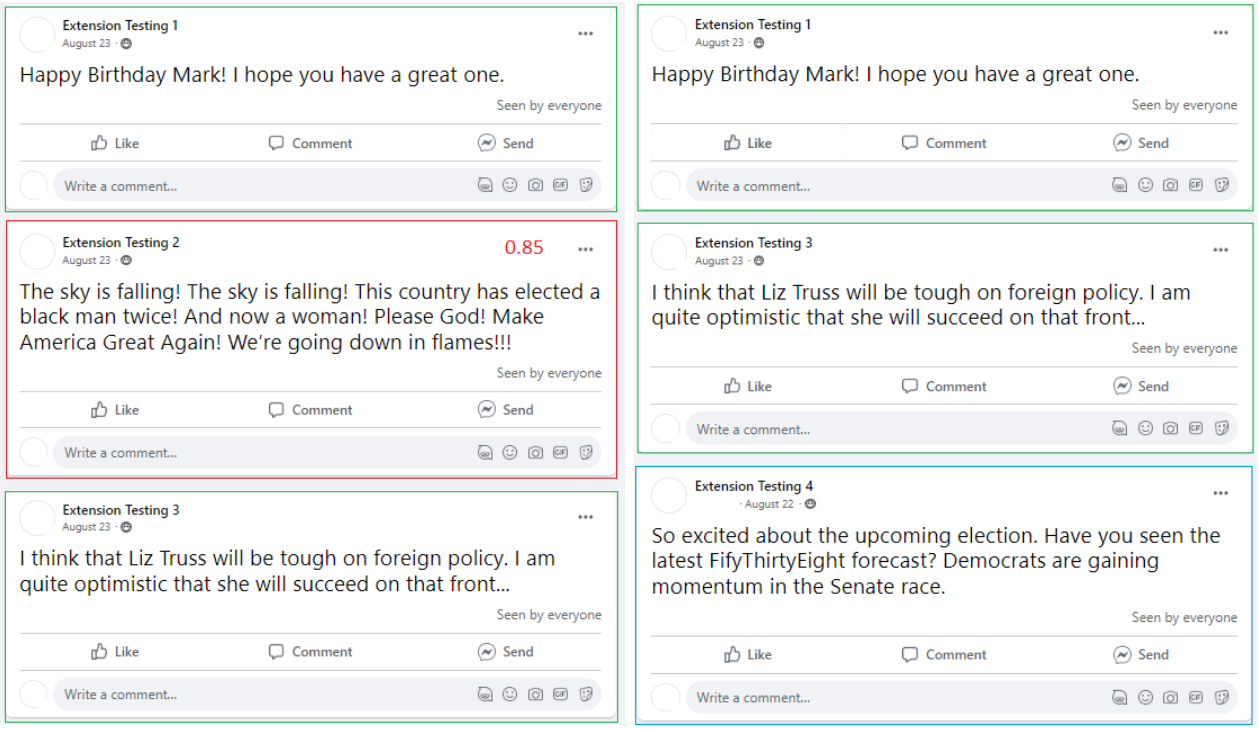


FIGURE A4: SPEED OF HIDING TOXIC CONTENT ON TWITTER

Note: The histogram depicts the distribution of the hiding speed for posts, comments, and replies on Twitter. The hiding speed is defined as the difference in the timestamp when an element was removed from the user's page by the extension and the timestamp when the element was first identified. The extension listened to changes in the DOM structure of the page (using Mutation Observer) in order to detect a new element appearing on the page. The hiding speed is reported in milliseconds. The histogram is truncated at 2000 milliseconds. We collected data on the hiding speed from August 22nd 2022 until the end of the study (end of September 2022).



(A) Original Feed (B) Moderated Feed

Figure A5: Hiding Intervention: Feed

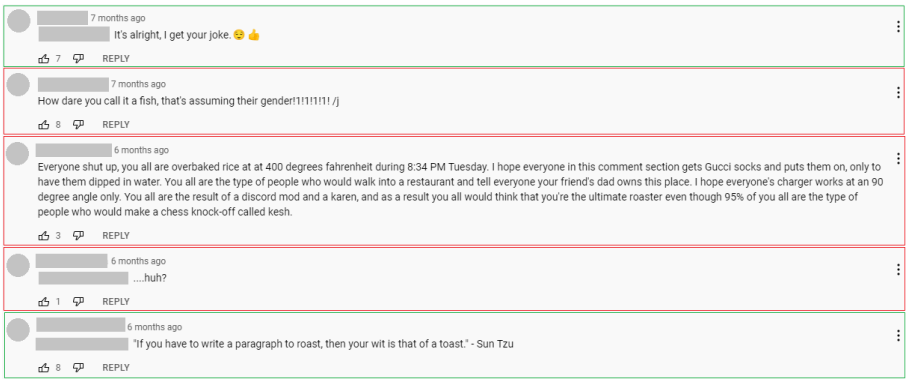
Note: Panel A shows an example of a Facebook group feed that we created for demonstrative purposes. Panel B depicts how this section would look for a user with the hiding intervention on. One post from Panel A (the element in a red frame) was removed, as it has a toxicity score of 0.85, above the hiding threshold of 0.3. The other two posts (green frames) were not classified as toxic. Panel B shows that the content below the hidden element is pulled up. Thus, the post by Extension Testing 3 is now directly below the one by Extension Testing 1. We also see a new element (blue frame), which was previously further below in the feed.



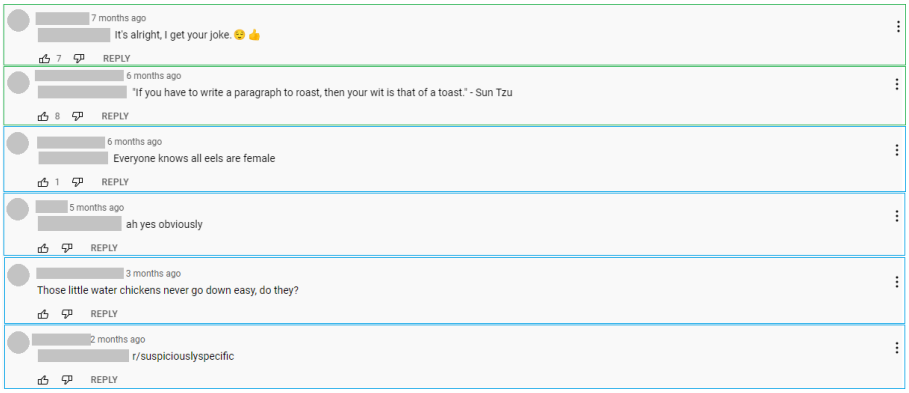
(A) Original Comments (B) Moderated Comments

Figure A6: Hiding Intervention: Comments

Note: Panel A shows an example of a comment section on Facebook that we created for demonstrative purposes. Panel B depicts how this section would look for a user with the hiding intervention on. Two comments were removed (red frames). The first one, "Come on, women are not as smart as men", has a toxicity score of 0.67. The second one, "Why does it matter? Your comments are pathetic...", has a score of 0.93. Note that replies are removed together with toxic comments (see the element in a gray frame). Panel B demonstrates that the content below the hidden elements is pulled up. We also see new elements (blue frames), which were previously further below.



(A) Original Comment Section



(B) Moderated Comment Section

Figure A7: Hiding Intervention on YouTube

Note: Panel A shows an example of a real comment section under a YouTube video. Panel B depicts how this section would look for a user with the hiding intervention on. Three comments from Panel A were removed (elements in red frames). Starting from the top, their toxicity scores were 0.42, 0.81, and 0.7, respectively. The last comment (not hidden) is just below the hiding threshold—with a score of 0.28. Overall, two of the comments from Panel A remained after the intervention was applied (elements in green frames). In Panel B, we see new elements (blue frames)—previously further below in the comment section—which replaced the hidden elements. The presented comments do not originate from our sample—they are publicly available online (as of 2022-10-31).



(A) Extension Disabled



(B) Extension Enabled

Figure A8: Show More Replies (Twitter)

Note: Panel A shows the bottom of the comments section on Twitter in the case when the extension is disabled—the user has to click "Show more replies" to load the remaining comments. Panel B depicts the same section in the case when the extension is enabled—the remaining comments are already loaded. The presented comments do not originate from our sample—they are publicly available online (as of 2022-10-31).

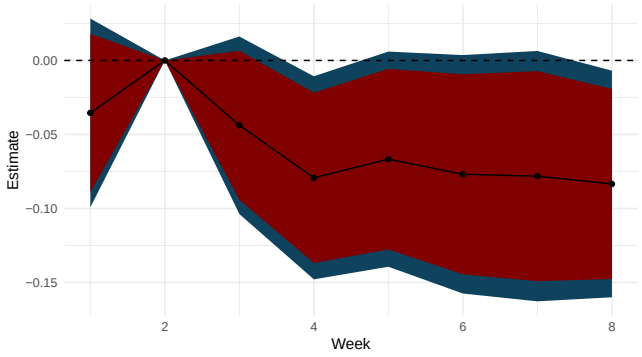


FIGURE A9: EVENT STUDY SPECIFICATION OF INDEX OF ENGAGEMENT

Note: This figure presents estimates from the event study version of Equation (1) for the preferred index of user engagement for all treated platforms (Facebook, Twitter, YouTube). The index is defined in Section 3.1.5. The unit of observation is the individual-week, where week is measured relative to the intervention date. We include users who were active on the last day of the study (day 56) or later. The blue area represents 95% confidence intervals. The red area represents 90% confidence intervals. Standard errors are clustered at the individual level.



