The Economics of Content Moderation: Theory and Experimental Evidence from Hate Speech on Twitter

Rafael Jiménez-Durán*

September 18, 2022

[Latest version]

Abstract

Social media platforms ban users and remove posts to moderate their content. This "speech policing" remains controversial because little is known about its consequences and the costs and benefits for different individuals. I conduct two field experiments on Twitter to examine the effect of moderating hate speech on user behavior and welfare. Randomly reporting posts for violating the rules against hateful conduct increases the likelihood that Twitter removes them. Reporting does not affect the activity on the platform of the posts' authors or their likelihood of reposting hate, but it does increase the activity of those attacked by the posts. These results are consistent with a model in which content moderation is a quality decision for platforms that increases user engagement and hence advertising revenue. The second experiment shows that changing users' perceived content removal does not change their willingness to pause using social media, a measure of consumer surplus. My results imply that content moderation does not necessarily moderate users, but it marginally increases advertising revenue. It can be consistent with both profit- and welfare-maximization if out-of-platform externalities are small.

JEL codes: C93, D12, D85, D90, I31, J15, L82, L86, Z13

Keywords: social media, moderation, report, hate speech, experiment, welfare

*University of Chicago, Kenneth C. Griffin Department of Economics and Booth School of Business. Email: rafaeljjd@uchicago.edu. I'm deeply grateful to Leonardo Bursztyn, Pietro Tebaldi, Ali Hortaçsu, and Kevin Murphy for their guidance and support. I also thank Michael Dinerstein, Michael Greenstone, Justin Holz, Diego Jiménez, John List, Casey Mulligan, David Novgorodsky, Javier Pérez, Marta Prato, and many more for helpful conversations. The Political Economics Initiative from the Becker Friedman Institute at the University of Chicago funded this research. The University of Chicago IRB approved the experiments (IRB18-1586, IRB21-1067), and I pre-registered them at the AEA RCT Registry (0005131, 0008189). Any errors are mine.

1 Introduction

Social media is the "modern public square," according to the U.S. Supreme Court[1]—a place where speech happens among individuals with different backgrounds and ideologies. Yet, the biggest strengths of platforms—their size and diversity—also represent their greatest challenges. Forty percent of people have experienced online harassment (Anti-Defamation League, 2021), and studies document real-world consequences of online speech on hate crimes (Bursztyn et al., 2019; Müller and Schwarz, 2020a), and election outcomes (Fujiwara et al., 2021). Despite these consequences, few governments have crafted laws or regulation of online content (Carlson, 2021).

As a result, social media companies self-regulate and issue community guidelines that forbid not only illegal content but also some combination of hate speech, misinformation, harassment, spam, sexual content, and graphical content (Gillespie, 2018). Platforms moderate content by enforcing these guidelines with sanctions such as removing posts or accounts. Still, even if it is widespread, "policing speech" remains controversial (Kaye, 2019), in part due to scarce data and studies about this practice. The debate oscillates between arguments about freedom of expression (Strossen, 2018) and the harms that can be caused by online content (Waldron, 2012).

This paper contributes to the discussion by providing theory and experimental evidence of how moderation works, how it changes online behavior, and how to weigh its welfare gains and losses to different users. Guided by a model, I run two large-scale field experiments to document the consequences of content moderation on user behavior and welfare. The focus is on hate speech on Twitter as a prominent example. One quarter of U.S. adults use this platform, and hate speech is its most sanctioned violation (Pew Research Center, 2021; Twitter, 2020b).

I begin by modeling a platform on which users spend time and interact. The platform maximizes profits by choosing its prices—the frequency at which it displays ads—and its content removal rate, which reduces spillovers between users. As in Weyl (2010), the pricing policy is what allows the platform to effectively choose the amount of time that users spend on it. The moderation policy is a quality decision that maximizes the users' willingness to engage with ads. When setting its moderation rate, the platform trades off the change in censored and non-censored users' engagement with ads. Because moderation is costly, it is only profitable if it

[1]Packingham v. North Carolina, 137 S. Ct. 1730, 1737 (2017).

increases the activity of at least some users. In other words, it makes sense for profit-maximizing platforms to restrict the content of some of their users if this increases the overall engagement with ads. Thus, a key parameter of the platform's decisions is how users change their time on the platform in response to moderation.

The first experiment provides information about this parameter by leveraging the reporting tool of the platform that allows flagging content that violates the rules. Twitter combines these reports with algorithms to detect violations and enforce its guidelines. It then chooses from a wide range of sanctions at the Tweet or user level, such as reducing Tweet or user visibility (also called shadowbanning), temporarily locking accounts, removing Tweets, or suspending (banning) users. Because reports increase the likelihood that Twitter detects content and, plausibly, do not affect users directly, they instrument moderation as long as sanctions are perfectly observed. Thus, reporting overcomes the challenge of moderation not being randomly assigned.

Over the course of two months, I sampled 6,000 Tweets containing slurs about disability, which constitute 98% of the sample, or that deny the Holocaust. These slurs are covered by Twitter's hateful-conduct policy.[2] The sample included different spellings of the slurs to capture attempts to evade detection algorithms, excluded bots, inactive accounts, and other quality filters. Users enter the sample once, so there is one Tweet per user.

The day after they were posted, I randomly reported half of the Tweets for violating the rules against hateful conduct. I then collected daily server-level data of users' sanctions, their behavior and their followers' behavior, and the behavior of the users that the Tweets replied to, if any. The data comes from Twitter's Application Programming Interface (API) and other sources such as Google's Perspective API (Wulczyn et al., 2017), Botometer's API (Yang et al., 2020), and shadowban.eu's API (Merrer et al., 2020).

The first set of results show that reporting has a first-stage impact on sanctions. Reported Tweets are 66% (1.4 percentage points or 0.08 standard deviations) more likely to be removed within three weeks by Twitter than non-reported Tweets, with an F-statistic of 11. The treatment does not significantly change user suspensions and

[2]The policy mentions the Holocaust and slurs that reinforce negative stereotypes about a protected category, including disability. These slurs are only a subset of hate speech, but most other slurs are appropriated by minorities (Bianchi, 2014) and led to high false positives in a pilot study. Another option was to sample Tweets with a detection algorithm from the computer science literature, but even state-of-the-art methods suffer from low external validity (Arango et al., 2019).

shadowbans, the other observable sanctions.[3] However, I also find evidence of "unobservable" sanctions, such as temporarily locking users' accounts, which I obtained from the updates that Twitter sent me after I reported users.[4] Hence, reports remain a valid instrument for all sanctions even if they violate the exclusion restriction for observed sanctions.

The second set of results concern the reduced-form impact of reports on user behavior on the platform. This estimation is possible because accounts do not disappear after reporting. I find that reports do not reduce the users' Twitter activity or their likelihood of reposting hate. A proxy of the hours spent on Twitter, constructed with the daily number of Tweets and likes, increases by 7.5% (0.042 standard deviations) in the three weeks after reports, but it is not statistically significant.[5] The fraction of hateful Tweets that users post in the three weeks after the treatment, measured using Google's toxicity score, decreases by an insignificant 1.8%.

The third set of results show that reporting has significant spillovers on other users. The main measure of spillover is the activity of the users to whom the Tweets in my sample are replying, which I call "replied users"—86% of Tweets reply or quote a post from another user. Reports increase the time the replied users spend on the platform over the course of three weeks by 10%, or 10 minutes per week. Furthermore, the estimate is stronger among Tweets that attack the other user, rather than, for example, those that are just replies among friends. The effect is 13.4% among those Tweets that were labeled as attacks by human annotators.

Results are robust to alternative measures of user activity and hatefulness, dropping outliers, and specifications with different sets of controls. Together, these findings imply that sanctions induced by reporting do not change the behavior of those who posted the Tweets; content moderation does not seem to moderate users. Reports, however, increase the activity of those attacked by the hateful posts. Hence, the evidence supports the model's prediction that content moderation in a profit-maximizing

[3]There is no evidence of users self-censoring (deleting their posts or accounts, or locking their accounts from public view) in response to reports. There is also no evidence that reports induce their other Tweets to disappear.

[4]No observable sanctions were implemented in 6.6% of the accounts I reported, but Twitter provided an update that it had found rule violations. This number is likely biased downward because Twitter does not always send updates, even for reports for which a sanction is observed.

[5]Moreover, the impact on time spent might be biased downward, because Twitter restricts some accounts temporarily (Twitter, 2021d). In these cases, the number of Tweets and likes will be mechanically lower, even if users do not change their behavior.

platform marginally increases the advertising revenue from some users.

Does this evidence mean that moderation increases welfare? Not necessarily. Following Spence (1975), another result from my model is that a platform can, in principle, remove too much or too little content relative to a surplus-maximizing planner. Intuitively, the monopolist caters to the marginal consumer, whereas the planner caters to the average consumer. From the consumers' point of view, a utilitarian test of whether the platform underprovides or overprovides moderation is whether a small change in censorship, all else equal, increases or decreases consumer surplus.[6] Even if this test ignores externalities, many costs and benefits associated with moderation, such as free speech and direct harms from hateful expressions, occur inside platforms.

I conducted the test in a survey of 3,000 U.S. Twitter users sampled through Luc.id, a widely used online survey panel provider.[7] I shift users' beliefs about the likelihood that Twitter moderates hate speech, and I elicit their willingness to accept (WTA) to stop using social media. I vary the perceived likelihood of moderation using an information-provision design with an active control group (Haaland et al., 2020). I randomize survey participants into two treatment arms that receive different information about the likelihood of moderation among hateful Tweets.

The information provided comes from a random sample of 10,000 Tweets that I collected in August 2020 and classified as hate speech with the help of human annotators. I vary the likelihood of moderation without deception by using different rules to classify hate speech. Under a majority decision rule, in which a post is hateful if most annotators label it as such, Twitter removes 3.6% of hateful Tweets or suspends their authors within one month. Under a consensus decision rule, in which a post is hateful if all annotators label it as such, the likelihood of moderation is 9.1%. Under both rules, the prevalence of hate—that is, the fraction of hateful Tweets—is less than 1%. Both treatment arms receive the same information about hate prevalence, which allows isolating the effect of moderation.

The survey first elicits beliefs about the prevalence of hate speech and the likelihood of moderation with incentives for accuracy and then provides participants with the randomized information. Respondents are told that some of them will be randomly selected for a small follow-up study that compensates participants to stop

[6] This test can be generalized to a model of multiple platforms by measuring the change in users' social-media valuation, not just their valuation of a given platform.

[7] I reweight observations to match a representative sample of Twitter users based on gender, age, race or ethnicity, region, and political orientation. I also present unweighted results.

using social media (Twitter, Facebook, Instagram, Snapchat, YouTube, Reddit, and TikTok) for one week. I then elicit the WTA to participate in this follow-up, using an incentive-compatible procedure in the form of an iterative multiple price list (iMPL).[8]

I find large misperceptions about hate speech and moderation. Most users overestimate the prevalence of hate speech on Twitter and the likelihood of sanctions. Ninety-six percent of users believe the prevalence of hate speech is above the observed value of less than 1%, with a median of 33%. Eighty-four percent of respondents believe the likelihood of moderation is above 9.1%, with a median of 36%.[9]

Informing participants of a higher likelihood of moderation does not change their valuation of social media. The WTA falls by 15 cents (0.5% or 0.004 standard deviations), from $33.7 to $33.6. This result is robust to different specifications and measures of WTA, and I find no evidence that it is explained by inattention or experimenter demand effects.[10] At the end of the survey, I asked participants to repeat the information about the percent of Tweets that get sanctioned. The treatment effect on this recollection was 5.6 percentage points, significantly different from zero (F-statistic = 36) and not statistically different from 5.5, which is the gap between the information provided in both arms, 9.1% and 3.6%. The treatment also significantly shifted the posterior beliefs about the likelihood of moderation on Facebook and there is suggestive evidence that it increased the time that minorities spent on Twitter one week after the survey.

Overall, my results suggest that moderation on Twitter is consistent with profit maximization, and they rule out large moderation distortions from the consumers' point of view, holding constant the prevalence of hate speech. These findings have two policy implications. First, cost-benefit analyses of online moderation can emphasize its offline consequences, such as hate crimes. Second, authorities might want to deal with hate speech on social media not by directly regulating moderation, but by supervising platforms' pricing (advertising) policies; Twitter could still be setting its

[8] In this procedure, participants have to choose if they are willing to participate in the follow-up for different compensation offers. The sequence of offers starts at $50, and subsequent amounts decrease or increase depending on whether participants accept or reject. The sequence stops until the WTA is classified in 11 intervals, which I then convert into a continuous measure following Allcott and Kessler (2019).

[9] Platforms' lack of transparency could be driving these misperceptions. The likelihood of moderation on Facebook remained unknown until a whistleblower revealed internal documents some weeks after my survey (Giansiracusa, 2021).

[10] The experiment was ex-ante powered to detect effects of 0.1 standard deviations, and the sample size is more than double what Haaland et al. (2020) recommend for information-provision designs.

advertising loads suboptimally, leading to inefficient amounts of hate speech.

The paper makes four contributions to a multi-disciplinary literature. First, it provides evidence of the online effects of moderation. A growing body of work in economics focuses on the offline consequences of online content and social-media penetration (Enikolopov et al., 2020; Müller and Schwarz, 2020a,b; Bursztyn et al., 2019; Braghieri et al., 2021). Other work studies government online censorship (Hobbs and Roberts, 2018; Roberts and Roberts, 2018; Chen and Yang, 2019). The computer science literature documents the relationship between content moderation and online behavior (Ali et al., 2021; Rauchfleisch and Kaiser, 2021; Jhaver et al., 2021; Zannettou, 2021), but most of these exercises are non-causal. A challenge with observational studies is isolating the effect of moderation from confounders. For example, Chandrasekharan et al. (2017) find that banning groups on Reddit decreased their former members' activity on the platform, but this could happen both because they are sanctioned or because they find the platform less attractive after the group closures.

Experimentally varying moderation, however, is also challenging due to limited cooperation with platforms. Thus, a second contribution is using social media's infrastructure experimentally, as Levy (2021) did on Facebook, which is useful for independent research. The reporting treatment is similar to other exercises by academics (Carlson and Rousselle, 2020), Governmental organizations (Jourová, 2016; Reynders, 2020), and non-profits (Matias et al., 2015; Center for Countering Digital Hate, 2021) who report content to monitor platforms' responsiveness. However, these exercises are non-experimental (they contain no control group) and do not analyze the impact on other outcomes beyond the platform's response. Experimental interventions include counter-speech treatments (Munger, 2017, 2021; Siegel and Badaan, 2020), reminders of Twitter suspensions (Yildirim et al., 2021), and censoring hate speech in the lab (Álvarez-Benjumea and Winter, 2018).

A third contribution is combining an information-provision design with a welfare elicitation of social media. Haaland et al. (2020) and Bursztyn and Yang (2021) provide overviews of information-provision designs, and Bottan and Perez-Truglia (2017) and Bursztyn et al. (2020) are some applications. The WTA that I elicit is in the ballpark of other social-media welfare studies such as Mosquera et al. (2020), and Allcott et al. (2020); the median and mean WTA per week were \$15 and \$34, respectively.[11]

[11] This is after reweighting my sample to match representative U.S. Twitter users. Allcott et al. (2020) find a median and average WTA of \$25 and \$45 per week, respectively. These estimates were

Providing information required computing other basic statistics, surprisingly scarce in the literature, such as the prevalence of hate speech in a random sample of posts (0.1%-5.6% depending on the measure) and the occurrence of Tweet deletions and user suspensions (2.5%-9.1% among hateful Tweets, within one month).[12] As other surveys find (Anti-Defamation League, 2021), minorities experience more harassment online, but I also find that they are more likely to be sanctioned by Twitter.

The fourth contribution is to develop a simple model of user behavior and platform moderation decisions that captures spillovers between users, using the two-sided market framework of Weyl (2010). Prices allow the platform to determine its amount of hateful and non-hateful content, which clarifies the separation between pricing distortions and moderation distortions as in Spence (1975).[13] Liu et al. (2021) are among the first to model moderation decisions and to discuss the implications of different revenue models on platform incentives. One difference with their model is that, in my framework, users respond to the pricing policy (advertising frequency) of the platform.[14] Acemoglu et al. (2021) model online misinformation and show that engagement-maximizing platforms have incentives to create filter bubbles and propagate extremist content.

The next section develops the model. Section 3 provides background information about hate speech, moderation, Twitter, and the data sources that this study uses. Sections 4 and 5 describe the experimental design of both experiments and present their results. Section 6 concludes.

---

for deactivating Facebook over four weeks.

[12] Relia et al. (2019) find that 0.5% of Tweets in a sample of 73.42 million posts contained hate speech keywords. Founta et al. (2018) found a 4% prevalence in a random sample of 10,000 Tweets. Facebook (2021) reports a prevalence of 0.05% of hate speech among all views. Few studies report the occurrence of sanctions. An exception is Merrer et al. (2020), who document that 2.34% of accounts are shadowbanned. Seyler et al. (2021) find that 5.1% of accounts from a 2009 sample are suspended.

[13] There is evidence that consumers respond to platforms' advertising policies; Huang et al. (2018) traced out a downward-sloping demand curve for a music platform by randomizing ad-loads across consumers.

[14] In both frameworks (under an advertising business model), platforms use moderation as a tool to increase revenue. In Liu et al. (2021), moderation increases revenue through increases in the consumer base. However, in my model, prices determine the customer base and moderation increases revenue through the willingness of users to engage with ads, given the customer base.

2 Model

Users and platform. Users can be one of two types, $\theta \in \{A, H\}$. "Acceptable" users ($\theta = A$) post content that is not subject to content moderation. "Hateful" users ($\theta = H$) post content that is censored by the platform with probability c, the censorship or moderation rate. Users who join the platform experience utility that increases on the time that they spend consuming or posting content.[15]

The utility of spending t minutes on the platform for user i of type $\theta_i = \theta$ is

$$\underbrace{U_i^\theta(t; \mathbf{T}, c)}_{\text{Utility from consuming content}} - \underbrace{t \times w_i(1 + p^\theta)}_{\text{Time cost}}, \tag{1}$$

where $U_i^\theta(0; \mathbf{T}, c) = 0$ for all i. $\mathbf{T} = (T^A, T^H)$ is the aggregate content of the platform and captures spillovers and network effects, where T^θ is the total content posted by θ users. The sign of spillovers is flexible; users can be positively or negatively affected by each type of content. For instance, A users could dislike encountering hate speech, but haters might like trolling A users. The censorship rate c enters the utility function because it reduces spillovers from hateful content (since users see less of it), but users can also obtain direct utility or disutility from c, for example, haters might dislike having their account locked.

The time-cost of t minutes spent enjoying content is proportional to the value of time $w_i > 0$. Moreover, the "price" that users pay is the advertising load p^θ; the time they have to spend watching ads per minute of content consumed. Following Weyl (2010), the platform can set a different price for each type of user.

The time that user i spends enjoying content is t_i^*, which maximizes (1) with respect to $t \geq 0$. The aggregate content demand $\mathbf{T}$ is then computed setting

$$T^\theta \equiv \int_{\{i:\theta_i=\theta\}} t_i^* \mathrm{d}i, \quad \text{for each } \theta. \tag{2}$$

Since the time spent on the platform by any user is decreasing in the advertising load, one can define the inverse demand functions $P^A(\mathbf{T}, c)$ and $P^H(\mathbf{T}, c)$, where

[15] No difference exists between consuming or producing content. In practice, however, users differ substantially in the amount of content they post. On Twitter, few users post the majority of Tweets (Wojcik and Hughes, 2019). Yet, it is not obvious whether users who like posts are less responsible for their diffusion than those who write them. For instance, sometimes Twitter alerts the followers of a user when she likes a post.

$P^A(\mathbf{T}, c)$ is equal to the p^A inducing T^A given $\mathbf{T}$ and c; similarly for $P^H(\mathbf{T}, c)$.[16] In other words, the pricing policy—not moderation—is what allows the platform to choose the amount of content of both types of user.

The platform maximizes profits solving

$$\max_{\mathbf{T},c} \underbrace{a \times \left(\underbrace{P^A(\mathbf{T}, c)\, T^A + P^H(\mathbf{T}, c)\, T^H}_{\text{Time spent watching ads}} \right)}_{\text{Advertising revenue}} - \underbrace{\phi(\mathbf{T}, c)}_{\text{Cost of moderation}}, \tag{3}$$

where $a > 0$ is the price per unit of advertising,[17] and ϕ is a function describing the costs of censorship. For instance, Gillespie (2018) documents that moderation is a labor-intensive task, and Kaye (2019) argues that regulatory fines push platforms to remove borderline content.[18]

The first-order condition (FOC) with respect to quantity T^θ is similar to a standard monopoly problem. The FOC with respect to c requires that

$$a \times \left(\underbrace{\frac{\partial P^A(\mathbf{T}, c)}{\partial c} T^A + \frac{\partial P^H(\mathbf{T}, c)}{\partial c} T^H}_{\text{Change in time spent watching ads}} \right) = \underbrace{\frac{\partial \phi(\mathbf{T}, c)}{\partial c}}_{\text{Marginal cost of moderation}}. \tag{4}$$

This condition is analogous to the quality decision in Spence (1975); the platform moderates such that the marginal benefit—the value of the marginal increase in the willingness to watch ads—equals the marginal cost. The left-hand side of equation (4) clarifies the main trade-off faced by the platform when choosing its moderation policy. Consistently with the observations in Kaye (2019) regarding controversial pages,

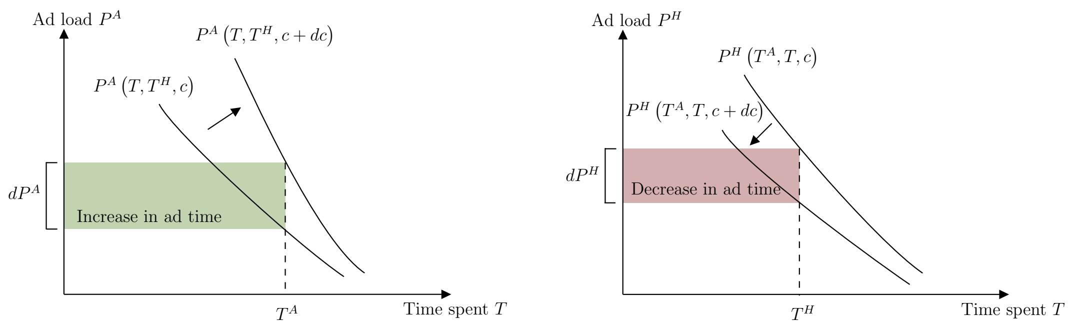
> These kinds of pages seem to put Facebook in a no-win position: If they leave up the page, they anger opponents who see hateful content or disinformation; if they take it down, they offend free-expression advocates who do not think the rules very clearly articulate hate speech standards.

