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

George Beknazar-Yuzbashev[†] **Rafael Jiménez-Durán**[‡] **Jesse McCrosky**[§]
Mateusz Stalinski (Job Market Paper)[¶]

January 21, 2023, Click for the latest version

Abstract

As much as forty percent of social media users have been harassed online, but there is scarce causal evidence of how toxic content impacts user engagement and whether it is contagious. In a pre-registered field experiment, we recruited participants to install a browser extension, and randomly assigned them to either a treatment group where the extension automatically hides toxic text content on Facebook, Twitter, and YouTube, or to a control group without hiding. As the first stage, 6.6% of the content displayed to users was classified as toxic by the extension relying on state-of-the-art toxicity detection tools, and duly hidden in the treatment group during a six-week long period. Lowering exposure to toxicity reduced content consumption on Facebook by 23% relative to the mean – beyond the mechanical effect of our intervention. We also report a 9.2% drop in ad consumption on Twitter (relative to the mean), where this metric is available. Additionally, the intervention reduced the average toxicity of content posted by users on Facebook and Twitter, evidence of toxicity being contagious. Taken together, our results suggest a trade-off faced by platforms: they can curb users' toxicity at the expense of their content consumption.

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

Keywords: toxic content, hate speech, moderation, social media, user engagement, field experiment

[*] This research is funded by the Becker Friedman Institute at the University of Chicago. Mateusz Stalinski would like to thank his doctoral thesis advisors, Leonardo Bursztyn, Alex Imas, and John List for invaluable support an guidance. We are also grateful to Joshua Dean, Sota Ichiba, Andrea Prat, Andrey Simonov, Avner Strulov-Shlain, Konstantin Sonin, Haruka Uchida, and numerous seminar participants for helpful comments and suggestions. Furthermore, we are indebted to the Mozilla Foundation for their help in promoting the study. We would also like to thank Onar Alili, an anti-abuse engineer, for his tremendous help in developing the browser extensions for this project. We are grateful to Andreas Wagner who provided feedback on our Firefox add-on and the onboarding process prior to the store submission. The study 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 for randomized control trials under trial numbers AEARCTR-0009628 (main study) and AEARCTR-0010138 (endline survey).

[†] Columbia University gb2683@columbia.edu.

[‡] Social Science Research Council and Chicago Booth Stigler Center rjimenez@ssrc.org.

[§] Mozilla Foundation jessed@mozillafoundation.org.

[¶] University of Chicago mstalinski@uchicago.edu.

> [As] long as your goal is creating more engagement, optimizing for likes, reshares and comments, you're going to continue prioritizing polarizing, hateful content.
>
> Frances Haugen, WSJ

1 Introduction

More than seven in ten Americans are active on social media, with as many as forty percent of users experiencing some form of harassment online.[1] Due to the links between inflammatory content and violence (Müller and Schwarz, 2020a,b; Bursztyn et al., 2019), and the effect of social media on mental health (Allcott et al., 2020; Mosquera et al., 2020; Allcott et al., 2021; Braghieri et al., 2021), platforms' incentives to curb toxic content have been under public scrutiny.

Yet, the multifaceted debate lacks evidence about the link between toxicity and user engagement – an important performance metric,[2] which is key to understand social media's incentives to self-regulate. From the point of view of profit-maximizing sites, two competing forces affect their optimal level of toxicity. On the one hand, for some users, a high prevalence of toxicity could increase their cost of participating in conversations, a worry exacerbated by research documenting the negative psychological effects of exposure to offensive materials (Schmitt et al., 2014).[3] On the other, toxic content is something that we react to, protest, or argue against; thus its presence might spike up engagement (Crockett, 2017; Kosmidis and Theocharis, 2020).[4] The picture gets even more complex with the possibility that toxicity is contagious (Rydgren, 2005; Mathew et al., 2019; Ziems et al., 2020), catalyzing the process. The relative magnitude of these channels is an open question for stakeholders, spanning from regulators scrutinizing tech giants to platform managers.

We attempt to fill these gaps by conducting a field experiment targeting three leading sites: Facebook, Twitter, and YouTube, and prepared in cooperation with the Mozilla Foundation.[5] We study the impact of toxic social media materials on user engagement. Specifically, we ask whether a lower

[1] See https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ and https://www.adl.org/resources/report/online-hate-and-harassment-american-experience-2022, accessed: 2022-10-20.

[2] For example, Meta's Q2 2022 report opens with "positive trajectory on our engagement trends". Furthermore, ad impressions, an engagement-dependent outcome, are listed among financial highlights. https://s21.q4cdn.com/399680738/files/doc_financials/2022/q2/Meta-06.30.2022-Exhibit-99.1-Final.pdf, accessed 2022-10-19.

[3] Elon Musk referred to this issue in his recent address to the employees of Twitter, underscoring that "people need to 'like' being on the platform, and if they feel 'harassed or uncomfortable', they would leave", https://www.washingtonpost.com/technology/2022/06/16/elon-musk-twitter-employee-meeting/, accessed: 2022-10-02.

[4] This notion is well-illustrated by Frances Haugen's testimony (see the epigraph), https://www.wsj.com/articles/facebook-whistleblower-frances-haugen-says-she-wants-to-fix-the-company-not-harm-it-11633304122, accessed: 2022-10-01.

[5] The Mozilla Foundation kindly aided us in promoting our study during the recruitment stage by retweeting a tailored recruitment post. Furthermore, Jesse McCrosky joined our team, advising on user experience and facilitating the choice of toxicity detection tools for the project. Lastly, we received high-quality feedback on our browser extension and user onboarding, which was critical in ensuring that users have a safe and pleasant experience.

exposure to toxicity affects users' content consumption, content production, and the time they spend on the platforms. Additionally, we investigate whether toxicity is contagious, that is, whether lowering individuals' exposure to it decreases their propensity to spread further toxicity. Lastly, we explore a potential mechanism for why toxic content might be contagious – it could change users' evaluations of what is toxic – and we analyze toxicity's effect on measures of users' well-being.

To address the research questions, we conducted a framed field experiment with 836 individuals recruited on social media, collecting over 15 million posts and comments displayed to users based on more than 20,000 hours of social media activity throughout the study. We asked all participants to install a custom-built browser extension that hides toxic content. To divert attention from the true purpose of the study, we advertised the extension more generally – as potentially improving user experience on social media. One of the limitations of introducing experimental interventions through browser extensions is that they operate on desktop devices, potentially weakening the induced variation as a share of consumed content (due to high mobile device usage). We anticipated this issue by targeting users on desktop devices with recruitment ads. As a result, our sample reported a mean desktop share of social media usage to be 57% on Facebook and 62% on Twitter.

The extension relied on state-of-the-art machine learning toxicity detection algorithms, trained on a large dataset of online comments categorized by human coders, to assign toxicity scores to all posts, comments, and replies displayed to the user on Facebook, Twitter, and YouTube, with 17 languages supported.[6] We exogenously varied users' exposure to toxicity by randomly hiding toxic text content on the three social media platforms for the duration of six weeks. We randomized the participants into two conditions: treatment – in which we hid all content with a toxicity score exceeding a common threshold of 0.3, and control – where no hiding occurred.[7] Our hiding intervention directly informs policy, as it is similar to current platform solutions that deprioritize some forms of content (Le Merrer et al., 2021) and use toxicity thresholds to trigger content moderation actions (Katsaros et al., 2022; Ribeiro et al., 2022). Moreover, we designed the extension to make hiding as seamless as possible; indeed, the experimental variation did not give rise to any data patterns indicative of differential attrition in our sample.

[6] Algorithms by various providers, trained on the dataset of online comments from Wikipedia and Civil, competed in machine learning challenges to most accurately predict toxicity scores of additional statements rated by human coders. We selected one of the high achievers, Unitary's Detoxify library, as our main toxicity detection tool.

[7] The score of 0.3 means that 3 out of 10 human raters labeled the statement "toxic" or "very toxic". The coders were provided with industry-standard definitions of toxicity. A "toxic" statement was described as "a rude, disrespectful, or unreasonable comment that is somewhat likely to make you leave a discussion or give up on sharing your perspective." Moreover, a "very toxic" statement was defined as "a very hateful, aggressive, or disrespectful comment that is very likely to make you leave a discussion or give up on sharing your perspective." It is important to note that the presence of elements such as "leaving a discussion" in the definitions of toxicity could make it more likely that lower exposure to toxicity would increase user engagement. As we argue below, this means that our results are actually conservative.

The six-week intervention was preceded by a two-week baseline period, where we collected data on users' social media activity. The baseline period enabled us to employ a difference-in-differences identification approach. We pre-registered this empirical strategy anticipating that it is essential to ensure sufficient statistical power. The extension collected multiple outcomes measuring various forms of user engagement: time spent on the platform, content consumption, ad consumption, as well as users' own posts, replies, and reactions (such as likes). We additionally collected toxicity scores of text content created by users to illuminate the contagion hypothesis. Lastly, we elicited further outcomes, such as measures of well-being and toxicity ratings of online comments, in an endline survey.

During the intervention period, the extension recorded that the average toxicity score of text content displayed to users in the treatment group was 73.2% lower than in the control.[8] This stark drop in exposure to toxic content was enabled by the efficiency of the extension's hiding functionality, with our logs indicating that we successfully hid 97.7% of text elements with toxicity scores exceeding 0.3. Overall, the intervention resulted in the hiding of 6.6% of posts, comments, and replies displayed in the browser across the three platforms for users in the treatment arm. Specifically, we hid about 5% of content on Facebook and 7% on Twitter. Given the heavy browser usage of our sample, we conclude that our intervention led to a substantial reduction in exposure to toxicity on social media, even taking into account the mobile app usage.

We now proceed to report our main findings.[9] The first group of results concerns content and ad consumption on social media. The hiding intervention, which lowered exposure to toxicity, significantly reduced content consumption on Facebook. The conclusion was reached from the treatment effect on the quantity of content that the platform intended to display, i.e., including the hidden elements. The effect on this conservative measure of consumption, which we refer to as content offered, cannot be explained by the mechanical effect of hiding, and thus indicates a genuine reduction in this form of user engagement.[10] Specifically, the intervention reduced content consumption by at least 17.9 elements a day, a 23% change in comparison to the mean quantity of content throughout the study. Social media strive to encourage people to engage by viewing many posts in a short time, as it enables the

[8] Importantly, the average toxicity scores of content that the platforms intended to display (before the hiding applied) were almost identical by group: 0.064 in treatment and 0.063 in control.

[9] In the main text of the paper, we focus on reporting the results for Facebook and Twitter. YouTube significantly differs from Facebook and Twitter in the context of our intervention, as text exchanges are only a secondary reason (behind watching videos) for visiting the platform. This was crystallized by the data on the average quantity of content per minute spent on each of the platforms: 6.73 for Twitter, 3.27 for Facebook, and only 1.52 for YouTube (22.6% of the value for Twitter and 46.5% of the value for Facebook). We report YouTube results in Appendix H.

[10] An alternative measure is the quantity of content actually displayed to users. While we use the content offered as our main outcome, in order to dispel the criticism that the result may have been a trivial consequence of the intervention, it is important to note that the effect on the content displayed is not necessarily mechanical. In the feed and in long comment sections, hidden elements are instantly replaced by the content below – they are pulled up. We provide more discussion in Section 4.

platforms to input more ads in between posts. In this context, it is notable that the intervention led to a significant reduction both in the consumption of posts (user feed) and comments on Facebook, the former of which is critical for ad impressions. We conclude that while we cannot directly assess ad impressions on Facebook, our results are consistent with fewer impressions. On Twitter, on the other hand, it is possible to identify ads. We find that the hiding intervention significantly reduced ad consumption on the platform, even though we do not detect an effect on the conservative measure of content consumption. Taken together, the evidence indicates a likely reduction in profitability, at least given the ad policy focusing on impressions.

As our second main outcome, we report evidence in favor of the contagion hypothesis. The treatment significantly reduced the average toxicity of posts and comments published by users both on Facebook and Twitter – respectively by 35% and 20% relative to the mean. Additionally, exploring a pre-registered angle of heterogeneity, we report an effect in the same direction, though with a slightly higher magnitude, for the subsample of users with above-median baseline exposure to toxicity.[11] Taken together, we provide broad evidence that lower exposure to toxic content reduces the toxicity of own posts and comments. Despite the overall strong effect, we find no evidence of normalization of toxicity – the ratings of seven toxic statements evaluated by participants in the endline survey did not vary by treatment group.

Finally, we summarize the results on the remaining outcomes. On both Facebook and Twitter, the intervention did not alter the time spent on the platforms, which is notable given the previously discussed evidence indicating a negative impact on content consumption. Furthermore, reducing exposure to toxicity on the treated platforms led to positive spillover effects on the combined total time spent on 38 related websites where the intervention did not take place. In addition, the hiding intervention reduced content production on Facebook, which is in line with lower content consumption on the platform. We do not find a similar effect for Twitter. Lastly, we report no significant effects of the intervention on the remaining survey outcomes, including the index of well-being.

Our results are robust to considering difference-in-differences specifications other than the two-way fixed effects regression, which we use as our main specification. In particular, we account for the fact that the participants experienced the intervention period during different calendar dates, even though the recruitment period spanned only three weeks. To that end, we verified that our results are robust to using the stacked regression specification with start date × individual and start date × period fixed effects. Furthermore, we provide evidence that our findings are unlikely to be driven by differential attrition. According to our most conservative measure, the last day seen in the study, 85.2% of users

[11] It is important to note that the baseline level of toxicity can serve as a proxy for preference for toxic content if we are willing to assume that the platforms optimize the exposure to toxicity to match user's tastes at least to some extent.

active during the intervention survived until its conclusion (i.e., they were seen on day 56 or later), with the difference in the proportion of survivors not significantly different by the treatment group.[12] Additionally, we investigate the dropout dynamics throughout the study, and find no evidence that it varied by group. Moreover, we consider two likely channels that could have led to differential attrition regarding types of individuals leaving. Given the character of our hiding intervention, we worried that people with a preference for toxic content or those with high levels of social media activity were more likely to drop out of the treatment group. We refute these conjectures by verifying that the baseline level of activity and the baseline average toxicity of consumed content are not significant predictors of attrition by group.

The paper contributes to the debate on the consequences of online toxicity in three different ways. First, we provide novel evidence on the impact of exposure to text-based toxicity on social media, focusing in particular on content consumption and the contagious character of toxicity. Second, we offer a method of studying the effects of exposure to toxic content in isolation. This provides an additional level of granularity to the analysis, with the usual focus being the effects of social media censorship, a category that is broader and harder to interpret. Third, we hope to inform policy. Both platforms and regulators are likely to consider moderation tools that are less severe than outright removal of content. One option is reducing prominence of toxic content on the platform (reducing its visibility), which is akin to our hiding intervention. Moreover, responding to the interest resulting from the rising volume of online content that requires scrutiny, we offer evidence on the effects of employing automated toxicity detection tools, a class of algorithms that will shape the future of content moderation.

This paper is related to three main strands of the literature. First, there is a growing body of work in economics that studies the effects of social media penetration, usage, and advertising on a variety of outcomes, including political participation and persuasion (Fergusson and Molina, 2019; Enikolopov et al., 2020; Fujiwara et al., 2021; Zhuravskaya et al., 2020; Coppock et al., 2022; Beknazar-Yuzbashev and Stalinski, 2022), polarization (Sunstein, 2017; Allcott and Gentzkow, 2017; Boxell et al., 2019; Levy, 2021; Melnikov, 2021), hate crimes (Müller and Schwarz, 2020a,b; Bursztyn et al., 2019; Jiménez-Durán et al., 2022), and mental health (Allcott et al., 2020; Mosquera et al., 2020; Allcott et al., 2021; Braghieri et al., 2021). 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.

This paper is also part of a rapidly-growing literature that studies online engagement, its deter-

[12] We rely on the last day seen in the study because information on whether a user uninstalled the browser extension is not available to developers, nor can it be obtained on the extension level (i.e., uninstallation event).

minants, and its connection to platforms' decisions and content moderation policies. Theoretically, Acemoglu et al. (2021) argue that, in homophilic networks, polarizing and divisive content is more likely to spread virally. Liu et al. (2021); Jiménez-Durán (2021), and Madio and Quinn (2021) model platforms' optimal content moderation decisions. Our paper contributes to this strand of literature by providing field evidence illuminating a key parameter of platforms' decisions – the responsiveness of user engagement to exposure to toxic content. Empirically, Jiménez-Durán (2021); Ribeiro et al. (2022); Katsaros et al. (2022) provide experimental evidence of the impact of content moderation on user behavior. Our paper contributes to this work by directly manipulating the toxicity in user feeds and comment sections.

This paper is also related to the literature that studies how media can impact the diffusion of hateful attitudes. Yanagizawa-Drott (2014); DellaVigna et al. (2014); Adena et al. (2015); Wang (2021) find that propaganda in traditional media can help spread violence and extremism, and Blouin and Mukand (2019) report that government propaganda can manipulate the salience of ethnic identity. Several studies document the spread of toxic attitudes online (Rydgren, 2005; Mathew et al., 2019; Ziems et al., 2020; Velasquez et al., 2021) and hypothesize that toxicity is contagious. Yet, survey-based evaluation of the contagion hypothesis has yielded mixed evidence (Kim et al., 2021). We contribute to this line of work by providing field evidence of toxicity's contagiousness.

The paper is organized as follows. Section 2 provides background information, including a description of toxicity detection algorithms. Section 3 outlines the experiment design and characterizes the intervention. Section 4 provides a discussion of the results and addresses potential concerns. Section 5 concludes.

2 Background

2.1 Supported Platforms

Our hiding intervention encompasses three leading social media platforms: Facebook, YouTube and Twitter. As of January 2022, the former two can boast of the top highest global number of users – 2.9 billion (rank 1) and 2.6 billion (rank 2) respectively, with the latter's user base being 436 million.[13] The platforms are equally impactful in the United States as they are worldwide. According to Pew Research, in 2021, 69% of US adults reported using Facebook, with 48.4% of Americans doing it daily. The proportion was equal to 81% for YouTube (43.7% accessing daily) and 23% for Twitter (10.6% accessing daily).[14] With Facebook and YouTube selected for their sheer size and overall influence, we

[13] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/, accessed: 2022-08-28.