(A) INITIAL DISPLAY (B) COMMENTS ON CLICK

FIGURE A10: IMAGE OF PRACTICE POST 1

Note: The figure demonstrates the first practice post, which is displayed to all participants in the survey experiment. **Panel A** provides the post as initially shown to the participants. **Panel B** demonstrates the post with the comment section that appears if the user clicks "View 2 comments."

A.2 Additional Tables

TABLE A1: DESCRIPTIVE STATISTICS

Panel A: User demographics

	Main Sample (mean)	Representative (mean)	Difference (t)
Age 18-29 (%)	30.80	30.99	0.06
Age 30-49 (%)	36.29	39.84	1.27
Age 50-64 (%)	22.57	20.76	-0.81
Male (%)	52.28	54.17	0.65
Democrat (%)	52.58	35.35	-6.13
Independent (%)	38.14	43.81	1.96
White (%)	64.94	69.24	1.53

Panel B: Twitter accounts

	Main Sample (mean)	Representative (mean)	Difference (t)
Account years	7.1	5.2	-9.34
Number of followers	1,715.1	4,803.9	4.58
Accounts followed	1,204.8	1,071.3	-1.42

Panel C: Baseline Outcomes

	Facebook, N= 638			Twitter, N= 742			YouTube, N= 724		
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Content shown/day	89	17	194	233	108	357	45	7	97
Feed elements/day	58	12	132	164	83	231	-	-	-
Comments/day	31	4	81	69	18	170	45	7	97
Toxicity/content shown	0.03	0.03	0.02	0.07	0.06	0.04	0.05	0.04	0.05
Reactions & posts /day	7	1	16	15	3	34	1	0	2
Toxicity/content produced	0.04	0.01	0.07	0.07	0.02	0.13	0.12	0.01	0.24
Minutes spent/day	17.0	3.8	32.4	29.5	12.4	45.7	10.5	2.5	20.1

Note: Panel A compares means of user characteristics in the main experimental sample (Main Sample) relative to a representative sample of Twitter users from the American Trends Panel (ATP) of September 2020 (Representative). It also presents t-statistics from tests of difference in means between both samples. Panel B compares Twitter accounts in our main sample relative to a random sample of 200,000 English Tweets collected in August 2020 from the 1% random sample of Twitter's API (Jiménez Durán, 2022). Panel C displays the mean, median, and standard deviation of some of our outcomes on Facebook and Twitter during the 14-day baseline period.

Table A2: Balance Table

		Control (N=351)		Treatment (N=391)			
		Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	p
Qualtrics		0.718	0.451	0.691	0.463	−0.027	0.414
Twitter API		0.880	0.325	0.847	0.361	−0.034	0.180
Use Facebook		58.211	27.248	54.359	29.134	−3.852	0.188
Use Twitter		60.434	26.120	60.585	25.408	0.151	0.949
Age		42.551	16.914	39.426	15.584	−3.125	0.038
Male		0.509	0.501	0.535	0.500	0.027	0.559
Democrat		0.550	0.499	0.504	0.501	−0.046	0.309
Independent		0.371	0.484	0.391	0.489	0.019	0.660
White		0.649	0.478	0.650	0.478	0.000	0.991
Private		0.091	0.288	0.066	0.249	−0.024	0.258
Followers		2246.469	16 812.660	1219.124	10 734.035	−1027.345	0.361
Friends		1352.874	2989.165	1066.471	1512.232	−286.402	0.131
Listed		32.786	239.926	21.082	107.892	−11.705	0.432
Years on Twitter		7.238	4.872	6.884	4.965	−0.354	0.363
Likes		17 310.625	40 054.627	16 703.601	41 853.629	−607.023	0.851
Tweets		13 354.515	45 482.157	8956.710	20 513.705	−4397.805	0.120
		N	Pct.	N	Pct.		
Region	Midwest	47	13.4	48	12.3		
	Northeast	39	11.1	48	12.3		
	Outside the US	4	1.1	4	1.0		
	South	80	22.8	87	22.3		
	West	58	16.5	67	17.1		
	NA	123	35.0	137	35.0		

Note: This table compares characteristics of users assigned to the treatment and control arms, for the main experimental sample. The top panel presents means, standard deviations, difference in means, and the p-value from a test of difference in means. We report the following variables in order: (1) a dummy equal to one if a user was matched to an intake survey response and zero otherwise, (2) a dummy equal to one if they were matched to Twitter API data, (3) desktop share of Facebook usage, (4) desktop share of Twitter usage, (5) age, in years, (6) a dummy equal to one if a person is male, (7) a dummy equal to one if a person is a Democrat, (8) a dummy equal to one if a person is an independent, (9) a dummy equal to one if a person reported white/Caucasian ethnicity, (10) a dummy equal to one if user's Twitter account was private, (11) the number of followers on Twitter, (12) the number of friends on Twitter, (13) the number of objects the user listed on Twitter, (14) the number of years since registering an account on Twitter, (15) the total number of posts and comments liked by the user over the account's lifetime, (16) the total number of tweets posted by the user over the account's lifetime. Variables 3-9 are based on responses to the intake survey. Variables 10-16 are based on Twitter API data. The bottom panel presents the distribution of users per region in both treatment arms.

Table A3: Effect of Intervention on Toxicity of Content Shown to Users

	Facebook		Twitter		YouTube	
	Mean Toxicity	Proportion Toxic	Mean Toxicity	Proportion Toxic	Mean Toxicity	Proportion Toxic
Treated	−0.020***	−0.033***	−0.048***	−0.069***	−0.034***	−0.053***
	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
Mean	0.0249	0.0209	0.0487	0.0459	0.035	0.0312
SD	0.0308	0.0416	0.0469	0.0592	0.0376	0.0499
N	12 044	12 044	19 311	19 311	10 248	10 248

Note: This table reports estimates from Equation (1) for our main experimental sample. The dependent variables are the average toxicity of the content shown to users (Mean Toxicity) and the proportion of content shown to users that is toxic, i.e., with a toxicity score exceeding 0.3 (Proportion Toxic). The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active on the last day of the study (day 56) or later. Driscoll-Kraay standard errors are parenthesized. *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A4: Heterogeneity by Baseline Exposure to Toxicity

	Preferred Engagement Index				Contagiousness			
	All Platforms	Facebook	Twitter	YouTube	All Platforms	Facebook	Twitter	YouTube
Treated	−0.018	0.005	0.017	0.010	−0.005	−0.003	−0.004	0.038
	(0.011)	(0.016)	(0.016)	(0.021)	(0.006)	(0.006)	(0.007)	(0.023)
Interaction	−0.076***	−0.149***	−0.092***	−0.056	−0.030***	−0.012***	−0.025**	−0.108**
	(0.012)	(0.033)	(0.014)	(0.035)	(0.010)	(0.004)	(0.010)	(0.044)
Mean	0.0245	0.108	0.0111	0.0664	0.0782	0.0373	0.0748	0.0816
SD	0.76	0.945	0.725	0.607	0.156	0.0851	0.145	0.20
N	37 072	26 936	36 456	28 000	12 997	6337	8955	1313

Note: This table reports estimates from Equation (1) for our main experimental sample. The dependent variables are the preferred index of user engagement (Preferred Engagement Index) and the average toxicity of the published content, conditional on posting (Contagiousness). The preferred index is precisely defined in Section 3.1.5. The independent variables are the treatment dummy and the treatment dummy interacted with a dummy equal to one if the individual had an above-median exposure to toxic content during the baseline period and zero otherwise. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active on the last day of the study (day 56) or later. Driscoll-Kraay standard errors are parenthesized. *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A5: Attrition Regressions

	(1)	(2)	(3)	(4)
Treatment	0.036		0.037	0.047
	(0.023)		(0.023)	(0.046)
Baseline Toxicity		−0.583	−0.556	−0.222
		(0.358)	(0.362)	(0.440)
Baseline Time on Social Media		−0.008	−0.008	−0.015*
		(0.005)	(0.005)	(0.008)
Baseline Toxicity × Treatment				−0.787
				(0.719)
Baseline Time on Social Media × Treatment				0.015
				(0.010)
Mean	0.894	0.894	0.894	0.894
SD	0.309	0.307	0.307	0.307
N	742	739	739	739

Note: This table reports estimates of an OLS regression on treatment assignment for our main experimental sample. The dependent variable is a dummy equal to one if a user completed 56 days of the study and zero otherwise. Column 2 includes the average toxicity score of content displayed to the user during the baseline (Baseline Toxicity), and its interaction with the treatment dummy. It also includes the average time spent on social media during the baseline (Baseline Time on Social Media), and its interaction with the treatment dummy. The unit of observation is the individual user. We include respondents who answered the endline survey. Robust standard errors are parenthesized. *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A6: Robustness to Alternative Attrition Threshold

	Preferred Engagement Index				Contagiousness			
	All Platforms	Facebook	Twitter	YouTube	All Platforms	Facebook	Twitter	YouTube
Treated	−0.045***	−0.056***	−0.025**	−0.009	−0.019***	−0.011**	−0.018***	−0.037
	(0.009)	(0.012)	(0.012)	(0.006)	(0.005)	(0.005)	(0.006)	(0.035)
Mean	0.0228	0.0367	0.005 41	0.0212	0.0793	0.037	0.0756	0.0877
SD	0.762	0.868	0.727	0.544	0.159	0.0844	0.146	0.209
N	38 864	33 600	38 864	38 080	13 502	6556	9340	1388

Note: This table reports estimates from Equation (1) for our main experimental sample. The dependent variables are the preferred index of user engagement (Preferred Engagement Index) and the average toxicity of the published content, conditional on posting (Contagiousness). The preferred index is precisely defined in Section 3.1.5. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active on day 42 of the study or later. Driscoll-Kraay standard errors are parenthesized. *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A7: Robustness to Using Stacked Regression