[16] Formally, imposing rational expectations $\widetilde{T^\theta} = T^\theta(p^\theta, c, \widetilde{T^A}, \widetilde{T^H})$, one can invert T^θ in an interior equilibrium point, where $\widetilde{T^\theta} > 0$. This procedure requires demands T^θ to be strictly decreasing in p^θ, which results from imposing Inada conditions on utilities or full-support assumptions as in Weyl (2010).

[17] Incorporating different prices for haters and non-haters or market power on the digital advertising market is possible, but this extension adds little to the results.

[18] Platforms might worry about future fines, even if current ones are small; e.g., in 2019, Germany fined Facebook for 2 million euros for violating the NetzDG law (Bundesamt für Justiz, 2019).

Figure 1 illustrates this trade-off for the case in which $\partial T^A/\partial c > 0$, and $\partial T^H/\partial c < 0$. For a fixed amount of time that users spend on the platform, moderation changes the number of ads they are willing to watch. The platform increases revenue from A users, who dislike hateful content, while it looses revenue from H users, who do not like to be censored. The optimal level of content moderation balances the net change in revenue with the marginal cost of increasing the censorship rate.



(a) Acceptable users **(b)** Hateful users

Figure 1: Graphical intuition of the platform's moderation decision

Notes: These figures plot the change in the inverse demands of acceptable and hateful users in response to an increase in moderation, dc, holding quantities fixed, and assuming that the moderation elasticity is positive for A and negative for H. The colored areas are the change in time spent watching ads, which equals the change in ad load dP^θ times the time spent T^θ. The net revenue gains equal the green minus the red area, multiplied by ad prices.

The FOC is, equivalently,[19]

$$-\frac{\partial T^A/\partial c}{\partial T^A/\partial p^A} a T^A - \frac{\partial T^H/\partial c}{\partial T^H/\partial p^H} a T^H = \frac{\partial \phi(\mathbf{T}, c)}{\partial c}. \tag{5}$$

We know that the right-hand side is strictly positive (by assumption), and that demand decreases in prices, $\partial T^\theta/\partial p^\theta < 0$. Therefore, at the optimal level of c for the platform it must be that either $\partial T^A/\partial c > 0$, or $\partial T^H/\partial c > 0$, or both. In words, for at least one type of user, moderation must increase their platform activity, holding constant the aggregate quantities. The derivatives of T^θ with respect to c, one for each type, are the main parameters of interest of my first experiment.

[19]This applies the implicit function theorem. For example, letting $\widetilde{T^A} = T^A(p^A, c, \widetilde{T^A}, \widetilde{T^H})$, taking the total derivative implies $0 = \frac{\partial T^A}{\partial p^A} dp^A + \frac{\partial T^A}{\partial c} dc$, so that $\frac{dp^A}{dc} = -\frac{\partial T^A/\partial c}{\partial T^A/\partial p^A}$.

Welfare. The platform-optimal level of censorship could differ from the socially-optimal level. Similarly to Spence (1975), the platform in my model optimizes moderation with respect to the marginal users. The social planner, however, chooses the level of censorship that maximizes total welfare, which includes the impact of moderation on inframarginal consumers. I formalize this argument in Appendix A; the platform can moderate more or less than a surplus-maximizing social planner, holding quantities $\mathbf{T}$ fixed.[20] Hence, two distortions exist: the usual monopolist pricing distortion that leads to inefficient quantities and an additional quality distortion.

The goal of my second experiment is to test the second distortion—to evaluate whether Twitter provides too little or too much content moderation from the perspective of the user. I follow the approach of Mosquera et al. (2020) and Allcott et al. (2020) to measure consumer surplus. In practice, I quantify the impact of different levels of censorship on the willingness to accept a monetary reward to pause the use of social media. I ask users to pause the use of social media, not just a single platform, to allow for substitution between platforms as argued in Appendix A.

3 Background and Data Sources

3.1 Twitter and Moderation of Hate Speech

Twitter is a microblogging social media platform. Users of this platform create profiles that display self-reported information such as their name, a short biography, and a profile picture. They also post messages to their profiles called Tweets, which contain a combination of text of up to 280 characters, photos, and videos. Users can follow other accounts to see their Tweets more readily, but they can interact with others without following them. They interact with others' Tweets by giving them a like (or favorite), replying to them, Retweeting (reposting) them, or quoting them.

Like all social media platforms in the Surface Web, Twitter has rules that delimit the content that users are allowed to post. Besides illegal activity, the rules tend to cover hate speech (as well as misinformation, harassment, spam, sexual content, and

[20] Liu et al. (2021) argue that platforms undermoderate in an ad-based business model and overmoderate in a subscription-based business model. However, in their ad-based business model there are no prices, so the platform has to use moderation as the only tool to adjust its quantity. In their subscription-based case there are prices, but the sign of the Spence distortion is determined by their extremeness aversion assumption, which in practice implies that the demand curve becomes steeper in response to a small increase in moderation.

graphic content). Hate speech has no single legal definition of hate speech (Waldron, 2012; Strossen, 2018). Still, most platforms define it in their rules using common elements such as the concept of protected categories from U.S. anti-discrimination law (Gillespie, 2018). Twitter's hateful-conduct policy (Twitter, 2021b) says, "You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease".

Twitter enforces these rules and moderates content by sanctioning users. Figure 2 illustrates the process of content moderation. Twitter can detect content by algorithms, or by the "flagging" mechanism that allows users to report Tweets or accounts for violating the rules. After the content is detected, a team of human moderators or an algorithm decide whether to enforce the rules by imposing post-level or account-level sanctions. The range of sanctions include a combination of removing Tweets from the platform, shadowbanning (reducing the visibility of) users or Tweets,[21] and suspending or banning users (that is, deleting their accounts). Other sanctions, such as locking accounts, prevent users from posting or liking content and can last from 12 hours to seven days. See Twitter (2021d) for the full list of sanctions.

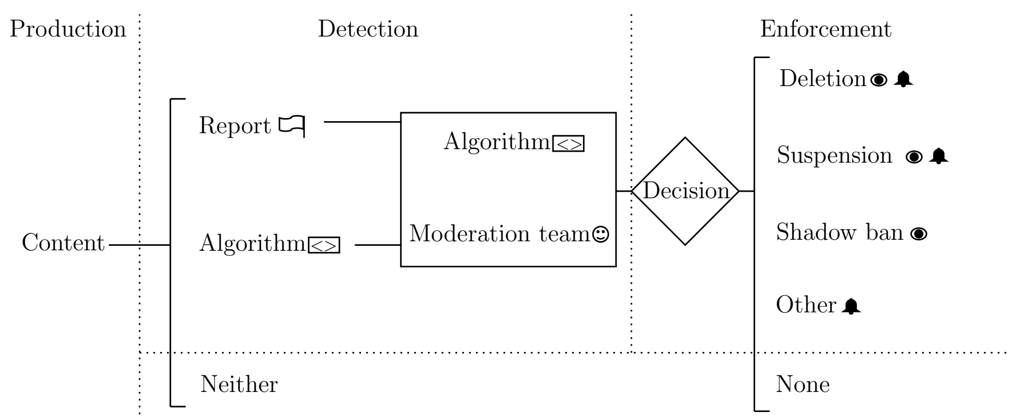


Figure 2: Content moderation process

Notes: Eye icons indicate that the sanction is observable to others besides the user. Bell icons indicate that the user receives a notification. This diagram omits interventions at the production stage, such as recent tests in which Twitter asked users if they wanted to review offensive Tweets before posting. It also omits the appealing process, in which users can contest a sanction.

[21]Twitter has stated it does not shadowban users (Gadde and Beykpour, 2018), even if it ranks Tweets "to create a more relevant experience" (Twitter, 2021a). However, Merrer et al. (2020) document evidence of shadowbanning.

Sanctions differ by their observability, that is, whether they are privately observed by the user or publicly observable, and whether the user is notified about them. When Twitter removes a Tweet, users are notified that they violated the rules, and must remove the Tweet to be able to use the platform again. Twitter replaces the Tweet with a notice that indicates it violated the rules. Anyone with access to the Tweets can see the notice, starting from the moment that Twitter sanctions the Tweet and up to 14 days after the user agrees to remove the post. When Twitter suspends a user, she can no longer log in to the account, and her profile and Tweets are replaced by suspension notices, which seem to last indefinitely. In principle, suspended users cannot create new accounts, but in practice they do. Users are not notified when they are shadowbanned, but Twitter sometimes hides their Tweets behind a notice—especially those that reply to another user. Twitter notifies users when their accounts are locked, but whether others can observe this sanction is unknown.[22] Figure B.1 shows examples of public notices and the notifications that users receive.

3.2 Measuring Hate Speech

Platforms rely to some extent on algorithms to detect hate speech and enforce their rules. Most of the detection algorithms in the computer science literature share the following procedure (see Fortuna and Nunes (2018) for a review of the literature). The first step is to obtain a training dataset, consisting of a sample of texts—usually social media posts—paired with labels, for example, hate speech or not hate speech. Often, these labels or "ground truth" result from aggregating the opinions of multiple humans or "annotators" into a single category. For example, Davidson et al. (2017) ask three or more crowd workers to annotate each Tweet as "hate," "offensive,", or "neither." Then, they aggregate these annotations into a single label with the majority decision rule, that is, the category chosen by most annotators. The second step is to convert the text into vectors of features with text analysis, reviewed in Gentzkow et al. (2019). The final step is to use machine learning to predict the labels with the features.

One challenge in the hate-speech-detection literature is the algorithms' low external validity; see Arango et al. (2019) and Fortuna et al. (2021). For this reason,

[22] Twitter (2021c) shows examples of notices of locked accounts, but anecdotal evidence suggests accounts are locked without any notice. For instance, in 2020 Twitter locked actor James Woods and his account did not show any notice (Whalen, 2020).

this study uses three approaches to classify hate speech and limit measurement error. First, for large-scale tasks, I use the Perspective toxicity score developed by Google. This score is widely used in the industry and as a benchmark in academic articles. It is a number between 0 and 1 that reflects the likelihood that a text is an attack or harassment.[23] Many studies classify posts as hate speech if their toxicity is higher than a 0.8 cutoff (ElSherief et al., 2018; Han and Tsvetkov, 2020; Vidgen et al., 2020). Second, I sample hate speech using keywords, instead of an algorithm, to minimize false positives in the reporting experiment. Third, I use human annotation by MTurk workers to account for measurement error in the first experiment and to increase the interpretability of the information treatment of the second experiment.

3.3 Data Sources

Most of the variables analyzed in this paper come from Twitter's API. This data source provides publicly available information about all Twitter users, such as their number of followers, number of accounts they follow, date of account creation, total number of Tweets and likes, and biography. The API provides additional information about users who do not restrict their profile visibility, such as their list of followers and accounts followed, and a collection of up to 3,200 of their most recent Tweets. The API returns detailed information for these Tweets, such as their timestamp, text and media, likes, and Retweets. This source also allows me to sample Tweets by searching for specific keywords or sampling at random 1% of all Tweets. Lastly, I also use this API to detect whether Twitter removes specific Tweets or suspends users, following the procedure outlined in Appendix C.1.

Besides the API, I also collect some information manually from the website. Twitter occasionally notifies users when it sanctions an account they previously reported, even if the sanction might not correspond to the reported content.[24] Figure B.3 has a screenshot of some of these updates. I collect this information for the reporting experiment, because it provides evidence of "unobservable" sanctions.

I also use other APIs. I retrieve the toxicity score of posts from Google's Perspective API. I also obtain a measure of the likelihood that users are bots from the

[23] The algorithm is a convolutional neural network trained on Wikipedia Talk Pages; see Wulczyn et al. (2017) and Dixon et al. (2018).

[24] Twitter says: "You will receive an in-product notification if an action is taken on an account that you recently reported. This action may or may not be related to your report" (Twitter, 2021e).

Botometer API (see Yang et al. (2020)). Finally, I retrieve measures of shadowbans from the API of Shadowban.eu, because Twitter does not give an official shadowban measure. This API measures different forms of shadowbanning, for example, whether Twitter hides accounts, Tweets, or replies from search results (see Merrer et al. (2020) for more details). I combine the different measures into a single indicator of whether users are shadowbanned.

Another data source is human annotation; I ask MTurk workers to annotate posts. For example, I follow the approach in Davidson et al. (2017) and ask workers to classify posts as "Hate speech," "Offensive but not hate speech," and "Neither offensive nor hate speech." I assign three workers to annotate each post. I give them Twitter's definition of hate speech for reference, offer a \$20 bonus to the five most accurate workers (measured by the inter-annotator agreement), and include attention checks to improve the quality of annotations. Figure D.1 in the Appendix includes screen shots of the instructions. Then, I aggregate workers' annotations into a single label using either the majority decision rule, in which a post is hate speech if two or three workers label it a such, or the consensus decision rule, in which all three workers have to agree.

Lastly, I obtain demographics of representative Twitter users from the American Trends Panel (ATP) of September 2020. The Pew Research Center conducts this nationally representative panel of randomly selected U.S. adults.

3.4 Summary Statistics

Accounts and Tweets. According to the ATP, 25% of adults in the U.S. use Twitter. Table 1 displays selected summary statistics of Twitter users and their accounts. Twitter users are younger, more educated, and more likely to be Democrats than the general population. Thirty-one percent of them are between 18 and 29 years old, 40% are at least college graduates, and 35% are Democrats, compared to 18%, 33%, and 30%, respectively, in the overall ATP respondents. The table also shows statistics from a sample of 200,000 Tweets that I collected in August 2020 from the 1% random sample of Twitter's API. On average, the accounts in this sample were five years old, posted 12 Tweets per day, gave 13 likes per day, followed 1,000 users, and had 4,800 followers. Ten percent of these accounts are bots; that is, they have a Botometer score of 0.5 or more.

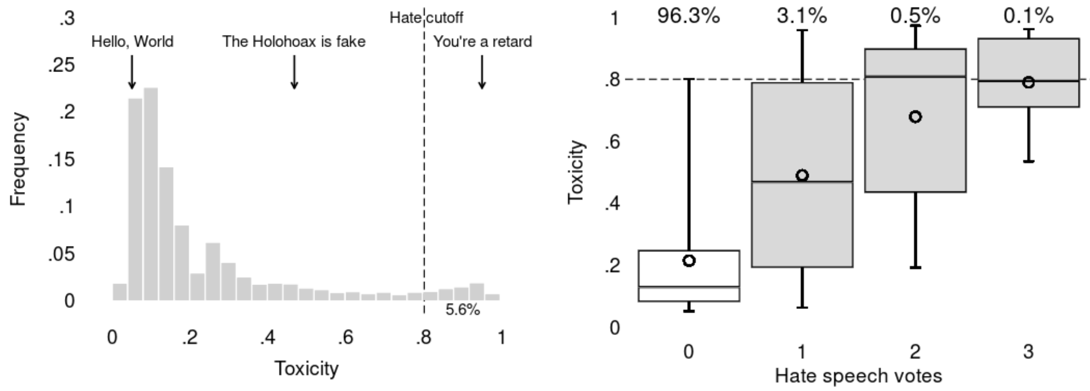
Table 1: Summary statistics

	Mean	Std. Dev	p10	Median	p90	Obs.	Sample
Accounts							
Account years	5.24	3.82	1.17	4.04	11.21	191,835	Random
Tweets per day	12.02	39.26	0.23	4.35	29.62	191,835	Random
Likes per day	13.03	24.08	0.07	3.98	36.16	191,835	Random
Followers	4,804	169,343	15	340	3,167	191,835	Random
Followed	1,071	6,755	45	381	2,078	191,835	Random
Is bot (%)	9.90	29.88	0.00	0.00	0.00	1,000	Random
Age 18-29 (%)	30.99	46.25	0.00	0.00	100.00	2,463	ATP
Male (%)	53.20	49.91	0.00	100.00	100.00	2,464	ATP
White (%)	57.92	49.38	0.00	100.00	100.00	2,464	ATP
College graduate (%)	40.47	49.09	0.00	0.00	100.00	2,464	ATP
Republican (%)	20.72	40.54	0.00	0.00	100.00	2,464	ATP
Democrat (%)	35.16	47.76	0.00	0.00	100.00	2,464	ATP
Tweets							
Is reply or quote (%)	62.53	48.40	0.00	100.00	100.00	201,038	Random
Is toxic (%)	5.61	23.01	0.00	0.00	0.00	201,038	Random
Is hate (%, majority)	0.56	7.47	0.00	0.00	0.00	9,991	MTurk
Is hate (%, consensus)	0.11	3.32	0.00	0.00	0.00	9,991	MTurk

Notes: The random sample indicates a random extraction of 201,308 Tweets from Twitter's API on August 2020. The bot score was computed on a subsample of 1,000 accounts from the random sample of Tweets, due to rate limits from the Botometer API. The ATP sample is a subsample of Twitter users from the Pew Research Center's ATP. The MTurk sample is a random subsample of Tweets that I annotated on MTurk following Davidson et al. (2017).

Prevalence of hate speech. The random sample of Tweets allows me to quantify the percent of Tweets that are hate according to different measures. Using the 0.8 toxicity cutoff, I find 5.6% of Tweets are hate. To compare this number with human annotation, I annotated a subsample of 10,000 Tweets from the random sample. As Table 1 shows, less than 1% of Tweets are considered hate speech using human annotation, under both the majority decision rule and the consensus rule. Thus, hate speech is a low-probability event.

The long-tailed nature of hate is more evident in Figure 3a, which plots a histogram of the toxicity score in the random sample of Tweets. The figure also includes the toxicity scores of three example texts: the neutral phrase "Hello, World" (toxicity = 0.05), one phrase related to disability (toxicity = 0.95), and one that denies the Holocaust (toxicity = 0.47). These examples, which are relevant for the reporting experiment, illustrate that the toxicity cutoff adequately separates some slurs from neutral expressions, but it fails to identify more subtle hate. Still, toxicity is closely correlated with human annotation. Figure 3b shows the distribution of toxicity scores shifts to the right as more workers label Tweets as hate speech.



(a) Toxicity histogram **(b)** Toxicity by hate votes

Figure 3: Toxicity scores and annotation in a random sample of Tweets

Notes: Panel (a) displays a histogram of toxicity scores based on a random sample of 201,038 Tweets from August 2020. The dashed line is the 0.8 toxicity cutoff to classify hate speech; 5.6% of Tweets have a toxicity above that cutoff. The phrases "Hello, World," "The Holohoax is fake," and "You're a retard" have toxicities of 0.05, 0.47, and 0.95, respectively. Panel (b) has toxicity box plots by the number of workers who voted that a Tweet is hateful. The data is from a subsample of 10,000 Tweets annotated by MTurk workers. The boxes indicate percentiles 25, 50, and 75; the circles indicate the means; and the lines indicate percentiles 5 and 95. The percentages at the top indicate the fraction of Tweets by number of votes.

Occurrence of sanctions and reports. Table 2 presents the fraction of removals and suspensions in the random sample of Tweets and the different subsamples of hate speech. Depending on the measure of hate, the fraction of Tweets that Twitter removed or suspended within one month is 2.6% to 9.1%—higher than the 2% in a random sample. These numbers match the statistics recently revealed in Facebook's whistleblower event, that the platform removes 3% to 5% of hateful content (Giansiracusa, 2021). From this table, we can also see that removals are a rare event. I did not measure shadowbans in this sample, but Merrer et al. (2020) document that 2.3% of accounts are shadowbanned. Figure C.1 in the Appendix plots the fraction of sanctions by the type of rule violation; hateful conduct and harassment are the most commonly sanctioned violations in the platform.

In the second half of 2020, 11% of active accounts were reported according to offi-

Table 2: Likelihood of sanctions by subsample

	Random	Hate speech		
		Toxicity ≥0.8	MTurk annotation	
			Majority	Consensus
Removal	0.01	0.1	0	0
Suspension	1.9	2.5	3.6	9.1

Notes: This table shows the fraction of Tweets or accounts that get removed from the platform within 1 month of posting hate speech by each subsample. The random sample of posts is based on 201,038 Tweets and the MTurk annotation is based on a subsample of 9,991 annotated Tweets.

cial Twitter data,[25] and 1% of accounts concentrate the majority of reports (Twitter, 2018). Recently, Twitter's CEO reported that algorithms detect 51% of the content that the platform finds in violation of the rules and that the company's goal is to increase this percentage to 90% (Melendez, 2020). Users can report content even if they are not its targets; in a small study by a nonprofit, 57% of reports were filed on behalf of someone else (Matias et al., 2015).