[14] https://www.pewresearch.org/internet/fact-sheet/social-media/, accessed: 2022-08-28.

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 the treatment did not take place.[15]

2.2 Browser Extension

All participants installed our dedicated browser extension called *Social Media Research*, which enables the hiding intervention and records the 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, the supported browsers account for 93% of the global market share for desktop browsers.[16] Extensions constitute a well-established element of browsing in 2022, with more than 65 million users of the iconic AdBlock.[17] Therefore, we expect that many prospective participants were familiar with the environment in which the study took place. Just as extensions became prominent tools enhancing browsing experience for web users, they also gained popularity among researchers, who can use them to collect data on user activity and displayed content (e.g. Levy, 2021; Beknazar-Yuzbashev and Stalinski, 2022; Aridor, 2022). In addition to tracking users' online behavior, we extended this methodology by relying on the add-on to introduce exogenous variation in exposure to toxic content.

A major advantage of toxicity hiding implemented through a browser 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 feature minimizes the risk that any algorithm-induced adjustment in the content presented to the user could have occurred as a reaction to the intervention in the time span of six weeks. Our data corroborates this point.[18]

[15] Our platform choices warrant a question of why to stop at three. One reason is that the hiding intervention requires that the extension code is 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.

[16] https://kinsta.com/browser-market-share/, accessed: 2022-08-28.

[17] https://getadblock.com/en/, accessed: 2022-08-28.

[18] For example, the average toxicity scores of content that the platforms intended to display to users (before any hiding

Lastly, we have been aware that 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 to be of paramount importance, we ensured that the extension onboarding followed Firefox's best practices[19] and was vetted by their add-on reviewer. Moreover, all data was encrypted when stored in our database, with the decryption key only known to the researchers. Details on the installation, onboarding, and privacy policy are provided in Appendix C. 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 considerations placing constraints on human moderation, the algorithms must play a central role in toxicity detection efforts. With that in mind, we evaluate the impact of hiding toxic content on social media as detected by state-of-the-art tools available.

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 social media such as Taringa!, a large platform in South America, and major publishers like *Le Monde* or *The Financial Times*.[20] The need for constant improvement of the algorithms' precision led to the creation of Jigsaw challenges, hosted by Kaggle, a machine learning company. These were toxicity detection competitions for machine learning solutions supplied by independent developers and companies. 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.[21] 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 98.64 score (the top score was 98.86). In addition, Unitary supplied a successful "multilingual" model.[22] Owing to the high quality performance

applied) did not differ by treatment and over time, indicating that the platforms' learning about users' preferences for toxicity was limited.

[19] https://extensionworkshop.com/documentation/develop/best-practices-for-collecting-user-data-consents/, accessed: 2022-08-30.

[20] https://perspectiveapi.com/case-studies/, accessed 2022-09-07.

[21] https://www.scientificamerican.com/article/can-ai-identify-toxic-online-content/, accessed 2022-09-03.

[22] https://github.com/unitaryai/detoxify (section Description), accessed 2022-09-07.

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 render toxicity scores corresponding to the probability that a text is considered toxic. This way, we could think of scores exceeding 0.7 as cases where the algorithm is quite confident that a statement is toxic, whereas values ranging from 0.3 to 0.7 would represent "suspect" cases, where the algorithm is uncertain.[23] 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, Kaggle describes the target levels of toxicity in the training and evaluation samples 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 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".[24] 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 4) are conservative.

2.3.3 Limitations

While the tools enabling our intervention are a sign of a 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

[23] https://developers.perspectiveapi.com/s/about-the-api-score, accessed 2022-09-03.

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

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.[25]

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.

3 Experiment Design

3.1 Experiment 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 for an endline survey, where we collected additional outcomes.

3.2 Sample

3.2.1 Recruitment

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.[26] To attract a broad subject pool, we relied on a variety of ad designs, including video ads with social media themed animations (Figure 3a in the appendix), static ads drawing attention to our gift card raffle (Figure 3b in the appendix), and ads offering a report on user’s social media

[25] We verified that our extension would assign this sentence a toxicity score of 0.63, which would lead to its hiding in the treatment group. See the following article for a further discussion of the limitations: https://medium.com/unitary/how-well-can-we-detoxify-comments-online-bfffe5f716d7, accessed: 2022-09-07.

[26] 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 using the other two social media sites.

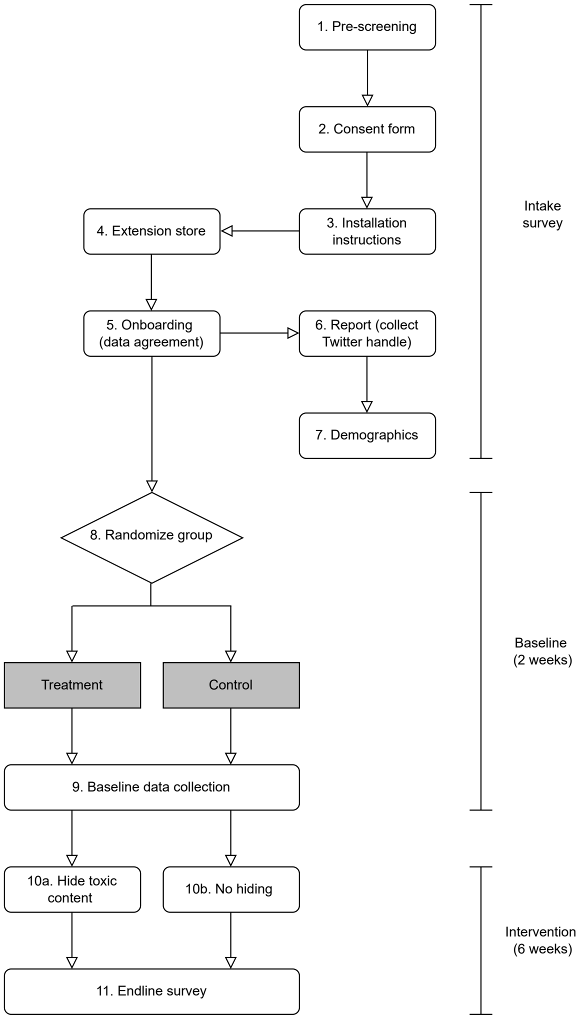


Figure 1: The Study Flow

stats (Figure 3c in the appendix). Individuals who clicked on the link in the ads were directed to a Qualtrics environment for the intake survey. 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 4 in the appendix) tailored to their followers (278.3 thousand as of August 2022).[27] 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.[28]

A consequential choice that we made when planning the ad campaigns was targeting 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

[27] The account's profile can be accessed by clicking https://twitter.com/mozilla, accessed: 2022-08-28.

[28] Figure 1 outlines the components of the intake survey as experienced by those who were recruited through ads posted by the researches – an overwhelming majority of participants. Individuals who enrolled through the post retweeted by the Mozilla Foundation faced a simplified intake survey, composed of only two screens. See Appendix E for details.

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). 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 (see Section 4.1.1).

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 5 in the appendix). As a part of the procedure, the participants could explore the extension store listing, followed by onboarding. The 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 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, though we acknowledge that some users might have guessed our interests.[29]

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 the browser notification (a new tab opened) sent to all users through the extension. We supplemented this process by sending emails to the 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. We instructed the participants on how to check whether they won a prize in the endline survey. Additionally, everyone who was eligible for a raffle entry was also entitled to a report on their social media activity.

3.2.2 Treatment Assignment

Individuals who installed the browser extension and agreed to data collection were randomly assigned either a treatment or a control condition, as well as attached to 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 the id and treatment assignment were performed at the browser level, in the

[29] The wording of the store listing and user onboarding is provided in Appendix C. We used a similar strategy in our pre-screening tasks (intake survey), where the prospective participants learned about the extension functionality and were asked to indicate if they are willing to keep it installed until the end of September. See Figure 19a in the appendix for an example.

intake survey we instructed participants to only install the extension for 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 the user ids provide a reliable system of identifying participants. After a minimal cleaning procedure,[30] we detected 836 extension users, 439 in the treatment and 397 in the control.[31]

3.2.3 Main Sample

Out of 836 initial users, 775 (92.7%) individuals were still using the extension during the intervention period – after the baseline period concluded (14 days). From now on, we refer to these participants as the main sample. Note that there were no differences in user experience between the two groups during baseline. As expected, attrition at this stage did not vary by group, with 410 individuals (93.4%) remaining in the treatment, and 365 (91.9%) in the control.

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.[32] The extension retrieved at least one Twitter handle for 89.9% of users, based on whether the handle was available on the page while the participant was browsing. We expect the 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 the 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 563 individuals (72.6%) to their responses. Our ability to match the records depends on whether users correctly and truthfully reported their handle in the intake survey. We collected Twitter handles before treatment assignment, therefore, the matched individuals should constitute an as-if random subset of the main sample.[33]

[30] We discarded ids in a handful of 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.

[31] We performed randomization using JavaScript's Math.random() function, with both groups being equally likely.

[32] To find more information about the Twitter API, visit https://developer.twitter.com/en/docs/twitter-api, accessed 2022-09-07.

[33] This conclusion should not be undermined by the fact that we recruited some users through promotion by the Mozilla Foundation, where we do not collect Twitter handles in the intake survey. The reason is that this way of recruitment constituted a very minor proportion of all installations – we only recorded 36 responses in which the user declared their willingness to participate (and of those not everyone necessarily installed the extension).

Sample Balance Data from the Twitter API and the intake survey allowed us to create a rich balance table, depicted in Table 7 in the appendix. None of the sixteen covariates indicates significant differences by treatment assignment at the 5% level. Furthermore, the geographical distributions of users (categorized into five regions: Midwest, Northeast, South, West, and Other) in both groups mirror each other. We conclude that the main sample is well-balanced.

3.2.4 Survey Sample

Based on matching the endline survey responses to the extension data by Twitter handles, we identified 364 participants – 43.5% of the initial users (assigned treatment) – who completed the endline survey. This includes 189 individuals in the treatment group (43.1%) and 175 in the control group (44.1%). Table 8 in the appendix provides the survey sample balance. Only one out of sixteen listed covariates – the number of Twitter followers – is significantly different by treatment group at the 5% level.

3.3 Treatment

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 before it was displayed to the user on the three sites. Based on each element's text, a probabilistic toxicity score between 0 and 1 was assigned. The extension hid all content with the score exceeding a fixed threshold – the same for all participants in the treatment group.

Analyzing Text Content The extension sent text content to be evaluated to our server.[34] 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 for details). Otherwise, we applied one of the multilingual models, which together support 16 additional languages.[35] 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 and linguistic backgrounds. Once the toxicity analysis on the server was concluded, the toxicity score was sent back to the extension, where a decision was reached if an element should be hidden or not.

[34] Our main toxicity detection tool is a Python library. For this reason, we created a Flask app, which allows running Python on a web server. The app was stored on Digital Ocean for the duration of the study.

[35] 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.

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. The raters worked with the industry standard definition of toxicity, understood as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion." If 3 out of 10 people would agree that a statement can be characterized by this strongly worded description of toxicity, we considered it 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.[36] 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.

Speed of Hiding Immediate hiding of toxic content was of paramount importance to the project. To ensure it, the app on our server – processing all statements and assigning them toxicity scores – was deployed on Digital Ocean and 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 time at which a toxic element was hidden and the time when it appeared on the page. The median hiding speed was equal to 407 milliseconds. The histogram, presented in Figure 6 in the appendix, and the ECDF, depicted in Figure 7 in the appendix, confirm 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 loaded in batches ahead of where the user is on the page (e.g., if a user scrolls down to see posts 4-6, posts 7-9 are already being loaded), so the extension hides toxic content before users could see it.

Style of Hiding In addition, we minimized traces left on the page by our hiding intervention. Figure 8 in the appendix demonstrates user experience in the Facebook feed (and on group pages) – with hidden posts seamlessly replaced by the content below. Furthermore, Figure 9 in the appendix offers an example of a comment section under a post in the original state and with the intervention

[36] Our data suggests that our choice of a low threshold level was less consequential than anticipated. In particular, Figure 12 in the appendix demonstrates that the distribution of toxicity of above-the-threshold content is skewed to the right.

provided by our extension. In general, the hiding of posts in the feed and comments under posts across the three platforms should not have been easily noticeable to a casual user.[37] 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. On YouTube, we were hiding toxic posts together with "Show replies" button (if unwrapped) and nested replies (if visible). Figure 10 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.[38] The functionality was enabled both in the treatment and the control group 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.

3.4 Outcomes

3.4.1 Extension Outcomes

The main questions addressed by the paper concern the impact of exposure to toxic content on various form of user engagement – spanning both what the user views and what they post – and the exposure's impact on the propensity to spread further toxicity. In this context, it is natural to divide our outcomes into two categories: content consumption and content production. Within each category, we separately focus on toxicity and quantity. Additionally, we describe time that users spent on social media.

Content Consumption (Toxicity) 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. 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 the platforms (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.

[37] Despite our best efforts, there were some minor exceptions, e.g., in the case of toxic replies on Twitter when they were marked with a vertical line connecting elements of the thread – we could not entirely remove the line from an element preceding the hidden one. Importantly, our data suggests that the hiding intervention did not have a negative effect on user experience. For example, 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.

[38] Typically, to see this content, the user needs to click a button – Figure 11 in the appendix shows the difference in user experience.

Content Consumption (Quantity) As a basic measure, we record the quantity of content displayed to users on each platform as a proxy for content consumption. However, when evaluating the effect of exposure to toxicity on content consumption, we rely on the quantity of content offered by the platform – inclusive of the hidden elements, to ensure that any negative treatment effect is not driven by the mechanical consequences of our intervention. To illuminate the impact of the intervention on ad impressions, we distinguish between posts and comments/replies. 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. Moreover, we directly report on the number of ads displayed to the participants – we can do so on Twitter, where we are able to credibly identify ads appearing in the feed.

Content Production First, in order to capture more intensive forms of user engagement, i.e., ones requiring a visible action, we compute the total number of posts and comments published by users on the platforms. Additionally, we report the number of reactions, such as likes. Second, to investigate the contagiousness of toxicity, we provide the average toxicity scores of posts, comments, and replies published by each participant.[39]

Time on Social Media We record the total time that participants spend on Twitter, Facebook, and YouTube, with one minute precision. This outcome encapsulates both the time spent on active consumption and production of social media content as well as passive time, when one of the sites is open in the current browser tab. We also consider potential spillover effects to platforms where the intervention did not take place. The extension measured time spent by users on 38 related websites (the list is provided in Appendix F).

Heterogeneity As indicated in the pre-registration, we explored two angles of heterogeneity. First, we split the sample into two parts according to the toxicity of content consumed on the three platforms during the baseline period. To that end, we ranked individuals by the average toxicity score, and categorized the participants relative to the median person. Considering the above-the-median individuals – henceforth referred to as the *toxic sample* – 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

[39] Please note that we cannot include likes and retweets in our analysis of contagion. This is because any effect for these endpoints would be explained by a mechanical effect; content with toxicity exceeding 0.3 that users could have shared or reacted to would be hidden in the treatment group. This only allows for any meaningful difference to occur for elements with the scores below the threshold.

about each participant. The second angle of heterogeneity is by platform. Due to the fundamental differences between Facebook, Twitter, and YouTube – some of which became apparent during preliminary data analysis – we focus mostly on platform-specific investigation, reporting our results for each website separately.

3.4.2 Survey Outcomes

We collected additional outcomes in the endline survey (see Appendix D for the exact wording of all questions). In particular, 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 pertained to positive emotions and behavior: happiness, life satisfaction, being absorbed in doing something worthwhile. The other three focused 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 its mean, and σ_i the standard deviation,[40] 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.[41] Moreover, to analyze whether a lower exposure to toxic content reduces users' normalization of hateful attitudes, we asked the participants to read seven online comments (see Appendix D.2.3), 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.

3.5 Descriptive Statistics

Panels A and B of Table 1 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

[40] The proposed solution to aggregating outcomes illuminating the same phenomenon has been widely used by researchers. Examples include Kling et al. (2007) and Bursztyn et al. (2017).

[41] This approach mirrors Allcott et al. (2020) who underscored in their survey questions that the period of interest is the last 4 weeks, the duration of their Facebook deactivation intervention.

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. Moreover, the 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.

Table 1: Descriptive statistics

Panel A: User demographics

	Main Sample (mean)	Representative (mean)	Difference (t)
Age 18-29 (%)	30.90	30.99	0.03
Age 30-49 (%)	36.01	39.84	1.43
Age 50-64 (%)	22.12	20.76	-0.64
Male (%)	52.24	54.17	0.70
Democrat (%)	53.56	35.35	-6.79
Independent (%)	36.83	43.81	2.53
White (%)	64.40	69.24	1.79

Panel B: Twitter accounts

	Main Sample (mean)	Representative (mean)	Difference (t)
Account years	7.0	5.2	-9.38
Number of followers	1,643.0	4,803.7	4.95
Accounts followed	1,211.7	1,071.3	-1.59

Panel C: Baseline outcomes

	Facebook, N= 579			Twitter, N= 747		
	Mean	Median	SD	Mean	Median	SD
Content shown/day	96.7	23.1	198.5	237.4	106.0	361.9
Posts/day	46.6	11.2	106.8	159.4	79.7	225.3
Comments/day	50.1	11.4	105.4	75.7	20.1	173.2
Toxicity/content shown	0.03	0.03	0.02	0.07	0.07	0.04
Content produced/day	2.3	0.3	5.7	2.3	0.2	6.0
Toxicity/content produced	0.04	0.02	0.08	0.08	0.03	0.13
Minutes spent/day	29.5	5.4	72.6	36.3	14.1	67.9

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.