	Preferred Engagement Index				Contagiousness			
	All Platforms	Facebook	Twitter	YouTube	All Platforms	Facebook	Twitter	YouTube
Treated	−0.056***	−0.061***	−0.037***	−0.011*	−0.020***	−0.011**	−0.021***	−0.073
	(0.009)	(0.012)	(0.012)	(0.006)	(0.006)	(0.005)	(0.007)	(0.044)
Mean	0.0232	0.0432	0.003 99	0.0188	0.0782	0.0371	0.0748	0.0805
SD	0.76	0.875	0.72	0.54	0.156	0.0848	0.145	0.198
N	37 184	32 312	37 184	36 456	12 997	6411	8962	1341

Note: This table reports estimates from Equation (1) with start date × period fixed effects for our main experimental sample. The dependent variables are the preferred index of user engagement (Preferred Engagement Index) and the average toxicity of the published content, conditional on posting (Contagiousness). The preferred index is precisely defined in Section 3.1.5. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active on the last day of the study (day 56) or later. Driscoll-Kraay standard errors are parenthesized. *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A8: Robustness to Alternative Standard Errors

	Preferred Engagement Index				Contagiousness			
	All Platforms	Facebook	Twitter	YouTube	All Platforms	Facebook	Twitter	YouTube
Treated	−0.054*	−0.063**	−0.030	−0.015	−0.020***	−0.011**	−0.019***	−0.037
	(0.030)	(0.030)	(0.033)	(0.020)	(0.007)	(0.005)	(0.007)	(0.041)
Mean	0.0232	0.0432	0.003 99	0.0188	0.0782	0.0371	0.0748	0.0805
SD	0.76	0.875	0.72	0.54	0.156	0.0848	0.145	0.198
N	37 184	32 312	37 184	36 456	12 997	6411	8962	1341

Note: This table reports estimates from Equation (1) for our main experimental sample. The dependent variables are the preferred index of user engagement (Preferred Engagement Index) and the average toxicity of the published content, conditional on posting (Contagiousness). The preferred index is precisely defined in Section 3.1.5. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active on the last day of the study (day 56) or later. Individually-clustered standard errors are parenthesized. *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A9: Descriptive Statistics (Survey Experiment)

	Summary
N	4,120
Age (years)	41.608 (12.537)
Male	0.494 (0.500)
Minority Status	0.457 (0.498)
Religious	0.640 (0.480)
College $\geq$ 4 Years	0.567 (0.496)
Household Income $\geq$ \$70k	0.488 (0.500)
Days on Social Media per Week	6.261 (1.512)
WTA Comprehension	0.617 (0.486)
Full Comprehension	0.523 (0.500)
Survey Duration (minutes)	9.757 (7.167)
Comments Post 1	0.417 (0.493)
Comments Post 2	0.343 (0.475)
WTA 100 Posts (\$)	13.378 (6.737)

Note: The table presents descriptive statistics for the sample of participants who were assigned a treatment condition. For each variable, we report the sample mean and the standard deviation (in parentheses). We report the following variables in order: (1) age, in years, (2) a dummy equal to one if a person is male and zero otherwise, (3) a dummy equal to one if they have a minority status, i.e., they selected an answer other than "white" in a question about ethnicity, or they identify as a Hispanic, Latino, or a person of Spanish origin, or they reported a different sexual orientation than heterosexual, (4) a dummy equal to one if they declared that they have a religion, (5) a dummy equal to one if they have completed a 4 year degree or a post-graduate degree, (6) a dummy equal to one if they have a household income greater than \$70,000, (7) the number of days in a week that they consume social media, (8) a dummy equal to one if they correctly answered comprehension questions related to the second price auction procedure, (9) a dummy equal to one if they correctly answered all comprehension questions, (10) the duration of the survey in minutes, (11) a dummy equal to one if they clicked to uncover the comments under the first practice post (NFL post), (12) a dummy equal to one if they clicked to uncover the comments under the second practice post (the treatment post), (13) willingness to accept for the task of transcribing 100 social media posts, in US dollars.

A.3 Additional Analysis: Browser Experiment

In this section, we present additional analysis and robustness results for the browser experiment.

A.3.1 Sensitivity Analysis for Selected Outcomes

We present sensitivity analysis related to two outcomes—active time spent on social media and ad clicks. We do so to address concerns that the specific way in which we defined them drives the results.

Time Spent on Social Media. Recall that when computing active time spent on a social media platform, we define the end of the session as the user either (1) switching to another platform (indicated by the extension recording a new element from another platform or a ping related to another platform) or (2) becoming inactive, i.e., there have been no new elements loaded for at least 3 minutes. The choice of 3 minutes without loading new content as the time required to consider a user inactive is based on our estimation of the average amount of content on a desktop screen and the average reading speed of a human (i.e., after 3 minutes the user has likely consumed everything there is on the page). Table A10 demonstrates robustness to alternative inactivity thresholds (1, 2, 4, 5, and 10 minutes). In the paper, we report that the intervention has negative effects on the time spent on social media on both Facebook and YouTube. The effect on Facebook is robust to all five alternative inactivity thresholds at the 5% level. The effect on YouTube is robust to 1-, 2-, and 4-minute inactivity thresholds at the 5% level and to a 5-minute inactivity threshold at the 10% level. Overall, we demonstrate that the results reported in the paper are not driven by choosing a specific inactivity threshold.

Ad Clicks. In the paper, ad clicks are measured by observing pings that indicate that a user accessed a new website, which is not among the platforms that we track, shortly after an ad was displayed to them on social media. We define "shortly after" as before another 12 new elements on the page load and within 2 minutes. These criteria reflect the fact that elements are loaded in batches—the ad is likely to be recorded with the same or a very similar timestamp as posts next to it. Table A11 offers robustness to alternative criteria—before another 8, 10, or 15 elements load on the page and within 3 or 4 minutes. The negative effect of the intervention on ad clicks on Twitter is robust to 11 out of 12 combinations of the number of elements and time breaks. The negative effect on Facebook is robust to extending the time break from 2 to 3 minutes at all levels of the number of elements. However, it is not robust to the time break of 4 minutes, likely due to a long time elapsed since the ad was shown, which inflates the noise. We conclude that our results on ad clicks are robust to alternative definitions of the outcome, especially on Twitter.

TABLE A10: ACTIVE TIME: ROBUSTNESS TO ALTERNATIVE INACTIVITY THRESHOLDS

	Inactivity Threshold					
	1	2	3	4	5	10
Facebook						
Treated	-1.180***	-1.278***	-1.314***	-1.328***	-1.292***	-1.239**
	(0.292)	(0.365)	(0.402)	(0.437)	(0.466)	(0.586)
Mean	9.18	12.36	14.31	15.74	16.89	20.72
SD	24.20	31.33	35.53	38.63	41.14	49.69
N	32312	32312	32312	32312	32312	32312
Twitter						
Treated	-0.319	-0.460	-0.351	-0.246	-0.169	-0.161
	(0.765)	(0.909)	(1.001)	(1.058)	(1.132)	(1.348)
Mean	17.14	21.38	23.76	25.45	26.76	31.07
SD	35.93	44.34	48.53	51.58	53.97	62.07
N	37184	37184	37184	37184	37184	37184
YouTube						
Treated	-0.512***	-0.541**	-0.611**	-0.632**	-0.618*	-0.363
	(0.189)	(0.237)	(0.269)	(0.305)	(0.342)	(0.482)
Mean	4.32	7.04	9.04	10.71	12.15	17.42
SD	15.28	21.77	25.32	28.46	31.26	42.01
N	36456	36456	36456	36456	36456	36456

Note: This table reports estimates from Equation (1) for our main experimental sample. The dependent variable is the total active time in minutes spent on a particular social media platform. Each column corresponds to a different inactivity threshold (1, 2, 3, 4, 5, and 10 minutes) used in computing the total active time (see Section 3.1.5 in the paper for details on how the outcome is defined). The top panel presents the results on Facebook, the middle panel on Twitter, and the bottom panel on YouTube. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active on the last day of the study (day 56) or later. Driscoll-Kraay standard errors are parenthesized. *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

A.3.2 Endline Survey

Outcomes As discussed in the paper, we collected additional outcomes in the endline survey to complement those recorded by the browser extension. First, without revealing any further information about the study, we used a dynamic MPL to elicit participants' willingness to pay (or accept) for keeping our extension installed for another month. We asked participants whether they would prefer to "keep our browser extension installed for one more month" or "receive \$X", with the possible values of $X \in \{0, 0.5, 1, 1.5, 2, 4, 6\}$. In addition, we asked whether they would prefer to "keep our browser extension installed for one more month AND receive \$Y" or "receive \$0", with the possible values of $Y \in \{0, 0.5, 1, 1.5, 2, 4, 6\}$. We randomly selected 10 participants for whom one of the MPL choices was implemented.