4 Reporting Experiment

4.1 Experimental Design

Sample. I sampled 6,148 Tweets containing hateful keywords during July and August 2020. I collected the Tweets every day with an algorithm that uses the search function of Twitter's API, which queries a subset of recent English-language Tweets excluding Retweets.[26] I searched posts containing two slurs: one that denies the Holocaust (Holohoax), and a disability slur (retard), the latter constituting 98% of the sample. Both terms are prevalent on social media and considered by many to be hate speech; see Guhl and Davey (2020) and Sherry (2019). Even if some people use the disability slur frequently (Albert et al., 2016), it is precisely the removal of this type of slurs that is controversial and policy-relevant. Moreover, Twitter's hateful-

[25] This number results from dividing the total number of accounts reported, from the Rules Enforcement Report (Twitter, 2020b), by the monetizable daily active usage published on the letter to shareholders from Q4 2020 (Twitter, 2020a).

[26] The algorithm conducted the search every 20 minutes. This timing allowed the data processing to be spread throughout the day to comply with the API's rate limits.

conduct policy covers the Holocaust and slurs that reinforce negative stereotypes about a protected category, which includes disability (Twitter, 2021b).[27]

Because the disability slur has alternative meanings, for example, to retard the progress of something, I refine the search with sentence structures such as, "You are a retard." This refinement captures directed hate speech (ElSherief et al., 2018) and facilitates identifying the targets of hate speech. I also consider multiple misspellings and word distortions to sample Tweets that attempt to bypass detection algorithms. Table B.1 in the Appendix contains the full list of queries used to search Tweets.

After my search algorithm detects a Tweet, it filters users to increase the quality of the sample and to reduce false positives, that is, Tweets that are not hate speech even if they contain the slurs. The filter drops users who self-report being under 18 in their profile biographies, those with new accounts (opened less than 2 weeks before the Tweet), inactive users or non-English speakers (with less than 10 posts in English and more than 50% of posts in another language), and bots (those with a Botometer score higher than 0.5). I also exclude users who display their preferred pronouns on their profile biographies,[28] Tweets that enclose the slurs in quotation marks (to capture users who are only referring to the slur), and those in the Holocaust sample who self-report being Jewish in their biographies. Users enter the sample once, so the filter also drops Tweets from duplicate users. This way, every observation in the sample is a user-Tweet pair, and I report users at most once.

At midnight every day, immediately before randomization into treatment, my algorithm checks whether the users or Tweets collected the previous day were removed from the platform; only those that have not been removed at this point enter the final sample. Table 3 compares descriptive statistics between the experimental sample and the random sample of Tweets from section 3. These samples are quite different. Experimental subjects have more recent accounts, give more likes per day, and are more likely to have posted toxic Tweets in the past. Tweets in the experimental sample are more toxic, as expected. The Holocaust and disability samples are also different; for example, the Tweets and timelines of users from the Holocaust sample

[27] In a pilot study, I included a broader list of slurs about race, ethnicity, religion, gender and sexual orientation. However, the sample contained many false positives, because most slurs are used by the members of the group that they target. Bianchi (2014) refers to this practice as appropriated or reclaimed uses of slurs. The two keywords that I use seem, anecdotally, to have lower false positives.

[28] Arguably, these users might be more empathetic and more likely to refer to the slurs rather than use them to attack.

have a lower toxicity. Figure B.5 in the Appendix plots the most common topics in each subsample, which I obtained by annotating the Tweets on MTurk. Some common topics include politics, religion, sports, and COVID-19.

Table 3: Characteristics of the reporting experiment sample

	Means			Difference t-statistic	
	Full Sample	Holocaust	Disability	Random-Full	Hol.-Disab.
Observations	6,148	123	6,025		
Accounts					
Account years	3.22	3.29	3.22	40.2	0.2
Tweets per day	11.62	19.69	11.46	2.2	3.7
Likes per day	24.17	33.64	23.98	-32.3	2.1
Followers	634.85	1,436.41	618.49	2.1	1.7
Followed	433.75	554.98	431.27	7.6	1.6
Initial shadow ban	0.71	0.71	0.71		0.1
Tweets					
Word count	15.98	23.98	15.81	-14.1	6.8
Is toxic	0.80	0.06	0.82	-244.3	-22.0
Is hate (MTurk)	0.30	0.43	0.30	-63.6	3.1
Is reply	0.84	0.56	0.84	-48.4	-8.4
Is attack (MTurk)	0.78	0.24	0.79		-14.8
Is quote	0.07	0.02	0.07	7.9	-2.0
Is mention	0.85	0.67	0.85	-42.5	-5.5
Tweet from phone	0.79	0.49	0.80	-9.0	-8.3
Timelines					
Previous toxicity	0.93	0.69	0.94	-28.2	-11.1
Previous disability	0.39	0.15	0.40	-179.1	-5.6
Previous Holocaust	0.10	0.66	0.09	-6.0	21.9

Notes: This tables presents means of characteristics in the reporting experiment sample and subsamples. It also presents t-statistics from tests of difference in means between the random and the experimental samples and between the Holocaust and disability subsamples.

Treatment. Figure 4 summarizes the experimental design and the timing of the algorithms involved. Every day at midnight, my algorithm randomly splits users or Tweets sampled in the previous 24 hours, who have not been removed from the platform, into a control or a treatment arm. The assignment is stratified by sampling date and slur; every day, half of the Tweets using each slur enter each experimental arm. Users in the control arm do not receive any intervention. The treatment consists of reporting Tweets for violating Twitter's rules against hateful conduct on the next day after they enter the sample, so Tweets can be reported between five and 48 hours after they are posted. Every day, my algorithm assigns the Tweets in the reporting arm evenly to one out of the 11 accounts that I use for reporting. Table B.4 in the Appendix displays summary statistics of the accounts that I used for reporting and Figure 5 includes screenshots of the reporting process.[29]

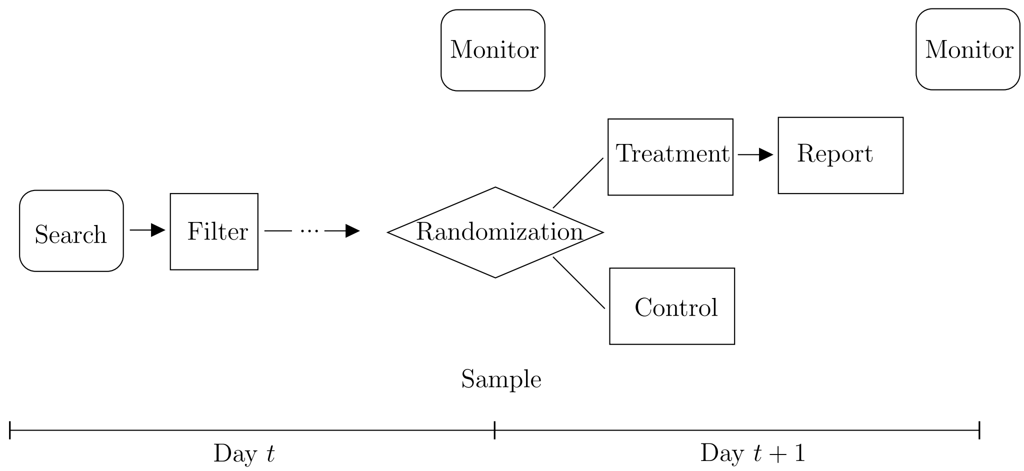


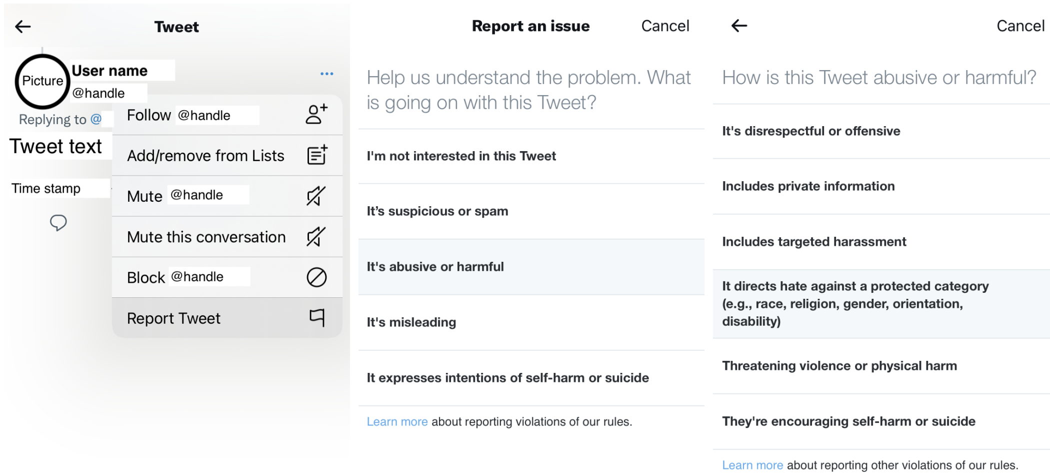
Figure 4: Design of the reporting experiment

Notes: Two main programs collect the sample and outcomes. The search program looks for hateful posts every 20 minutes and the monitor program keeps track of user activity every day. Randomization takes place at the beginning of every day with the sample of users collected the previous day.

Table B.3 in the Appendix shows that the two experimental arms are balanced in pre-treatment characteristics. Normalized differences—for each characteristic and all of them jointly—are well below the 0.25 value that Imbens and Rubin (2015) suggest. This balance confirms randomization was successful. I did not report 3.26% of the Tweets that were assigned to the reporting arm; that is, there is one-sided non-compliance. Three percent of Tweets in the reporting arm disappeared after treatment assignment and before I could report them, because users deleted them or deleted or protected their account, or because Twitter deleted the Tweets or suspended the users. Additionally, I did not report eight Tweets (0.26% of the Tweets in the reporting arm) that were clearly not hate speech.[30] Because of this one-sided

[29] When reporting Tweets, I click "It's abusive or harmful," then "It directs hate against a protected category (e.g., race, religion, gender, orientation, disability)." Due to logistics, 1% of the reported subjects were reported using a different account than the one that was assigned at the moment of randomization.

[30] For example, one user quoted some people using the disability slur to refer to him or her. Other users posted the Holocaust-denial term quoting a study that was published around those dates



(a) Tweet options **(b)** Reporting screen 1 **(c)** Reporting screen 2

Figure 5: Procedure to report Tweets

non-compliance, the estimates can be interpreted as an intention-to-treat (ITT), but I also give instrumental variable estimates that account for non-compliance.

Reports are an instrument for content moderation, that is, for receiving any sanction from Twitter. First, Twitter uses reports to detect content and enforce its rules, which implies the relevance and monotonicity conditions of instrumental variables hold. Second, reports only affect user behavior through their effect on sanctions, so the exclusion restriction holds if sanctions are perfectly observed. To the best of my knowledge, Twitter does not notify users they have been reported.[31]

Outcomes. I measure two types of outcomes: first-stage outcomes are the sanctions that Twitter enforces on users and second-stage outcomes are the users' activity on Twitter, their hatefulness, and spillovers to the activity of others. These outcomes allow testing whether reports influence moderation, whether Twitter's sanctions moderate users, and whether sanctions affect other users. I construct these outcomes with data that my algorithm collects every day. I gather users' cumulative number of Tweets, likes, accounts followed, and followers. I also collect the 100 most recent

(Center for Countering Digital Hate, 2021).

[31]Some users have received notifications from Twitter saying their posts were reported. According to the survey of section 5, 9% of users have received a notification that someone reported their Tweets. However, users seem to receive these notifications only when an account from Germany reports content, due to the Network Enforcement Act. Figure B.2 in the Appendix has a screenshot of one of these notifications.

Tweets of each user (posted within 24 hours), and select 20 Tweets at random per user to compute their toxicity score by calling Perspective's API.

I also measure whether Twitter sanctions the Tweets in the sample or their authors. Three sanctions are observable: Tweet removals or deletions, user suspensions, and shadowbans. I measure these outcomes as an absorbing state; that is, once users receive a sanction, they remain sanctioned. By construction, at the time of entering the sample, none of the Tweets have been removed by Twitter and none of the users are suspended; however, 71% of users are initially shadowbanned.

I measure activity on Twitter as the time that users spend posting or liking Tweets, which corresponds to t in the model of section 2. I do not directly observe time spent, but I construct a proxy using the number of Tweets that users post (that is, the statuses count object from the API) and the number of likes that they give (that is, the favorites count of the API). I then approximate the total number of words that users wrote and read during the period, by multiplying the Tweets and likes times the average number of words per Tweet in the random sample of Tweets, which is 13.81. Then, I convert words into time by using the average reading and typing speeds that have been documented in the literature.[32]

The main measure of hatefulness is the fraction of Tweets with a toxicity score higher than 0.8, but I consider alternative measures for robustness. Spillovers focus on the time spent by the users to whom the Tweets in the sample are replying ("replied users"); 86% of Tweets in the sample are replies to others. I focus on replied users because, arguably, users mentioned in a Tweet are more likely to notice sanctions related to the Tweet than others. Figure B.4 illustrates a reply to another Tweet.

Empirical strategy. This paper reports cross-sectional estimates of the effect of reporting users on different outcomes, three weeks after treatment assignment. I focus on first-stage and ITT estimates because, as the next subsection shows, I find evidence of unobservable sanctions which means that reports violate the exclusion restriction. In other words, reports affect outcomes not only through their impact

[32]I use the words per Tweet from the random sample, as opposed to the value from the experimental sample, because this is the value that I pre-registered, before the experimental sample existed. The average typing speed on a desktop computer is 51.56 words per minute (WPM) according to Dhakal et al. (2018). The average typing speed on a mobile device is 36.2 WPM (Palin et al., 2019). Elliott et al. (2019) estimate a reading speed of 179 WPM that is constant across different devices and screen sizes. I obtain the device of a user from the source object of Tweets; I consider the device to be a desktop when the source is "Twitter Web App" and mobile for all the other sources.

on observable sanctions, but also through unobservable sanctions. Thus, I estimate regressions of the form:

$$Y_i = \alpha + \beta Z_i + \delta X_i + \varepsilon_i, \tag{6}$$

where i indexes user-Tweet pairs, Y_i denotes first-stage or second-stage outcomes, Z_i denotes treatment assignment (reports), and X_i is a vector of controls. I estimate specifications without controls, controlling for stratum—sampling date and slur—fixed effects, and adding controls from the rich set of pre-treatment characteristics of Table B.3. I select controls with a two-step method using lasso as suggested in Urminsky et al. (2016) with the methodology of Belloni et al. (2014). Regressions use robust standard errors unless noted otherwise.

I also estimate dynamic treatment effects, which is possible because my algorithm collects outcomes every day after users enter the sample. I use the efficient estimator proposed by Roth and Sant'Anna (2021), which is robust to heterogeneous treatment effects. Because reports are randomized every day, the design satisfies their assumptions of random treatment timing and no anticipation.[33] I use their method to obtain event-study estimates, in which the event date is the number of days since a report. I construct the estimates on balanced panels but also report the treatment effect on attrition. The results use their Neyman-style pointwise confidence intervals and the sup-t confidence bands of Montiel Olea and Plagborg-Møller (2019).

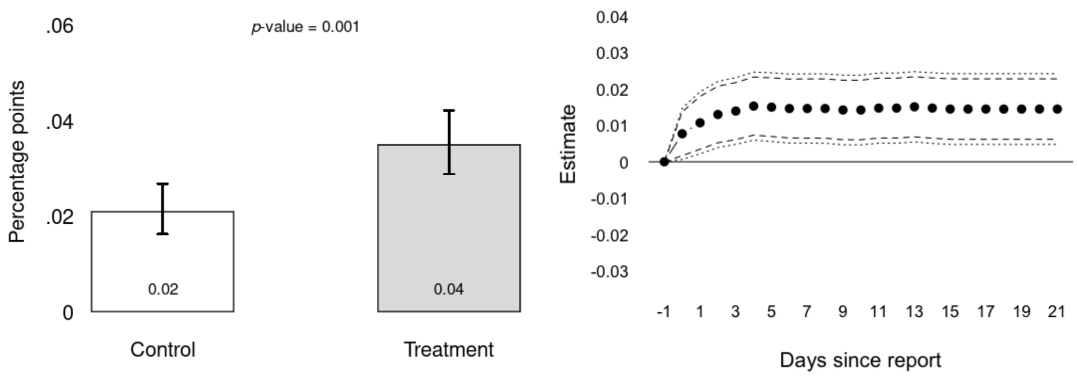
4.2 Results

Sanctions. Reporting Tweets increases the likelihood that Twitter deletes them. Figure 6a shows the impact of assignment to treatment on the likelihood that Twitter removes the Tweets in the sample. Twitter removed 2.1% of the Tweets in the control arm within three weeks (21 days) after they entered the sample, and it removed 3.5% of them in the treatment arm. The treatment effect is 1.4 percentage points, which is 0.08 standard deviations, or a 66% increase.[34] The p-value of the difference in proportions is 0.001, and the F statistic from a regression of deletions on treatment

[33]These assumptions hold within each stratum (sampling date and slur). However, since the Holocaust denial slur has few observations per day, I compute the estimators within each sampling date, pooling observations from both slurs.

[34]These numbers include cases in which Twitter required the removal of a Tweet but the user did not remove it within the three weeks. Eleven percent of Tweets were not removed by users in the control arm, and 5% were not removed in the treatment arm.

assignment is 11.01.[35] Figure 6b displays dynamic treatment effects over event time; that is, the number of days since assignment to treatment. The dependent variable indicates whether Twitter removed Tweets at or before each event date. This figure shows that reports induce Twitter to remove Tweets within the first four days.



(a) Means **(b)** Cumulative dynamic treatment effects

Figure 6: Likelihood that Twitter removes a post

Notes: Panel (a) displays the mean and 95% confidence intervals of the likelihood that Twitter removes a Tweet in the three weeks after reporting by treatment arm. The p-value is from a test of proportion differences. Panel (b) presents cumulative dynamic treatment effects of the likelihood of deletions, pointwise confidence intervals (dashed), and sup-t simultaneous confidence bands (dotted). Dynamic effects use the estimator from Roth and Sant'Anna (2021).

Table C.1 in the Appendix shows estimates of the effect of reporting on other Twitter sanctions and user self-censorship. Reporting does not significantly influence the other observable sanctions; that is, suspensions or shadowbans. The table also displays insignificant effects on the likelihood of users deleting their own posts or accounts, or protecting their accounts (making them private) within three weeks after reporting. Moreover, it shows reporting does not change the likelihood that other Tweets in the users' profiles go missing, which includes self-removals and Twitter removals.[36] The null effects persist after adding strata fixed effects and other controls, and the size of all estimates is below 0.033 standard deviations.

[35]These estimates keep all users, even those whose accounts were deleted. Results are unchanged if we drop them. Results from a two-stage least-squares regression that uses treatment assignment as an instrument for reports are the same.

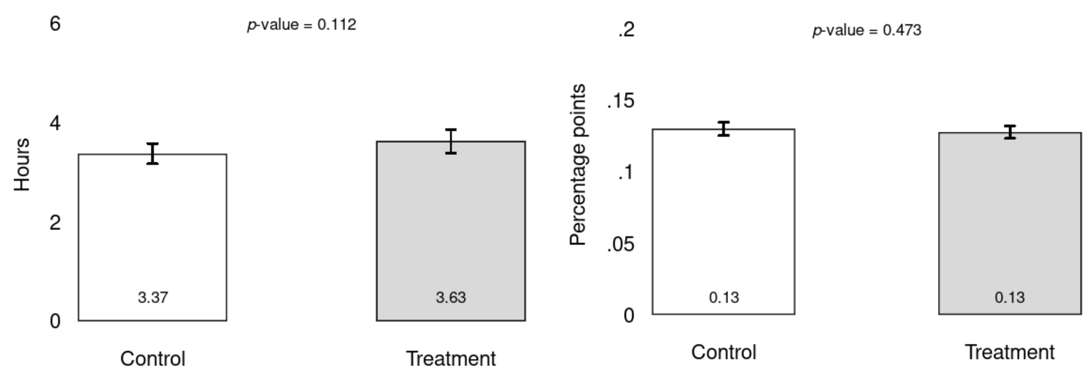
[36]These numbers include the Tweets that users post after the sampling date and up to three weeks after the end of the sampling period. For these Tweets, distinguishing user deletions from Twitter deletions was not possible due to the API's rate limits.

Evidence of unobservable sanctions exists, however. Twitter sent updates informing me it found that 270 (8.8%) out of the 3,074 accounts on the reporting arm violated the rules, within three weeks of the reports (see Figure B.3 for an example). One hundred fourteen (42%) of these updates were not accompanied by Tweet deletions, user suspensions, or shadowbans, which means 6.2% of reports led to an update but not an observable sanction. Some unobservable sanction is likely in these cases, such as accounts being temporarily locked (see Figure B.1f for an example). Moreover, that percentage likely understates the true number of unobservable sanctions, because Twitter does not always send updates whenever it imposes a sanction. For instance, Twitter sent me updates only for 13.4% of the 1,162 accounts in the reporting arm that received an observable sanction. Overall, I received updates on 12.52% of my reports. As a benchmark, an exercise conducted by the European Commission (Reynders, 2020) observed that Twitter sent an update on 26% of reports filed by general users.

Figure C.2 provides additional evidence of unobservable sanctions; it plots daily treatment effects on the number of hours since the last post, computed at midnight. The treatment effect is positive and pointwise significant around day 10 after reporting, although not significant with the simultaneous confidence bands. This figure suggests that reporting slightly increases the gap in between posts, which indicates that users might have had their accounts locked, although the daily number of hours since last post might not reflect locking periods of less than 24 hours.