Panel C of Table 1 reports summary statistics for a subset of our outcomes based on the 14-day baseline period for our two main platforms: Facebook and Twitter. On average, users spend roughly half an hour per day on both platforms. They consume 1.4 times more content on Twitter, despite producing 2.3 elements of content on both platforms. Comments constitute half of Facebook content but only one-third in the case of Twitter. The average toxicity score per unit of content (both consumed and produced) is almost double on Twitter. Throughout the study we collected a total of 15,287,908 posts, comments, and replies shown to our participants, including 10,281,107 on Twitter, 3,065,231 on Facebook, and 1,941,570 on YouTube. Despite platforms' existing content moderation efforts, the extension recorded a remarkable total of 1,053,787 toxic elements – posts, comments, and replies with a toxicity score exceeding 0.3. During the study, the highest proportion of toxic content was collected on Twitter (7.84%), followed by YouTube (5.98%), and Facebook (4.28%). Even more surprisingly, the distribution of the toxicity scores for the above-the-threshold content strongly skewed to the right, as depicted in Figure 12 in the appendix. In fact, we observed that as much as 52.94% of toxic elements had a score above 0.7.

3.6 Empirical Strategy

At the core of our identification strategy is the use of the baseline period to establish 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 precision. 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 activity equal to 3449 minutes. The first quartile values were 12 and 1644, 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 date. 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 alternative standard errors.[42]

Even though our recruitment period was very short (about 3 weeks), one may be concerned that our participants enrolled in the study on different days, and thus, in terms of calendar days, their presence in the experiment did not perfectly overlap. 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 (proposed by Cengiz et al., 2019; Gardner, 2022), which address this problem. This involves extending specification 1 by including start date × individual and start date × period fixed effects.

Lastly, it is important to note that browser extensions do not allow developers to observe 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 day 46 or later (at least 10 days before the end of the intervention). This assumes that 10 days of inactivity is a reliable signal of attrition. As a robustness check, we explore different attrition thresholds and show that our results are not driven by the particular choice of day 46.

4 Results

In this section we present the findings of the paper. We categorize the results into four strands: content consumption, content production, time spent on the platforms (including spillover effects), and users' well-being. We address each category in turn. For platform-specific outcomes, in the main text of the paper we focus on presenting the findings for Twitter and Facebook – social media where text-based exchanges are the primary reason for visiting the site.[43] We outline the results pertaining to YouTube activity in Appendix H. The section concludes with a discussion of robustness and potential concerns.

Chiefly among the concerns is the risk of differential attrition. As a result, we devote Section 4.5.1 to providing evidence that our findings are unlikely to be driven by differential attrition. There, we evaluate survival patterns throughout the study, highlighting parallel trends in dropout dynamics. Additionally, we report regression analysis indicating that the hiding intervention did not significantly affect the proportion of survivors at the end of the study, and that the average number of days in the study was uncorrelated with treatment group. Lastly, we discuss regression analysis refuting two likely

[42] Driscoll and Kraay (1998) provide a non-parametric 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 units per individual.

[43] YouTube significantly differs from Facebook and Twitter in the context of our intervention, as text exchanges are only a secondary reason (behind watching videos) for visiting the platform. This was crystallized by the extension data on the average quantity of content per minute spent on each of the platforms: 6.73 for Twitter, 3.27 for Facebook, and only 1.52 for YouTube (22.6% of Twitter's value and 46.5% of Facebook's value).

channels that could fuel differential attrition – related to users' activity and "preference" for toxicity (with the baseline exposure to toxicity as a proxy).

4.1 Content Consumption

4.1.1 Toxicity

During the intervention period, the extension automatically removed a daily average of 23.7 pieces of toxic text content per treated user on the three supported platforms, including 15.8 on Twitter and 4.6 on Facebook. To put these quantities in context, we report hiding 6.55% of such content displayed to users in their browser – 7.10% on Twitter, and 4.94% on Facebook. Given that our participants reported spending, on average, 62% of their Twitter time and 57% of their Facebook time on a desktop device, a back-of-the-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 toxic content on social media.

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 content *offered* by the platforms, i.e., inclusive of the elements that were hidden (dashed line), and the toxicity of content *shown* , i.e., displayed to users (solid line). For participants in the control group, the figure plots the toxicity of content shown. 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 to participants did not differ by treatment arm – it was 0.063 in the control and 0.064 in the treatment. These 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 73.2% across the three platforms.

Table 9 in the appendix demonstrates that the treatment lowered the average toxicity score by about 2 pp. on Facebook (p-value < 0.001) and 5 pp. on Twitter (p-value < 0.001). Respectively, these can be interpreted as a 58.6% and a 70.5% reduction relative to the mean. Lastly, the lack of difference in the average toxicity of content that the platforms intended to display to users between the experimental groups and across time indicates that, at least on average, the platforms' algorithms did not learn anything about the participants' preferences for toxic content as a result of the intervention.

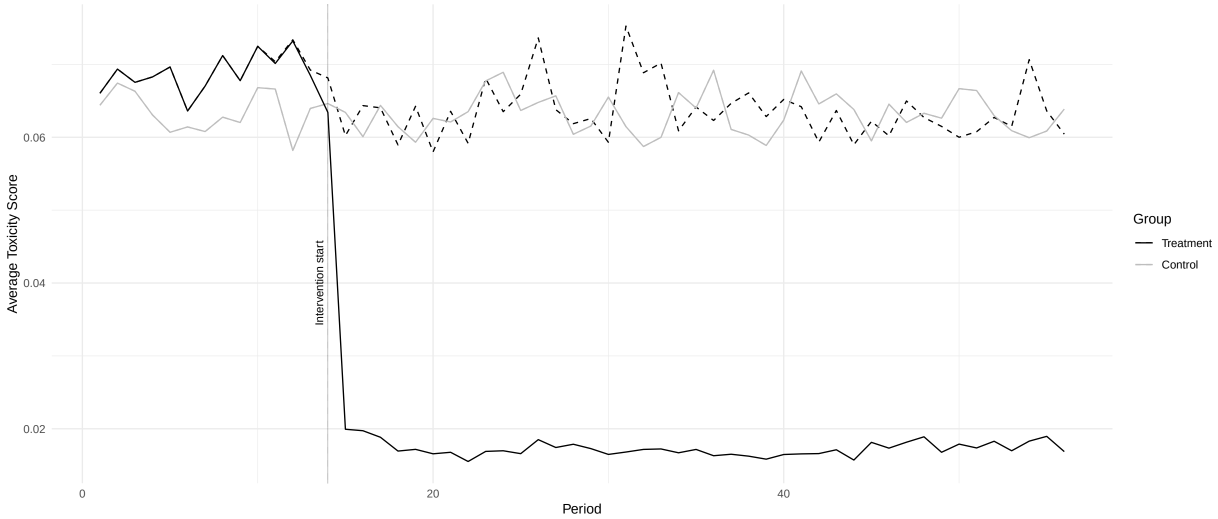


Figure 2: Average Toxicity of Content Shown to Users During the Study

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.

4.1.2 Quantity

Table 2 summarizes the main results on our key measure of engagement – the quantity of posts and comments that users consume. The hiding intervention, which lowered exposure to toxicity, significantly reduced content consumption on Facebook. The conclusion was reached from the treatment effect on the posts and comments 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 17.6 elements a day (p-value $<$ 0.001). This magnitude represents a 23% decrease relative to the mean quantity of content throughout the study, or 0.08 standard deviations. The effect is similar for the sample of all Facebook users and for those in the toxic sample, although it is slightly stronger for the latter. On Twitter, the effect on content consumption, measured using content offered, is statistically (p-value $=$ 0.912) and economically insignificant (-0.8 elements per day or 0.002 standard deviations).

Figure 13a and 13b in the appendix present estimates from the event study specification for Facebook and Twitter respectively, which allow us to analyze the dynamics of the treatment effects. The drop in content consumption on Facebook occurs after the first week of the intervention, and appears to persist throughout most of the study. Hence, the effects are not driven by period-outliers.

Table 2: Effect of Intervention on Offered Content

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	-17.6***	-0.8	-21***	4.1
	(2.753)	(7.177)	(3.888)	(13.882)
	p < 0.001	p = 0.912	p < 0.001	p = 0.77
N	31 864	38 472	15 120	19 320
Mean	76.15	204.53	60.84	274.32
SD	218.08	424.28	193.66	508.85

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is 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. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

To further shed light on our main estimates, we split our data by whether a piece of content in question is a post or a comment – Table 10 in the appendix presents the results. It is notable that the intervention reduced consumption of both posts (user feed) and comments on Facebook, the former of which is consequential for ad impressions – as the platform places ads in between posts. For the main sample, the consumption of posts fell by 9.7 per day (a 26% change). The reduction for comments was equal to 8.5 (a 21% change). This is mirrored in the toxic sample, with the numbers being, respectively, 9.8 and 11.7. On Twitter, the results are inconclusive, although suggestive of a slightly negative effect on comments.

Our conservative measure of content consumption – content offered – adds the elements hidden by the browser to the number of elements actually displayed to users. This provides a lower bound (in absolute value) on any negative treatment effect, and ensures that the results are not driven by the mechanical effect of hiding. While we use content offered as our main 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. Overall, the intervention led to a reduction in content displayed to users, as presented in Table 11 in the appendix, by 21.3 posts per day (p-value < 0.001) on Facebook and 14.6 (p-value $= 0.04$) on Twitter.

4.1.3 Advertising

The hiding intervention resulted in a reduction of ad consumption on Twitter. In particular, Table 3 indicates that the average number of ads displayed to users per day fell by 1.8 (p-value = 0.002), a 9.2% difference. Similarly to our main measure of content consumption, we count all ads including those that are hidden, to ensure that the result is not driven by the mechanical effect of the intervention. In this context, it is important to note that only a very small fraction of ads (0.59%) was identified as toxic, so relying on the conservative approach has a minimal impact on the point estimate. Unlike in the case of Twitter, we are unable to precisely identify ads on Facebook. However, as previously discussed, we can use consumption of posts (excluding comments and replies) as a proxy for the number of ads displayed to users on Facebook. We find a negative effect of the intervention on that measure, which is consistent with a drop in ad consumption on Twitter.

Table 3: Effect of Intervention on the Number of Twitter Ads

	Main Sample	Toxic Sample
	(1)	(2)
Treated	-1.8*** (0.534) p = 0.002	-0.3 (0.925) p = 0.708
N	20 617	11 464
Mean	19.54	20.76
SD	24.17	24.18

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of ads offered to users on Twitter; those displayed on their feeds plus the ads mechanically hidden by the extension. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

4.2 Content Production

4.2.1 Toxicity

Having discussed content consumption, we now turn our attention to content production. Table 4 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 the content they publish – conditional on posting. The effect is quantitatively similar for both platforms: -0.014 on Facebook (p-value = 0.016) and -0.016 on Twitter (p-value = 0.006). These can be interpreted as a 35% and a 20% reduction in the content toxicity relative to the mean, or a decrease of 0.15 and 0.106 standard deviations, respectively, for Facebook and Twitter. The pattern is similar for users in the main sample and those in the toxic sample, although the magnitude is slightly stronger for the latter. Taken together, we find broad evidence consistent with the hypothesis that toxicity on social media is contagious.

Table 4: Effect of Intervention on Toxicity of Produced Content

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	-0.014** (0.006) p = 0.016	-0.016*** (0.006) p = 0.006	-0.023** (0.009) p = 0.013	-0.02*** (0.008) p = 0.01
N	6658	9621	2737	5968
Mean	0.04	0.08	0.05	0.1
SD	0.09	0.15	0.1	0.17

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic 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 up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Next, we investigate a potential mechanism behind the contagion hypothesis, namely, that exposure to toxic content contributes to normalization of toxic behavior, which then increases the likelihood that users engage in this type of behavior. Despite the overall strong effect of the exposure on toxicity of own content, we find no evidence of normalization of toxicity. The results are presented in Table 15 in the appendix. We report an insignificant effect (p-value = 0.491) on the index summarizing users' evaluations of seven toxic statements in the endline survey, which offers suggestive evidence that exposure to toxicity does not change their opinions on what is considered toxic. Additionally, the table provides regression analysis for each statement separately, and shows that the intervention resulted in a significant difference in toxicity evaluation only for 1 out of 7 statements, with comment C7 (see Appendix D.2.3) considered more toxic in the treatment group by 11 pp. (p-value = 0.028).

4.2.2 Quantity

Besides an effect on the toxicity of users' posts, we also present a result on the quantity of content they write. As Table 5 demonstrates, the intervention reduced the publishing of posts and comments by 0.7 per day (p-value = 0.022) in the main Facebook sample. The effect for Twitter is in the opposite direction and insignificant (0.1 more daily content, p-value = 0.443). Taking into account the previously discussed findings, we conclude that our intervention results in a reduction of both content consumption and content production on Facebook. We do not provide similar evidence for Twitter.

The effects on consumption and production diverge among users in the toxic sample, suggesting a relative shift towards production. On Facebook, individuals did not change the quantity of content they produce but reduced their consumption. On Twitter, they increased content production but consumption did not rise (some measures, e.g. ad consumption went down).

Table 5: Effect of Intervention on Total Number of (Own) Posts

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	-0.7**	0.1	0.2	0.6***
	(0.272)	(0.099)	(0.335)	(0.19)
	p = 0.011	p = 0.443	p = 0.591	p = 0.003
N	27 720	30 352	12 600	16 296
Mean	2.51	2.42	1.93	3.28
SD	14.16	7.56	7.92	8.89

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of elements of content posted by users. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

4.2.3 Reactions

Separately, we investigate the effects of the intervention on the number of reactions (such as likes). Table 12 in the appendix provides the results. Overall, we report no significant effect on both Facebook and Twitter. However, the hiding intervention led to a reduction in the number of reactions on Facebook for the toxic sample (p-value = 0.018), a result consistent with a drop in other types of engagement (such as lower content consumption).

4.3 Time Spent

We now present the results on the time users spend on the platforms – both those undergoing the intervention and otherwise. First, we find no evidence of a significant change in the amount of time individuals spend either on Facebook or on Twitter. This is so despite being ex-ante powered to rule out effects larger than 0.03 standard deviations around the point estimate in the main sample. Table 6 presents the regression details. A caveat here is that the effect for Twitter in the toxic sample is positive and marginally significant, at the 10% level. This finding is interesting in the context of the previously reported results regarding the toxic sample – the intervention increased their content production as well as the number of reactions. Taken together, these indicate that there is potential in exploring whether for users with high exposure to toxicity hiding toxic content could improve some forms of engagement. We hope that this type of heterogeneity will be a subject of further academic work.

Table 6: Effect of Intervention on Social Media Consumption Time

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	1.5 (1.384) p = 0.284	-0.4 (1.316) p = 0.777	1.3 (1.641) p = 0.417	3.9* (2.158) p = 0.079
N	32 760	38 752	15 512	19 320
Mean	22.88	30.58	22.08	42.76
SD	80.69	85.35	91.44	106.54

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of minutes spent on the platform. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Finally, somewhat surprisingly, reducing exposure to toxicity on the treated platforms led to positive spillover effects on the combined total time spent on 38 *other* social media platforms, where the intervention did *not* take place. Table 13 in the appendix presents the results of the regression of time spent on social media platforms other than Facebook, Twitter, and YouTube. In particular, the hiding intervention resulted in users spending, on average, 1.8 more minutes per day on the not treated platforms (p-value = 0.029).

4.4 Measures of Well-Being

In this subsection we focus 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 self-reported individual well-being (p-value = 0.634). Moreover, Table 14 indicates that the treatment had no significant impact on any components of the index considered in isolation: happiness (p-value = 0.488), life satisfaction (p-value = 0.710), anxiety (p-value = 0.730), depression (p-value = 0.629), doing something worthwile (p-value = 0.303), and boredom (p-value = 0.523). These findings provide suggestive evidence that exposure to toxicity may not be the main driving force behind the negative effects on social media on well-being, a relationship well-documented in the literature. This point should be treated with caution given the small sample size (N=388 individuals). 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 penetration affects individual welfare.

4.5 Robustness and Potential Concerns

4.5.1 Attrition

A major threat to identification in the paper is a risk of differential attrition. We devote this section to documenting that the hiding intervention, lasting six weeks, did not result in differential dropout of participants throughout the experiment.

We begin the analysis by inspecting attrition trends using Figure 14 in the appendix. For each day of the study (relative to when a participant started), the figure visualizes the proportion of individuals who were active on that day or later. In other words, it plots the proportion of users who certainly remained extension users in each period. The graph provides no indication of differential dropout by treatment group at any point of the experiment. Given that the user experience did not vary by group during the baseline, it is insightful to consider attrition specifically during the intervention period. We report that 84.9% of those who remained until the start of the intervention were active on day 56 (the last day of the study) or later, with the following split by group: 86.6% in the treatment and 83.0% in the control. The difference in the proportion of survivors between the groups is not statistically significant (p-value = 0.166). Together with the earlier graphical evidence, this implies that the hiding intervention did not lead to differential dropout of participants, at least in terms of their number.

Another way of using the survival rates to inspect the issue of attrition is to compare its pace between the intervention period and the baseline, where there were no differences across the groups. The average attrition pace per week was similar during the former (2.5 pp.) and the latter (3.6 pp.). The symmetry in dropout by treatment group is even more clear in Figure 15 in the appendix, where we depict the number (rather than the proportion) of remaining participants, computed using the same method. We deliberately set the beginning of the y-axis to 300 to zoom in on the attrition trends. By inspecting the graph, we can conclude that they were almost exactly parallel. Lastly, Figure 16 in the appendix depicts the distribution of users' last active day by the treatment group. This figure offers an alternative depiction of the patterns of attrition throughout the study.

More formally, Table 16 in the appendix shows a regression of the last day seen in the study on the treatment dummy (Column 1). The coefficient is insignificant with a p-value of 0.55. Furthermore, we considered two likely channels that could have led to differential attrition regarding types of individuals leaving. Given the character of our hiding intervention, we worried that people with preference for toxic content or those with high levels of social media activity were more likely to drop out of the treatment group. We extend the regression analysis to refute these conjectures. First, we include the average toxicity score of content displayed to the user during the baseline, and its interaction with the treatment dummy (Column 2). This specification is motivated by the possibility that the platforms might optimize on the toxicity of content shown; thus, the added covariate could be a proxy for the tolerance (or preference) for toxicity. Second, we provide a specification with the average time spent on social media during the baseline, and its interaction with the treatment dummy (Column 3). All of the interaction coefficients are insignificant.

4.5.2 Robustness: Attrition Thresholds

As previously discussed, due to our inability to observe unistallation events, which is a general issue with browser extensions, 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 day 46 or later, thus requiring 10 days of inactivity to determine that someone dropped out of the study. In this subsection, we explore different attrition thresholds and show that our results are robust to applying them.