Second, we elicited the impact of the intervention on participants' self-reported well-being. Here, we followed the methodology proposed by Allcott et al. (2020) by selecting six of their survey questions encapsulating subjective well-being. Three measures pertain to positive emotions and behavior: happiness, life satisfaction, being absorbed in doing something worthwhile. The other three focus on the negative aspects: depression, anxiety, and boredom. To evaluate the outcome, we created an index aggregating the answers to the six questions. For each individual, we computed $\frac{1}{6}\sum_{i=1}^{6}\frac{y_i-\bar{y}_i}{\sigma_i}$, where y_i is the numerical answer to the i^{th} question, $\bar{y}_i$ is

Table A11: Ad Clicks: Robustness to Alternative Definitions

	Elements							
	Facebook				Twitter			
	8	10	12	15	8	10	12	15
2-Minute Break								
Treated	-0.037**	-0.037**	-0.037**	-0.040**	-0.043**	-0.050***	-0.051***	-0.057***
	(0.015)	(0.015)	(0.015)	(0.016)	(0.016)	(0.017)	(0.018)	(0.019)
Mean	0.187	0.194	0.197	0.201	0.257	0.274	0.285	0.297
SD	0.876	0.903	0.913	0.925	0.962	1.03	1.07	1.13
N	33152	33152	33152	33152	38360	38360	38360	38360
3-Minute Break								
Treated	-0.042**	-0.046**	-0.046**	-0.050**	-0.039*	-0.050**	-0.053**	-0.057**
	(0.018)	(0.018)	(0.019)	(0.020)	(0.022)	(0.022)	(0.022)	(0.023)
Mean	0.27	0.283	0.289	0.296	0.342	0.368	0.387	0.407
SD	1.06	1.10	1.11	1.13	1.10	1.17	1.23	1.29
N	33152	33152	33152	33152	38360	38360	38360	38360
4-Minute Break								
Treated	-0.030	-0.033	-0.032	-0.038	-0.041	-0.051**	-0.054**	-0.060**
	(0.020)	(0.021)	(0.022)	(0.023)	(0.025)	(0.025)	(0.025)	(0.026)
Mean	0.319	0.336	0.345	0.354	0.39	0.421	0.445	0.471
SD	1.18	1.22	1.24	1.26	1.17	1.25	1.32	1.39
N	33152	33152	33152	33152	38360	38360	38360	38360

Note: This table reports estimates from Equation (1) for our main experimental sample. The dependent variables are the total active time in minutes spent on social media platforms (Time) and the number of separate browsing sessions (Sessions). Precise definitions of these outcomes are provided in Section 3.1.5. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active on the last day of the study (day 56) or later. Driscoll-Kraay standard errors are parenthesized. *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

its mean, and σ_i the standard deviation, with the negative measures (a higher value indicates lower well-being) re-scaled by -1. In each question, we emphasized the period of interest—the last six weeks, focusing attention on the intervention time.

Lastly, to analyze whether a lower exposure to toxic content reduces users' normalization of hateful attitudes, we asked the participants to read seven online comments and indicate to what extent they consider each of them toxic. The statements were displayed in random order. We selected the texts from the training dataset for the Jigsaw challenges. The chosen statements represent different degrees of toxicity, with Jigsaw's toxicity scores ranging from 0.4 to 0.93. We provided the survey participants with the same definitions of toxicity and the same comment evaluation scale as the ones faced by Jigsaw's annotators. We computed the proportion of people who reported each statement to be "Toxic" or "Very Toxic" to maintain the original fractional interpretation of toxicity scores. Then, we averaged the proportions across the statements to report the final outcome.

Results Table A12 shows that the hiding intervention had an insignificant effect on the willingness to pay/accept for using the browser extension for an additional month. The sample size of 375 users is small, especially given that we perform between-subjects analaysis, without being able to rely on difference-in-differences estimation. Hence, we treat the null result as

inconclusive.

Table A12: Effect of Intervention on Users' WTA/WTP for the Extension

	Willingness to Pay/Accept
Treated	0.125
	(0.289)
Mean	4.80
SD	2.78
N	375

Note: This table reports estimates of an OLS regression on treatment assignment for our main experimental sample. The dependent variable is the willingness to pay/accept for using the browser extension for an extra month. The unit of observation is the individual user. We include respondents who answered the endline survey. Robust standard errors are in parentheses. *significant at 10%; ** significant at 5%; *** significant at 1%.

Table A13 focuses on measures of user well-being collected in the endline survey. Overall, we do not detect a significant effect of the hiding intervention on the index of individual well-being. Moreover, we find that the treatment had no significant impact on any components of the index considered in isolation. These findings provide suggestive evidence that exposure to toxicity may not be the main driving force behind the negative effects of social media on well-being, a relationship well-documented in the literature. This point should be treated with caution given the small sample size. We hope that our design will be applied in the future to investigate toxicity's impact on well-being with a larger group of users, an important step in understanding the mechanisms through which social media affects individual welfare.

Table A13: Effect of Intervention on Users' Well-Being

	Index	Happiness	Satisfaction	Depression	Anxiety	Worthwile	Boredom
Treated	-0.027	-0.148	-0.073	0.019	0.091	-0.065	-0.043
	(0.075)	(0.141)	(0.159)	(0.091)	(0.091)	(0.092)	(0.103)
Mean	-0.00	4.79	4.97	-1.92	-2.26	2.77	-1.95
SD	0.72	1.36	1.52	0.87	0.87	0.88	0.99
N	370	373	370	370	370	370	370

Note: This table reports estimates of an OLS regression on treatment assignment for our main experimental sample. The dependent variables are an index of well-being and its components. The wording of survey measures for each component are provided in Section A.5.3. The unit of observation is the individual user. We include respondents who answered the endline survey. Robust standard errors are in parentheses. *significant at 10%; ** significant at 5%; *** significant at 1%.

Lastly, we investigate a potential mechanism behind contagiousness of toxicity, namely, that exposure to toxic content contributes to normalization of toxic behavior, which then increases the likelihood that users engage in such behavior. Despite the overall strong effect of the exposure on toxicity of own content (as reported in the paper), we find no evidence of normalization of toxicity. The results are presented in Table A14. The effect on the index summarizing users' ratings of the seven toxic statements is insignificant, which offers suggestive evidence that ex-

posure to toxicity does not change their opinions on what is considered toxic. We also do not detect significant effects on toxicity ratings for any of the seven statements individually.

TABLE A14: EFFECT OF INTERVENTION ON USERS' RATINGS OF TOXIC STATEMENTS

	Index	C1	C2	C3	C4	C5	C6	C7
Treated	-0.001	-0.058	0.021	-0.007	-0.005	-0.004	-0.029	0.077
	(0.023)	(0.050)	(0.041)	(0.049)	(0.027)	(0.052)	(0.043)	(0.051)
Mean	0.59	0.65	0.81	0.68	0.93	0.45	0.22	0.40
SD	0.22	0.48	0.40	0.47	0.26	0.50	0.41	0.49
N	367	367	367	367	367	367	367	367

Note: This table reports estimates of an OLS regression on treatment assignment for our main experimental sample. The dependent variables are an index of users' evaluation of the toxicity of 7 social media posts and its components (C1-C7). The statements are provided in Section A.5.3. The unit of observation is the individual user. We include respondents who answered the endline survey. Robust standard errors are in parentheses. *significant at 10%; ** significant at 5%; *** significant at 1%.

A.3.3 Topic Analysis

In Section 3.3.2 of the paper, we report results of classifying posts and comments shown to users during the study into 26 different topic categories.[27] We relied on gpt-4o with temperature set to zero. Below, we provide the exact wording of prompts used for the analysis.

System prompt *You'll be asked to say what percent of social media posts talk about 26 topics. Reply with ONLY a comma-separated list of 26 numbers. Each number should be between 0 and 1 and have 2 decimals, representing the % of posts that talk about each topic. Each post can talk about multiple topics.*

User prompt *What percent of posts below talk about each of the following 26 topics?: 1) Automotive, 2) Beauty, 3) Books and literature, 4) Business, 5) Careers, 6) Education, 7) Events, 8) Family and parenting, 9) Food and drink, 10) Gaming, 11) Health, 12) Hobbies and interests, 13) Home and garden, 14) Law, government, and politics, 15) Life stages, 16) Movies and television, 17) Music and radio, 18) Personal finance, 19) Pets, 20) Science, 21) Society, 22) Sports, 23) Style and fashion, 24) Technology and computing, 25) Travel, 26) Other.* Subsequently, we provided user's posts on a given day.

A.4 Additional Analysis: Survey Experiment

In this section, we present additional analysis and robustness results for the survey experiment.

[27] These are the topic categories that Twitter relies on when allowing advertisers to target specific users with their ads.

A.4.1 Robustness to Excluding Pilot Observations

Following our pre-registration, and in order to maximize statistical power, the sample in the paper (N=4,120) includes 414 pilot observations. Below, we reproduce regression tables for our primary outcomes excluding the pilot observations. Table A15 shows the treatment effects on user engagement, measured as the likelihood of clicking to uncover the comment section below the treatment post. All significant results on user engagement reported in the paper are robust to excluding the pilot observations. This includes the main specification comparing the pooled treatments against the pooled controls ($p<0.001$), as well as the effects of different types of toxicity (hate speech and profanity).

Table A15: User Engagement (Pilot Excluded)

	Pooled		Hate		Profanity	
	All	Comp.	All	Comp.	All	Comp.
Toxic	0.059***	0.083***	0.093***	0.103***	0.025	0.063**
	(0.016)	(0.022)	(0.022)	(0.032)	(0.022)	(0.032)
Mean	0.34	0.42	0.35	0.42	0.34	0.41
SD	0.47	0.49	0.48	0.49	0.47	0.49
N	3638	1921	1816	961	1822	960

Note: The table is based on a sample of all people who completed the second practice post, excluding the pilot observations (N=3,638). Column 1 shows the results of a regression of a dummy variable equal to one if a participant clicked to uncover the comment section below the second practice post and zero otherwise on a dummy variable equal to one if they were assigned one of the toxic treatments (either the hate speech treatment or the profanity treatment) and zero otherwise. Column 2 provides the same regression but with a sample restricted to individuals who passed all comprehension checks. Columns 3-4 pertain to specifications analogous to those in Columns 1-2 but with the sample restricted to individuals who were in the hate speech treatment or the hate speech control. Columns 5-6 pertain to specifications analogous to those in Columns 1-2 but with the sample restricted to individuals who were in the profanity treatment or the profanity control. Robust standard errors are in parentheses. *significant at 10%; ** significant at 5%; *** significant at 1%.