Activity. Reporting does not significantly decrease user activity on Twitter. Figure 7a displays the treatment effect on the number of hours that users spend posting and liking Tweets in the three weeks after reporting. Both treatment and control spent around three and a half hours, and the treatment effect is 0.25 hours (5 minutes per week), which is a 7.5% increase or .042 standard deviations. This effect, however, is not significant at conventional levels, because the p-value of the difference in means is 0.11. Figure C.3a in the Appendix shows that treatment effects remain flat throughout the period. Table C.2 shows regression estimates using alternative measures of activity: Tweets and likes separately, a winsorized measure of time spent online removing the top and bottom percentiles, and an extensive-margin measure (the fraction of days that users post, like, or follow someone). The results remain unchanged using these alternative measures; if anything, the effect on Tweets is positive and significant at the 10% level under some specifications. Moreover, these estimates are mechanically biased downward since Twitter might temporarily lock user accounts.

Hatefulness. Reporting does not significantly decrease the likelihood of posting hate on Twitter. Figure 7b shows that the fraction of hateful Tweets (toxicity bigger than 0.8) that users post in the three weeks after the treatment is the same for both experimental arms. The treatment effect is -0.02 percentage points of hateful Tweets, which is a 1.7% decrease (-0.02 standard deviations). Figure C.3b shows a decrease in hatefulness in the first three to five days after reporting (pointwise significant), but the effect returns to zero by the end of the three weeks. Table C.3 considers other measures of hatefulness; two extensive-margin measures (whether users post any Tweet with toxicity ≥ 0.8 or they repeat the slur), the average toxicity, and the average severe toxicity (another measure developed by Google). None of these measures yield significant effects, and the treatment effect is less than 0.011 standard deviations across all variables using different specifications.



(a) Hours spent on Twitter **(b)** Fraction of hateful posts

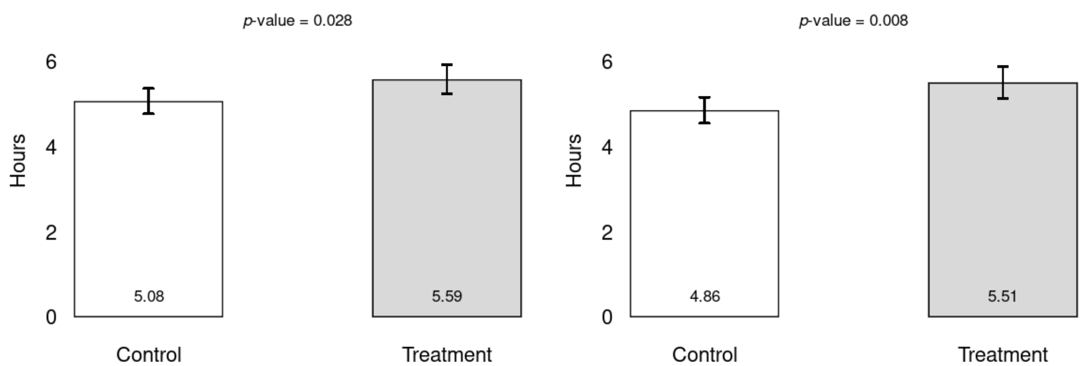
Figure 7: Hours spent on Twitter and fraction of hateful posts

Notes: This figure displays means and 95% confidence intervals of outcomes in the three weeks after reporting by treatment arm. Hours spent is calculated using statuses and favorites. Hateful posts are those with toxicity higher than 0.8. The *p*-value is from a test of difference in means.

Spillovers. Even if reporting does not seem to moderate the authors of the Tweets, that is, decrease their activity or hatefulness, it impacts other users. Figure 8a shows that reporting increases the time the replied users spend Tweeting and liking by 0.51

hours, which is 10 minutes per week, 10%, or 0.064 standard deviations (p-value = 0.028). The treatment effect seems persistent; Figure C.4 shows the cumulative effect increases continuously after the reporting day.

Although many of these replies are attacks, some are replies between social media friends. As specified in the pre-analysis plan, I asked MTurk workers to read the context of both posts and classify whether the replies in my sample were attacks on the replied user. Under the majority decision rule, in which Tweets are attacks if the majority of workers agree, 87% of replies were attacks on others. Figure 8b shows that the effect of reporting is stronger among attacks; it is 0.65 hours (13 minutes per week, a 13.4% increase or 0.08 standard deviations, p-value = 0.008). Table C.5 shows estimates using the same alternative measures of activity as above. Results remain significant at the 5% level across specifications considering Tweets and likes separately or winsorizing time spent.[37] Hence, reporting seems to increase the activity of those users that are attacked by the Tweets in the sample.



(a) All replies **(b)** Attacks

Figure 8: Spillover on the time spent of users replied by the posts

Notes: This figure displays means and 95% confidence intervals of the time spent on Twitter in the three weeks after reporting by treatment arm. Panel (a) includes all users replied by the Tweets. Panel (b) includes users that were attacked by the Tweets, according to MTurk annotators. The p-value is from a test of difference in means.

Attrition. In this experiment, attrition occurs because accounts go missing after treatment assignment; 7% of them were missing after three weeks. Attrition happens

[37] The findings are also robust to dropping those users who were replies to more than one Tweet in the hateful sample. More than 93% of the replied users corresponded to a single user in the hateful sample, so concerns about SUTVA violations are minimal.

when users delete their own accounts or Twitter suspends them. Given that the previous results showed no treatment effect on account suspensions or on the likelihood that users delete their accounts, finding no differential attrition by treatment arm is not surprising. Table C.6 shows insignificant treatment effects on the likelihood that users leave the sample at the end or on any day of the three weeks after users enter the sample. Figure C.5 shows dynamic treatment effects on attrition; the effect is not significant pointwise or with the simultaneous bands.

Heterogeneity. The pre-analysis plan specified two dimensions for the heterogeneity analysis, besides attacks: by slur (Holocaust vs disability) and by human annotation (among the hate sample). I report these results in Figure C.6 in the Appendix due to their low informational content.[38]

4.3 Interpretation

The previous results indicate that reports instrument for sanctions, particularly Tweet removals and, potentially, unobservable sanctions. Moreover, the treatment did not decrease user activity or the likelihood of posting hate within three weeks; reports did not moderate users. The effect on the users' activity is insignificant, which I interpret as a low elasticity of time spent with respect to moderation among the users in my sample; $\partial T^H/\partial c \approx 0$ in the notation of the model of section 2.

Yet, reports spill over to other users; they increased the amount of time that the attacked users spent posting and liking. I interpret these findings as evidence of a positive elasticity of the time spent of some users in my sample with respect to moderation; $\partial T^A/\partial c > 0$ in the notation of the model of Section 2.

Three main mechanisms may explain why reports impacted the replied users. First, the reported users could have changed their behavior or their interactions with the replied users. Second, if Twitter removed the Tweets, the replied users could have noticed the legends that Twitter placed on the Tweets, as in Figure B.1a. Third, if

[38]The experiment is not powered to detect the effect on the small Holocaust sample. The heterogeneity analysis by human annotation was intended to capture measurement error (false positives) in the sampling of hate speech. More than false positives, these labels seem to capture heterogeneity due to the subjective nature of hate speech. Thirty percent of Tweets in the sample were labeled as hate speech by the majority of annotators, 61% were considered offensive, 1.6% were not considered offensive or hate, and the remaining did not have a majority label. Hence, splitting the sample between "hate" and "not hate," as preregistered, captures the difference between hateful and offensive Tweets.

the replied users also reported these Tweets, Twitter could have sent them an update on their reports, as in Figure B.3.

Regarding the first mechanism, the results in subsection 4.2 rule out that the reported users substantially changed their behavior. Additionally, Figure C.7 shows an insignificant effect on the likelihood that the users in the sample mention the replied users again within three weeks. Hence, the evidence in favor of this mechanism is weak. The same is true for the second mechanism. Table C.7 shows that the treatment effect on deletions is smaller in the sample of replies relative to the full sample, and insignificant in the sample of Tweets that attack others.

As for the third mechanism, the percentage of reports for which I received an update and found no observable sanction is similar in the full sample, among replies, and among attacks (6.2%, 6.4%, and 6.5%, respectively). Hence, Twitter may have imposed an unobservable sanction (e.g., locking accounts) on the users who attacked others, and the attacked users who reported these Tweets may have received an update about the sanction.[39]

How does reporting affect monetization? I obtain a back-of-the-envelope estimate as follows. The treatment increased by 10-15 minutes per week the time that reported users and replied users spent liking and posting. The advertising load on a small sample of 50 Tweets was one ad per four regular Tweets. Assume this number translates into an ad load of 0.25 minutes per minute of content consumed. Twitter's Ad website has a default bid of $0.21 per six-second video advertisement.[40] Ignoring effects on others, the treatment amounts to a $5.25-$7.88 increase in ad revenue per week per report.[41]

[39]This hypothesis is difficult to test without access to internal data, because user reports are unobservable. Moreover, whether users would reveal that they reported a particular Tweet in a survey is unclear, even if reporting is common (indeed, the next section shows that one-third of users have reported content).

[40]This price is for the general audience of U.S. adults. The ad price did not change when I tried targeting an ad to the list of users in the sample.

[41]Besides being a rough estimate, this calculation is based on a selected sample and ignores equilibrium effects, so it does not imply that Twitter would like to increase reports. Moreover, these numbers do not consider the marginal costs of moderating.

5 A Test of Overprovision or Underprovision

5.1 Experimental Design

Sample. I recruited 3,027 respondents in September 2021 through Luc.id, a widely used online marketplace that matches researchers with survey providers (Coppock and McClellan, 2019; Bursztyn et al., 2020). I pre-screened participants to select English speakers who live in the U.S., are over 18 years old, are willing to provide their email, self-report using Twitter, and pass a basic attention check. After the pre-screen, participants entered the online survey and had to answer demographic questions. The survey also asked them for their Twitter handle (optionally), which I used to get their account creation date, Tweet counts, and like counts. Sixty-four percent of participants provided a handle, and 74% of the handles were valid. This results in a sample size of 1,427 respondents, which satisfies the recommendation of Haaland et al. (2020) of 700 respondents per treatment arm.

Table 4 compares the characteristics of the sample with representative adult Twitter users from the ATP survey and with accounts from the random sample of Tweets. My survey undersamples users in the 18-29 age range, college graduates, and politically Independents, and oversamples white respondents and Democrats.[42] Users who provided their Twitter handle have an older account, and fewer Tweets and likes per day relative to accounts in a random sample of Tweets.

Afterward, the survey asked questions about social media use, online harassment, hate speech, and Twitter sanctions. These questions provide further insights about the previous experiment. Figure C.8 shows that the API-based measure of time spent on Twitter correlates closely with users' self-reported hours, so it is a good proxy measure. Table C.8 includes additional statistics. For instance, 32% of users have reported content for violating the rules, 10% have had a Tweet removed, and 5% have been suspended. Moreover, the experience on the platform differs by minority status, which I define based on religion (Jewish, Muslim, Buddhist, Hindu, or other), sexual preference (not heterosexual), gender (other than man or woman), and race (other than white). Consistent with other surveys (Anti-Defamation League, 2021), minorities are more likely to experience harassment online, to self-report seeing more

[42] I pre-registered introducing quotas to match representative Twitter users on gender, age, race or ethnicity, region, and political orientation, but relax the quotas to obtain the desired sample size was necessary.

Table 4: Characteristics of the welfare experiment sample

Panel A: Demographics, N= 3,027		
	Means (Survey)	ATP-Survey t-stat.
Age 18-29 (%)	24.48	3.01
Male (%)	53.88	-0.34
White (%)	68.19	-5.04
College graduate (%)	31.68	4.89
Republican (%)	22.76	-1.34
Democrat (%)	52.89	-9.41
Panel B: Twitter accounts, N= 1,427		
	Means (Survey)	Random-Survey t-stat.
Account years	7.93	-23.65
Likes per day	2.34	29.37
Tweets per day	1.54	41.76

Notes: This tables presents means of characteristics in the welfare experiment sample. It also presents t-statistics from tests of difference in means between the ATP or the random sample of Tweets, and the experimental samples.

hate speech in their feed, and to report content. However, they also receive more sanctions and reports, which, to the best of my knowledge, has not been documented before.[43]

Treatment. Figure 9 summarizes the experimental design. I use an information-provision treatment with an active control group (Haaland et al., 2020). After the baseline questions, I randomize survey participants into two treatment arms that receive different information about the likelihood of moderation among hateful Tweets. The information provided comes from the annotated random sample of 10,000 Tweets. To vary the likelihood of moderation without deception, I use different decision rules to classify hate speech. As Table 2 shows, 3.6% of hateful Tweets are removed or their authors are suspended within one month of the post under the majority decision rule, that is, if most annotators agree. That percentage changes to 9.1% under the consensus rule, that is, if all annotators agree. Half of participants are randomized into the low-moderation arm (3.6%) and half into the high-moderation arm (9.1%). The treatment is stratified by whether respondents are male, minorities, and have

[43]This finding is related to the literature on racial biases in detection algorithms; see, for example, Cowgill and Tucker (2019).

been sanctioned by Twitter, and whether they provided a Twitter handle.

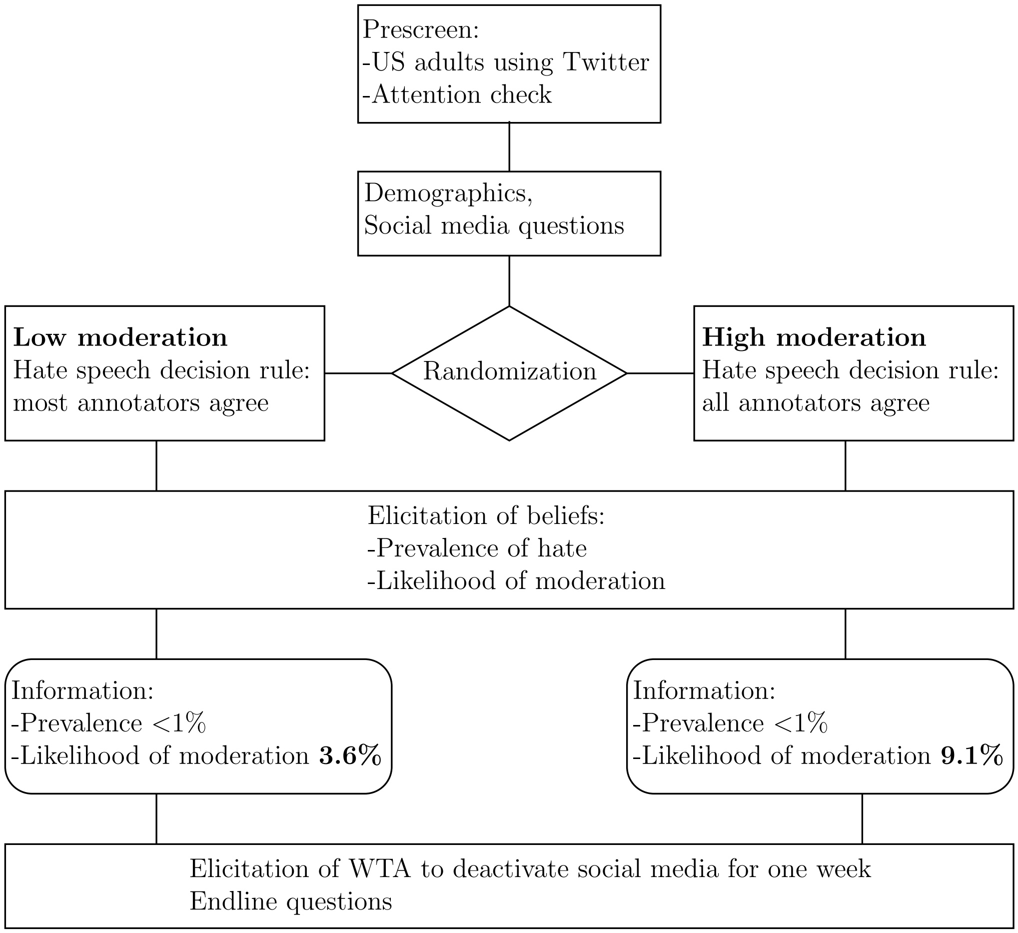
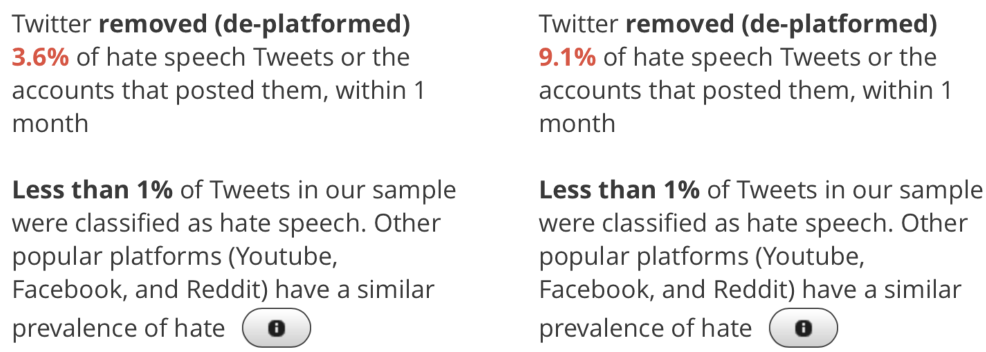


Figure 9: Design of the welfare experiment

After randomizing participants, I inform them, for transparency, of the rule that I use to classify hate. As pictured in Figure B.6, I tell them that a crowd-sourced team of annotators identified hate speech using 10,000 Tweets, and that a Tweet is hate speech if [most/all] annotators label it as hateful. I then elicit their beliefs about (1) the prevalence of hate speech in this sample and (2) the fraction of Tweets that are removed or suspended within one month. These elicitations are incentivized, because they know that one participant with the closest guess will get a $50 Amazon gift card.

After the elicitation, I provide information about the likelihood of moderation, as displayed in Figure 10. I also hold constant the prevalence of hate speech in both arms, by telling respondents that less than 1% of Tweets are classified as hate (recall that 0.56% Tweets are hate under the majority rule and 0.11% are hate under the consensus rule). The message also shows that other popular platforms, such as YouTube, Facebook, and Reddit, have a similar prevalence of hate, according to different sources (Kennedy et al., 2020; Vidgen et al., 2020; Facebook, 2021). They can consult the sources by clicking a button on this screen.



(a) Low moderation **(b)** High moderation

Figure 10: Information provision by treatment arm

Table B.5 shows that both experimental arms are balanced on pre-treatment characteristics. The table also rules out that changing the decision rule to classify hate influences the participants' concept of hate; the treatment has no effect on the belief about the prevalence of hate or the likelihood of moderation.[44]

Outcomes. There are two outcomes of interest. Based on the results of section 2, the main outcome is the willingness to accept (WTA) to stop using social media, that is, Twitter, Facebook, Instagram, YouTube, Snapchat, TikTok, and Reddit, for one week. I first tell participants that the research team will conduct a small follow-up study that compensates some participants to deactivate their social media for one week. I inform them that similar studies have been conducted in the past (Hunt et al., 2018; Mosquera et al., 2020; Allcott et al., 2020). I then elicit their WTA with an iterative multiple price list (iMPL, see Harrison et al. (2005); Andersen et al. (2006)).[45] Subjects have to decide whether they are willing to stop using social media for different Amazon gift card offers. The first offer is for $50, and subsequent amounts

[44]This finding is similar to what other studies obtain, such as Bottan and Perez-Truglia (2017), who argue that changing the source of information does not have an impact on participants who do not have expertise on the data.

[45]The iMPL has two advantages over a regular multiple price list. First, it induces monotonicity on responses by construction. Second, it saves time by omitting reduntant questions.

increase or decrease until the WTA is placed in intervals that go from $(-\infty, 0]$ to $[100, \infty)$ and increase by \$10, as Figure B.7 illustrates. I transform these intervals into a continuous measure using the triangular distribution procedure from Allcott and Kessler (2019).

This elicitation is incentivized. I inform respondents that a computer will randomly choose some eligible participants whom the research team will contact for the follow-up.[46] If the participant is selected, the computer will also choose one of her answers at random. If the answer is "yes," the research team will ask her to stop using social media for one week and pay the offered amount. If the answer is "no," the participant will not be asked to stop using social media. This information is truthful; I recontacted 50 participants at random in October 2021 and implemented the follow-up study.[47]

The second outcome of interest is the API-based time spent on Twitter one week after the survey, which I compute for the participants who provided valid Twitter handles following the procedure outlined in section 4. At the end of the survey, I ask questions to measure attention, experimenter demand effects, and posterior beliefs.

Empirical strategy. The empirical strategy consists of OLS regressions of outcomes on an indicator of treatment status. All estimates use robust standard errors. I run regressions without controls, controlling for stratum fixed effects, and a specification adding controls as in Urminsky et al. (2016). As pre-registered, I report estimates of the main outcomes reweighting observations to match the ATP on first moments of gender, age, race or ethnicity, region, and political orientation, but I also report unweighted estimates. I obtain the weights using the maximum entropy approach of Hainmueller (2012).

[46] Following Allcott et al. (2020), I did not tell participants the likelihood of being selected into the follow-up; previous research has shown that, at least on Becker-DeGroot-Marschak elicitations, informing participants can bias WTA estimates.

[47] Thirteen participants replied to the recontact email. Seven of them had been randomized into the deactivation treatment, and six to the control group. I asked participants in the deactivation arm to upload screenshots of the time-tracking app of their phones as proof of deactivation, as in Hunt et al. (2018). Five out of seven participants self-reported that they had stopped using social media, and four submitted the screenshots.