First, Appendix G.1 presents the main regression tables for the panel of users who were seen on 56 or later – this applies to 84.9% of participants who were active at any point of the intervention period. The advantage of this specification is that for every single period we have certainty that we do not mistake the lack of a particular activity (e.g. time spent on a platform or posting) for missing data. The regression tables demonstrate that all of our main significant results – the effects on content consumption on Facebook (content offered), both overall as well as considering posts and comments separately, ad consumption on Twitter, production and toxicity of own posts, and time spillover effects, are robust to applying day 56 attrition threshold. Additionally, using this specification we find a significant effect on comments consumption on Twitter (using the conservative measure) both in the main sample (p-value = 0.017) and in the toxic sample (p-value = 0.006).

Second, the panels of day 46 and day 56 survivors have a drawback of excluding initial data for participants who likely dropped out in the middle of the intervention. To address the issue, in Appendix G.2 we consider the regression specifications where for each individual we include data up until the point of their last day of browser activity. We find that the main results (as listed above) are robust to applying this definition of attrition, with the caveat that the effect of the intervention on the time spent on untreated platforms related to social media is now significant only at the 10% level. Taking this together with the previous set of robustness checks, we conclude that our results are not driven by an arbitrary choice of the attrition threshold, and do not hinge on wrongly imputing zeros in lieu of missing data.

4.5.3 Robustness: Alternative Specifications

In this subsection we discuss robustness of our results to alternative regression specifications. In Appendix G.3, we replicate the main regression tables using a stacked regression specification with start date × individual and start date × period fixed effects. All of the significant results for the main sample reported in Section 4 are robust to applying this specification.

Finally, we move on to discuss robustness to clustering standard errors at the individual level rather than relying on Driscoll and Kraay standard errors. The relevant regression analysis is presented in Appendix G.4. We report that our finding that the intervention reduced content consumption on Facebook (content offered) is robust to such clustering both in the sample of all users (p-value $= 0.029$) and the toxic sample (p-value $= 0.007$). Furthermore, the results on the consumption of comments are robust at the 5% level, whereas the effects on the consumption of posts are robust at the 10% level. We have insufficient power to detect an effect of the hiding intervention on ad consumption on Twitter with clustered standard errors – the p-value increases to 0.112. Similarly, the results on production of own posts and spillovers to the untreated platforms are not robust to clustering at the individual level. At the same time, we find that the evidence in favor of the contagion hypothesis is robust such clustering – both on Twitter (p-value $= 0.014$) and on Facebook (p-value $= 0.014$).

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 the quantity of content that users consume on some platforms. We also discuss the potential impact on ad impressions and profitability by showcasing that the intervention led to lower consumption of posts on Facebook's feeds (where ads are inserted by the platform) and, more directly, a lower ad consumption on Twitter. Lastly, we find evidence supporting the concerns that toxic online behavior can be contagious.

These findings confirm that policymakers and companies attempting to incentivize healthier conversations on the online agora face no easy task. The contagiousness of toxicity suggests that reducing its prevalence can bring welfare improvements due to the presence of spillovers between users (and in light of the well-documented real-world effects of toxic content). However, a revealed-preference approach that takes users' consumption patterns at face value might suggest that curbing toxic content is not unambiguously welfare-improving, and that users may have a degree of preference for it. 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. Additionally, our finding that exposure to toxicity can reduce content consumption holding the time spent on the platform constant suggests that it might be important to study the impact of toxicity on the attention paid to social media content and information more broadly.

References

ACEMOGLU, D., A. OZDAGLAR, AND J. SIDERIUS (2021): "Misinformation: Strategic sharing, homophily, and endogenous echo chambers," Tech. rep., National Bureau of Economic Research.

ADENA, M., R. ENIKOLOPOV, M. PETROVA, V. SANTAROSA, AND E. ZHURAVSKAYA (2015): "Radio and the Rise of The Nazis in Prewar Germany," *The Quarterly Journal of Economics*, 130, 1885–1939.

ALLCOTT, H., L. BRAGHIERI, S. EICHMEYER, AND M. GENTZKOW (2020): "The welfare effects of social media," *American Economic Review*, 110, 629–76.

ALLCOTT, H. AND M. GENTZKOW (2017): "Social Media and Fake News in the 2016 Election," *Journal of Economic Perspectives*, 31, 211–36.

ALLCOTT, H., M. GENTZKOW, AND L. SONG (2021): "Digital addiction," Tech. rep., National Bureau of Economic Research.

ARIDOR, G. (2022): "Drivers of Digital Attention: Evidence from a Social Media Experiment," *Available at SSRN 4069567*.

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, 370–395.

BEKNAZAR-YUZBASHEV, G. AND M. STALINSKI (2022): "Do social media ads matter for political behavior? A field experiment," *Journal of Public Economics*, 214, 104735.

BLOUIN, A. AND S. W. MUKAND (2019): "Erasing ethnicity? Propaganda, nation building, and identity in Rwanda," *Journal of Political Economy*, 127, 1008–1062.

BOXELL, L., M. GENTZKOW, AND J. M. SHAPIRO (2019): "Cross-Country Trends in Affective Polarization," *Working Paper.*

Braghieri, L., R. Levy, and A. Makarin (2021): "Social Media and Mental Health," *Available at SSRN.*

Bursztyn, L., G. Egorov, R. Enikolopov, and M. Petrova (2019): "Social media and xenophobia: evidence from Russia," Tech. rep., National Bureau of Economic Research.

Bursztyn, L., T. Fujiwara, and A. Pallais (2017): "'Acting Wife': Marriage Market Incentives and Labor Market Investments," *American Economic Review*, 107, 3288–3319.

Cameron, A. C. and D. L. Miller (2015): "A practitioner's guide to cluster-robust inference," *Journal of human resources*, 50, 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, 1405–1454.

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

Coppock, A., D. P. Green, and E. Porter (2022): "Does Digital Advertising Affect Vote Choice? Evidence from a Randomized Field Experiment," *Research & Politics*, forthcoming.

Crockett, M. J. (2017): "Moral outrage in the digital age," *Nature human behaviour*, 1, 769–771.

DellaVigna, S., R. Enikolopov, V. Mironova, M. Petrova, and E. Zhuravskaya (2014): "Cross-Border Media and Nationalism: Evidence from Serbian Radio in Croatia," *American Economic Journal: Applied Economics*, 6, 103–32.

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

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

Fergusson, L. and C. Molina (2019): "Facebook causes protests," *Documento CEDE.*

Fujiwara, T., K. Müller, and C. Schwarz (2021): "The effect of social media on elections: Evidence from the United States," Tech. rep., National Bureau of Economic Research.

Gardner, J. (2022): "Two-stage differences in differences," *arXiv preprint arXiv:2207.05943.*

Jiménez-Durán, R. (2021): "The Economics of Content Moderation: Theory and Experimental Evidence from Hate Speech on Twitter," *Job Market Paper.*

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," Tech. rep., CEPR Press Discussion 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*, vol. 16, 477–487.

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, 922–946.

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

Kosmidis, S. and Y. Theocharis (2020): "Can social media incivility induce enthusiasm? Evidence from survey experiments," *Public Opinion Quarterly*, 84, 284–308.

Le Merrer, E., B. Morgan, and G. Trédan (2021): "Setting the record straighter on shadow banning," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, IEEE, 1–10.

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

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

Madio, L. and M. Quinn (2021): "Content moderation and advertising in social media platforms," *Available at SSRN 3551103*.

Mathew, B., R. Dutt, P. Goyal, and A. Mukherjee (2019): "Spread of hate speech in online social media," in *Proceedings of the 10th ACM conference on web science*, 173–182.

Melnikov, N. (2021): "Mobile Internet and Political Polarization," *Available at SSRN 3937760*.

Mosquera, R., M. Odunowo, T. McNamara, X. Guo, and R. Petrie (2020): "The economic effects of Facebook," *Experimental Economics*, 23, 575–602.

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

——— (2020b): "From hashtag to hate crime: Twitter and anti-minority sentiment," *Available at SSRN 3149103.*

Ribeiro, M. H., J. Cheng, and R. West (2022): "Automated Content Moderation Increases Adherence to Community Guidelines," *arXiv preprint arXiv:2210.10454.*

Rydgren, J. (2005): "Is extreme right-wing populism contagious? Explaining the emergence of a new party family," *European journal of political research*, 44, 413–437.

Schmitt, M. T., N. R. Branscombe, T. Postmes, and A. Garcia (2014): "The consequences of perceived discrimination for psychological well-being: a meta-analytic review." *Psychological bulletin*, 140, 921.

Sunstein, C. R. (2017): *# Republic: Divided Democracy in the Age of Social Media*, Princeton University Press.

Velasquez, N., R. Leahy, N. J. Restrepo, Y. Lupu, R. Sear, N. Gabriel, O. Jha, B. Goldberg, and N. Johnson (2021): "Online hate network spreads malicious COVID-19 content outside the control of individual social media platforms," *Scientific reports*, 11, 1–8.

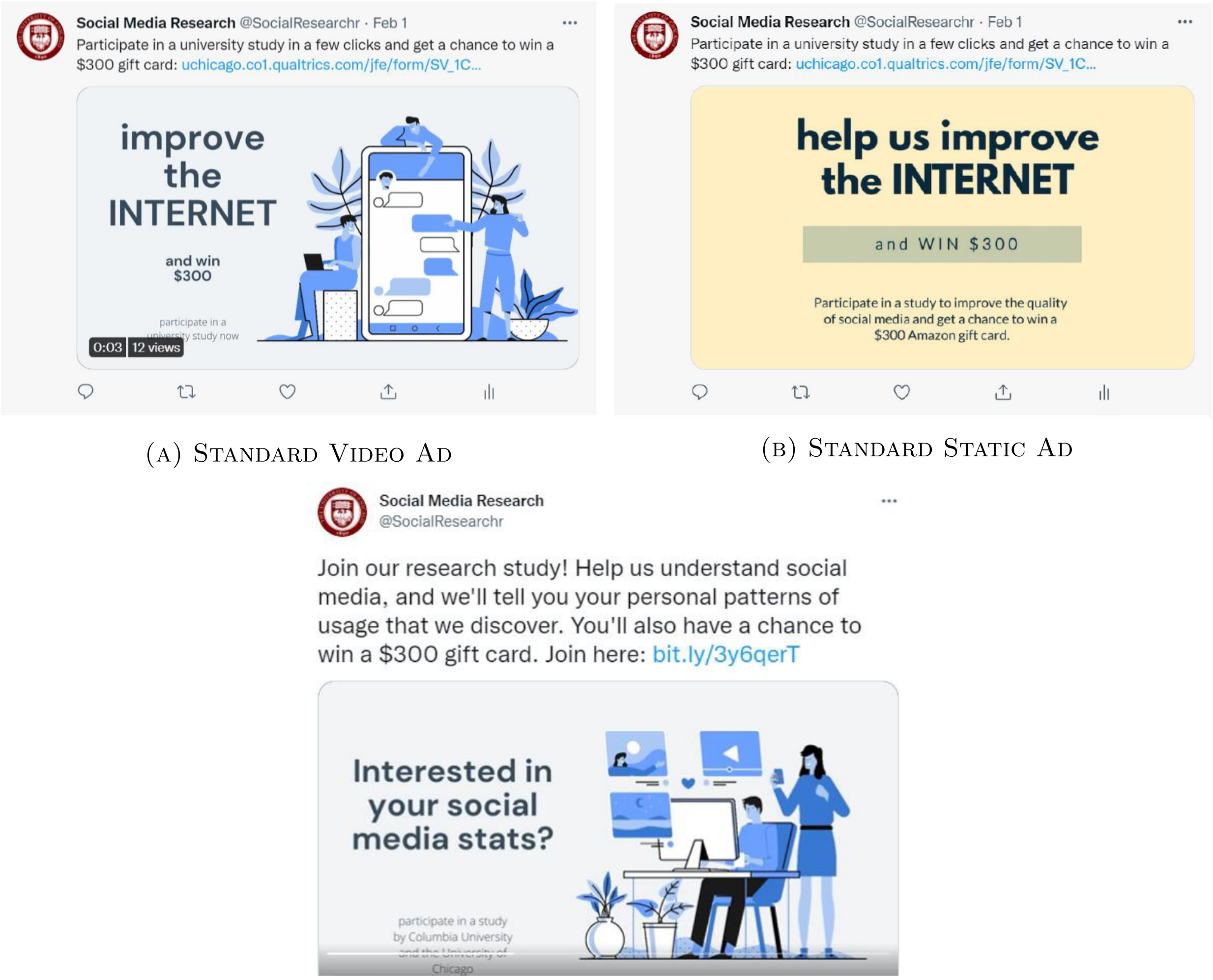
Wang, T. (2021): "Media, pulpit, and populist persuasion: evidence from father Coughlin," *American Economic Review*, 111, 3064–92.

Yanagizawa-Drott, D. (2014): "Propaganda and Conflict: Evidence from the Rwandan Genocide," *The Quarterly Journal of Economics*, 129, 1947–1994.

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

Ziems, C., B. He, S. Soni, and S. Kumar (2020): "Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis," *arXiv preprint arXiv:2005.12423.*

A Figures



Social Media Research @SocialResearchr · Feb 1

Participate in a university study in a few clicks and get a chance to win a $300 gift card: uchicago.co1.qualtrics.com/jfe/form/SV_1C...

improve the INTERNET
and win $300
participate in a university study now
0:03
12 views

Social Media Research @SocialResearchr · Feb 1

Participate in a university study in a few clicks and get a chance to win a $300 gift card: uchicago.co1.qualtrics.com/jfe/form/SV_1C...

help us improve the INTERNET
and WIN $300
Participate in a study to improve the quality of social media and get a chance to win a $300 Amazon gift card.

Social Media Research
@SocialResearchr

Join our research study! Help us understand social media, and we'll tell you your personal patterns of usage that we discover. You'll also have a chance to win a $300 gift card. Join here: bit.ly/3y6qerT

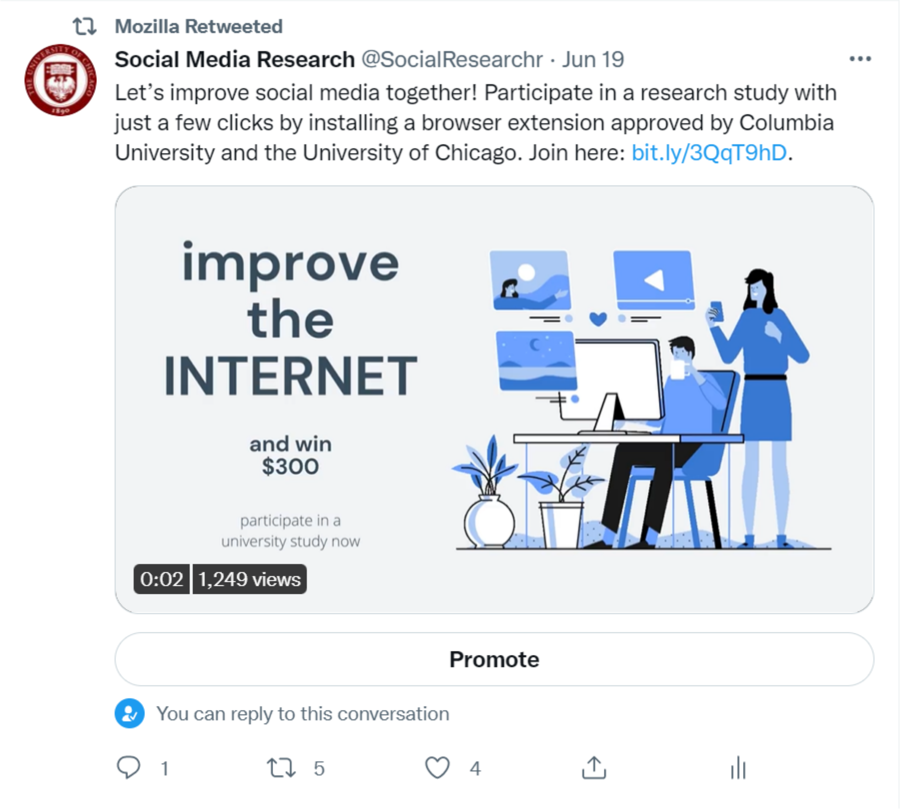
Interested in your social media stats?
participate in a study by Columbia University and the University of Chicago

(a) Standard Video Ad

(b) Standard Static Ad

(c) Learning One's Social Media Stats

Figure 3: Examples of Recruitment Ads on Twitter

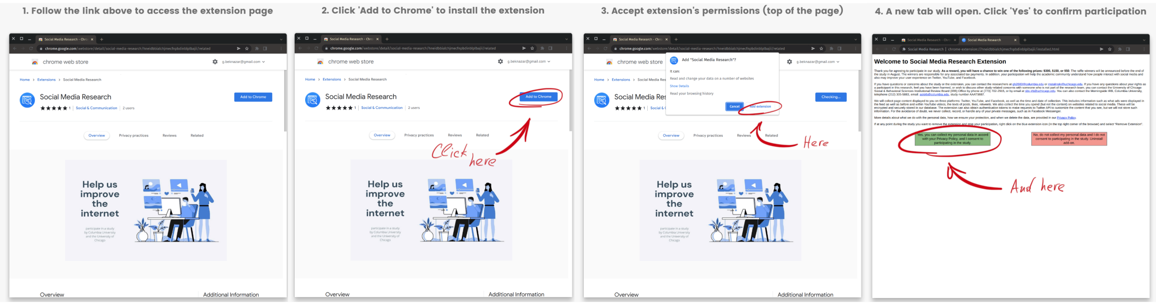


Mozilla Retweeted

Social Media Research @SocialResearchr · Jun 19

Let's improve social media together! Participate in a research study with just a few clicks by installing a browser extension approved by Columbia University and the University of Chicago. Join here: bit.ly/3QqT9hD.