Table A16 shows treatment effects on the WTA to transcribe 100 social media posts, excluding the pilot observations. As highlighted in the paper, the result that hate speech increases the WTA for the transcription task is not robust to the exclusion, although the relevant coefficients remain positive. The results on profanity's impact are virtually unchanged, with coefficients very close to zero. We interpret the results as an indication that the engagement findings cannot be explained by people having *positive* utility of consuming toxic content.

A.4.2 Heterogeneity Analysis

As discussed in the paper, we test for heterogeneity of the treatment effects by gender, age, minority status, and religiosity. Table A17 indicates no heterogeneous effects of hate speech, against the associated control, on user engagement. Similarly, Table A18 reveals no heterogeneous effects of profanity, against the associated control, on gender, minority status, or religiosity. However, we report an interesting result by age. For older individuals (35+), the effect

Table A16: WTA for Transcribing (Pilot Excluded)

	Hate			Profanity		
	All	WTA Comp.	Comp.	All	WTA Comp.	Comp.
Toxic	0.455	0.455	0.300	0.125	0.076	-0.012
	(0.319)	(0.406)	(0.439)	(0.314)	(0.388)	(0.421)
Mean	13.67	13.44	13.32	13.27	12.91	12.89
SD	6.79	6.82	6.81	6.71	6.49	6.52
N	1815	1128	961	1822	1123	960

Note: The table is based on a sample of all people who provided the willingness to accept for transcribing 100 social media posts, excluding the pilot observations (N=3,637). Column 1 shows the results of a regression of the willingness to accept (WTA) for transcribing 100 social media posts on a dummy variable equal to one if the participant was assigned the hate speech treatment and zero if they were assigned the hate speech control. Column 2 provides the same regression but with a sample restricted to individuals who passed the comprehension checks about the WTA elicitation. Column 3 provides the same regression as Column 1 but with a sample restricted to individuals who passed all comprehension checks. Column 4 shows the results of a regression of the WTA for transcribing 100 social media posts on a dummy variable equal to one if the participant was assigned the profanity treatment and zero if they were assigned the profanity control. Column 5 provides the same regression but with a sample restricted to individuals who passed the comprehension checks about the WTA elicitation. Column 6 provides the same regression as Column 4 but with a sample restricted to individuals who passed all comprehension checks. Robust standard errors are in parentheses. *significant at 10%; ** significant at 5%; *** significant at 1%.

of profanity on user engagement is strong (5.9 pp). Yet, the overall effect is dampened by a significantly lower effect for young people (p=0.027)—in fact, it has the opposite sign (−4.1 pp).

Table A19 reveals two potential moderators for the effects of hate speech on the WTA to transcribe 100 social media posts at the 10% level. Both religiosity and belonging to a minority group *weaken* positive effects on the WTA. This suggests the following association: those more likely to be targeted by hate speech (members of minority groups) adjust the WTA by less in the presence of hate speech. Lastly, Table A20 indicates no significant moderators for the effects of profanity on the WTA.

Table A17: Heterogeneous Effects of Hate Speech on User Engagement

	(1)	(2)	(3)	(4)	(5)
Toxic	0.094***	0.088***	0.085***	0.082***	0.114***
	(0.021)	(0.030)	(0.026)	(0.028)	(0.036)
Toxic×Male		0.013			
		(0.042)			
Toxic×Young			0.029		
			(0.044)		
Toxic×Minority				0.027	
				(0.042)	
Toxic×Religious					-0.030
					(0.044)
Mean	0.35	0.35	0.35	0.35	0.35
SD	0.48	0.48	0.48	0.48	0.48
N	2021	2021	2021	2021	2021

Note: Column 1 shows the results of a regression of a dummy variable equal to one if a participant clicked to uncover the comment section below the second practice post and zero otherwise on a dummy variable equal to one if they were assigned the hate speech treatment and zero if they were assigned the hate speech control. Specifications in Columns 2-5 extend specification (1) by including an additional dummy variable and its interaction with the hate speech dummy. In Column 2, the additional variable is a dummy equal to one if the individual is male. In Column 3, it is a dummy equal to one if the individual is younger than 35 years old. In Column 4, it is a dummy equal to one if the individual is a member of a minority group (based on ethnicity or sexual orientation). In Column 5, it is a dummy equal to one if the individual reports being religious. Robust standard errors are in parentheses. *significant at 10%; ** significant at 5%; *** significant at 1%.

Table A18: Heterogeneous Effects of Profanity on User Engagement

	(1)	(2)	(3)	(4)	(5)
Toxic	0.028	0.029	0.059**	0.027	0.013
	(0.021)	(0.030)	(0.025)	(0.028)	(0.035)
Toxic×Male		-0.006			
		(0.042)			
Toxic×Young			-0.100**		
			(0.045)		
Toxic×Minority				0.002	
				(0.042)	
Toxic×Religious					0.022
					(0.044)
Mean	0.34	0.34	0.34	0.34	0.34
SD	0.47	0.47	0.47	0.47	0.47
N	2028	2028	2028	2028	2028

Note: Column 1 shows the results of a regression of a dummy variable equal to one if a participant clicked to uncover the comment section below the second practice post and zero otherwise on a dummy variable equal to one if they were assigned the profanity treatment and zero if they were assigned the profanity control. Specifications in Columns 2-5 extend specification (1) by including an additional dummy variable and its interaction with the profanity dummy. In Column 2, the additional variable is a dummy equal to one if the individual is male. In Column 3, it is a dummy equal to one if the individual is younger than 35 years old. In Column 4, it is a dummy equal to one if the individual is a member of a minority group (based on ethnicity or sexual orientation). In Column 5, it is a dummy equal to one if the individual reports being religious. Robust standard errors are in parentheses. *significant at 10%; ** significant at 5%; *** significant at 1%.

Table A19: Heterogeneous Effects of Hate Speech on WTA to Transcribe

	(1)	(2)	(3)	(4)	(5)
Toxic	0.624**	0.314	0.543	1.151***	1.364***
	(0.301)	(0.402)	(0.380)	(0.405)	(0.508)
Toxic×Male		0.631			
		(0.604)			
Toxic×Young			0.322		
			(0.609)		
Toxic×Minority				-1.165*	
				(0.605)	
Toxic×Religious					-1.142*
					(0.631)
Mean	13.61	13.61	13.61	13.61	13.61
SD	6.77	6.77	6.77	6.77	6.77
N	2020	2020	2020	2020	2020

Note: Column 1 shows the results of a regression of the willingness to accept for the task of transcribing 100 social media posts on a dummy variable equal to one if they were assigned the hate speech treatment and zero if they were assigned the hate speech control. Specifications in Columns 2-5 extend specification (1) by including an additional dummy variable and its interaction with the hate speech dummy. In Column 2, the additional variable is a dummy equal to one if the individual is male. In Column 3, it is a dummy equal to one if the individual is younger than 35 years old. In Column 4, it is a dummy equal to one if the individual is a member of a minority group (based on ethnicity or sexual orientation). In Column 5, it is a dummy equal to one if the individual reports being religious. Robust standard errors are in parentheses. *significant at 10%; ** significant at 5%; *** significant at 1%.

Table A20: Heterogeneous Effects of Profanity on WTA to Transcribe

	(1)	(2)	(3)	(4)	(5)
Toxic	0.033	-0.072	0.112	0.022	-0.568
	(0.297)	(0.400)	(0.367)	(0.407)	(0.509)
Toxic×Male		0.190			
		(0.596)			
Toxic×Young			-0.034		
			(0.619)		
Toxic×Minority				0.063	
				(0.595)	
Toxic×Religious					0.948
					(0.627)
Mean	13.15	13.15	13.15	13.15	13.15
SD	6.70	6.70	6.70	6.70	6.70
N	2028	2028	2028	2028	2028

Note: Column 1 shows the results of a regression of the willingness to accept for the task of transcribing 100 social media posts on a dummy variable equal to one if they were assigned the profanity treatment and zero if they were assigned the profanity control. Specifications in Columns 2-5 extend specification (1) by including an additional dummy variable and its interaction with the profanity dummy. In Column 2, the additional variable is a dummy equal to one if the individual is male. In Column 3, it is a dummy equal to one if the individual is younger than 35 years old. In Column 4, it is a dummy equal to one if the individual is a member of a minority group (based on ethnicity or sexual orientation). In Column 5, it is a dummy equal to one if the individual reports being religious. Robust standard errors are in parentheses. *significant at 10%; ** significant at 5%; *** significant at 1%.

A.5 Supplementary Materials

This section contains supplementary materials accompanying both the browser (Section 3 of the paper) and the survey (Section 4 of the paper) experiments. In Section A.5.1, we present additional information about the browser extension, such as the onboarding process and the privacy policy. In Section A.5.2, we discuss our secondary method of recruitment—promotion by the Mozilla Foundation. In Section A.5.3, we provide the wording of all demographic questions as well as questions used to elicit outcomes in the browser experiment (intake survey and endline survey). Section A.5.4 contains the corresponding materials for the survey experiment.

A.5.1 Browser Extension

We provide more information about our custom-built browser extension *Social Media Research*. In particular, we outline the installation sequence, onboarding, and our privacy policy.

Store Listing

During the intake survey, we provided each individual with a link to the store compatible with their browser. On clicking the link, users accessed our extension's store listing page (Figure A11), which outlined the core functionality, our privacy policy, and contact details of the researchers and the IRBs.