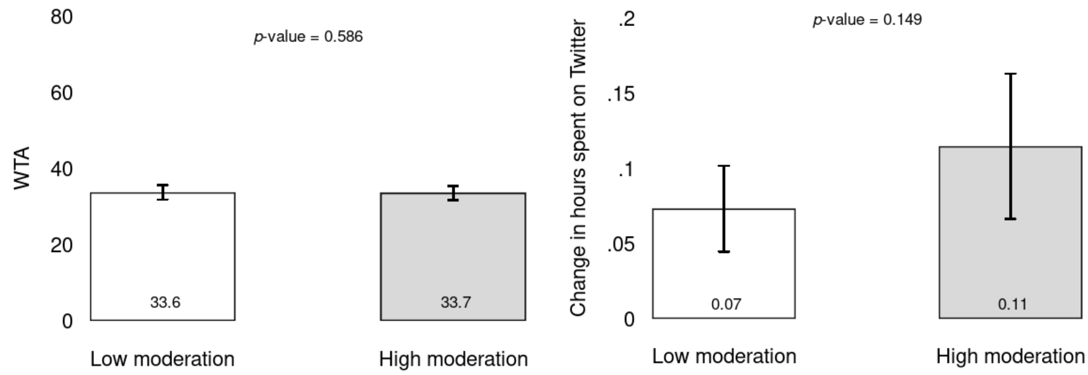
5.2 Results

Misperceptions about hate speech and moderation. Most users overestimate the prevalence of hate speech on Twitter and the likelihood that Twitter sanctions hateful content. Figure C.9 displays histograms of beliefs among respondents. Ninety-six percent of Twitter users overestimate the prevalence of hate speech, that is, their belief is above 1%, and 84% guess a moderation rate above the higher 9.1% value. These results add another example to the literature on misperceptions about others (Bursztyn and Yang, 2021).

There are several explanations for these facts. An "echo-chamber" argument is that users might not notice what happens outside their curated feeds, which they personalize with the help of Twitter's algorithms. Consistent with this argument, I find that 74% of users believe that the prevalence of hate in the random sample of Tweets is higher than what they see in their feed. Platforms' lack of transparency might also contribute to misperceptions. Even Facebook, which publishes a substantial amount of information (Facebook, 2021), informs only about the prevalence of hate speech but not about the likelihood of moderation (Bradford et al., 2019). The only information about the likelihood of moderation, between 3 to 5% of hateful content, was revealed thanks to the recent whistleblower incident (Giansiracusa, 2021).

WTA to stop using social media. Providing information about a higher likelihood of moderation has little effect on the users' social-media valuation. Figure 11a displays the treatment effect on the WTA to stop using social media during one week. The average WTA was \$33.6 in the low moderation arm, and \$33.7 in the high moderation arm. The treatment effect is -15 cents per week, which is 0.004 standard deviations, or a 0.5% decrease. The null effect is not just on average; Figure C.10 in the Appendix shows that the cumulative distribution function of WTA is the same for both arms. Table C.9 presents regression estimates with alternative measures of social-media valuation. As in Allcott and Kessler (2019), I assume a uniform distribution of WTA beyond the endpoints instead of the triangular distribution. I also use -\$50 and \$150 for the endpoints as benchmarks, or a take-it-or-leave-it dummy for the first \$50 offer. The results remain unchanged using these alternative measures.

Activity. The information provision treatment has an positive but insignificant effect on the time that users spent on Twitter one week after the survey. Figure 11a plots the effect on the number of hours spent by users who provided their Twitter handle. The effect is 0.04 hours, which is 2.4 minutes (57% increase relative to the low-moderation arm, or 0.077 standard deviations).[48]



(a) Means of WTA

(b) Means of hours spent

Figure 11: WTA to stop using social media and hours spent on Twitter

Notes: Panel (a) displays the mean and 95% confidence intervals of the WTA to stop using social media one week by treatment arm. Panel (b) presents the mean and 95% confidence intervals of the hours spent on Twitter one week after the survey. The p-values are from a test of difference in means and observations are reweighted to match Twitter users from the ATP on observables.

Posterior beliefs, attention, attrition, and experimenter demand. Respondent inattention cannot explain the previous null results; providing information significantly shifts participant's recollection of the information provided and their posterior beliefs about moderation. At the end of the survey, I asked participants to repeat the moderation rate that I gave them, and I incentivized the closest answer with a $50 gift card. Figure 12 plots the effect on the respondents' recollection of the moderation information. Sixty percent of participants recalled a number within one percentage point of the true value.[49] The treatment effect on this recollection is 5.6 percentage points (53% or 0.425 standard deviations, with an F-statistic of 36), not statistically different from 5.5 (p-value = 0.907), which is the gap between the high moderation rate (9.1%) and the low moderation rate (3.6%).

[48]Table C.10 shows estimates using the same alternative measures of activity as in section 4; Tweets and likes separately, winsorized hours, and an extensive-margin measure of the fraction of days in which users post or like. The effect remains insignificant with these measures across specifications. Figure C.11 confirms that dynamic treatment effects remain flat throughout the

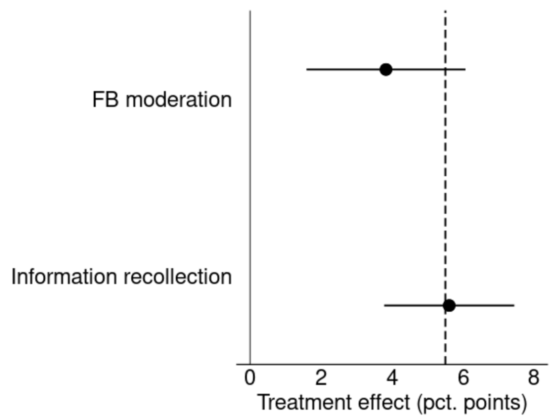


Figure 12: Posterior beliefs about moderation on Facebook and attention check

Notes: This figure presents coefficients and 95% confidence intervals of OLS regressions on an indicator of the high-moderation arm. FB moderation is the users' beliefs about the fraction of posts or users that Facebook moderates. The attention check is the participants' recollection, at the end of the survey, of the information provided about the moderation rate on Twitter. The dashed line is at 5.5 percentage points, the difference between the high moderation rate (9.1%) and the low moderation rate (3.6%). Observations are reweighted to match Twitter users from the ATP on observables.

Figure 12 also plots the treatment effect of information on users' beliefs about the likelihood of moderation on Facebook. This follows the recommendation of Haaland et al. (2020), of measuring posteriors by asking post-treatment beliefs about a related but different variable. The average belief of the moderation rate on Facebook was 19% for users in the low-moderation arm and 22.9% in the high-moderation arm. The treatment effect was 3.8 percentage points (20% or 0.16 standard deviations, with an F-statistic of 11.2).

Additionally, there is no evidence of differential inattention, attrition, or experimenter demand effects by treatment arm. Table C.11 presents insignificant treatment effects on inattention, measured as the absolute difference between participants' recollection and the information provided. Thirty-four participants (1.1% of the sample) completed the prescreening questions but did not finish the survey, and Table C.11 shows null treatment effects on attrition under different specifications. Following Allcott et al. (2020), the last part of the survey asked a question to test for experimenter demand effects: "Do you think the researchers in this study had an agenda?" Similar to that study, 57% of respondents in both arms thought I had no particular agenda

week post-survey.

[49]Because of left-digit bias, many participants in the low-moderation arm remembered 3%.

or were not sure. Figure C.12 shows insignificant treatment effects on the responses to that question.

Heterogeneity. I do not find substantial heterogeneity of the effect on the WTA across most of the pre-registered covariates, including minority status (as defined above), whether participants have experienced a sanction on Twitter, and whether their beliefs are above or below the median moderation belief of 33% of hateful Tweets. The exception is the time spent on Twitter after the survey. Figure C.13 in the Appendix shows suggestive evidence that minorities spend more time on Twitter when they receive the high moderation information. The treatment effect in this subsample is 0.054 hours (three minutes, 100% increase relative to the control group, 0.17 standard deviations, p-value = 0.03).[50]

5.3 Interpretation

The previous results indicate that providing information about a higher moderation rate shifted users' beliefs, but had little impact on their social-media valuation. Taken at face value, these results mean that Twitter does not moderate too much or too little from the consumers' point of view, for a fixed prevalence of hate speech. One explanation for this finding is that Twitter internalizes the impact of moderation on users' willingness to pay for the platform, which requires that marginal and inframarginal users respond similarly to sanctions.

Another option is that users do not directly care about moderation, holding constant the hate they encounter. Indeed, it is possible that the experiment did not change users' perceptions about hate in their own feed. In that case, users could have differentially updated their beliefs about how effective the algorithms are at hiding content without moderating. This is consistent with platforms providing a wide range of tools that allow users to customize their experience. For instance, Twitter allows users to mute and block accounts and words, and to hide sensitive content from their feeds.

One challenge to the interpretation of these findings comes from the welfare discussion in Allcott et al. (2020). They argue that users might misperceive Facebook's

[50] The treatment effect among minorities is significant at the 10% without reweighting observations. Figure C.13 also shows large point estimates on the subsample of users who have been sanctioned and those with high prior beliefs, although these are noisily estimated.

value, and thus the WTA might overstate consumer surplus. These value misperceptions could explain why increasing perceived moderation did not impact users' WTA. Another challenge is that the treatment not only shifted users' beliefs about moderation on Twitter; it also impacted beliefs about moderation on other platforms (at least Facebook). Based on Appendix A, the correct measure of the change in consumer surplus is to consider current social media users, not just current Twitter users. Table C.12 in the Appendix shows that results are unchanged after reweighting observations to match representative social media users, or without reweighting.

There is also suggestive evidence that the treatment increased minorities' time spent on Twitter. Given that these individuals are more likely to experience harassment online (Table C.8), this is consistent with the finding from the previous experiment that reporting increases the activity of the targets of hate speech.

6 Conclusions

Simple economics explain why it makes sense for profit-maximizing social media companies to ban some of their customers or restrict their content: because this increases the willingness to pay of marginal users. In an advertising-driven business model, platforms remove content only if this increases the time that some users spend consuming content, and hence interacting with ads. I find evidence consistent with this implication, by running a natural field experiment in which I report content that violates Twitter's rules against hateful conduct. Reports increase Tweet removals and, potentially, unobservable sanctions, and they do not decrease user activity or hatefulness. Yet, the targets of hateful posts increase their activity after the reports. While this treatment provides some evidence of the behavioral effects of moderation, further work is needed to understand repeated sanctions, different classes of platform interventions, or the effects of moderation on other types of content.

In terms of policy, both sides in the discussion of how to regulate platforms often mention a tension between profit maximization and optimality of content moderation. While platforms can, in theory, remove too little or too much content relative to a surplus-maximizing planner, this study finds no evidence of distortions from the consumers' point of view. There are, however, two caveats to these findings. First, consumer surplus ignores the costs that hate speech imposes outside platforms. Hence, an avenue for future research is to examine the costs and benefits of the real-world consequences of content moderation. Second, even without moderation distortions, imperfect competition between platforms likely leads to pricing distortions, so they might be setting the ad loads of haters or non-haters suboptimally. These distortions can be empirically confirmed by future work.

References

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

Albert, A. B., H. E. Jacobs, and G. N. Siperstein (2016). Sticks, stones, and stigma: Student bystander behavior in response to hearing the word "retard". *Intellectual and developmental disabilities 54*(6), 391–401.

Ali, S., M. H. Saeed, E. Aldreabi, J. Blackburn, E. De Cristofaro, S. Zannettou, and G. Stringhini (2021). Understanding the effect of deplatforming on social networks. In *13th ACM Web Science Conference 2021*, pp. 187–195.

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

Allcott, H. and J. B. Kessler (2019). The welfare effects of nudges: A case study of energy use social comparisons. *American Economic Journal: Applied Economics 11*(1), 236–76.

Álvarez-Benjumea, A. and F. Winter (2018). Normative change and culture of hate: An experiment in online environments. *European Sociological Review 34*(3), 223–237.

Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutström (2006). Elicitation using multiple price list formats. *Experimental Economics 9*(4), 383–405.

Anti-Defamation League (2021). Online hate and harassment. the american experience 2021. *Center for Technology and Society*. Accessed: 2021-10-23.

Arango, A., J. Pérez, and B. Poblete (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pp. 45–54.

Becker, G. S. and K. M. Murphy (1993). A simple theory of advertising as a good or bad. *The Quarterly Journal of Economics 108*(4), 941–964.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*(2), 608–650.

Berry, S., A. Gandhi, and P. Haile (2013). Connected substitutes and invertibility of demand. *Econometrica 81*(5), 2087–2111.

Bianchi, C. (2014). Slurs and appropriation: An echoic account. *Journal of Pragmatics 66*, 35–44.

Bottan, N. L. and R. Perez-Truglia (2017). Choosing your pond: Revealed-preference estimates of relative income concerns. *Available at SSRN 2944427*.

Bradford, B., F. Grisel, T. L. Meares, E. Owens, B. L. Pineda, J. N. Shapiro, T. R. Tyler, and D. E. Peterman (2019). Report of the facebook data transparency advisory group. *Yale Justice Collaboratory*.

Braghieri, L., R. Levy, and A. Makarin (2021). Social media and mental health. *Available at SSRN*.

Bundesamt für Justiz (2019). Federal office of justice issues fine against facebook. `https://www.bundesjustizamt.de/DE/Presse/Archiv/2019/20190702_EN.html`. Accessed: 2021-09-30.

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

Bursztyn, L., A. L. González, and D. Yanagizawa-Drott (2020). Misperceived social norms: Women working outside the home in saudi arabia. *American economic review 110*(10), 2997–3029.

Bursztyn, L., I. K. Haaland, A. Rao, and C. P. Roth (2020). Disguising prejudice: Popular rationales as excuses for intolerant expression. Technical report, National Bureau of Economic Research.

Bursztyn, L. and D. Y. Yang (2021). Misperceptions about others. Technical report, National Bureau of Economic Research.

Carlson, C. R. (2021). *Hate Speech*. MIT Press.

Carlson, C. R. and H. Rousselle (2020). Report and repeat: Investigating facebook's hate speech removal process. *First Monday*.

Center for Countering Digital Hate (2021). Failure to protect: how tech giants fail to act on user reports of antisemitism. Technical report.

Chandrasekharan, E., U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and

E. Gilbert (2017). You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction 1*(CSCW), 1–22.

Chen, Y. and D. Y. Yang (2019). The impact of media censorship: 1984 or brave new world? *American Economic Review 109*(6), 2294–2332.

Chowdhury, F. A., L. Allen, M. Yousuf, and A. Mueen (2020). On twitter purge: A retrospective analysis of suspended users. In *Companion Proceedings of the Web Conference 2020*, pp. 371–378.

Coppock, A. and O. A. McClellan (2019). Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & Politics 6*(1), 2053168018822174.

Correia-da Silva, J., B. Jullien, Y. Lefouili, and J. Pinho (2019). Horizontal mergers between multisided platforms: Insights from cournot competition. *Journal of Economics & Management Strategy 28*(1), 109–124.

Cowgill, B. and C. E. Tucker (2019). Economics, fairness and algorithmic bias. *preparation for: Journal of Economic Perspectives*.

Davidson, T., D. Warmsley, M. Macy, and I. Weber (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, Volume 11.

Dhakal, V., A. M. Feit, P. O. Kristensson, and A. Oulasvirta (2018). Observations on typing from 136 million keystrokes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–12.

Dixon, L., J. Li, J. Sorensen, N. Thain, and L. Vasserman (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 67–73.

Elliott, L. J., M. Ljubijanac, and D. Wieczorek (2019). The effect of screen size on reading speed: A comparison of three screens to print. In *International Conference on Applied Human Factors and Ergonomics*, pp. 103–109. Springer.

ElSherief, M., V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Volume 12.

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

Facebook (2021). Community standards enforcement report, second quarter 2021. https://about.fb.com/news/2021/08/community-standards-enforcement-report-q2-2021/. Accessed: 2021-10-23.

Fortuna, P. and S. Nunes (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR) 51*(4), 1–30.

Fortuna, P., J. Soler-Company, and L. Wanner (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management 58*(3), 102524.

Founta, A. M., C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Fujiwara, T., K. Müller, and C. Schwarz (2021). The effect of social media on elections: Evidence from the united states. Technical report, National Bureau of Economic Research.

Gadde, V. and K. Beykpour (2018). Setting the record straight on shadow banning. https://blog.twitter.com/en_us/topics/company/2018/Setting-the-record-straight-on-shadow-banning. Accessed: 2021-10-13.

Gentzkow, M. (2007). Valuing new goods in a model with complementarity: Online newspapers. *American Economic Review 97*(3), 713–744.

Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as data. *Journal of Economic Literature 57*(3), 535–74.

Giansiracusa, N. (2021). Facebook uses deceptive math to hide its hate speech problem. *Wired*.

Gillespie, T. (2018). *Custodians of the Internet*. Yale University Press.

Guhl, J. and J. Davey (2020). Hosting the 'holohoax': A snapshot of holocaust denial across social media. *The Institute for Strategic Dialogue*.

Haaland, I., C. Roth, and J. Wohlfart (2020). Designing information provision experiments.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis 20*(1), 25–46.

Han, X. and Y. Tsvetkov (2020). Fortifying toxic speech detectors against veiled toxicity. *arXiv preprint arXiv:2010.03154*.

Harrison, G. W., M. I. Lau, E. E. Rutström, and M. B. Sullivan (2005). Eliciting risk and time preferences using field experiments: Some methodological issues. In *Field experiments in economics*. Emerald Group Publishing Limited.

Hobbs, W. R. and M. E. Roberts (2018). How sudden censorship can increase access to information. *American Political Science Review 112*(3), 621–636.

Huang, J., D. Reiley, and N. Riabov (2018). Measuring consumer sensitivity to audio advertising: A field experiment on Pandora internet radio. *Available at SSRN 3166676*.

Hunt, M. G., R. Marx, C. Lipson, and J. Young (2018). No more fomo: Limiting social media decreases loneliness and depression. *Journal of Social and Clinical Psychology 37*(10), 751–768.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jhaver, S., C. Boylston, D. Yang, and A. Bruckman (2021). Evaluating the effectiveness of deplatforming as a moderation strategy on twitter.

Jourová, V. (2016). Code of conduct on countering illegal hate speech online: First results on implementation. Technical report, European Commission, Directorate-General for Justice and Consumers.

Kaye, D. (2019). *Speech Police: The Global Struggle to Govern the Internet*. Columbia Global Reports.

Kennedy, C. J., G. Bacon, A. Sahn, and C. von Vacano (2020). Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

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

Liu, Y., P. Yildirim, and Z. J. Zhang (2021). Implications of revenue models and technology for content moderation strategies. *Implications of Revenue Models and Technology for Content Moderation Strategies (November 23, 2021)*.

Matias, J., A. Johnson, W. E. Boesel, B. Keegan, J. Friedman, and C. DeTar (2015). Reporting, reviewing, and responding to harassment on twitter. *Available at SSRN 2602018*.

Melendez, S. (2020). Twitter automatically flags more than half of all tweets that violate its rules. *Fast Company*. Accessed: 2021-10-13.

Merrer, E. L., B. Morgan, and G. Trédan (2020). Setting the record straighter on shadow banning. *arXiv preprint arXiv:2012.05101*.

Montiel Olea, J. L. and M. Plagborg-Møller (2019). Simultaneous confidence bands: Theory, implementation, and an application to svars. *Journal of Applied Econometrics 34*(1), 1–17.

Mosquera, R., M. Odunowo, T. McNamara, X. Guo, and R. Petrie (2020). The economic effects of facebook. *Experimental Economics 23*(2), 575–602.

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

Müller, K. and C. Schwarz (2020b). From hashtag to hate crime: Twitter and anti-minority sentiment. *Available at SSRN 3149103*.

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

Munger, K. (2021). Don't@ me: Experimentally reducing partisan incivility on twitter. *Journal of Experimental Political Science 8*(2), 102–116.

Palin, K., A. M. Feit, S. Kim, P. O. Kristensson, and A. Oulasvirta (2019). How do people type on mobile devices? observations from a study with 37,000 volunteers. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, pp. 1–12.

Pew Research Center (2021). Social media use in 2021. `https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/`. Accessed: 2021-10-24.

Rauchfleisch, A. and J. Kaiser (2021). Deplatforming the far-right: An analysis of YouTube and BitChute. *Available at SSRN*.

Relia, K., Z. Li, S. H. Cook, and R. Chunara (2019). Race, ethnicity and national origin-based discrimination in social media and hate crimes across 100 us cities. In *Proceedings of the International AAAI Conference on Web and Social Media*, Volume 13, pp. 417–427.

Reynders, D. (2020). Countering illegal hate speech online: 5th evaluation of the code of conduct. Technical report, European Commission, Directorate-General for Justice and Consumers.

Roberts, M. and M. E. Roberts (2018). *Censored.* Princeton University Press.

Roth, J. and P. H. Sant'Anna (2021). Efficient estimation for staggered rollout designs. *arXiv preprint arXiv:2102.01291*.

Seyler, D., S. Tan, D. Li, J. Zhang, and P. Li (2021). Textual analysis and timely detection of suspended social media accounts.

Sherry, M. (2019). *Disability Hate Speech: Social, Cultural and Political Contexts*, Chapter Disablist hate speech online. Routledge.

Siegel, A. A. and V. Badaan (2020). # no2sectarianism: Experimental approaches to reducing sectarian hate speech online. *American Political Science Review 114*(3), 837–855.

Spence, A. M. (1975). Monopoly, quality, and regulation. *The Bell Journal of Economics*, 417–429.

Strossen, N. (2018). *Hate: Why we should resist it with free speech, not censorship.* Oxford University Press.