Figure 4: Our Recruitment Post Retweeted by The Mozilla Foundation



1. Follow the link above to access the extension page

2. Click 'Add to Chrome' to install the extension

3. Accept extension's permissions (top of the page)

4. A new tab will open. Click 'Yes' to confirm participation

Figure 5: Screens from Installation Instructions GIF (Chrome Browser)

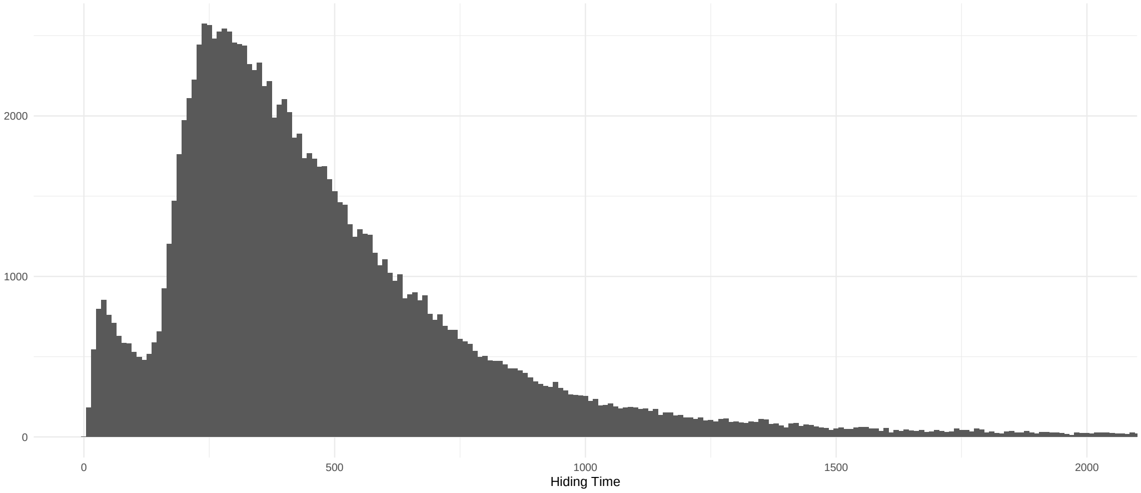


Figure 6: Speed of Hiding Toxic Content on Twitter

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 until the end of the study (end of September).

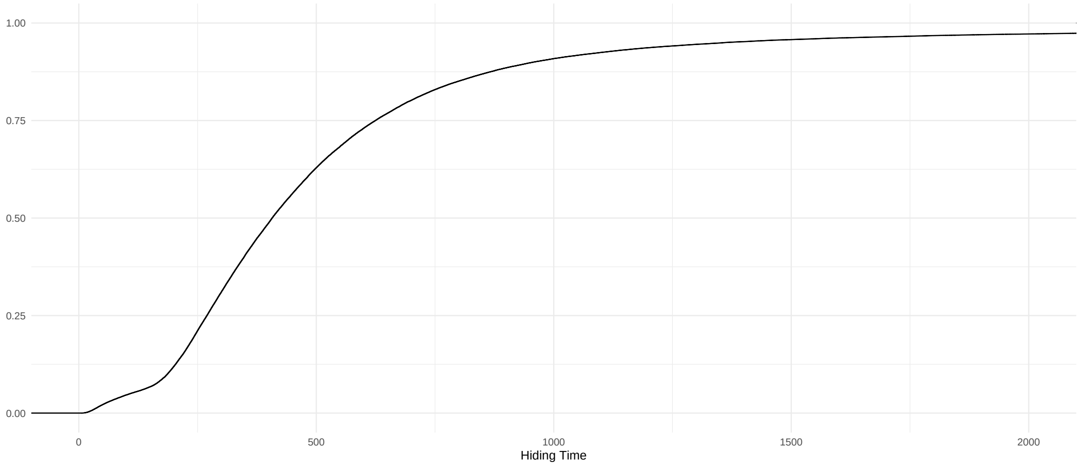
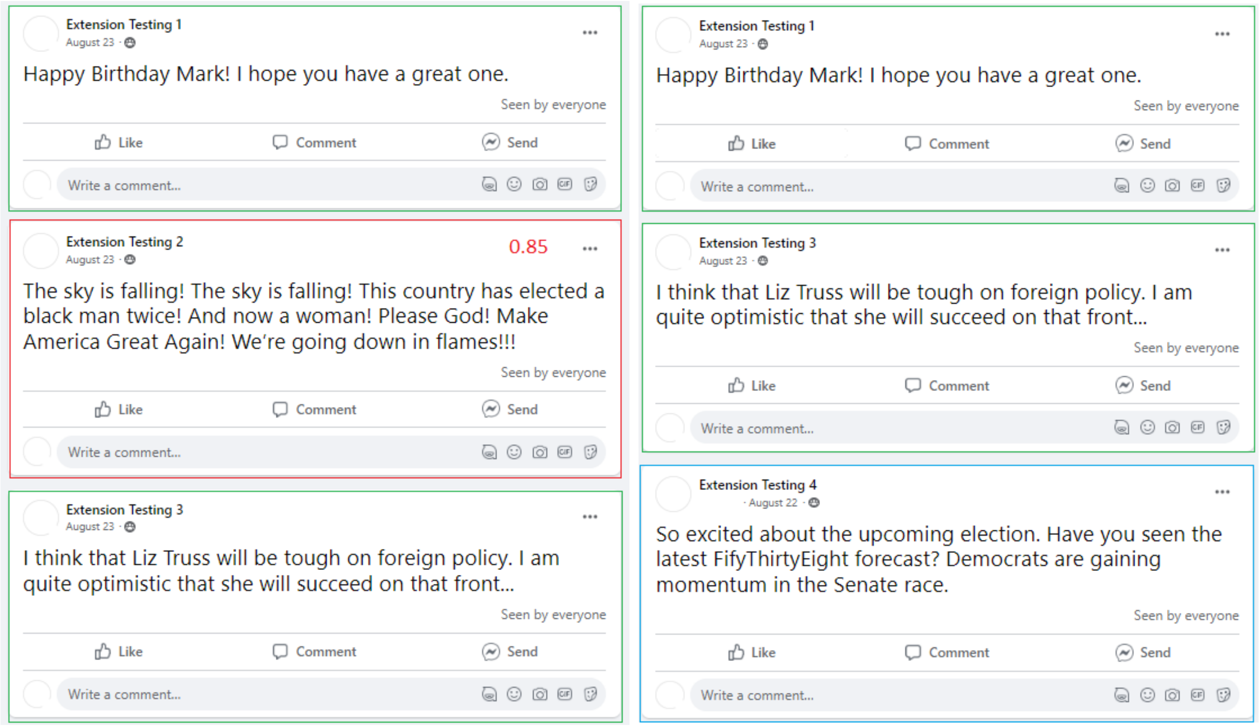


Figure 7: ECDF of the Hiding Speed on Twitter

Note: The figure depicts the cumulative empirical 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. We collected data on the hiding speed from August 22nd until the end of the study (end of September).



Extension Testing 1
August 23 ·

Happy Birthday Mark! I hope you have a great one.

Seen by everyone

Extension Testing 2
August 23 ·

0.85

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!!!

Seen by everyone

Extension Testing 3
August 23 ·

I think that Liz Truss will be tough on foreign policy. I am quite optimistic that she will succeed on that front...

Seen by everyone

Extension Testing 1
August 23 ·

Happy Birthday Mark! I hope you have a great one.

Seen by everyone

Extension Testing 3
August 23 ·

I think that Liz Truss will be tough on foreign policy. I am quite optimistic that she will succeed on that front...

Seen by everyone

Extension Testing 4
· August 22 ·

So excited about the upcoming election. Have you seen the latest FifyThirtyEight forecast? Democrats are gaining momentum in the Senate race.

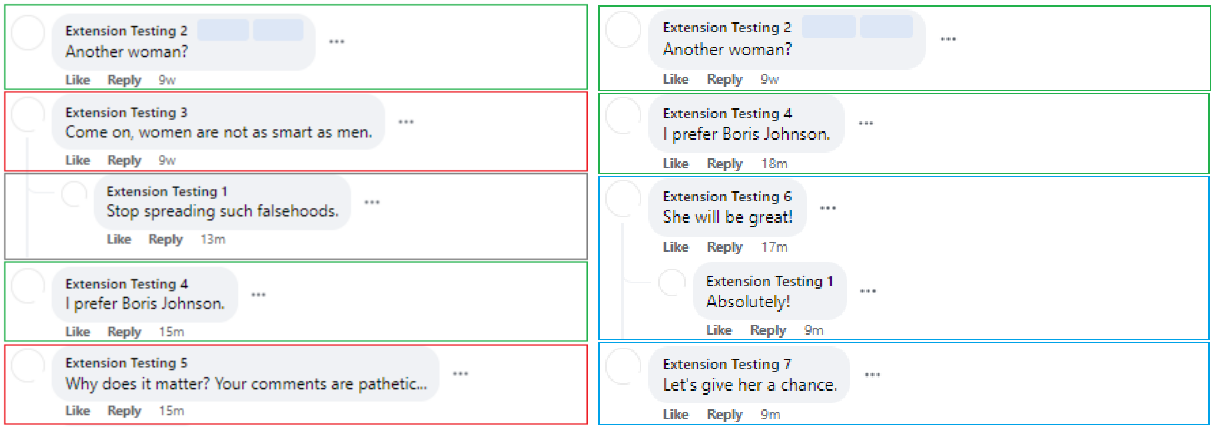
Seen by everyone

(A) Original Feed

(B) Moderated Feed

Figure 8: 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 demonstrates that the content below the hidden element is pulled up. This means that 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.



Extension Testing 2
Another woman?
Like Reply 9w

Extension Testing 3
Come on, women are not as smart as men.
Like Reply 9w

Extension Testing 1
Stop spreading such falsehoods.
Like Reply 13m

Extension Testing 4
I prefer Boris Johnson.
Like Reply 15m

Extension Testing 5
Why does it matter? Your comments are pathetic...
Like Reply 15m

(A) Original Comment Section

Extension Testing 2
Another woman?
Like Reply 9w

Extension Testing 4
I prefer Boris Johnson.
Like Reply 18m

Extension Testing 6
She will be great!
Like Reply 17m

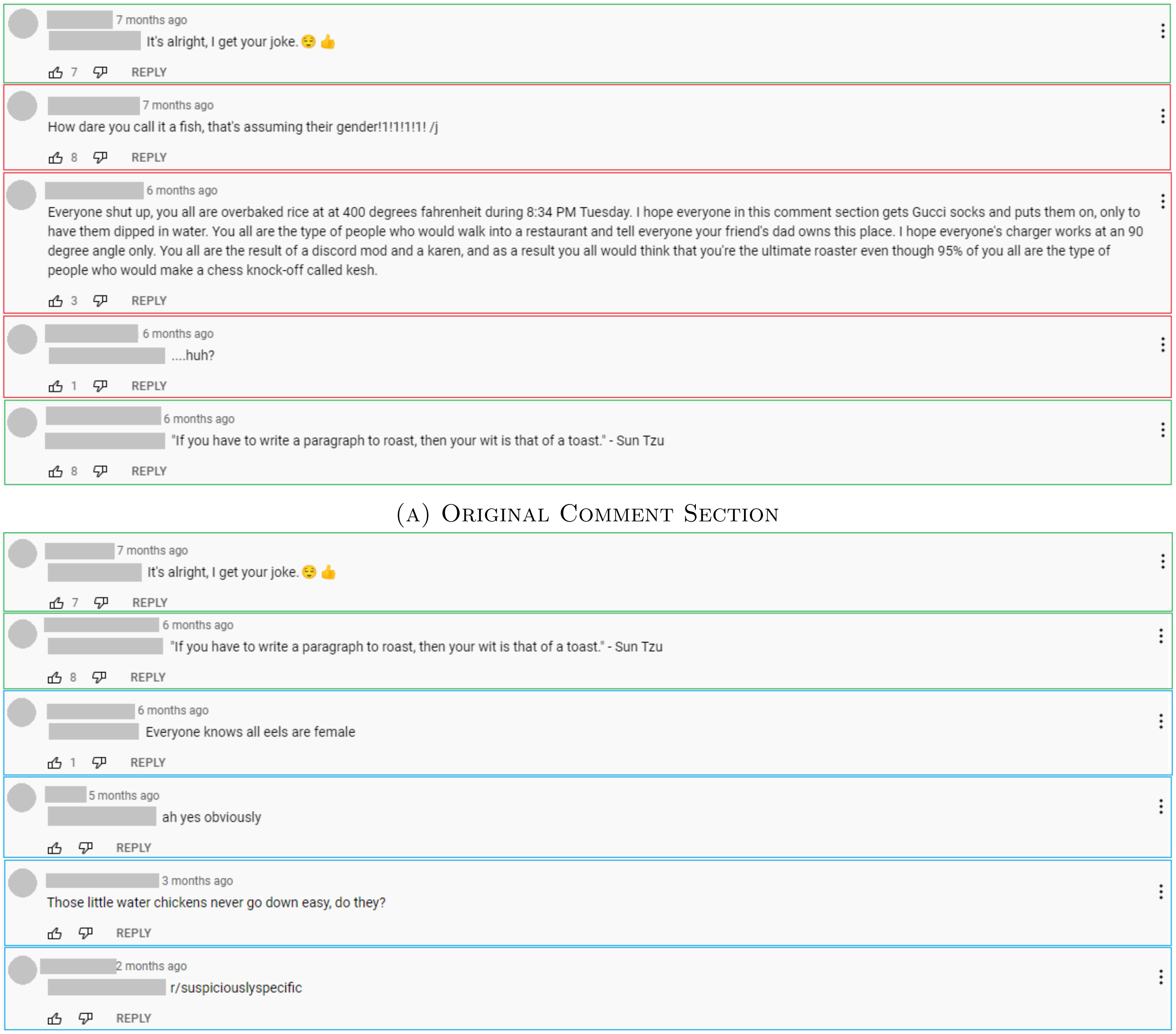
Extension Testing 1
Absolutely!
Like Reply 9m

Extension Testing 7
Let's give her a chance.
Like Reply 9m

(B) Moderated Comment Section

Figure 9: 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. This means that the comment by Extension Testing 4 is now directly below the one by Extension Testing 2. We also see new elements (blue frames), which were previously further below in the comment section.



[redacted] 7 months ago
[redacted] It's alright, I get your joke. 😌 👍
7 REPLY

[redacted] 7 months ago
How dare you call it a fish, that's assuming their gender!1!1!1!1! /j
8 REPLY

[redacted] 6 months ago
Everyone shut up, you all are overbaked rice at at 400 degrees fahrenheit during 8:34 PM Tuesday. I hope everyone in this comment section gets Gucci socks and puts them on, only to have them dipped in water. You all are the type of people who would walk into a restaurant and tell everyone your friend's dad owns this place. I hope everyone's charger works at an 90 degree angle only. You all are the result of a discord mod and a karen, and as a result you all would think that you're the ultimate roaster even though 95% of you all are the type of people who would make a chess knock-off called kesh.
3 REPLY

[redacted] 6 months ago
[redacted] ....huh?
1 REPLY

[redacted] 6 months ago
[redacted] "If you have to write a paragraph to roast, then your wit is that of a toast." - Sun Tzu
8 REPLY

(a) Original Comment Section

[redacted] 7 months ago
[redacted] It's alright, I get your joke. 😌 👍
7 REPLY

[redacted] 6 months ago
[redacted] "If you have to write a paragraph to roast, then your wit is that of a toast." - Sun Tzu
8 REPLY

[redacted] 6 months ago
[redacted] Everyone knows all eels are female
1 REPLY

[redacted] 5 months ago
[redacted] ah yes obviously
REPLY

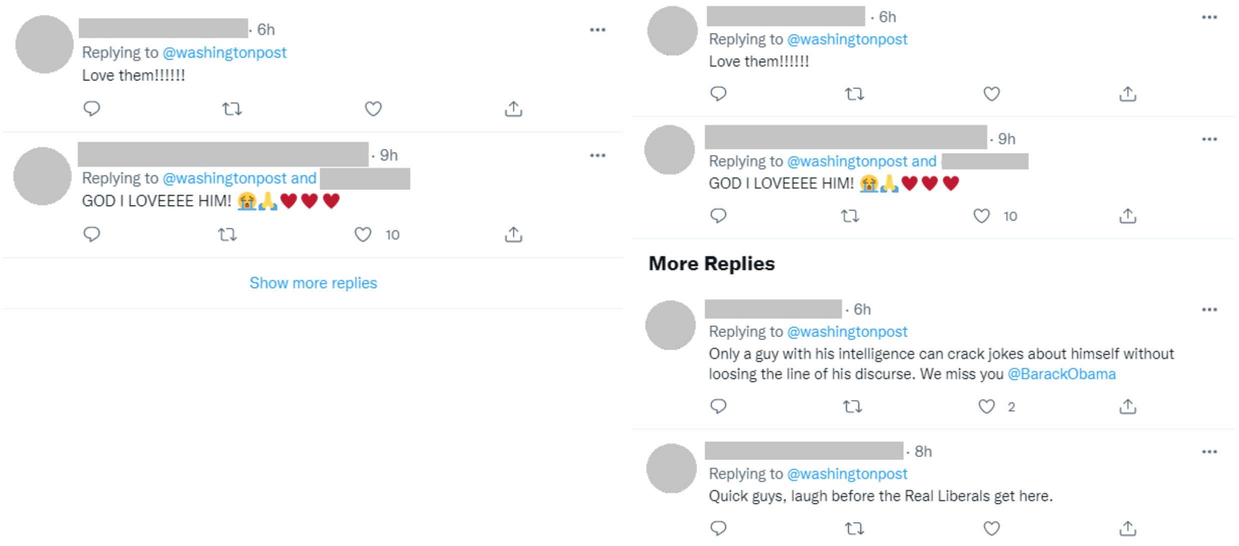
[redacted] 3 months ago
Those little water chickens never go down easy, do they?
REPLY

[redacted] 2 months ago
[redacted] r/suspiciouslyspecific
REPLY

(b) Moderated Comment Section

Figure 10: 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).



[redacted] · 6h

Replying to @washingtonpost

Love them!!!!!!

[redacted] · 9h

Replying to @washingtonpost and [redacted]

GOD I LOVEEEE HIM! 😭🙏❤️❤️❤️

(A) Extension Disabled

[redacted] · 6h

Replying to @washingtonpost

Love them!!!!!!