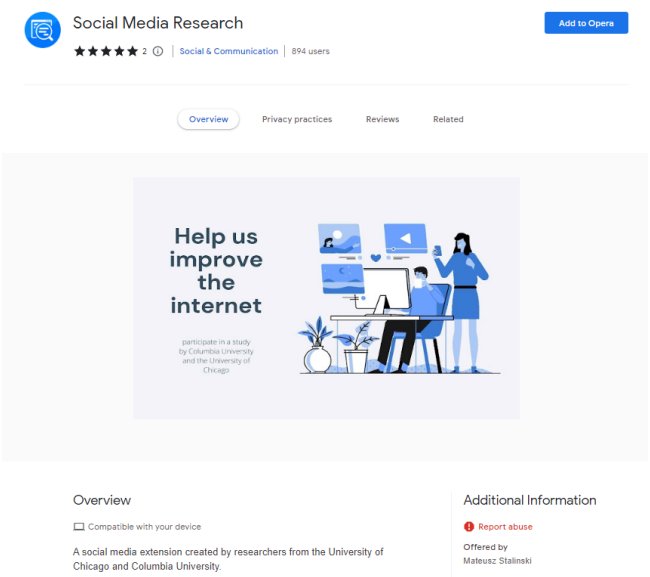


FIGURE A11: OUR EXTENSION'S STORE LISTING PAGE

Prospective users could read that their participation in the study helps "the academic community understand how people interact with social media." and that the extension "can improve [their] user experience on Twitter, YouTube, and Facebook". Furthermore, we informed them

that the extension "may optimize [their] Twitter, YouTube, and Facebook pages by changing page content". The store description did not directly reference hate speech or moderation of toxic content. In an attempt to obfuscate the exact purpose of the study, we chose to describe the functionality in general high-level terms that among other things could include hiding toxic content. Following the advice from the IRBs overseeing the study, we provided a more precise description of the purpose of the study in a debriefing script disseminated after the project's conclusion.

The privacy policy on the store listing page explained what types of data can be collected by the extension: "We will collect page content displayed to you on three platforms: Twitter, YouTube, and Facebook, as well as the time and date of collection". We highlighted that this includes information such as "the texts of posts, likes, retweets" and that "we will also collect the time [they] spend (but not the content) on websites related to social media". Additionally, we assured the participants that the collected data is encrypted when stored in our database. The decryption key is known only to the research team, thus reducing the risk of confidentiality breach even in the unlikely event that the database is accessed by an unauthorized person.

Installation and Onboarding

The installation process was uncomplicated, and likely familiar to many users. First, it required clicking a blue "Add" button in the top right corner of the store listing page (Figure A11), which prompted a confirmation screen where the user had to accept the required permissions for the extension. Second, upon completing the previous step, the extension opened a new tab with the onboarding screen (Figure A12).

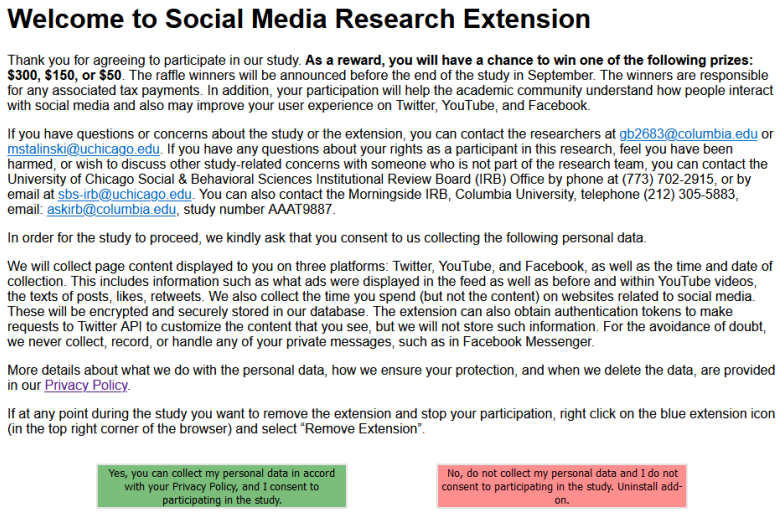


FIGURE A12: ONBOARDING PAGE

The main purpose of onboarding was obtaining affirmative consent for data collection. A

description of the types of data that the extension records was repeated on the page alongside with information about compensation (gift card raffle) and contact details (of the research team and the IRBs). The user had two options to choose from: (1) "Yes, you can collect my personal data in accord with your Privacy Policy, and I consent to participate in the study" and (2) "No, do not collect my personal data, and I do not consent to participating in the study. Uninstall the add-on." The extension was programmed in a way that prevented any data recording unless the user clicked option (1). While most of the content on the onboarding screen was duplicating either the information from the intake survey or the store listing, it was an essential part of the process. In particular, we wanted to ensure that it was crystal clear to the participants what data is being obtained (especially the PII), and that an explicit authorization was given for it. Our onboarding process follows Firefox's best practices for collecting user data and was scrutinized by a Firefox add-on reviewer prior to the extension's publication.

Privacy Policy

Below, we provide the exact text of the extension's privacy policy.

Protecting the privacy of our users is of paramount importance both to us and our universities. The study has been approved by the internal review boards of the University of Chicago and Columbia University under numbers IRB22-0073 and AAAT9887.

We will collect page content displayed to you on three platforms: Twitter, YouTube, and Facebook, as well as the time and date of collection. This includes information such as what ads were displayed in the feed as well as before and within YouTube videos, the texts of posts, likes, retweets. We will also collect the time you spend (but not the content) on websites related to social media. These will be encrypted and securely stored in our database. The extension can also obtain authentication tokens to make requests to Twitter API to customize the content that you see, but we will not store such information. For the avoidance of doubt, we never collect, record, or handle any of your private messages, such as in Facebook Messenger.

Data are being collected exclusively for the purposes of this study. Data collected by the extension will be securely stored, and no identifiable information will be shared outside the research team. Furthermore, any such information will be deleted after the project concludes. If you would like us to delete your identifiable information at an earlier stage, please contact us and we will do so promptly.

Tracking Website Activity

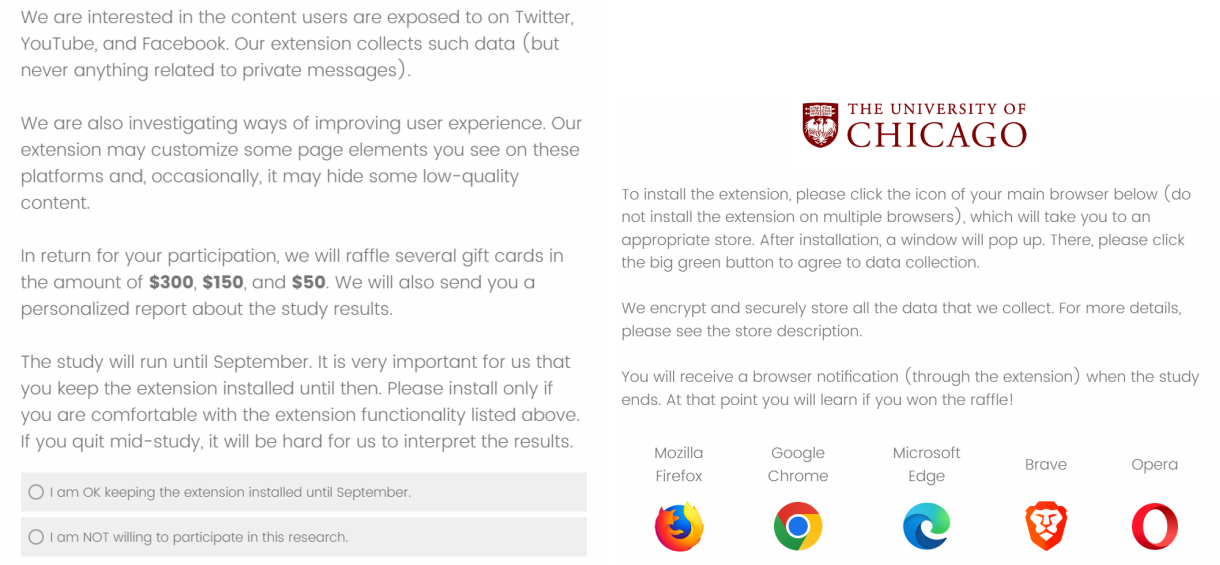
Below, we provide the list of 38 additional pre-registered platforms where the extension was tracking time spent by users. We rely on this data to understand spillover effects to related platforms where the hiding intervention did not take place.

- instagram.com,
- tiktok.com,
- wechat.com,
- whatsapp.com,
- mewe.com,
- tumblr.com,
- linkedin.com,
- snapchat.com,
- pinterest.com,
- telegram.com,
- meetup.com,
- medium.com,
- twitch.tv,
- discord.com,
- steemit.com,
- vk.com,
- quora.com,
- vimeo.com,
- zoom.us,
- reddit.com,
- houseparty.com,
- tapereal.com,
- qq.com,
- weibo.com,
- nextdoor.com,
- 4chan.org,
- blogger.com,
- livejournal.com,
- substack.com,
- zello.org,
- signal.org,
- messenger.com,
- spotify.com,
- clouthub.com,
- rumble.com,
- parler.com,
- gettr.com,
- gab.com.

A.5.2 Recruitment by the Mozilla Foundation

As indicated in the paper, the Mozilla Foundation promoted our study by retweeting a tailored recruitment post.