Tan, G. and J. Zhou (2021). The effects of competition and entry in multi-sided markets. *The Review of Economic Studies 88*(2), 1002–1030.

Twitter (2018). Serving healthy conversation. `https://blog.twitter.com/official/en_us/topics/product/2018/Serving_Healthy_Conversation.html`. Accessed: 2021-10-19.

Twitter (2020a). Q4 2020 letter to shareholders. `https://s22.q4cdn.com/826641620/files/doc_financials/2020/q4/FINAL-Q4'20-TWTR-Shareholder-Letter.pdf`. Accessed: 2021-10-13.

Twitter (2020b). Rules enforcement report. `https://transparency.twitter.com/en/reports/rules-enforcement.html#2020-jul-dec`. Accessed: 2021-10-13.

Twitter (2021a). Debunking twitter myths. `https://help.twitter.com/en/using-twitter/debunking-twitter-myths`. Accessed: 2021-10-13.

Twitter (2021b). Hateful conduct policy. `https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy`. Accessed: 2021-06-20.

Twitter (2021c). Notices on twitter and what they mean. `https://help.twitter.com/en/rules-and-policies/notices-on-twitter`. Accessed: 2021-10-13.

Twitter (2021d). Our range of enforcement options. `https://help.twitter.com/en/rules-and-policies/enforcement-options`. Accessed: 2021-06-21.

Twitter (2021e). Report abusive behavior. `https://help.twitter.com/en/safety-and-security/report-abusive-behavior`. Accessed: 2021-06-21.

Twitter (2021f). Response codes. `https://developer.twitter.com/ja/docs/basics/response-codes`. Accessed: 2021-06-21.

Urminsky, O., C. Hansen, and V. Chernozhukov (2016). Using double-lasso regression for principled variable selection. *Available at SSRN 2733374*.

Vidgen, B., S. Hale, S. Staton, T. Melham, H. Margetts, O. Kammar, and M. Szymczak (2020). Recalibrating classifiers for interpretable abusive content detection. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pp. 132–138.

Waldron, J. (2012). *The harm in hate speech.* Harvard University Press.

Weyl, E. G. (2010). A price theory of multi-sided platforms. *American Economic Review 100*(4), 1642–72.

Whalen, A. (2020). What Did Twitter Do to James Woods? The Story Behind the Trend. `https://www.newsweek.com/james-woods-twitter-feed-locked-freejameswoods-return-andrew-gillum\protect\discretionary{\char\hyphenchar\font}{}{}1494188`. Accessed: 2021-10-13.

White, A. and E. G. Weyl (2016). Insulated platform competition. *Available at SSRN 1694317*.

Wojcik, S. and A. Hughes (2019). Sizing up twitter users. *Pew Research Center 24*.

Wulczyn, E., N. Thain, and L. Dixon (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pp. 1391–1399.

Yang, K.-C., O. Varol, P.-M. Hui, and F. Menczer (2020). Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 34, pp. 1096–1103.

Yildirim, M. M., J. Nagler, R. Bonneau, and J. A. Tucker (2021). Short of suspension: How suspension warnings can reduce hate speech on twitter. *Perspectives on Politics*, 1–13.

Zannettou, S. (2021). " i won the election!": An empirical analysis of soft moderation interventions on twitter. *arXiv preprint arXiv:2101.07183*.

A Formal Propositions and Model Extensions

A.1 Overprovision or underprovision of moderation

The following proposition reproduces Spence's result for the model of Section 2.[51]

Proposition 1. *For fixed quantities T^θ and assuming that second-order conditions hold, the platform can overprovide or underprovide moderation relative to a surplus-maximizing planner. A sufficient condition for underprovision is that $P^\theta_{T^\theta} < 0$, for overprovision is that $P^\theta_{T^\theta} > 0$, and for efficient provision is that $P^\theta_{T^\theta} = 0$.*

Proof. A social planner chooses $\mathbf{T}$ and c to maximize total surplus W, which equals:

$$\begin{aligned}
W(\mathbf{T}, c) =& \underbrace{w\left(\int_0^{t^A} p^A(t, T^H, c)dt + \int_0^{t^H} p^H(T^A, t, c)dt - p^A T^A - p^H T^H\right)}_{\text{Consumer surplus}} \\
&+ \underbrace{a(p^A(\mathbf{T}, c)T^A + p^H(\mathbf{T}, c)T^H) - \phi(\mathbf{T}, c)}_{\text{Producer surplus}} \\
=& w\left(\int_0^{T^A} p^A(t, T^H, c)dt + \int_0^{T^H} p^H(T^A, t, c)dt\right) \\
&+ (a - w)(p^A(\mathbf{T}, c)T^A + p^H(\mathbf{T}, c)T^H) - \phi(\mathbf{T}, c).
\end{aligned}$$

The first two terms in the second equality are the areas below the inverse demand curves and the last term is the cost function, but the third term is new. This new term appears because the platform collects time with an opportunity cost w and sells it to advertisers for a price a. To the best of my knowledge there are no analyses that compare the price of advertisements of social media to the opportunity cost of time, so the magnitude of $a - w$ is unknown.[52]

[51] The model differs from Spence's framework in two ways. First, the monopolist sells two "products" instead of one. Second, there is a gap between the opportunity cost of time (w) and the value of time spent watching ads (a).

[52] A no-arbitrage argument suggests that $a \approx w$. Suppose that ad prices were higher than the opportunity cost of time. This creates incentives for companies to pay users to watch advertisements. While Becker and Murphy (1993) argue that this might not be profitable, since consumers would "buy" a large number of ads and ignore as many as possible, current technology might facilitate this. Indeed, websites like `adwallet.com` reward consumers for watching ads. On the other hand, if $w > a$, platforms would find it more profitable to have consumers complete tasks rather than show them ads; e.g., "Fill out this survey in order to proceed to your feed".

The first-order condition with respect to c from this problem is:

$$w\left(\int_0^{T^A}\frac{\partial p^A}{\partial c}dt+\int_0^{T^H}\frac{\partial p^H}{\partial c}dt\right)+(a-w)\left(\frac{\partial p^A}{\partial c}T^A+\frac{\partial p^H}{\partial c}T^H\right)=\frac{\partial\phi}{\partial c} \tag{7}$$

Suppose that $p^\theta_{t^\theta}<0$.[53] Then $\partial p^A(t,T^H,c)/\partial c>\partial p^A(\mathbf{T},c)/\partial c$ for all $t<T^A$ and likewise for H. Then, the left-hand side of equation (7) satisfies:

$$\begin{aligned}
&w\left(\int_0^{T^A}\frac{\partial p^A(t,T^H,c)}{\partial c}dt+\int_0^{T^H}\frac{\partial p^H(T^A,t,c)}{\partial c}dt\right)\\
&+(a-w)\left(\frac{\partial p^A(\mathbf{T},c)}{\partial c}T^A+\frac{\partial p^H(\mathbf{T},c)}{\partial c}T^H\right)\\
&>w\left(\frac{\partial p^A(\mathbf{T},c)}{\partial c}T^A+\frac{\partial p^H(\mathbf{T},c)}{\partial c}T^H\right)\\
&+(a-w)\left(\frac{\partial p^A(\mathbf{T},c)}{\partial c}T^A+\frac{\partial p^H(\mathbf{T},c)}{\partial c}T^H\right)\\
&=a\left(\frac{\partial p^A(\mathbf{T},c)}{\partial c}T^A+\frac{\partial p^H(\mathbf{T},c)}{\partial c}T^H\right),
\end{aligned}$$

which is identical to the left-hand side of equation (4). Since equations (4) and (7) both have $\partial\phi/\partial c$ on the right-hand side, this means that the planner's first-order condition is above the monopolist's one for fixed t^θ and all c: $\partial W(\mathbf{T},c)/\partial c<\partial\pi(\mathbf{T},c)/\partial c$. Assuming that second-order conditions hold, this means that the root of the planner's first-order condition, $c^{planner}$, is higher than the root of the monopolist's condition, $c^{platform}$, so there is under-provision of moderation. Figure A.1 illustrates the proof.

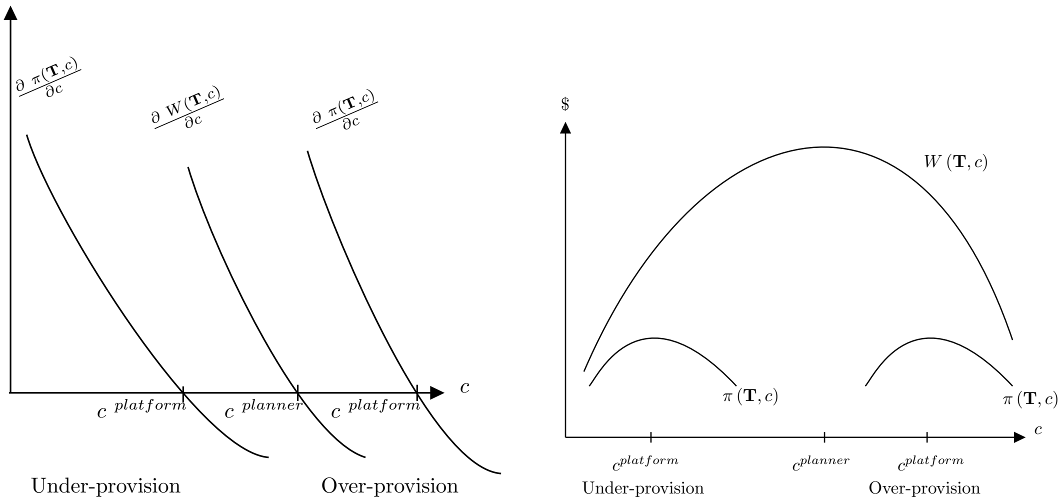
□

A.2 Generalization to Multiple Platforms

Assume without loss of generality that there are two platforms $j\in\{1,2\}$. The solution concept of the model is a Cournot equilibrium as in Correia-da Silva et al. (2019).[54] First, platforms simultaneously set the amount of content T^θ_j on each side of the market and the moderation rates c_j. Then, given the quantities and moderation

[53] The proof is analogous for the opposite case.

[54] Alternative solution concepts are flat pricing (Tan and Zhou, 2021) and insulating equilibrium (White and Weyl, 2016). See Correia-da Silva et al. (2019) for more discussion.



(a) First-order conditions

(b) Surplus and profits

Figure A.1: Illustration of moderation overprovision and underprovision

rates, prices adjust to equate demand and supply on each platform.

A fraction μ^θ of users are of type $\theta \in \{A, H\}$. Consumers are now characterized by the parameter vectors $\boldsymbol{\gamma}^\theta = (\gamma_1^\theta, \gamma_2^\theta, \delta_1^\theta, \delta_2^\theta)$. The γ's govern how utility responds to spillovers and the δ's govern membership benefits. As in (Weyl, 2010), the conditional density of membership benefits has full support. Users decide whether to join one of the platforms or neither. Below I discuss an extension to a multi-homing case. Once consumers join a platform, they decide how much time to spend on it. If they join platform j, they obtain membership benefits δ_j^θ and indirect utility:

$$v_j^\theta(\mathbf{T}_j, c_j, p_j^\theta, \gamma_j^\theta) = \max_{t \in [0,T]} \ u^\theta(t, \mathbf{T}_j, c_j; \gamma_j^\theta) - t \times w(1 + p_j^\theta),$$

where $\mathbf{T}_j = (T_j^A, T_j^H)$

Define the vectors $\mathbf{T} = (\mathbf{T}_1, \mathbf{T}_2)$, $\mathbf{T}^\theta = (T_1^\theta, T_2^\theta)$, $\boldsymbol{p}^\theta = (p_1^\theta, p_2^\theta)$, and $\boldsymbol{c} = (c_1, c_2)$, and the set of types that decide to use platform j as:

$$\overline{\boldsymbol{\gamma}}_j^\theta(\mathbf{T}, \boldsymbol{c}, \boldsymbol{p}^\theta) \equiv \left\{ \boldsymbol{\gamma}^\theta : v_j^\theta(\mathbf{T}_j, c_j, p_j^\theta, \gamma_j^\theta) + \delta_j^\theta \geq \max\{ v_{-j}^\theta(\mathbf{T}_{-j}, c_{-j}, p_{-j}^\theta, \gamma_{-j}^\theta) + \delta_{-j}^\theta, 0 \} \right\},$$

where $-j$ denotes the other platform. Let $t_j(\mathbf{T}_j, c_j, p_j^\theta, \gamma_j^\theta)$ be the optimal time spent on platform j. Aggregate demands are:

$$T_j^\theta(\mathbf{T}, \boldsymbol{c}, \boldsymbol{p}^\theta) = \mu^\theta \int_{\overline{\boldsymbol{\gamma}}^\theta(\mathbf{T}, \boldsymbol{c}, \boldsymbol{p}^\theta)} t_j(\mathbf{T}_j, c_j, p_j^\theta, \gamma_j^\theta) f^\theta(\boldsymbol{\gamma}^\theta) d\boldsymbol{\gamma}^\theta$$

The consumer equilibrium constraints are, for all j and θ:

$$t_j^\theta = T_j^\theta(\mathbf{T}, \boldsymbol{c}, \boldsymbol{p}^\theta)$$

Inverting the demand curves is not as straightforward as in Weyl (2010), since demands now depend on the other platform's prices. We can, however, use the global inverse function theorem from Berry et al. (2013) to obtain the twice-continuously differentiable inverse demands $P_j^\theta(\mathbf{T}, \boldsymbol{c})$.[55]

The problem of platform j is now:

$$\max_{T_j^A, T_j^H, c_j} \pi_j(\mathbf{T}, \boldsymbol{c}) \equiv a \left(P_j^A(\mathbf{T}, \boldsymbol{c})\, T_j^A + P_j^H(\mathbf{T}, \boldsymbol{c})\, T_j^H \right) - \phi_j(\mathbf{T}_j, c_j).$$

The first-order condition with respect to the moderation rate is identical to equation (4), but using residual inverse demands instead of the market inverse demand curve:

$$a \left(\frac{\partial P_j^A}{\partial c_j}\, T_j^A + \frac{\partial P_j^H}{\partial c_j}\, T_j^H \right) = \frac{\partial C_j}{\partial c_j}$$

Hence, the same intuition of the platform's moderation decision holds in a model with two platforms. Moderation is a quality decision that allows platforms to increase their advertising revenue. The increase in ad revenue is the weighted change in willingness to pay of both types of users.

The following proposition shows that it is sufficient to measure the change in surplus on a sample of existing consumers; one can ignore the change in marginal users since they get zero surplus by definition.

Proposition 2. *The derivative of consumer surplus with respect to the moderation rate of platform j equals the average derivative of consumer surplus among users of*

[55] Note that demands T_j^θ are twice-continuously differentiable and strictly decreasing in prices p_{-j}^θ and weakly decreasing in prices $p_{-j}^{-\mathcal{I}}$ and $p_j^{-\mathcal{I}}$, where $-\mathcal{I}$ denotes the other side. Moreover, the demand of the outside option is stricly increasing in all prices. Hence, this model satisfies all the conditions of Corollary 2 from Berry et al. (2013).

that platform:

$$\frac{\partial CS(\mathbf{T},\boldsymbol{c},\boldsymbol{p})}{\partial c_j} = \sum_\theta \mu^\theta \int_{\overline{\boldsymbol{\gamma}}_j^\theta(\mathbf{T},\boldsymbol{c},\boldsymbol{p}^\theta)} \frac{\partial v_j^\theta(\mathbf{T}_j, c_j, p_j^\theta, \gamma_j^\theta)}{\partial c_j} f^\theta(\boldsymbol{\gamma}^\theta) d\boldsymbol{\gamma}^\theta.$$

Proof. Define the membership benefit from joining platform $-j$ relative to the membership benefit from j as $\widetilde{\delta}_{-j}^\theta \equiv \delta_{-j}^\theta - \delta_j^\theta$. Define also the vector of network parameters of both platforms $\gamma^\theta \equiv (\gamma_1^\theta, \gamma_2^\theta)$, the vector of parameters $\widetilde{\boldsymbol{\gamma}}^\theta \equiv (\gamma^\theta, \delta_j^\theta, \widetilde{\delta}_{-j}^\theta)$ and the distribution of types $\widetilde{f}^\theta(\widetilde{\boldsymbol{\gamma}}^\theta) \equiv f^\theta(\gamma^\theta, \delta_j^\theta, \delta_{-j}^\theta + \delta_j^\theta)$. The membership benefits of those users who join platform j are bounded as follows:

$$\begin{aligned}
\delta_j^\theta &\geq -v_j^\theta(\mathbf{T}_j, c_j, p_j^\theta, \gamma_j^\theta), \\
\widetilde{\delta}_{-j} &\leq v_j^\theta(\mathbf{T}_j, c_j, p_j^\theta, \gamma_j^\theta) - v_{-j}^\theta(\mathbf{T}_{-j}, c_{-j}, p_{-j}^\theta, \gamma_j^\theta)
\end{aligned}$$

Likewise, the bounds of the membership benefits of those users who join platform $-j$ are:

$$\begin{aligned}
\delta_j^\theta &\geq -v_{-j}^\theta(\mathbf{T}_{-j}, c_{-j}, p_{-j}^\theta, \gamma_j^\theta) - \widetilde{\delta}_{-j}, \\
\widetilde{\delta}_{-j} &\geq v_j^\theta(\mathbf{T}_j, c_j, p_j^\theta, \gamma_j^\theta) - v_{-j}^\theta(\mathbf{T}_{-j}, c_{-j}, p_{-j}^\theta, \gamma_j^\theta)
\end{aligned}$$

Omitting the arguments of v_j^θ and v_{-j}^θ for brevity, the consumer surplus is:

$$\begin{aligned}
CS(\mathbf{T},\boldsymbol{c},\boldsymbol{p}) = \sum_\theta \mu^\theta \Bigg(& \underbrace{\int \int_{-\infty}^{v_j^\theta - v_{-j}^\theta} \int_{-v_j^\theta}^{\infty} \left(v_j^\theta + \delta_j^\theta\right) \widetilde{f}^\theta(\widetilde{\boldsymbol{\gamma}}^\theta) d\delta_j^\theta d\widetilde{\delta}_{-j}^\theta d\gamma^\theta}_{\text{Surplus of } j\text{'s consumers}} \\
& + \underbrace{\int \int_{v_j^\theta - v_{-j}^\theta}^{\infty} \int_{-v_{-j}^\theta - \widetilde{\delta}_{-j}^\theta}^{\infty} \left(v_{-j}^\theta + \widetilde{\delta}_{-j}^\theta + \delta_j^\theta\right) \widetilde{f}^\theta(\widetilde{\boldsymbol{\gamma}}^\theta) d\delta_j^\theta d\widetilde{\delta}_{-j}^\theta d\gamma^\theta}_{\text{Surplus of } -j\text{'s consumers}} \Bigg) \qquad (8)
\end{aligned}$$

Use the Leibniz integral rule to differentiate the first row after the equality sign from the previous expression with respect to c_j:

$$\begin{aligned}
&\int \frac{\partial v_j^\theta}{\partial c_j} \int_{-v_j^\theta}^{\infty} \left(v_j^\theta + \delta_j^\theta\right) \widetilde{f}^\theta(\gamma^\theta, \delta_j^\theta, v_j^\theta - v_{-j}^\theta) d\delta_j^\theta d\gamma^\theta \\
&+ \int \int_{-\infty}^{v_j^\theta - v_{-j}^\theta} \frac{\partial v_j^\theta}{\partial c_j} \left(\underbrace{v_j^\theta - v_j^\theta}_{=0} \right) \widetilde{f}^\theta(\gamma^\theta, -v_j^\theta, \widetilde{\delta}_{-j}^\theta) d\widetilde{\delta}_{-j}^\theta d\gamma^\theta \\
&+ \int \int_{-\infty}^{v_j^\theta - v_{-j}^\theta} \int_{-v_j^\theta}^{\infty} \frac{\partial v_j^\theta}{\partial c_j} \widetilde{f}^\theta(\widetilde{\boldsymbol{\gamma}}^\theta) d\delta_j^\theta d\widetilde{\delta}_{-j}^\theta d\gamma^\theta
\end{aligned} \tag{9}$$

Likewise, differentiating the second row of equation (8):

$$\int -\frac{\partial v_j^\theta}{\partial c_j} \int_{-v_j^\theta}^{\infty} \left(v_j^\theta + \delta_j^\theta\right) \widetilde{f}^\theta(\gamma^\theta, \delta_j^\theta, v_j^\theta - v_{-j}^\theta) d\delta_j^\theta d\gamma^\theta \tag{10}$$

The first row from equation (9) cancels with equation (10), so adding these two expressions together gives

$$\int \int_{-\infty}^{v_j^\theta - v_{-j}^\theta} \int_{-v_j^\theta}^{\infty} \frac{\partial v_j^\theta}{\partial c_j} \widetilde{f}^\theta(\widetilde{\boldsymbol{\gamma}}^\theta) d\delta_j^\theta d\widetilde{\delta}_{-j}^\theta d\gamma^\theta,$$

which equals the average derivative of consumer surplus among users of j, $\boldsymbol{\gamma}^\theta \in \overline{\boldsymbol{\gamma}}_j^\theta(\mathbf{T}, \boldsymbol{c}, \boldsymbol{p}^\theta)$. $\square$

The previous single-homing model can be extended to allow multi-homing and flexible substitution or complementarity patterns by considering bundles of platforms as different choices, by using the approach of Gentzkow (2007) as Berry et al. (2013) suggest. Proposition 2 still holds in such a model, but the indirect utility v_j^θ would represent the utility from using any bundle of platforms that includes j, and v_{-j}^θ would correspond to bundles without j. Hence, in that setting, $\partial v_j^\theta / \partial c_j$ would be the derivative of the surplus of using social media, not just platform j, for users who join any bundle of platforms that includes j.