[redacted] · 9h

Replying to @washingtonpost and [redacted]

GOD I LOVEEEE HIM! 😭🙏❤️❤️❤️

[redacted] · 6h

Replying to @washingtonpost

Only a guy with his intelligence can crack jokes about himself without loosing the line of his discurse. We miss you @BarackObama

[redacted] · 8h

Replying to @washingtonpost

Quick guys, laugh before the Real Liberals get here.

(B) Extension Enabled

Figure 11: 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).

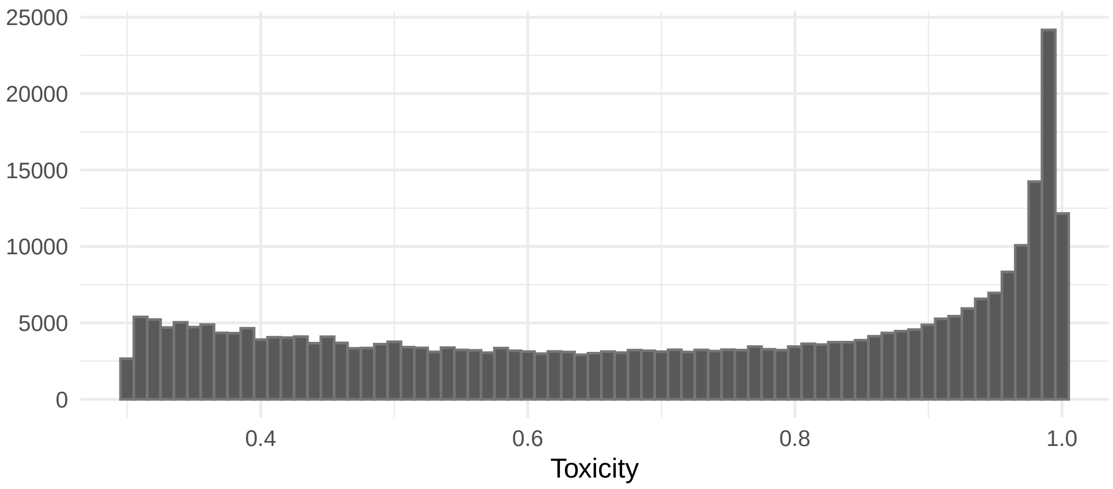
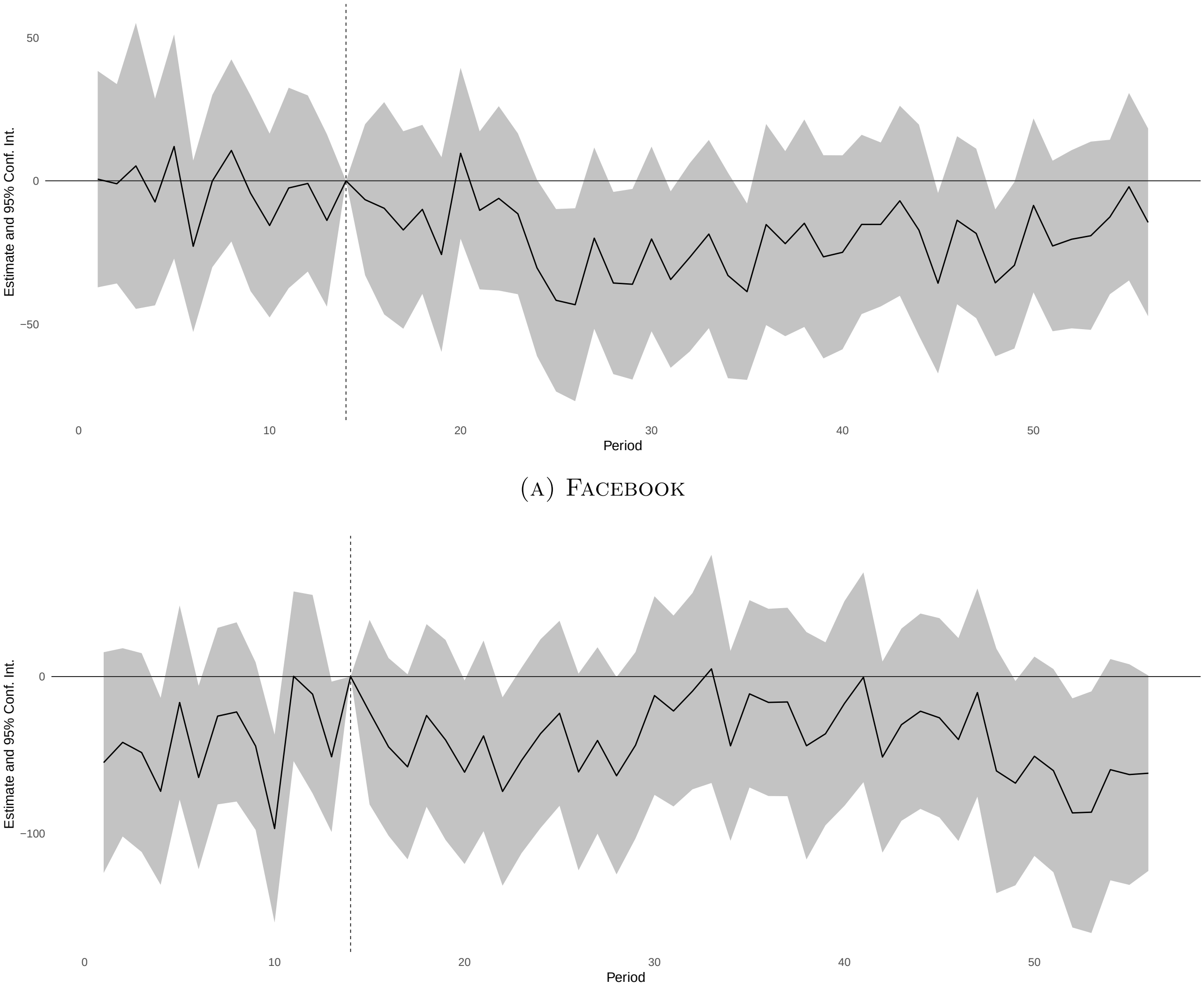


Figure 12: Toxicity of Elements above the Threshold (Histogram)

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.



(a) Facebook

(b) Twitter

Figure 13: Event Study Specifications of Offered Content

Note: This figure presents estimates from the event study version of Equation (1) for Facebook (panel A) and Twitter (panel B). The dependent variable is the number of posts and comments offered to users. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. The solid line represents point estimates and the gray area represents 95% confidence intervals. Standard errors are clustered at the individual level, given that the large-period assumption of Driscoll-Kraay is not plausible in the dynamic specification.

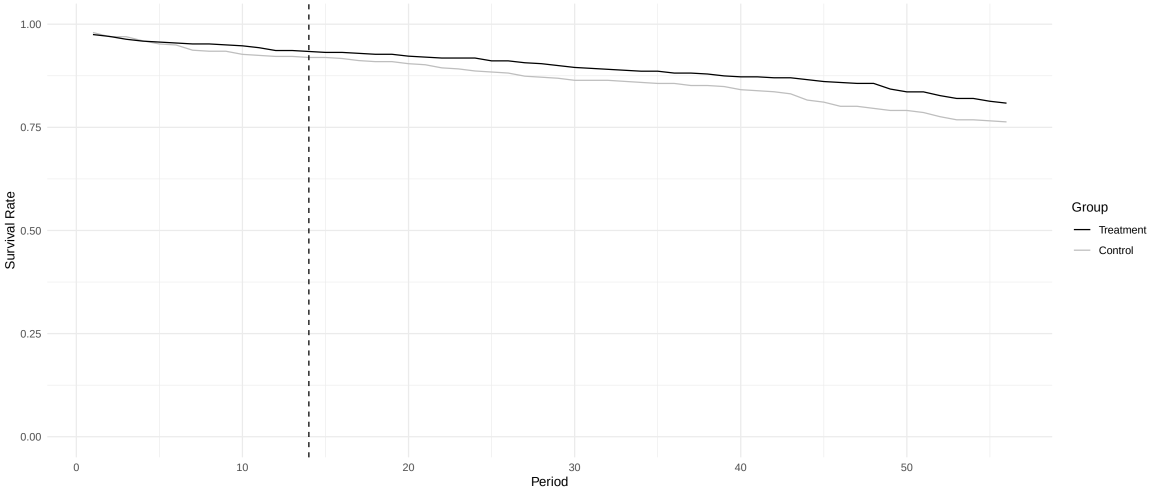


Figure 14: Proportion of Users Who Certainly Remained in the Study

Note: For each day in the study (relative to when a given participant started), the figure shows the proportion of users who were active on that day or later, separately for the treatment and the control group. The proportions reported in the figure provide the most conservative estimate of the survival rate – not being active on a given day (or even for several days) is an insufficient indication of attrition.

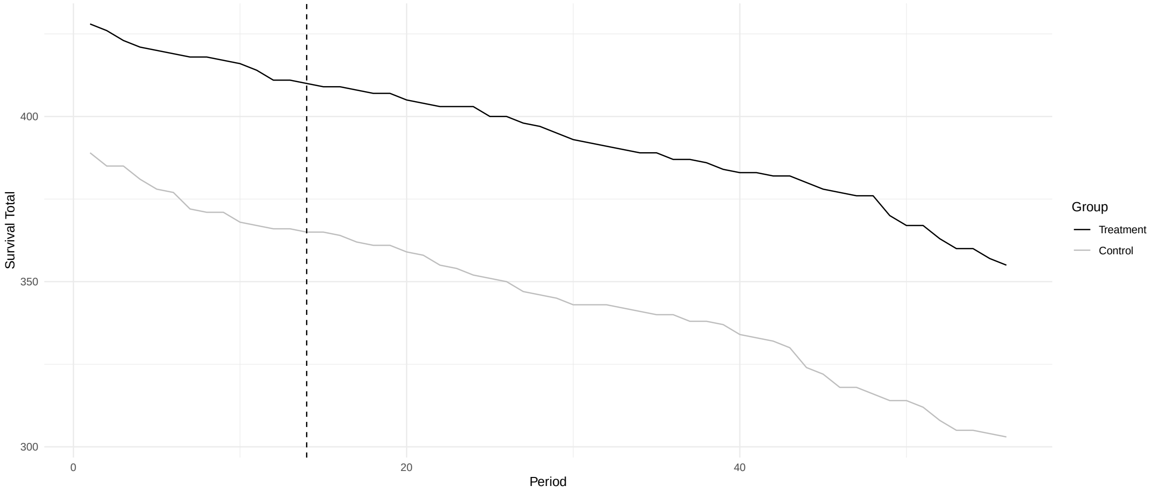


Figure 15: Number of Users Who Certainly Remained in the Study

Note: For each day in the study (relative to when a given participant started), the figure shows the number of users who were active on that day or later, separately for the treatment and the control group. The totals reported in the figure provide the most conservative estimate of number of people remaining in the study – not being active on a given day (or even for several days) is an insufficient indication of attrition.

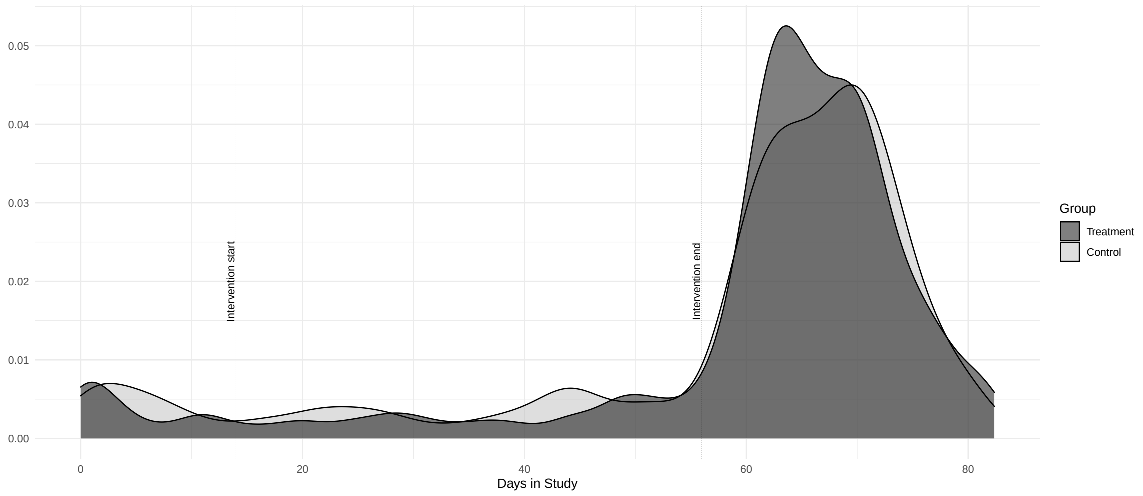


FIGURE 16: HISTOGRAM OF THE DAY OF USER'S LAST ACTIVITY

Note: The figure depicts the distribution of the last day on which the user was active according to the extension (relative to when they started), separately for the treatment and the control group. The dashed vertical line ("Intervention start") indicates day 15 – the first day of the intervention period. The distribution was plotted after the last person completed the intervention period.

B Tables

Table 7: Balance Table

		Control (N=365)		Treatment (N=410)			
		Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	p
Qualtrics		0.701	0.458	0.668	0.471	-0.033	0.323
Twitter API		0.868	0.338	0.824	0.381	-0.044	0.088
Days in Study		62.561	13.549	63.662	12.023	1.101	0.234
Use Facebook		58.621	26.873	55.356	28.852	-3.265	0.238
Use Twitter		61.337	25.478	61.159	25.608	-0.178	0.936
Age		42.036	16.884	39.684	15.816	-2.352	0.103
Male		0.510	0.501	0.539	0.499	0.029	0.507
Democrat		0.566	0.497	0.513	0.501	-0.054	0.217
Independent		0.355	0.480	0.385	0.487	0.029	0.489
White		0.639	0.481	0.653	0.477	0.014	0.739
Private		0.088	0.284	0.065	0.247	-0.023	0.266
Followers		2167.483	16567.270	1231.370	10639.782	-936.113	0.393
Friends		1282.899	2691.861	1081.544	1519.039	-201.355	0.243
Listed		32.625	236.941	22.083	109.136	-10.542	0.470
Years on Twitter		7.234	4.905	6.924	4.979	-0.309	0.423
Likes		15909.685	32821.184	16607.358	41449.953	697.673	0.811
Tweets		11836.785	37334.028	8854.642	20366.846	-2982.143	0.209
		N	Pct.	N	Pct.		
Region	Midwest	55	15.1	55	13.4		
	Northeast	48	13.2	51	12.4		
	Outside the US	4	1.1	5	1.2		
	South	82	22.5	90	22.0		
	West	66	18.1	70	17.1		
	NA	110	30.1	139	33.9		

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. The bottom panel presents the distribution of users per region in both treatment arms.

Table 8: Survey Balance Table

		Control (N=175)		Treatment (N=189)			
		Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	p
Qualtrics		0.926	0.263	0.937	0.244	0.011	0.686
Twitter API		1.000	0.000	0.995	0.073	-0.005	0.319
Days in Study		63.449	12.536	65.448	9.928	1.999	0.094
Use Facebook		58.900	25.989	54.833	27.376	-4.067	0.219
Use Twitter		62.586	25.730	60.165	24.402	-2.422	0.376
Age		40.509	15.947	38.240	14.252	-2.270	0.173
Male		0.519	0.501	0.537	0.500	0.019	0.733
Democrat		0.623	0.486	0.551	0.499	-0.072	0.178
Independent		0.321	0.468	0.352	0.479	0.031	0.544
White		0.642	0.481	0.629	0.485	-0.013	0.799
Private		0.091	0.289	0.069	0.254	-0.022	0.438
Followers		1444.886	5021.706	550.085	1002.812	-894.801	0.022
Friends		1264.149	2725.106	1154.894	1447.106	-109.255	0.637
Listed		22.497	88.161	17.574	78.826	-4.923	0.576
Years on Twitter		7.640	4.729	7.602	4.739	-0.038	0.940
Likes		15113.971	30990.165	23337.293	52859.072	8223.321	0.069
Tweets		11560.640	38363.506	10136.989	22342.608	-1423.651	0.669
		N	Pct.	N	Pct.		
Region	Midwest	37	21.1	35	18.5		
	Northeast	31	17.7	32	16.9		
	Outside the US	4	2.3	4	2.1		
	South	49	28.0	59	31.2		
	West	41	23.4	45	23.8		
	NA	13	7.4	14	7.4		

Note: This table compares characteristics of users assigned to the treatment and control arms, for the sample of users who completed our endline survey. The top panel presents means, standard deviations, difference in means, and the p-value from a test of difference in means. The bottom panel presents the distribution of users per region in both treatment arms.