Participants recruited this way completed a simplified version of the intake survey in comparison to the standard one—taken by prospective participants who clicked a link in one of the ads posted by the research team. In particular, the simplified survey contained only two screens: a pre-screening task (Figure A13a) and an installation screen (Figure A13b). The former outlined the extension functionality and elicited people's willingness to keep the extension installed until the end of September 2022. The latter provided links to the appropriate extension store for various browsers. Individuals who took this version of the survey did not answer the intake survey questions and did not provide their Twitter handle. This method of enrollment was supplementary to our main recruitment efforts, and constituted a minor proportion of all extension installations.



(A) Pre-Screening (B) Installation Screen

Figure A13: Simplified Intake Survey

Note: Users who enrolled through the post retweeted by the Mozilla Foundation faced a simplified intake survey, composed of only two screens. The first one contained a pre-screening task with a short explanation of extension functionality and compensation. The second one featured icons with logos of various supported browsers, which served as links to the appropriate stores.

A.5.3 Browser Experiment Questions

INTAKE SURVEY

Social Media Usage

How often would you say you use social media from your desktop computer, as opposed to your mobile device?

For each platform (Twitter and Facebook) respondents could pick an integer from 0-100 using a slider. We used five labels: Only mobile (0), Mostly mobile (25), About equally (50), Mostly desktop (75), Only desktop (100). There was also an option "Don't use". If the participant chose it, they did not have to report the proportion using the slider.

Demographics

A. What is your year of birth?

Text entry question. Only integers between 1900 and 2020 were allowed.

B. What is your sex?

- Male
- Female

C. In which state do you currently reside?
Participants had to choose one value from a drop-down list. The options included: 50 US states, District of Columbia, Puerto Rico, and "I do not reside in the United States".

D. Generally speaking, do you usually think of yourself as a Republican, a Democrat, or an Independent?

- Democrat
- Republican
- Independent

If Independent is selected in question D.
E. As an Independent, do you think of yourself as closer to Republicans or Democrats??

- Republicans
- Democrats

F. Which of the following best describe your race or ethnicity? You can select more than one option.

- African American/Black
- Asian/Asian American
- Caucasian/White
- Native American, Inuit or Aleut
- Native Hawaiian/Pacific Islander
- Other *(text entry)*

G. Are you of Hispanic, Latino, or Spanish origin?

- Yes
- No
- Prefer not to answer

ENDLINE SURVEY

Willingness to Pay

We are interested in how valuable the extension is to you.

To establish your valuation, we will offer you a series of choices between keeping our extension installed for another month vs. receiving various gift card amounts.

One of your choices will be randomly selected as the "choice that counts". We will then randomly determine 10 participants for whom their "choice that counts" will be implemented.

We asked participants a series of questions involving two options, one of which involves keeping the browser extension installed for another month. Each participant had to make the maximum of four choices—we eliminated redundant questions by assuming monotonicity.

A. Which of the following would you prefer? This is a real question: there is a chance that it will actually be implemented, so please answer carefully.

- You keep our **browser extension** installed for one more month.
- You receive **$6**.

B. Which of the following would you prefer? This is a real question: there is a chance that it will actually be implemented, so please answer carefully.

- You keep our **browser extension** installed for one more month.
- You receive **$4**.

C. Which of the following would you prefer? This is a real question: there is a chance that it will actually be implemented, so please answer carefully.

- You keep our **browser extension** installed for one more month.
- You receive **$2**.

D. Which of the following would you prefer? This is a real question: there is a chance that it will actually be implemented, so please answer carefully.

- You keep our **browser extension** installed for one more month.
- You receive **$1.5**.

E. Which of the following would you prefer? This is a real question: there is a chance that it will actually be implemented, so please answer carefully.

- You keep our **browser extension** installed for one more month.
- You receive **$1**.

F. Which of the following would you prefer? This is a real question: there is a chance that it will actually be implemented, so please answer carefully.

- You keep our **browser extension** installed for one more month.
- You receive **$0.5**.

G. Which of the following would you prefer? This is a real question: there is a chance that it will actually be implemented, so please answer carefully.

- You keep our **browser extension** installed for one more month.
- You receive **$0**.

H. Which of the following would you prefer? This is a real question: there is a chance that it

will actually be implemented, so please answer carefully.

- You keep our **browser extension** installed for one more month AND receive **$0.5**.
- You receive **$0**.

I. Which of the following would you prefer? This is a real question: there is a chance that it will actually be implemented, so please answer carefully.

- You keep our **browser extension** installed for one more month AND receive **$1**.
- You receive **$0**.

J. Which of the following would you prefer? This is a real question: there is a chance that it will actually be implemented, so please answer carefully.

- You keep our **browser extension** installed for one more month AND receive **$1.5**.
- You receive **$0**.

K. Which of the following would you prefer? This is a real question: there is a chance that it will actually be implemented, so please answer carefully.

- You keep our **browser extension** installed for one more month AND receive **$2**.
- You receive **$0**.

L. Which of the following would you prefer? This is a real question: there is a chance that it will actually be implemented, so please answer carefully.

- You keep our **browser extension** installed for one more month AND receive **$4**.
- You receive **$0**.

M. Which of the following would you prefer? This is a real question: there is a chance that it will actually be implemented, so please answer carefully.

- You keep our **browser extension** installed for one more month AND receive **$6**.
- You receive **$0**.

Measures of Well-Being

In this part of the survey, we would like to ask you several questions about your well-being.

Click the point on the scale below that you feel is most appropriate in describing you.

A. Over the past 6 weeks, I thought I was:

- 1 (not a very happy person)
- 2
- 3
- 4
- 5

- 6
- 7 (a very happy person)

Indicate your agreement with the statement below.

B. During the past 6 weeks, I was satisfied with my life.

- Strongly disagree
- Disagree
- Slightly disagree
- Neither agree nor disagree
- Slightly agree
- Agree
- Strongly agree

C. Below are some ways you might have felt or behaved in the past 6 weeks. Please tell us how much of the time during the past 6 weeks:

	1. None or almost none of the time	2.	3.	4. All or almost all of the time
... you felt depressed. ... you felt anxious. ... you were absorbed in doing something worthwile. ... you felt bored.				

Normalization of Toxicity

Lastly, we would like to ask you to rate toxicity of seven real comments posted online. Before you provide the ratings, please read the definitions of toxicity below.

Very Toxic: a very hateful, aggressive, or disrespectful comment that is very likely to make you leave a discussion or give up on sharing your perspective.

Toxic: a rude, disrespectful, or unreasonable comment that is somewhat likely to make you leave a discussion or give up on sharing your perspective.

Please rate the toxicity of the following comments. We kept the original spelling, including any mistakes.

The following questions were shown in random order.

C1. Comment: *So tired of all these Portlanders moving to California and ruining the culture. When will it stop?!?*

- Very Toxic

- Toxic
- Not Toxic
- I'm not sure

C2. Comment: *The sky is falling! The sky is falling! This country has elected a black man twice! And now a woman! Please God! Make America Great Again! We're going down in flames!!!*

- Very Toxic
- Toxic
- Not Toxic
- I'm not sure

C3. Comment: *Only an individual living in the bubble of Ottawa and "working" as a professor can come up with this nonsense.....wow!*

- Very Toxic
- Toxic
- Not Toxic
- I'm not sure

C4. Comment: *can we just exterminate this scumbag already? every miserable breath he takes is a slap in the face to people who understand the privilege of parenthood.*

- Very Toxic
- Toxic
- Not Toxic
- I'm not sure

C5. Comment: *If you want more EmX transit route, more bike lanes, and less traffic lanes, vote for either Joshua Skov or Chris Wig. If not Goggle the others and make sure they are not part of the arrogant progressive movement, which is ruining Eugene.*

- Very Toxic
- Toxic
- Not Toxic
- I'm not sure

C6. Comment: *FYI: "LOL" does not turn a pack of malicious lies into amusingly subversive wisdom.*

- Very Toxic
- Toxic
- Not Toxic

- I'm not sure

C7. Comment: *Sadly, you are probably right. Things in Africa have gone from bad to worse for gay people, thanks mostly to American evangelical homophobes like Scott Lively exporting their hate to foreign shores.*

- Very Toxic
- Toxic
- Not Toxic
- I'm not sure

A.5.4 Survey Experiment Questions

Demographics

A. What is your age (years)?

- I am [text entry]
- I'd prefer not to answer

B. What is your gender?

- Female
- Male
- Non-binary
- Other (please describe if you wish) [text entry]
- I'd prefer not to answer

C. What is your sexual orientation?

- Bisexual
- Gay man
- Gay woman / lesbian
- Heterosexual
- Asexual
- Other (please describe if you wish) [text entry]
- I'd prefer not to answer

D. What is your ethnicity?

- White
- Mixed / multiple ethnicity
- Asian / Asian American
- Black / African American / Caribbean

- Other ethnic group (please describe if you wish) [text entry]
- I'd prefer not to answer

E. Do you identify as a person of Hispanic, Latino, or Spanish origin?

- Yes
- No
- I'd prefer not to answer

F. The next question is about your interest in sports. In reality, this is an attention check. If you are reading carefully, please select the third and the fifth choices together and none other. Based on the text above, how interested are you in sports?

- Extremely interested
- Very interested
- A little bit interested
- Almost not interested
- Not at all interested

G. What is you religion or belief?

- Buddhist
- Christian
- Hindu
- Jewish
- Muslim
- Sikh
- Other (please specify if you wish) [text entry]
- No religion
- I'd prefer not to answer

H. What is the highest level of education you have completed?

- Less than high school
- High school graduate
- Some college but no degree
- 2 year degree
- 4 year degree
- Master's degree
- Doctorate degree
- Professional degree (JD, MD, etc.)