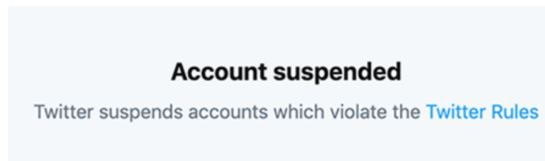
In a multi-platform model, the steps to test whether a platform under-provides or over-provides moderation, for fixed quantities, are as follows. First, introduce a small change in the moderation rate of the desired platform. Second, compute the average change in consumer surplus from using social media—not just the surplus from using the platform—among representative consumers of the platform.

B Experimental design

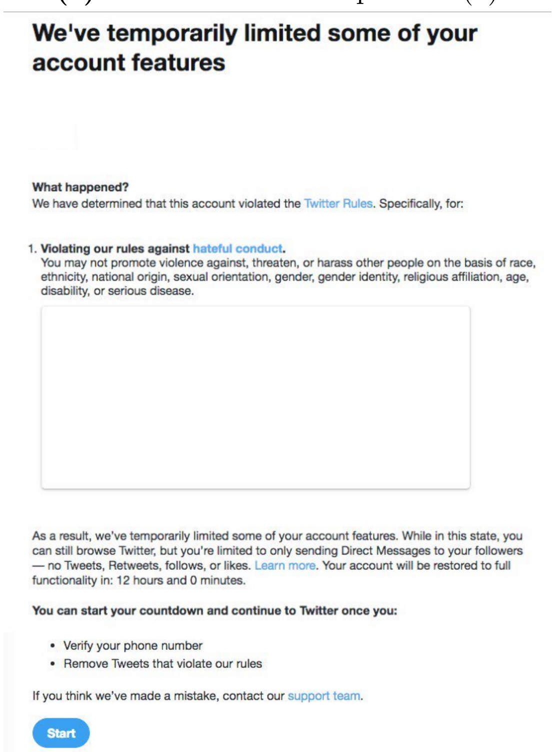
(a) Public notice of Tweet removal

(b) Public notice of shadowban

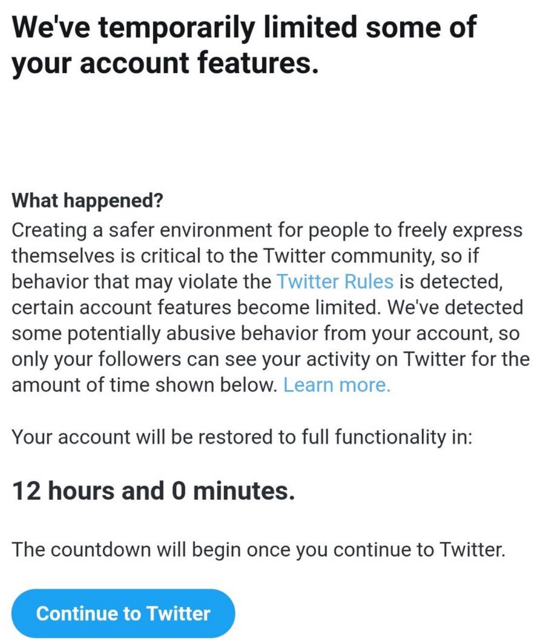
(c) Public notice of suspension (1)



(d) Public notice of suspension (2)



(e) Notification of a Tweet removal



(f) Notification of a locked account

Figure B.1: Public notices and notifications

Notes: This figure includes images of public notices of different sanctions and an example of a notification that a users receive when their account is locked.

Table B.1: Query list

Holocaust	holohoax, h0lohoax, hol0hoax, holoh0ax, holoho4x, holoho@x, holoho*x, h0l0h0ax, h0l0h04x, holo hoax, h0lo hoax, hol0 hoax, holo h0ax, holo ho4x, holo ho@x, holo ho*x, h0l0 h0ax, h0l0 h04x, holo-hoax, h0lo-hoax, hol0-hoax, holo-h0ax, holo-ho4x, holo-ho@x, holo-ho*x, h0l0-h0ax, h0l0-h04x
Disability	youre a retard, youre a r3tard, youre a r*tard, youre a ret*rd, youre a returd, youre a ret@rd, youre a ret4rd, youre a r3t4rd, youre a r3t@rd, your a retard, your a r3tard, your a r*tard, your a ret*rd, your a returd, your a ret@rd, your a ret4rd, your a r3t4rd, your a r3t@rd, you're a retard, you're a r3tard, you're a r*tard, you're a ret*rd, you're a returd, you're a ret@rd, you're a ret4rd, you're a r3t4rd, you're a r3t@rd, ure a retard, ure a r3tard, ure a r*tard, ure a ret*rd, ure a returd, ure a ret@rd, ure a ret4rd, ure a r3t4rd, ure a r3t@rd, ur a retard, ur a r3tard, ur a r*tard, ur a ret*rd, ur a returd, ur a ret@rd, ur a ret4rd, ur a r3t4rd, ur a r3t@rd, u're a retard, u're a r3tard, u're a r*tard, u're a ret*rd, u're a returd, u're a ret@rd, u're a ret4rd, u're a r3t4rd, u're a r3t@rd youre retarded, youre r3tarded, youre r*tarded, youre ret*rded, youre returded, youre ret@rded, youre ret4rded, youre r3t4rded, youre r3t@rded, you're retarded, you're r3tarded, you're r*tarded, you're ret*rded, you're returded, you're ret@rded, you're ret4rded, you're r3t4rded, you're r3t@rded, ure retarded, ure r3tarded, ure r*tarded, ure ret*rded, ure returded, ure ret@rded, ure ret4rded, ure r3t4rded, ure r3t@rded, u're retarded, u're r3tarded, u're r*tarded, u're ret*rded, u're returded, u're ret@rded, u're ret4rded, u're r3t4rded, u're r3t@rded youre a retarded, youre a r3tarded, youre a r*tarded, youre a ret*rded, youre a returded, youre a ret@rded, youre a ret4rded, youre a r3t4rded, youre a r3t@rded, your a retarded, your a r3tarded, your a r*tarded, your a ret*rded, your a returded, your a ret@rded, your a ret4rded, your a r3t4rded, your a r3t@rded, you're a retarded, you're a r3tarded, you're a r*tarded, you're a ret*rded, you're a returded, you're a ret@rded, you're a ret4rded, you're a r3t4rded, you're a r3t@rded, ure a retarded, ure a r3tarded, ure a r*tarded, ure a ret*rded, ure a returded, ure a ret@rded, ure a ret4rded, ure a r3t4rded, ure a r3t@rded, ur a retarded, ur a r3tarded, ur a r*tarded, ur a ret*rded, ur a returded, ur a ret@rded, ur a ret4rded, ur a r3t4rded, ur a r3t@rded, u're a retarded, u're a r3tarded, u're a r*tarded, u're a ret*rded, u're a returded, u're a ret@rded, u're a ret4rded, u're a r3t4rded, u're a r3t@rded

Hello,

Twitter is required by German law to provide notice to users who are reported by people from Germany via the Network Enforcement Act reporting flow.

We have received a complaint regarding your account, @handle , for the following content:

Tweet ID:
Tweet Text:

We have investigated the reported content and have found that it is not subject to removal under the Twitter Rules (https://support.twitter.com/articles/18311) or German law. Accordingly, we have not taken any action as a result of this specific report.

Sincerely,

Twitter

Figure B.2: Screenshot of a notification of a user report

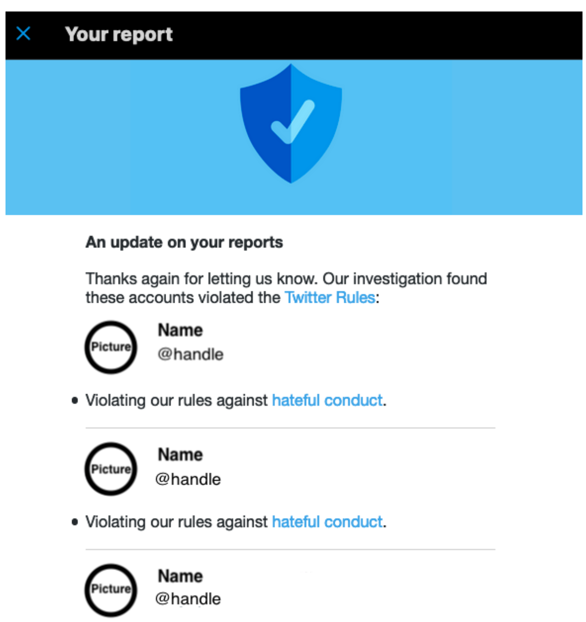


Figure B.3: Screenshot of an update on reports

Table B.2: Variable definition

Variable	Definition
Account years	Years from account creation date until measurement date
Tweets per day	Statuses count divided by days since account creation
Likes per day	Likes count divided by days since account creation
Followers	Number of accounts that follow a user
Followed	Number of accounts that the user follows
Bot score	Probability of being a bot, from Botometer API
Is bot	Indicates whether bot score ≥ 0.5
Initial shadow ban	Whether an account is shadow banned at the time of sampling
Word count	Number of words in a tweet
Is toxic	Indicates whether toxicity ≥ 0.8
Is hate (MTurk)	Indicates whether a majority of MTurkers label the post as hate
Is reply	Indicates whether the tweet is a reply to another user
Is attack (MTurk)	Indicates whether the majority of MTurkers consider the post to be an attack on another user
Is quote	Indicates whether the tweet is a quote to another user
Is mention	Indicates whether the tweet mentions another user
Has media	Indicates whether the tweet contains a video or picture
Disability key word	Indicates whether the tweet contains the expression "r*t*rd"
Holocaust key word	Indicates whether the tweet contains the expressions "h*l*h**x", "h*l*c**st", "jew"
Tweet from phone	Indicates whether the source of the tweet is Twitter for iPhone or Twitter for Android
Has description	Indicates whether a profile has a description
Has location	Indicates whether a profile has a location
Default picture	Indicates whether a profile has a default profile picture
Is verified	Indicates whether an account is verified
Has Instagram	Indicates whether a profile description, location or URL contains an Instagram handle
Has backup	Indicates whether a profile description, location or URL contains an alternative or backup Twitter handle
Has pronouns	Indicates whether a profile description or location contains pronouns or a `carrd.co` link
Under 18	Indicates whether a profile description or location contains numbers 13 to 17 (in number or word), years 2003 to 2008 or words like "minor" or "teen"
Previous toxicity	Indicates whether any of a user's most recent 50 tweets has toxicity ≥ 0.8
Previous disability	Indicates whether any of a user's most recent 50 tweets has the expression "r*t*rd"
Previous Holocaust	Indicates whether any of a user's most recent 50 tweets has the expressions "h*l*h**x", "h*l*c**st", "jew'

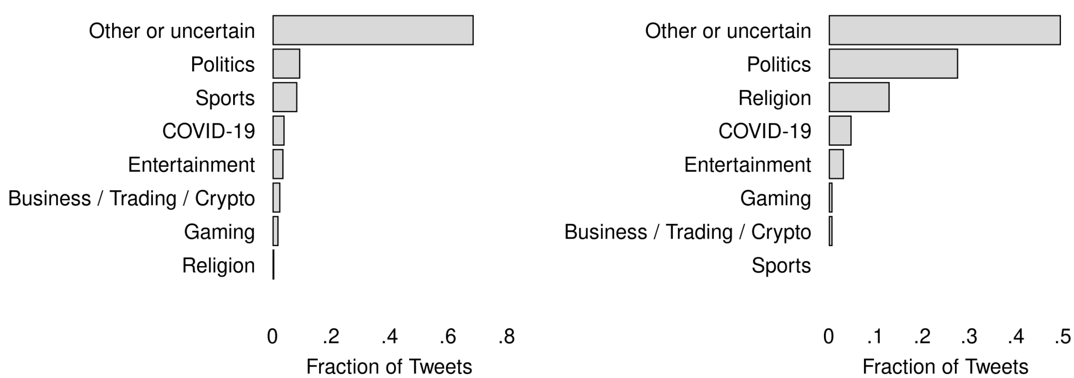
Table B.3: Balance in the reporting experiment

Characteristic	Control		Treatment		Difference	
	Mean	SD	Mean	SD	Normalized	p-value
Observations	3,074		3,074			
Accounts						
Account years	3.21	3.4	3.23	3.5	-0.01	0.77
Tweets per day	11.19	24.2	12.05	25.2	-0.03	0.20
Likes per day	23.39	50.2	24.95	51.6	-0.03	0.22
Followers	517.07	3,439.5	752.64	6,476.9	-0.05	0.08
Followed	426.84	751.4	440.65	946.0	-0.02	0.55
Bot score	0.24	0.1	0.24	0.1	0.01	0.54
Initial shadow ban	0.71	0.5	0.71	0.5	0.01	0.78
Tweets						
Word count	16.02	13.2	15.93	13.4	0.01	0.80
Is toxic	0.81	0.4	0.80	0.4	0.04	0.19
Is hate (MTurk)	0.31	0.5	0.30	0.5	0.00	0.81
Is reply	0.84	0.4	0.84	0.4	0.00	0.95
Is attack (MTurk)	0.78	0.4	0.78	0.4	-0.02	0.42
Is quote	0.07	0.3	0.07	0.3	0.02	0.53
Is mention	0.85	0.4	0.85	0.4	0.01	0.76
Has media	0.04	0.2	0.04	0.2	-0.04	0.13
Tweet from phone	0.80	0.4	0.78	0.4	0.03	0.25
Profiles						
Has description	0.82	0.4	0.82	0.4	-0.01	0.70
Has location	0.51	0.5	0.51	0.5	-0.01	0.53
Default picture	0.04	0.2	0.03	0.2	0.02	0.42
Is verified	0.00	0.0	0.00	0.0	-0.02	0.51
Has Instagram	0.01	0.1	0.02	0.1	-0.02	0.46
Has backup	0.01	0.1	0.01	0.1	0.04	0.15
Timelines						
Previous toxicity	0.94	0.2	0.93	0.2	0.01	0.83
Previous disability	0.39	0.5	0.39	0.5	0.01	0.74
Previous Holocaust	0.10	0.3	0.10	0.3	-0.01	0.98
Joint tests/differences						
F-test (p-value)						0.70
Multivariate normalized difference					0.12	

Notes: Columns 2 to 5 display means and standard deviations (SD). Column 6 displays normalized differences (Imbens and Rubin, 2015); all variables have differences below the recommended 0.25. Column 7 has p-values from regressions of characteristics on a treatment dummy and strata fixed-effects. F-tests are from regressions of a treatment indicator on pre-treatment variables.

Figure B.4: Screenshot of a reply

Notes: Some Tweets in my sample are replies or comments to other users' Tweets.



(a) Disability

(b) Holocaust

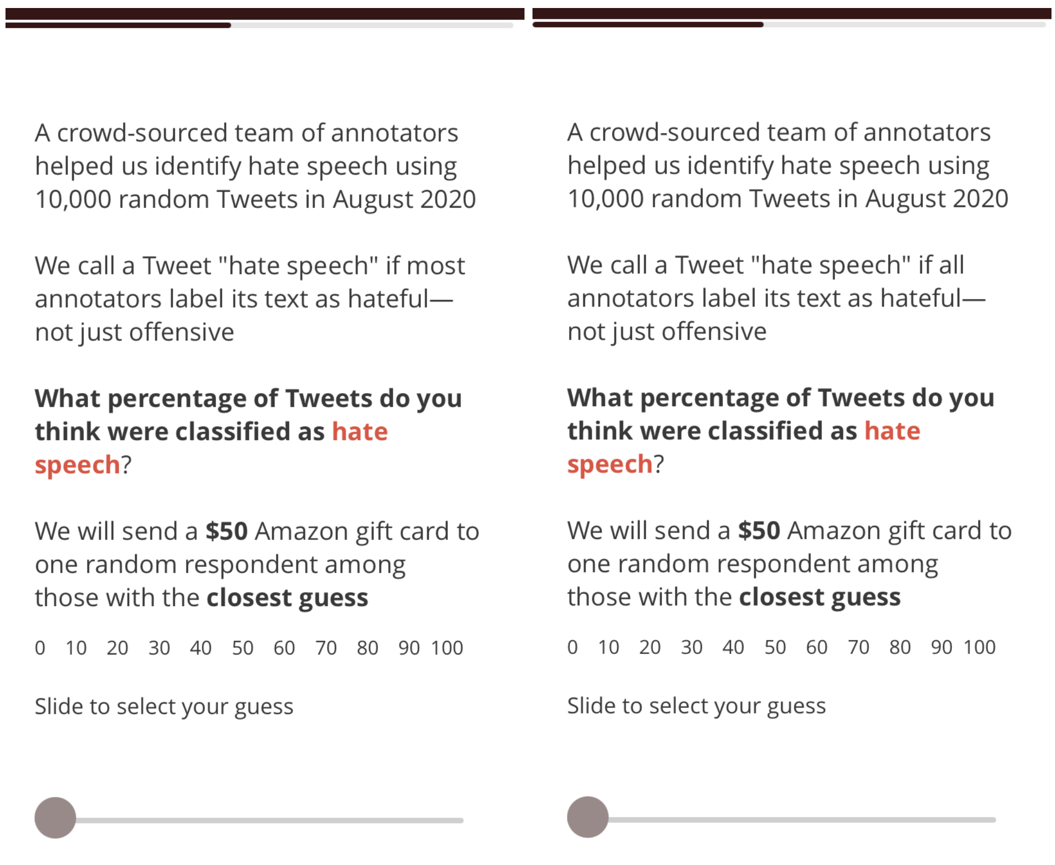
Figure B.5: Topic classification by slur

Notes: Each figure presents the distribution of Tweets by their main topic. Three MTurk workers read each Tweet and decided its most relevant topic among the eight options in the figures. The main topic is the one that two or three workers agreed upon. If there was no agreement, the topic of the Tweet is set to "Other or uncertain".

Table B.4: Reporting accounts summary statistics

	Account										
	1	2	3	4	5	6	7	8	9	10	11
Accounts											
City	CHI	CHI	NYC	MIA	LA	LA	DAL	SF	ATL	CHI	DC
Email	Yes	Yes	Yes	No	Yes	No	Yes	No	No	Yes	No
Phone	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mobile	Yes	No	No	No	No	Yes	No	No	Yes	No	Yes
App	Yes	No	No	No	No	No	Yes	No	No	Yes	No
Shadow ban	No	No	No	No	Yes	No	No	No	Yes	No	No
Account yrs	2.7	2.5	2.5	2.3	2.3	2.3	2.3	2.2	0.3	9.1	0.1
Tweets/mth	0	1.1	0.3	2.4	2.2	0.1	0.2	0.3	3.5	0.5	4.1
Likes/mth	0	1.0	0.6	1.8	1.8	0.8	1.1	0.5	4.7	1.2	8.2
Followers	0	18	3	2	1	0	0	0	0	168	1
Followed	6	22	28	48	43	25	19	16	14	134	17
Bot score	.	0.4	0.5	0.2	0.3	0.4	0.4	0.5	0.4	0.2	0.5
Profiles											
Description	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Location	Yes	Yes	No	Yes	Yes	No	No	Yes	No	Yes	Yes
Default pic	No	No	No	No	No	No	No	No	No	No	No
Verified	No	No	No	No	No	No	No	No	No	No	No
Protected	No	No	No	No	No	No	No	No	No	No	No

Notes: Each column corresponds to one of the 11 Twitter accounts used for the reporting treatment. City is the location of the virtual private network used for reporting. Email and Phone indicate whether the account had an associated email and phone number, respectively. Mobile indicates whether reporting was done using a phone or a computer. App indicates whether the account was accessed using the official Twitter app or a browser. Data gathered in August, 2021.



(a) Low moderation

(b) High moderation

Figure B.6: Instructions and elicitation of beliefs about prevalence

Table B.5: Balance in the welfare experiment

Characteristic	Control		Treatment		N. Dif.	p-value
	Mean	SD	Mean	SD		
Observations	1,515		1,512			
Demographics						
Age	38.05	12.8	38.10	12.3	-0.00	0.92
Female	0.45	0.5	0.45	0.5	0.01	0.44
College graduate +	0.32	0.5	0.31	0.5	0.03	0.35
Some college	0.33	0.5	0.33	0.5	-0.01	0.73
White non-Hispanic	0.67	0.5	0.69	0.5	-0.05	0.06
Black non-Hispanic	0.15	0.4	0.14	0.4	0.01	0.82
Hispanic	0.09	0.3	0.08	0.3	0.04	0.27
Asian non-Hispanic	0.03	0.2	0.02	0.2	0.06	0.10
Northeast	0.22	0.4	0.25	0.4	-0.07	0.07
Midwest	0.18	0.4	0.18	0.4	0.01	0.84
South	0.39	0.5	0.36	0.5	0.06	0.12
Republican	0.23	0.4	0.22	0.4	0.03	0.47
Democrat	0.52	0.5	0.54	0.5	-0.04	0.33
Christian	0.62	0.5	0.61	0.5	0.02	0.54
Jewish	0.02	0.1	0.02	0.1	-0.00	0.91
Muslim	0.04	0.2	0.04	0.2	-0.01	0.78
Buddhist or Hindu	0.01	0.1	0.01	0.1	0.02	0.65
Income	64.09	32.9	64.22	33.4	-0.00	0.90
Minority	0.48	0.5	0.48	0.5	-0.00	1.00
Twitter / Social media						
Daily hours on Twitter	1.52	2.3	1.52	2.2	-0.00	0.99
Provided handle	0.64	0.5	0.64	0.5	-0.00	1.00
User exists	0.47	0.5	0.47	0.5	-0.01	0.69
Tweets per day	1.29	6.4	1.79	10.7	-0.06	0.29
Likes per day	1.82	6.3	2.86	18.1	-0.08	0.14
Account years	8.06	4.3	7.80	4.2	0.06	0.25
Platforms other than Twitter	5.12	2.1	5.10	2.0	0.01	0.74
Has been harassed online	0.28	0.5	0.29	0.5	-0.01	0.85
Prevalence of hate in feed	20.06	23.1	20.85	24.5	-0.03	0.31
Moderation						
Has been sanctioned	0.23	0.4	0.23	0.4	0.00	1.00
Has reported	0.36	0.5	0.37	0.5	-0.00	0.88
Has been reported	0.12	0.3	0.12	0.3	0.03	0.36
Beliefs						
Prevalence of hate	36.73	25.7	36.12	26.2	0.02	0.51
Likelihood of moderation	39.73	28.6	40.91	28.7	-0.04	0.25
Joint tests/differences						
F-test (p-value)						0.33
Multivariate normalized difference					0.25	

Notes: Columns 2 to 5 display means and standard deviations (SD). Column 6 displays normalized differences (Imbens and Rubin, 2015). Column 7 has p-values from a regression of characteristics on treatment and strata fixed-effects. F-tests are from regressions of a treatment indicator on characteristics.