Table 9: Effect of Intervention on Toxicity of Content Shown

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	-0.019*** (0.001) p < 0.001	-0.049*** (0.002) p < 0.001	-0.028*** (0.002) p < 0.001	-0.063*** (0.002) p < 0.001
N	12 518	20 658	5176	11 493
Mean	0.02	0.05	0.03	0.06
SD	0.03	0.05	0.04	0.05

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the average toxicity of the content shown to users. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 10: Effect of Intervention on Offered Content, by Conversation Type

	Main Sample				Toxic Sample			
	Facebook		Twitter		Facebook		Twitter	
	Posts	Comments	Posts	Comments	Posts	Comments	Posts	Comments
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treated	-9.7*** (1.606) p < 0.001	-8.5*** (2.065) p < 0.001	3.4 (4.827) p = 0.484	-3.9 (2.944) p = 0.191	-9.8*** (1.591) p < 0.001	-11.7*** (2.922) p < 0.001	14.9* (8.492) p = 0.085	-10.7* (6.291) p = 0.096
N	31 472	30 912	38 360	37 688	14 896	14 784	19 320	19 152
Mean	36.99	40.83	136.45	68.22	25.82	36.2	174.63	98.48
SD	111.84	128.71	268.17	199.98	73.78	138.05	306.57	256.1

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependents variables are the number of posts and comments offered to users; those displayed to them plus the content mechanically hidden by the extension. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 11: Effect of Intervention on Shown Content

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	-21.3*** (2.653) p < 0.001	-14.6** (6.941) p = 0.04	-25.4*** (3.765) p < 0.001	-18.7 (13.443) p = 0.171
N	31 864	38 472	15 120	19 320
Mean	74.66	198.87	58.97	264.81
SD	212.63	410.7	185.25	489.32

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of posts and comments shown to users. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 12: Effect of Intervention on Liked Content

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	-0.3 (0.287) p = 0.246	-0.5 (0.423) p = 0.289	-1.1** (0.446) p = 0.018	1.1* (0.585) p = 0.07
N	25 984	35 168	11 424	17 920
Mean	4.42	8.6	4.1	12.84
SD	15.95	27.11	17	34.03

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of likes that users give. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 13: Effect of Intervention on Spillovers to Related Sites (Time Spent)

	Main Sample	Toxic Sample
	(1)	(2)
Treated	1.8** (0.808) p = 0.029	2.4** (0.953) p = 0.015
N	37 016	18 088
Mean	9.61	12.47
SD	41.46	48.21

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of minutes spent on 38 other platforms related to social media (listed in Appendix F). The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 14: Effect of Intervention on Users' Well-Being

	Index	Components					
	All	Happiness	Satisfaction	Depression	Anxiety	Worthwhile	Boredom
Treated	-0.035 (0.073) p = 0.634	-0.098 (0.141) p = 0.488	-0.058 (0.155) p = 0.710	0.031 (0.089) p = 0.730	0.043 (0.090) p = 0.629	-0.093 (0.090) p = 0.303	-0.065 (0.101) p = 0.523
N	388	391	388	388	388	388	388

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 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 15: Effect of Intervention on Users' Ratings of Toxic Statements

	Index	Individual Comments						
	C1-C7	C1	C2	C3	C4	C5	C6	C7
Treated	0.016 (0.023) p = 0.491	-0.045 (0.049) p = 0.354	0.043 (0.040) p = 0.284	-0.011 (0.048) p = 0.811	-0.000 (0.027) p = 0.986	0.014 (0.051) p = 0.788	0.001 (0.043) p = 0.973	0.110** (0.050) p = 0.028
N	384	384	384	384	384	384	384	384
Mean	0.591	0.643	0.807	0.682	0.924	0.448	0.227	0.406

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. 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 16: Attrition Regressions

	(1)	(2)	(3)
(Intercept)	52.786***	53.05***	52.569***
	(0.418)	(0.775)	(0.5)
	p < 0.001	p < 0.001	p < 0.001
Treatment	0.917	1.328	1.152*
	(0.574)	(1.047)	(0.691)
	p = 0.111	p = 0.205	p = 0.096
Baseline Toxicity		-6.53	
		(15.132)	
		p = 0.666	
Baseline Toxicity × Treatment		-6.286	
		(20.315)	
		p = 0.757	
Baseline PC Usage			0.003
			(0.004)
			p = 0.43
Baseline PC Usage × Treatment			-0.003
			(0.005)
			p = 0.534
N	775	767	775

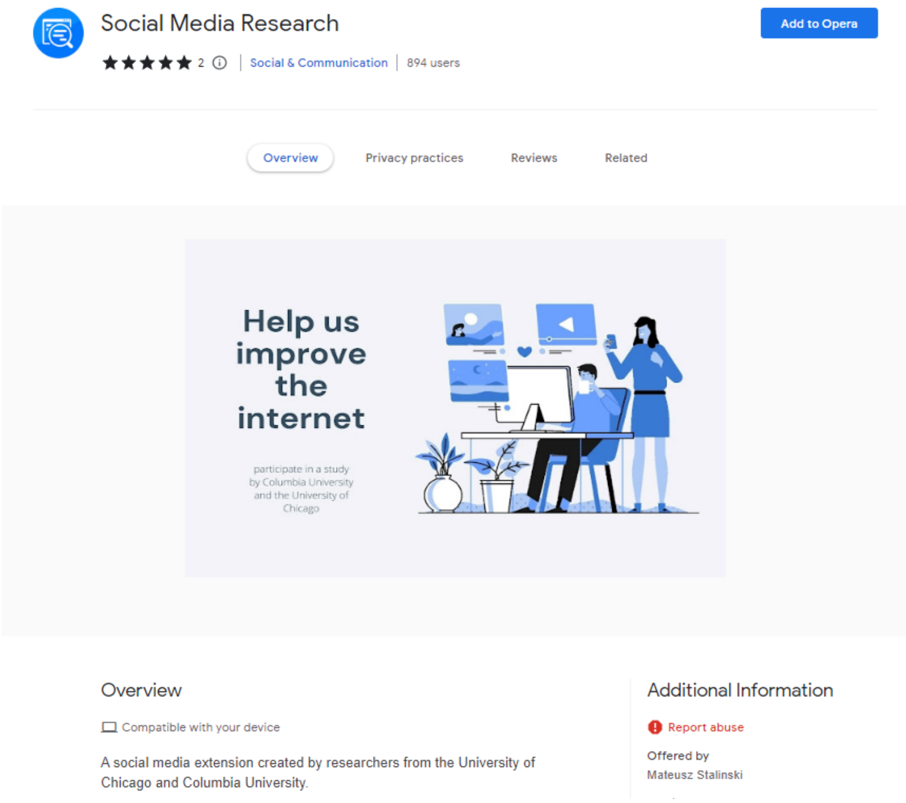
Note: This table reports estimates of an OLS regression on treatment assignment for our main experimental sample. The dependent variable is the last day that the extension registered user activity (capped at 56). Column 2 includes the average toxicity score of content displayed to the user during the baseline, and its interaction with the treatment dummy. Column 3 includes the average time spent on social media during the baseline, 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.

C Extension: Listing, Onboarding, and Privacy Policy

This appendix contains additional information about our browser extension *Social Media Research*. In particular, we outline the installation sequence, onboarding, and our privacy policy.

C.1 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 17), which outlined the core functionality, our privacy policy, and contact details of the researchers and the IRBs.


Social Media Research
Add to Opera
★★★★★ 2 ⓘ | Social & Communication | 894 users
Overview
Privacy practices
Reviews
Related
Help us improve the internet
participate in a study by Columbia University and the University of Chicago

Overview

Compatible with your device

A social media extension created by researchers from the University of Chicago and Columbia University.

Additional Information

Report abuse

Offered by
Mateusz Stalinski

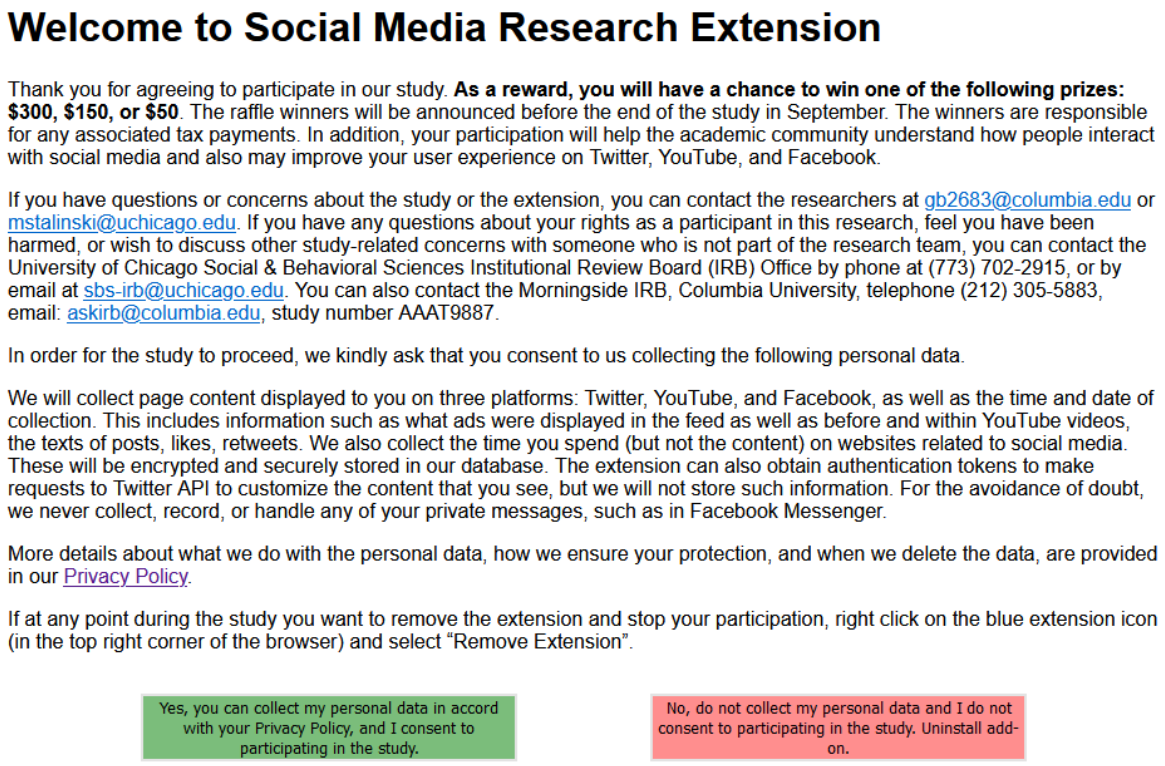
Figure 17: Our Extension's Store Listing Page (Opera Browser)

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.

C.2 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 17), 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 18).



Welcome to Social Media Research Extension

Thank you for agreeing to participate in our study. **As a reward, you will have a chance to win one of the following prizes: $300, $150, or $50**. The raffle winners will be announced before the end of the study in September. The winners are responsible for any associated tax payments. In addition, your participation will help the academic community understand how people interact with social media and also may improve your user experience on Twitter, YouTube, and Facebook.

If you have questions or concerns about the study or the extension, you can contact the researchers at gb2683@columbia.edu or mstalinski@uchicago.edu. If you have any questions about your rights as a participant in this research, feel you have been harmed, or wish to discuss other study-related concerns with someone who is not part of the research team, you can contact the University of Chicago Social & Behavioral Sciences Institutional Review Board (IRB) Office by phone at (773) 702-2915, or by email at sbs-irb@uchicago.edu. You can also contact the Morningside IRB, Columbia University, telephone (212) 305-5883, email: askirb@columbia.edu, study number AAAT9887.

In order for the study to proceed, we kindly ask that you consent to us collecting the following personal data.

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 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.

More details about what we do with the personal data, how we ensure your protection, and when we delete the data, are provided in our Privacy Policy.

If at any point during the study you want to remove the extension and stop your participation, right click on the blue extension icon (in the top right corner of the browser) and select “Remove Extension”.

Yes, you can collect my personal data in accord with your Privacy Policy, and I consent to participating in the study.

No, do not collect my personal data and I do not consent to participating in the study. Uninstall add-on.

Figure 18: Onboarding Page

The main purpose of the 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 in 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.

C.3 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.

D Survey Instruments

In this appendix, we provide the wording of all demographic questions as well as questions used to elicit survey outcomes described in the paper. We start by reporting the intake survey questions before moving to discuss the endline survey.

D.1 Intake Survey

D.1.1 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.

D.1.2 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

D.2 Endline Survey

D.2.1 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**.

D.2.2 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.				

D.2.3 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.*

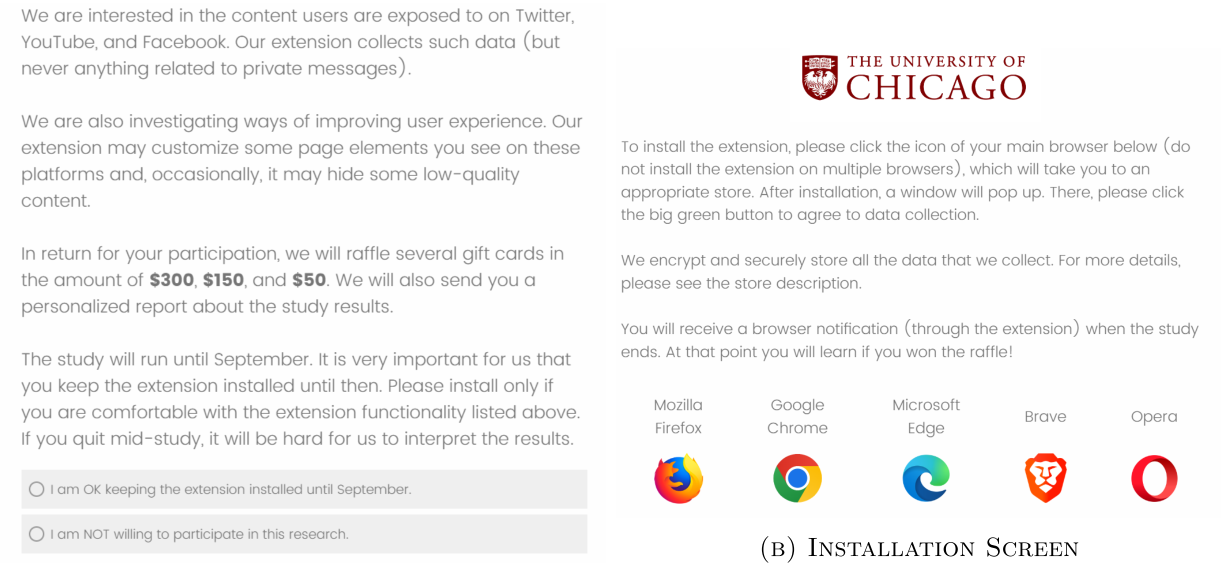
- Very Toxic
- Toxic
- Not Toxic
- I'm not sure

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

E Recruitment by the Mozilla Foundation

As indicated in Section 3.2.1, the Mozilla Foundation promoted our study by retweeting a tailored recruitment post (Figure 4). Participants recruited this way completed a simplified version of the intake survey in comparison to the standard one (see Figure 1) – 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 19a) and an installation screen (Figure 19b). 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 survey questions listed in Appendix D.1 and did not provide their Twitter handle.



We are interested in the content users are exposed to on Twitter, YouTube, and Facebook. Our extension collects such data (but never anything related to private messages).

We are also investigating ways of improving user experience. Our extension may customize some page elements you see on these platforms and, occasionally, it may hide some low-quality content.

In return for your participation, we will raffle several gift cards in the amount of **$300**, **$150**, and **$50**. We will also send you a personalized report about the study results.

The study will run until September. It is very important for us that you keep the extension installed until then. Please install only if you are comfortable with the extension functionality listed above. If you quit mid-study, it will be hard for us to interpret the results.

○ I am OK keeping the extension installed until September.

○ I am NOT willing to participate in this research.

THE UNIVERSITY OF CHICAGO

To install the extension, please click the icon of your main browser below (do not install the extension on multiple browsers), which will take you to an appropriate store. After installation, a window will pop up. There, please click the big green button to agree to data collection.

We encrypt and securely store all the data that we collect. For more details, please see the store description.

You will receive a browser notification (through the extension) when the study ends. At that point you will learn if you won the raffle!

(a) Pre-Screening

(b) Installation Screen

Figure 19: 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.

This method of enrollment was supplementary to our main recruitment efforts, and constituted a minor proportion of all extension installations – we only recorded 36 responses to the survey in which the user declared their willingness to participate (and of those not everyone necessarily installed the extension).

F Additional Platforms: Time Spent

Below, we provide the list of platforms which we used to compute the time spent by users on websites related to social media 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.

G Robustness Checks

G.1 Panel A: Day 56 Survivors

The regression analysis presented in this appendix relies on the panels of participants who were active on day 56 (the last day of the intervention) or later.

Table 17: Effect of Intervention on Toxicity of Content Shown

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	-0.019***	-0.048***	-0.028***	-0.063***
	(0.001)	(0.002)	(0.002)	(0.002)
	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$
N	12 062	19 712	5022	10 934
Mean	0.02	0.05	0.03	0.06
SD	0.03	0.05	0.04	0.05

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the average toxicity of the content shown to users. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 56 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 18: Effect of Intervention on Offered Content

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	-19.1***	-5.5	-23.7***	-6
	(2.964)	(6.442)	(4.442)	(11.741)
	$p < 0.001$	$p = 0.394$	$p < 0.001$	$p = 0.611$
N	30 464	36 400	14 392	18 200
Mean	77.8	205.96	63.16	274.03
SD	221.53	423.62	197.92	505.55

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is 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. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 56 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 19: Effect of Intervention on Offered Content, by Conversation Type

	Main Sample				Toxic Sample			
	Facebook		Twitter		Facebook		Twitter	
	Posts	Comments	Posts	Comments	Posts	Comments	Posts	Comments
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treated	-10.4***	-9.3***	1.2	-6.5**	-11.1***	-13.2***	10.1	-16.1***
	(1.69)	(2.217)	(4.493)	(2.645)	(1.756)	(3.274)	(7.318)	(5.613)
	p < 0.001	p < 0.001	p = 0.791	p = 0.017	p < 0.001	p < 0.001	p = 0.173	p = 0.006
N	30 184	29 512	36 344	35 616	14 224	14 056	18 200	18 032
Mean	37.69	41.76	138.05	67.96	26.63	37.72	175.84	97.08
SD	113.61	130.8	270.41	197.06	75.15	141.31	308.41	251.31

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependents variables are the number of posts and comments offered to users; those displayed to them plus the content mechanically hidden by the extension. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 56 of the study. Driscoll-Kraay standard errors are parenthesized. *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 20: Effect of Intervention on the Number of Twitter Ads

	Main Sample	Toxic Sample
	(1)	(2)
Treated	-1.6***	-0.4
	(0.557)	(0.87)
	p = 0.006	p = 0.661
N	19 673	10 906
Mean	19.64	20.69
SD	24.34	24.28

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of ads offered to users on Twitter; those displayed on their feeds plus the ads mechanically hidden by the extension. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 56 of the study. Driscoll-Kraay standard errors are parenthesized. *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 21: Effect of Intervention on Total Number of (Own) Posts

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	-0.8*** (0.28) p = 0.008	0.1 (0.101) p = 0.521	0.1 (0.359) p = 0.694	0.5*** (0.167) p = 0.002
N	26 600	28 840	12 152	15 400
Mean	2.54	2.4	1.95	3.24
SD	14.32	7.51	8.01	8.79

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of elements of content posted by users. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 56 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 22: Effect of Intervention on Toxicity of Produced Content

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	-0.015** (0.006) p = 0.014	-0.016*** (0.006) p = 0.006	-0.024** (0.009) p = 0.011	-0.02*** (0.007) p = 0.009
N	6415	9141	2635	5619
Mean	0.04	0.07	0.05	0.1
SD	0.09	0.15	0.1	0.17

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic 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 up to at least day 56 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 23: Effect of Intervention on Social Media Consumption Time

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	1.4 (1.31) p = 0.295	-0.3 (1.171) p = 0.801	0.8 (1.493) p = 0.576	3.7* (2.091) p = 0.084
N	31 192	36 680	14 728	18 200
Mean	23.51	30.09	22.82	42.7
SD	82.19	83.21	93.28	107.77

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of minutes spent on the platform. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 56 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 24: Effect of Intervention on Spillovers to Related Sites (Time Spent)

	Main Sample	Toxic Sample
	(1)	(2)
Treated	2.2** (0.885) p = 0.018	3*** (1.067) p = 0.007
N	35 000	17 024
Mean	9.27	11.86
SD	39.69	45.35

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of minutes spent on 38 other platforms related to social media (listed in Appendix F). The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 56 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

G.2 Panel B: Until Last Day Active

The regression analysis presented in this appendix, include, for each individual, all data up until the last day of their browser activity.