I. What was your gross household income in 2023 in US dollars?

- Less than $10,000
- $10,000 - $19,999
- $20,000 - $29,999
- $30,000 - $39,999
- $40,000 - $49,999
- $50,000 - $59,999
- $60,000 - $69,999
- $70,000 - $79,999
- $80,000 - $89,999
- $90,000 - $99,999
- $100,000 - $149,999
- More than $150,000
- I'd prefer not to answer

Social Media Usage

In this section, we will ask you some questions about your social media use.

A. On how many days did you use social media last week?

- 0
- 1
- 2
- 3
- 4
- 5
- 6
- 7

If the answer to the previous question is not zero:

B. On an average day that you used social media, how much time did you spend on it?

- Less than 30 minutes
- 30 minutes to 1 hour
- 1 to 2 hours
- 2 to 3 hours
- 3 to 4 hours
- 4 to 6 hours
- 6 to 10 hours
- More than 10 hours

Transcription and WTA Instructions

We **plan** to recruit participants to **transcribe 100 social media posts**, i.e., **write down the text of the posts displayed in image form**. This is helpful for training an algorithm that aids people in using correct spelling and grammar when posting on social media.

In a few moments, we will ask you about the minimum compensation that you would be willing to accept for this extra task.

[page break]

We will invite one participant who requests the smallest compensation to complete the extra task. If multiple people request the smallest amount, we will randomly select one (they will be invited to a separate survey for this task).

The selected participant will be paid the second smallest compensation.

For example, imagine that Alice, Bob, and Chloe select the following compensations to help us transcribe 100 posts:

- Alice requests $10. Bob requests $15. Chloe requests $20.
- In this case, we will recruit Alice and pay her $15 for the task.

Please note that under the above rules, it is always **best for you** to tell us the minimum amount that you are willing to accept to participate.

A. Suppose that Participant A requests $14 to complete the task of transcribing 100 posts, Participant B asks for $12, and Participant C requests $13. Which of the following statements is true?

- Participant A will be invited to complete the additional task.
- Participant B will be invited to complete the additional task.
- Participant C will be invited to complete the additional task.

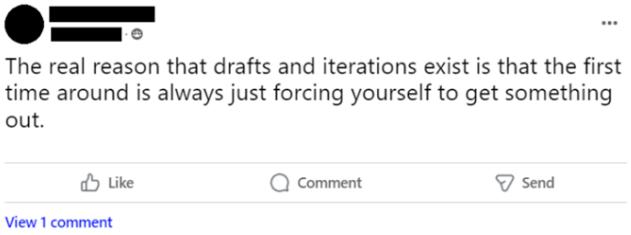
B. Based on the scenario outlined in the previous question, which of the following statements is true?

- The invited participant will be paid $10 for completing the extra task.
- The invited participant will be paid $12 for completing the extra task.
- The invited participant will be paid $13 for completing the extra task.
- The invited participant will be paid $14 for completing the extra task.
- The invited participant will be paid $15 for completing the extra task.

[page break]

To give you a sense of the type and complexity of posts in the extra task, we offer two **examples**. You will be required to transcribe them—write down the text of the post from an image of it.

Before we proceed, we will explain the transcription task. You will be shown an image like the one below with a social media post.



Your task is to write down the text of the post as displayed in the image in a text box. Note that the post starts with the words "The real reason" and ends with "something out." Below, we display an image of the correctly filled in text box.



When you are asked to transcribe posts, you will only be able to proceed after you enter the **minimum** number of characters, which will be close to the number of characters in the post. The minimum required will be specified and you will be able to see the character count.

Lastly, some posts in our pool have **comments** associated with them. You can view the comments by clicking the blue link in the post image. You will never be asked to transcribe the comments.

C. As mentioned above, links to comments are clickable. What is the text of the comment to the post shown above?

- Yes, 2nd pass is looking for whether it is actually worth something.
- I wish I never had to write any drafts.
- I am not so sure. My first drafts are always the best. The rest is overthinking.

D. Which of the following statement is true regarding the transcription task?

- I can proceed as soon as I enter 20 characters.
- I can view comments but I am not required to transcribe them.

- If I do not like a particular post, I can click a swap button and the post will be replaced.

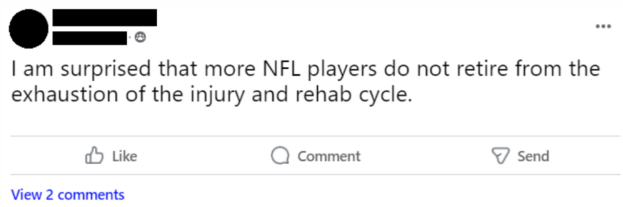
Please proceed to the next page to see the first example that you will be asked to transcribe.

Transcription Example 1

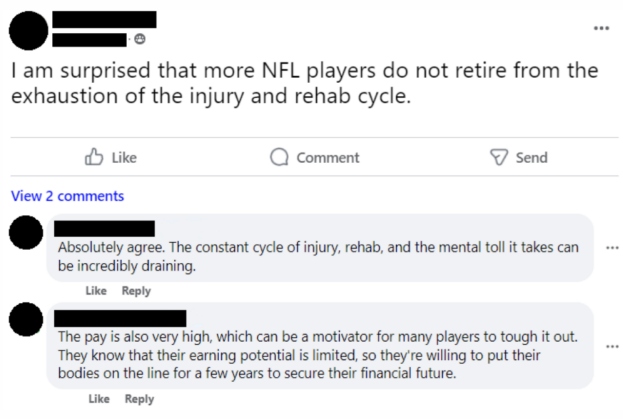
Please transcribe (write down) the text of the post presented in the image below.

If your transcription is correct, you will earn an extra bonus payment of £0.10 (approx. $0.13). You need to enter 95-105 characters (with spaces) to proceed.

[text entry]



On click of the blue link, the comments appear:



Transcription Example 2

Please transcribe (write down) the text of the post presented in the image below.

If your transcription is correct, you will earn an extra bonus payment of £0.10 (approx. $0.13). You need to enter [character range] characters (with spaces) to proceed.

[text entry]

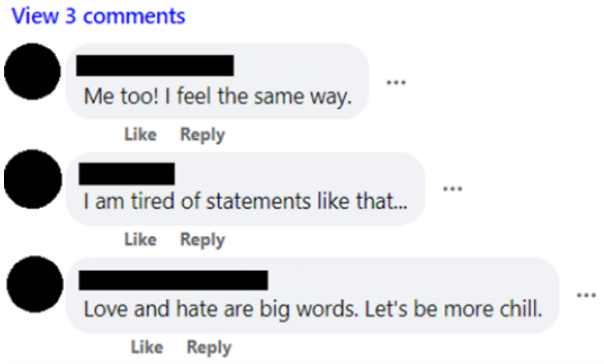
[image of the post, formatted in the same way as for Transcription Example 1]

The text of the post depends on the experimental condition. Note that any profane words were not censored when displayed to participants.

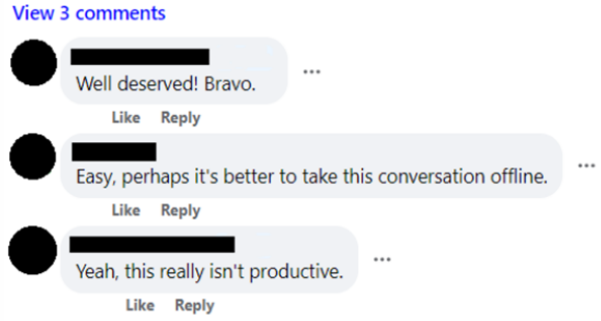
- *Hate speech treatment:* What is so difficult to understand? I hate Blacks.
- *Hate speech control:* What is so difficult to understand? I hate Mondays.
- *Profanity treatment:* You are a f***ing moron.
- *Profanity control:* You are very misguided.

On click of the blue link, the comments appear:

Comments for both hate speech conditions:



Comments for both profanity conditions:



WTA Elicitation

What is the minimum compensation that you would be willing to accept to help us transcribe 100 social media posts?

We will select one participant that requests the smallest compensation, and we will pay them the second smallest compensation for completing the task.

As a reminder, under this procedure **it is always best for you to tell us the minimum amount that you are willing to accept to participate—not more and not less**.

[A slider from 2 to 30 titled "Minimum compensation in US dollars ($)"]

Recall of Hate Speech

Did you consider any of the posts that you were asked to transcribe on the previous pages (Examples 1-2) to be hate speech?

Hate speech is a direct attack against people on the basis of protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease.

- Yes, there was at least one post containing hate speech on the previous pages.
- No, there were no posts containing hate speech on the previous pages.

Recall of Toxicity

Did you consider any of the posts that you were asked to transcribe on the previous pages (Examples 1-2) to be toxic?

A statement is toxic if it is a rude, disrespectful, or unreasonable comment that is somewhat likely to make you leave a discussion or give up on sharing your perspective.

- Yes, there was at least one toxic post on the previous pages.
- No, there were no toxic posts on the previous pages.

Recall of Treatment

In both hate speech conditions:
Which of the following posts were you asked to transcribe in one of the previous screens?

- What is so difficult to understand? I hate Mondays.
- What did you just say? You hate Mondays?
- What is so difficult to understand? I hate Blacks.
- What did you just say? You hate Blacks?

In both profanity conditions (profane words were not censored):
Which of the following posts were you asked to transcribe in one of the previous screens?

- You are very misguided.
- You are really misinformed.
- You are a f***ing moron.
- You are a f***ing idiot.

Entertainment Rating

On a scale from 0 to 100, how engaging were the posts on the previous pages (Examples 1-2)?

[A slider from 0 to 100 with five labels “Not entertaining at all”, “Little entertaining”, “Somewhat entertaining”, “Entertaining”, “Very entertaining”]