Table 25: Effect of Intervention on Toxicity of Content Shown

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	-0.019*** (0.001) p < 0.001	-0.05*** (0.002) p < 0.001	-0.027*** (0.002) p < 0.001	-0.064*** (0.002) p < 0.001
N	13 162	21 723	5514	12 187
Mean	0.02	0.05	0.03	0.06
SD	0.03	0.05	0.04	0.05

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the average toxicity of the content shown to users. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include an unbalanced panel with all users up until the last day of their browser activity. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 26: Effect of Intervention on Offered Content

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	-15.2*** (2.903) p < 0.001	1.1 (7.395) p = 0.88	-15.6*** (4.026) p < 0.001	0.6 (14.267) p = 0.967
N	33 452	40 329	15 925	20 346
Mean	76.6	209.96	62.67	281.95
SD	218.02	436.11	197.59	519.35

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is 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. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include an unbalanced panel with all users up until the last day of their browser activity. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 27: Effect of Intervention on Offered Content, by Conversation Type

	Main Sample				Toxic Sample			
	Facebook		Twitter		Facebook		Twitter	
	Posts	Comments	Posts	Comments	Posts	Comments	Posts	Comments
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treated	-8.3*** (1.804) p < 0.001	-7.5*** (2.008) p < 0.001	5.7 (5.087) p = 0.269	-4.3 (2.777) p = 0.13	-7.2*** (2.148) p = 0.002	-8.9*** (2.668) p = 0.001	12.5 (8.608) p = 0.151	-11.7* (6.217) p = 0.064
N	33 051	32 442	40 192	39 482	15 709	15 531	20 346	20 179
Mean	37.24	41.05	140.8	69.41	27.15	36.79	180.32	100.32
SD	112.17	128.01	281.77	201.57	79.71	137.41	317.1	257.82

Note: This table reports estimates from Equation (1) with start date × individual and start date × period fixed effects for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependents variables are the number of posts and comments offered to users; those displayed to them plus the content mechanically hidden by the extension. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include an unbalanced panel with all users up until the last day of their browser activity. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 28: Effect of Intervention on the Number of Twitter Ads

	Main Sample	Toxic Sample
	(1)	(2)
Treated	-1.4*** (0.5) p = 0.009	-0.1 (0.856) p = 0.898
N	21 693	12 158
Mean	19.9	21.15
SD	24.53	24.5

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of ads offered to users on Twitter; those displayed on their feeds plus the ads mechanically hidden by the extension. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include an unbalanced panel with all users up until the last day of their browser activity. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 29: Effect of Intervention on Total Number of (Own) Posts

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	-0.7**	0.1	0.3	0.5**
	(0.27)	(0.104)	(0.341)	(0.194)
	p = 0.013	p = 0.511	p = 0.392	p = 0.013
N	29 049	31 583	13 273	17 040
Mean	2.53	2.44	2.01	3.32
SD	14.22	7.58	9.21	8.93

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of elements of content posted by users. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include an unbalanced panel with all users up until the last day of their browser activity. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 30: Effect of Intervention on Toxicity of Produced Content

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	-0.017***	-0.016***	-0.032***	-0.02***
	(0.006)	(0.006)	(0.01)	(0.008)
	p = 0.004	p = 0.008	p = 0.003	p = 0.009
N	6998	10 078	2894	6285
Mean	0.04	0.08	0.05	0.1
SD	0.09	0.15	0.11	0.16

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic 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 an unbalanced panel with all users up until the last day of their browser activity. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 31: Effect of Intervention on Social Media Consumption Time

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	1.1	-0.7	1.6	3.2
	(1.245)	(1.243)	(1.561)	(1.971)
	p = 0.379	p = 0.587	p = 0.305	p = 0.109
N	34 419	40 609	16 347	20 346
Mean	22.8	31	21.96	43.28
SD	79.77	85.16	89.7	106.11

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of minutes spent on the platform. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include an unbalanced panel with all users up until the last day of their browser activity. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 32: Effect of Intervention on Spillovers to Related Sites (Time Spent)

	Main Sample	Toxic Sample
	(1)	(2)
Treated	1.5*	3***
	(0.827)	(0.939)
	p = 0.074	p = 0.002
N	38 721	19 059
Mean	9.83	12.63
SD	41.73	48.13

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of minutes spent on 38 other platforms related to social media (listed in Appendix F). The unit of observation is the individual-day, where day is measured relative to the intervention date. We include an unbalanced panel with all users up until the last day of their browser activity. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

G.3 Stacked Regression Specification

In this appendix, we replicate the main regression tables reported in the paper using a stacked regression specification, which extends the two-way fixed effects specification by including start date × individual and start date × period fixed effects.

Table 33: Effect of Intervention on Toxicity of Content Shown

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	-0.019*** (0.001) $p < 0.001$	-0.048*** (0.002) $p < 0.001$	-0.024*** (0.002) $p < 0.001$	-0.064*** (0.002) $p < 0.001$
N	12 518	20 658	5176	11 493
Mean	0.02	0.05	0.03	0.06
SD	0.03	0.05	0.04	0.05

Note: This table reports estimates from Equation (1) with start date × individual and start date × period fixed effects for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the average toxicity of the content shown to users. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 34: Effect of Intervention on Offered Content

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	-16.5*** (2.624) $p < 0.001$	-4.1 (7.25) $p = 0.57$	-21.6*** (4.815) $p < 0.001$	-14.7 (12.786) $p = 0.257$
N	31 864	38 472	15 120	19 320
Mean	76.15	204.53	60.84	274.32
SD	218.08	424.28	193.66	508.85

Note: This table reports estimates from Equation (1) with start date × individual and start date × period fixed effects for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is 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. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 35: Effect of Intervention on Offered Content, by Conversation Type

	Main Sample				Toxic Sample			
	Facebook		Twitter		Facebook		Twitter	
	Posts	Comments	Posts	Comments	Posts	Comments	Posts	Comments
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treated	-8.8*** (1.577) p < 0.001	-7.9*** (1.967) p < 0.001	1.9 (4.861) p = 0.703	-5.7* (3.101) p = 0.07	-10.7*** (1.949) p < 0.001	-11.2*** (3.473) p = 0.002	9.1 (7.405) p = 0.226	-23.3*** (6.7) p = 0.001
N	31 472	30 912	38 360	37 688	14 896	14 784	19 320	19 152
Mean	36.99	40.83	136.45	68.22	25.82	36.2	174.63	98.48
SD	111.84	128.71	268.17	199.98	73.78	138.05	306.57	256.1

Note: This table reports estimates from Equation (1) with start date × individual and start date × period fixed effects for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependents variables are the number of posts and comments offered to users; those displayed to them plus the content mechanically hidden by the extension. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 36: Effect of Intervention on the Number of Twitter Ads

	Main Sample	Toxic Sample
	(1)	(2)
Treated	-1.7*** (0.549) p = 0.003	0.3 (0.962) p = 0.755
N	20 617	11 464
Mean	19.54	20.76
SD	24.17	24.18

Note: This table reports estimates from Equation (1) with start date × individual and start date × period fixed effects for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of ads offered to users on Twitter; those displayed on their feeds plus the ads mechanically hidden by the extension. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 37: Effect of Intervention on Total Number of (Own) Posts

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	-0.8***	-0.1	0.4	0.4
	(0.274)	(0.096)	(0.311)	(0.221)
	p = 0.005	p = 0.473	p = 0.244	p = 0.117
N	27 720	30 352	12 600	16 296
Mean	2.51	2.42	1.93	3.28
SD	14.16	7.56	7.92	8.89

Note: This table reports estimates from Equation (1) with start date × individual and start date × period fixed effects for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of elements of content posted by users. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 38: Effect of Intervention on Toxicity of Produced Content

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	-0.014**	-0.019***	-0.031***	-0.023**
	(0.005)	(0.007)	(0.009)	(0.01)
	p = 0.012	p = 0.01	p = 0.001	p = 0.022
N	6658	9621	2737	5968
Mean	0.04	0.08	0.05	0.1
SD	0.09	0.15	0.1	0.17

Note: This table reports estimates from Equation (1) with start date × individual and start date × period fixed effects for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic 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 up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 39: Effect of Intervention on Social Media Consumption Time

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	1.9 (1.534) p = 0.211	-0.3 (1.398) p = 0.817	1.5 (1.87) p = 0.432	3.1 (2.162) p = 0.161
N	32 760	38 752	15 512	19 320
Mean	22.88	30.58	22.08	42.76
SD	80.69	85.35	91.44	106.54

Note: This table reports estimates from Equation (1) with start date × individual and start date × period fixed effects for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of minutes spent on the platform. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 40: Effect of Intervention on Spillovers to Related Sites (Time Spent)

	Main Sample	Toxic Sample
	(1)	(2)
Treated	2.2*** (0.808) p = 0.008	3*** (1.018) p = 0.004
N	37 016	18 088
Mean	9.61	12.47
SD	41.46	48.21

Note: This table reports estimates from Equation (1) with start date × individual and start date × period fixed effects for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of minutes spent on 38 other platforms related to social media (listed in Appendix F). The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

G.4 Clustered Standard Errors

In this appendix, we replicate the main regression tables reported in the paper with standard errors clustered at an individual level rather than with Driscoll and Kraay standard errors.

TABLE 41: EFFECT OF INTERVENTION ON TOXICITY OF CONTENT SHOWN

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	-0.019*** (0.002) $p < 0.001$	-0.049*** (0.002) $p < 0.001$	-0.028*** (0.003) $p < 0.001$	-0.063*** (0.003) $p < 0.001$
N	12 518	20 658	5176	11 493
Mean	0.02	0.05	0.03	0.06
SD	0.03	0.05	0.04	0.05

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the average toxicity of the content shown to users. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Individually-clustered standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

TABLE 42: EFFECT OF INTERVENTION ON OFFERED CONTENT

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	-17.6** (8.015) $p = 0.029$	-0.8 (16.448) $p = 0.961$	-21*** (7.66) $p = 0.007$	4.1 (29.818) $p = 0.891$
N	31 864	38 472	15 120	19 320
Mean	76.15	204.53	60.84	274.32
SD	218.08	424.28	193.66	508.85

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is 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. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Individually-clustered standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 43: Effect of Intervention on Offered Content, by Conversation Type

	Main Sample				Toxic Sample			
	Facebook		Twitter		Facebook		Twitter	
	Posts	Comments	Posts	Comments	Posts	Comments	Posts	Comments
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treated	-9.7*	-8.5**	3.4	-3.9	-9.8**	-11.7***	14.9	-10.7
	(5.016)	(3.7)	(9.156)	(9.526)	(4.198)	(4.49)	(15.102)	(18.225)
	p = 0.055	p = 0.023	p = 0.711	p = 0.683	p = 0.02	p = 0.01	p = 0.324	p = 0.559
N	31 472	30 912	38 360	37 688	14 896	14 784	19 320	19 152
Mean	36.99	40.83	136.45	68.22	25.82	36.2	174.63	98.48
SD	111.84	128.71	268.17	199.98	73.78	138.05	306.57	256.1

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependents variables are the number of posts and comments offered to users; those displayed to them plus the content mechanically hidden by the extension. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Individually-clustered standard errors are parenthesized. *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 44: Effect of Intervention on the Number of Twitter Ads

	Main Sample	Toxic Sample
	(1)	(2)
Treated	-1.8	-0.3
	(1.105)	(1.611)
	p = 0.112	p = 0.829
N	20 617	11 464
Mean	19.54	20.76
SD	24.17	24.18

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of ads offered to users on Twitter; those displayed on their feeds plus the ads mechanically hidden by the extension. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Individually-clustered standard errors are parenthesized. *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 45: Effect of Intervention on Total Number of (Own) Posts

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	-0.7 (0.638) p = 0.264	0.1 (0.34) p = 0.823	0.2 (0.362) p = 0.618	0.6 (0.599) p = 0.32
N	27 720	30 352	12 600	16 296
Mean	2.51	2.42	1.93	3.28
SD	14.16	7.56	7.92	8.89

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of elements of content posted by users. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Individually-clustered standard errors are parenthesized. ∗,∗∗ , and ∗∗∗ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 46: Effect of Intervention on Toxicity of Produced Content

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	-0.014** (0.006) p = 0.014	-0.016** (0.007) p = 0.014	-0.023* (0.012) p = 0.056	-0.02** (0.01) p = 0.043
N	6658	9621	2737	5968
Mean	0.04	0.08	0.05	0.1
SD	0.09	0.15	0.1	0.17

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic 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 up to at least day 46 of the study. Individually-clustered standard errors are parenthesized. ∗,∗∗ , and ∗∗∗ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 47: Effect of Intervention on Social Media Consumption Time

	Main Sample		Toxic Sample	
	Facebook	Twitter	Facebook	Twitter
	(1)	(2)	(3)	(4)
Treated	1.5 (2.268) p = 0.509	-0.4 (2.884) p = 0.897	1.3 (2.776) p = 0.629	3.9 (5.322) p = 0.468
N	32 760	38 752	15 512	19 320
Mean	22.88	30.58	22.08	42.76
SD	80.69	85.35	91.44	106.54

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of minutes spent on the platform. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Individually-clustered standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 48: Effect of Intervention on Spillovers to Related Sites (Time Spent)

	Main Sample	Toxic Sample
	(1)	(2)
Treated	1.8 (1.288) p = 0.16	2.4 (2.057) p = 0.247
N	37 016	18 088
Mean	9.61	12.47
SD	41.46	48.21

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose feeds and comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of minutes spent on 38 other platforms related to social media (listed in Appendix F). The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Individually-clustered standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

H YouTube Results

The intervention hid 6.19% of content displayed to users on YouTube in the treatment group. This includes comments about videos as well as replies to comments. Table 49 indicates a successful first stage, which reduced toxicity scores of elements that users are exposed to by 3.4 pp. (p-value < 0.001), a reduction of 65.1% in comparison to the control group.

Table 50 indicates that intervention had no impact on content consumption on YouTube, measured using content offered. Furthermore, Table 51 shows that it also did not alter consumption of content shown, a measure which could be affected by the mechanical effect of hiding. Taken together, we find no evidence that content consumption was affected by the intervention in any form. On the other hand, Table 52 shows that higher exposure to toxicity led to higher content production on YouTube, both in the main sample (0.2 per day, p-value $= 0.009$) and in the toxic sample (0.3 per day, p-value=0.005). Moreover, Table 53 shows an inconclusive result on the time spent on the platform, with an insignificant effect in the main sample (p-value $= 0.449$) and a marginally significant positive effect for the toxic sample (p-value $= 0.0053$).

We do not offer clear evidence in favor of the contagion hypothesis on YouTube. The point estimates in Table 54 point in the direction that higher exposure leads to higher toxicity of own content, which is consistent with the significant findings for Facebook and Twitter, but we have insufficient power to detect the effect on YouTube alone.

Table 49: Effect of Intervention on Toxicity of Content Shown

	Main Sample	Toxic Sample
	(1)	(2)
Treated	-0.034*** (0.002) p < 0.001	-0.042*** (0.003) p < 0.001
N	11 118	6643
Mean	0.04	0.04
SD	0.04	0.04

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the average toxicity of the content shown to users 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 up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 50: Effect of Intervention on Offered Content

	Main Sample	Toxic Sample
	(1)	(2)
Treated	0.2 (1.805) p = 0.925	2.3 (2.711) p = 0.406
N	35 392	18 032
Mean	43.21	54.49
SD	133.64	147.28

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of comments offered to users on YouTube; those displayed plus the comments mechanically hidden by the extension. The unit of observation is the individual-day, where day is measured relative to the intervention date. We include users who were active up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 51: Effect of Intervention on Shown Content

	Main Sample	Toxic Sample
	(1)	(2)
Treated	-2.5 (1.785) p = 0.167	-1.4 (2.674) p = 0.613
N	35 392	18 032
Mean	42.1	52.98
SD	130.29	143.28

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of comments shown to users 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 up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 52: Effect of Intervention on Total Number of (Own) Posts

	Main Sample	Toxic Sample
	(1)	(2)
Treated	0.2*** (0.069) p = 0.009	0.3*** (0.106) p = 0.005
N	11 984	6888
Mean	0.46	0.58
SD	2.48	2.91

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of elements of content posted by users 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 up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 53: Effect of Intervention on Social Media Consumption Time

	Main Sample	Toxic Sample
	(1)	(2)
Treated	-0.8 (1.042) p = 0.449	4* (2.008) p = 0.053
N	37 744	18 872
Mean	26.39	36.1
SD	85.76	102.29

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the number of minutes spent 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 up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table 54: Effect of Intervention on Toxicity of Produced Content

	Main Sample	Toxic Sample
	(1)	(2)
Treated	-0.035 (0.033) p = 0.289	-0.062 (0.041) p = 0.14
N	1458	951
Mean	0.09	0.1
SD	0.21	0.22

Note: This table reports estimates from Equation (1) for our main experimental sample (Main Sample) and the subsample of users whose comment sections had an average toxicity above median during the baseline period (Toxic Sample). The dependent variable is the average toxicity of the published content on YouTube, 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 up to at least day 46 of the study. Driscoll-Kraay standard errors are parenthesized. $*$,$**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.