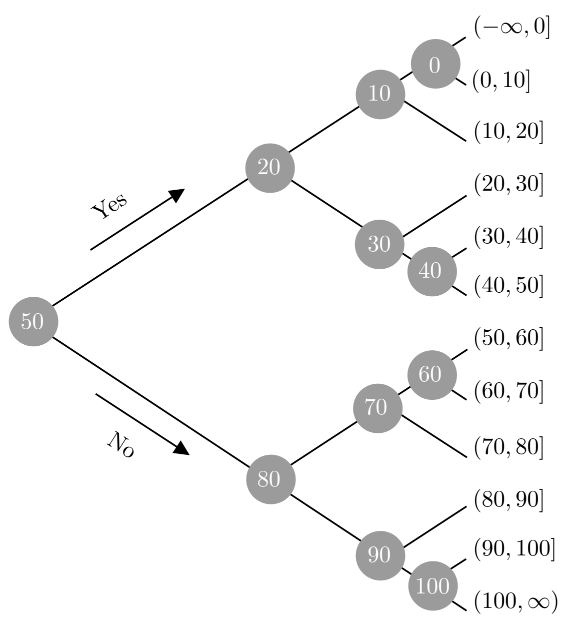


Figure B.7: Iterative multiple price list

Notes: The circles denote compensation (Amazon gift card) offers to deactivate social media. The intervals correspond to the willingness to accept.

C Data appendix

C.1 Measurement of sanctions

The "lookup statuses" or "lookup users" endpoints of the Twitter API indicate when a tweet or account go missing. Among missing accounts and statuses, the "show users" or "show statuses" endpoints of the API return an error code that details why they were missing (see Twitter (2021f) for a full list of error codes). With the error code information one can measure the following events:

- Twitter required the removal of a post, but it has not been removed by the user. This is reflected in a missing status with error code 421.
- Twitter required the removal of a post, and it has been removed by the user. This is reflected in a missing status with error code 422. After some days, the status transitions to code 144 (deleted status). Twitter claims that the notice will be available 14 days after the tweet is removed (Twitter, 2021d) but empirically it seems like this period varies.
- A post is missing because the user deleted it. This is reflected in a missing status with error code 144.
- A post is missing because the user protected their account or because the user blocked my developer account. This is reflected in a missing status with error code 179 or 136, respectively. Protected accounts are also detected with the lookup users endpoint. It is rare to encounter a user that blocks my developer account. Most likely is due to users mass-blocking all the followers of some famous account.
- A post and the account are missing because the user is suspended. This is reflected in a missing user and potentially missing status with code 63 (Chowdhury et al., 2020).
- A post and the account are missing because the user deleted their account. This is reflected in a missing user with code 50.

C.2 Figures and Tables

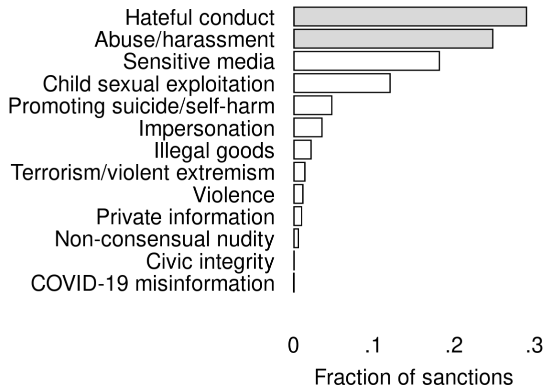


Figure C.1: Histogram of sanctions by rule violation

Notes: This figure plots the fraction of sanctions (actioned accounts) by the type of rule violation. It uses data from the second half of 2020 from Twitter's Transparency Rules Enforcement Report (Twitter, 2020b).

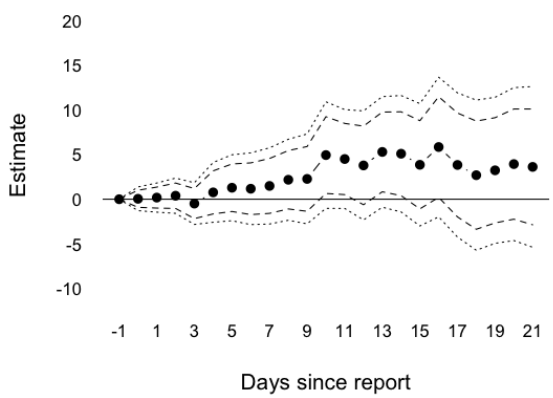


Figure C.2: Effect on hours since last post

Notes: This figure presents dynamic treatment effects on the number of hours since the last post at midnight of every day after reporting. Pointwise confidence intervals are dashed and sup-t confidence bands are dotted.

Table C.1: Effects of reporting on other observable sanctions and self-censorship

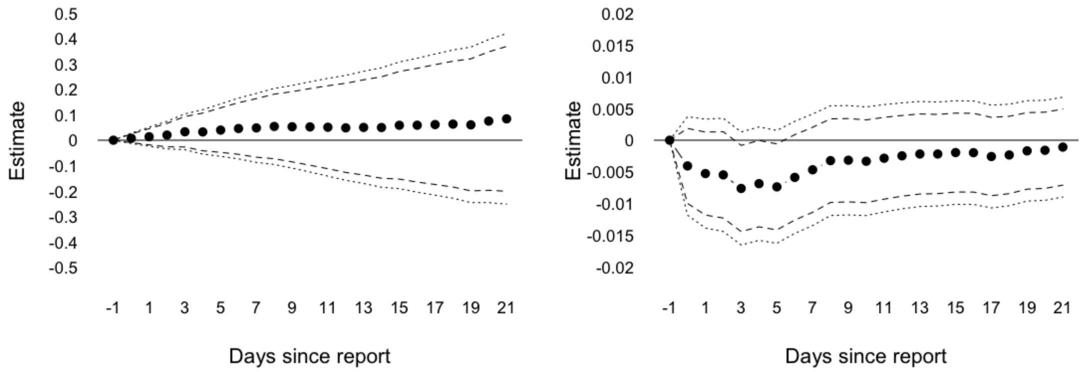
	Suspensions			Shadow-bans			Missing Other Tweets		
Panel A: other Twitter sanctions									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment	-0.000	-0.000	-0.000	0.001	0.001	-0.002	0.004	0.004	0.004
	(0.006)	(0.006)	(0.006)	(0.012)	(0.012)	(0.011)	(0.005)	(0.005)	(0.005)
y Mean	0.05	0.05	0.05	0.26	0.26	0.26	0.05	0.05	0.05
y SD	0.22	0.22	0.22	0.44	0.44	0.44	0.18	0.18	0.18
R^2	0.00	0.03	0.03	0.00	0.02	0.10	0.00	0.02	0.03
Obs.	6,148	6,134	6,134	5,692	5,675	5,675	5,381	5,360	5,360
Panel B: self-censorship									
	Tweet deletion			Account deletion			Protecting account		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment	0.006	0.005	0.005	0.001	0.001	0.001	0.000	0.000	-0.000
	(0.005)	(0.005)	(0.005)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)
y Mean	0.03	0.03	0.03	0.02	0.02	0.02	0.03	0.03	0.03
y SD	0.18	0.18	0.18	0.13	0.13	0.13	0.17	0.17	0.17
R^2	0.00	0.01	0.01	0.00	0.02	0.02	0.00	0.02	0.04
Obs.	6,148	6,134	6,134	6,148	6,134	6,134	6,148	6,134	6,134
Strata FE	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes	No	No	Yes

Notes: This table reports estimates from OLS regressions on treatment assignment. Robust standard errors are parenthesized. Controls are selected using the double-lasso method of Belloni et al. (2014) recommended by Urminsky et al. (2016). *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table C.2: Effects of reporting on other measures of activity

Panel A: Tweets and Likes						
	Tweets			Likes		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	35.882*	35.643*	22.946	26.416	27.580	4.863
	(20.694)	(20.753)	(15.419)	(41.243)	(41.467)	(29.651)
y Mean	405.47	405.89	405.89	846.49	847.86	847.86
y SD	782.50	783.58	783.58	1559.14	1560.95	1560.95
R^2	0.00	0.02	0.45	0.00	0.01	0.49
Obs.	5,717	5,697	5,697	5,717	5,697	5,697
Panel B: other activity measures						
	Winsorized time			Fraction of active days		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.192	0.192	0.126	-0.001	-0.001	-0.001
	(0.133)	(0.134)	(0.109)	(0.001)	(0.001)	(0.001)
y Mean	3.36	3.37	3.37	1.09	1.09	1.09
y SD	5.05	5.05	5.05	0.04	0.04	0.04
R^2	0.00	0.02	0.34	0.00	0.01	0.01
Obs.	5,717	5,697	5,697	5,727	5,708	5,708
Strata FE	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

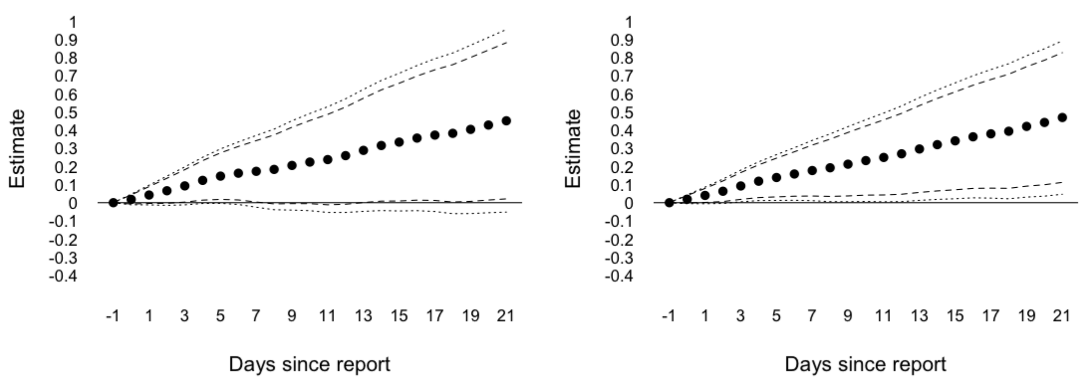
Notes: This table reports estimates from OLS regressions on treatment assignment. Robust standard errors are parenthesized. Controls are selected using the double-lasso method of Belloni et al. (2014) recommended by Urminsky et al. (2016). *,**, and *** denote significance at the 10%, 5%, and 1% levels, respectively.



(a) Hours spent on Twitter **(b)** Fraction of hateful posts

Figure C.3: Cumulative dynamic treatment effects on activity and hatefulness

Notes: This figure presents cumulative dynamic treatment effects, pointwise confidence intervals (dashed), and sup-t simultaneous confidence bands (dotted).



(a) All replies **(b)** Attacks

Figure C.4: Cumulative dynamic treatment effect on replies activity

Notes: This figure presents cumulative dynamic treatment effects, pointwise confidence intervals (dashed), and sup-t simultaneous confidence bands (dotted). The outcome variable is a measure of time spent of the users that the posts in the sample reply to. It is a linear combination of Tweets and Likes.

Table C.3: Effects of reporting on other measures of hatefulness

Panel A: extensive margin						
	Posting toxicity ≥ 0.8			Repeating the slur		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.004	0.004	0.003	0.001	0.001	0.001
	(0.008)	(0.008)	(0.008)	(0.013)	(0.013)	(0.011)
y Mean	0.90	0.90	0.90	0.62	0.61	0.61
y SD	0.30	0.30	0.30	0.49	0.49	0.49
R^2	0.00	0.01	0.03	0.00	0.02	0.34
Obs.	5,727	5,708	5,708	5,727	5,708	5,708
Panel B: average scores						
	Average toxicity			Average severe toxicity		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	-0.001	-0.001	0.000	-0.001	-0.001	-0.000
	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
y Mean	0.30	0.30	0.30	0.18	0.18	0.18
y SD	0.11	0.11	0.11	0.09	0.09	0.09
R^2	0.00	0.01	0.09	0.00	0.01	0.08
Obs.	5,631	5,616	5,616	5,631	5,616	5,616
Strata FE	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

Notes: This table reports estimates from OLS regressions on treatment assignment. Robust standard errors are parenthesized. Controls are selected using the double-lasso method of Belloni et al. (2014) recommended by Urminsky et al. (2016). *,**, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table C.4: Effects of reporting on other measures of replied users' activity

	Tweets			Likes		
Panel A: Tweets and Likes	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	59.739*	58.073*	50.849*	152.538***	148.888***	141.722**
	(30.756)	(30.903)	(30.824)	(56.932)	(57.615)	(57.611)
y Mean	656.20	657.39	657.39	1151.51	1151.96	1151.96
y SD	1060.33	1062.04	1062.04	1963.42	1964.77	1964.77
R^2	0.00	0.02	0.03	0.00	0.02	0.03
Obs.	4,752	4,733	4,733	4,752	4,733	4,733
Panel B: other activity measures	Winsorized time			Fraction of active days		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.420**	0.408*	0.362*	-0.019	-0.032	-0.033
	(0.210)	(0.211)	(0.210)	(0.113)	(0.113)	(0.113)
y Mean	5.21	5.21	5.21	20.42	20.42	20.42
y SD	7.23	7.24	7.24	3.91	3.91	3.91
R^2	0.00	0.02	0.04	0.00	0.02	0.02
Obs.	4,752	4,733	4,733	4,761	4,742	4,742
Strata FE	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

Notes: This table reports estimates from OLS regressions on treatment assignment. Robust standard errors are parenthesized. Controls are selected using the double-lasso method of Belloni et al. (2014) recommended by Urminsky et al. (2016). *,**, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table C.5: Effects of reporting on other measures of replied users' activity, sample of attacks

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Tweets and Likes						
	Tweets			Likes		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	77.136**	74.865**	71.894**	157.203**	156.095**	152.089**
	(32.018)	(32.165)	(32.037)	(61.304)	(62.240)	(62.025)
y Mean	635.01	635.59	635.59	1140.39	1142.13	1142.13
y SD	1035.93	1037.27	1037.27	1983.49	1986.15	1986.15
R^2	0.00	0.02	0.03	0.00	0.02	0.03
Obs.	4,171	4,155	4,155	4,171	4,155	4,155
Panel B: other activity measures						
	Winsorized time			Fraction of active days		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.512**	0.499**	0.478**	-0.012	-0.028	-0.028
	(0.221)	(0.223)	(0.222)	(0.123)	(0.123)	(0.123)
y Mean	5.07	5.08	5.08	20.35	20.35	20.35
y SD	7.14	7.15	7.15	3.97	3.97	3.97
R^2	0.00	0.02	0.03	0.00	0.02	0.02
Obs.	4,171	4,155	4,155	4,178	4,162	4,162
Strata FE	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

Notes: This table reports estimates from OLS regressions on treatment assignment. Robust standard errors are parenthesized. Controls are selected using the double-lasso method of Belloni et al. (2014) recommended by Urminsky et al. (2016). *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table C.6: Effects of reporting on attrition

	Attrition on day 21			Attrition on day ≤ 21		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.002	0.001	0.001	0.002	0.002	0.001
	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)
y Mean	0.07	0.07	0.07	0.08	0.08	0.08
y SD	0.25	0.25	0.25	0.28	0.28	0.28
R^2	0.00	0.02	0.03	0.00	0.02	0.04
Obs.	6,148	6,134	6,134	6,148	6,134	6,134
Strata FE	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

Notes: This table reports estimates from OLS regressions on treatment assignment. Robust standard errors are parenthesized. Controls are selected using the double-lasso method of Belloni et al. (2014) recommended by Urminsky et al. (2016). *,**, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

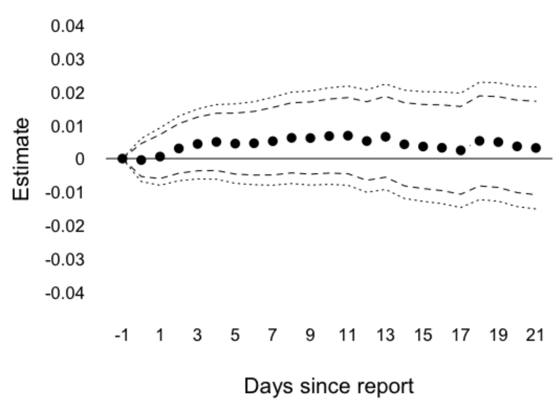
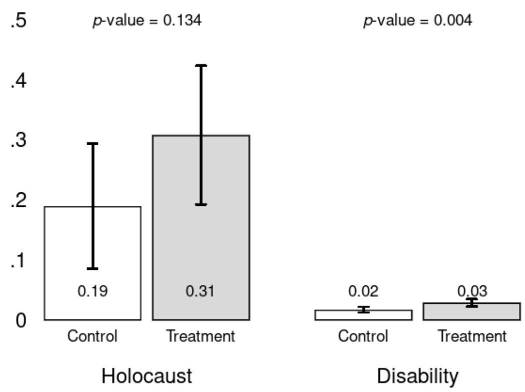
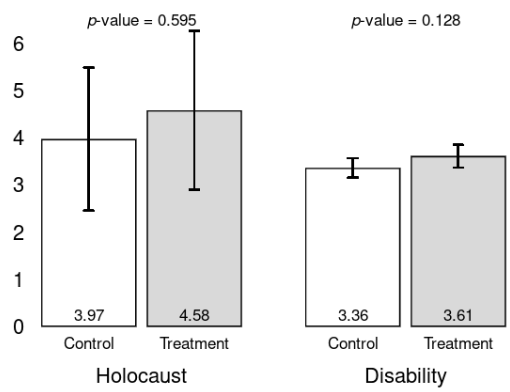


Figure C.5: Cumulative dynamic treatment effect on attrition

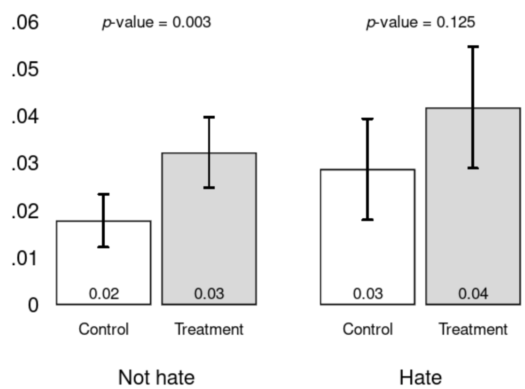
Notes: This figure presents dynamic treatment effects on an indicator of whether users drop from the sample at or before every day after reporting. Pointwise confidence intervals are dashed and sup-t confidence bands are dotted.



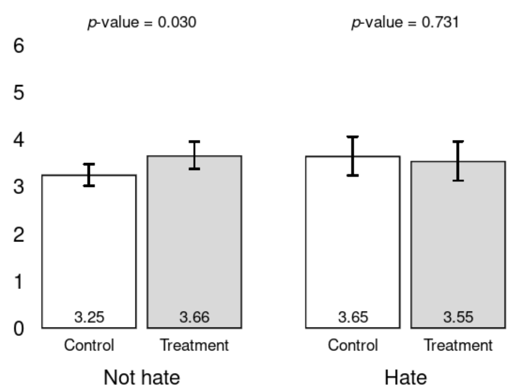
(a) Tweet removals by slur



(b) Time spent by slur



(c) Tweet removals by annotation



(d) Time spent by annotation

Figure C.6: Heterogeneity by slur and hate annotation

Notes: This figure reports estimates of reporting on Tweet removals and users' time spent posting and liking by slur and hate annotation.

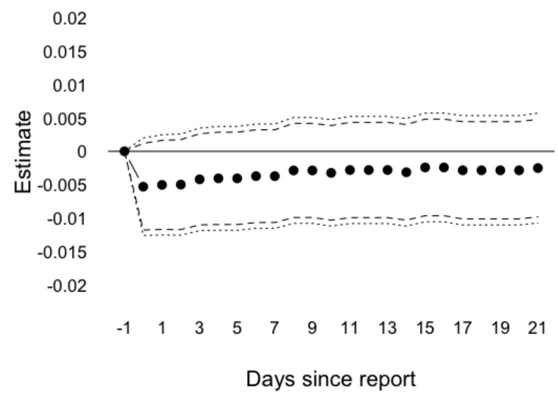


Figure C.7: Cumulative effect on the likelihood of mentioning the replied user

Notes: This figure presents dynamic treatment effects on an indicator of whether the users in the sample mention the replied users again. Pointwise confidence intervals are dashed and sup-t confidence bands are dotted.

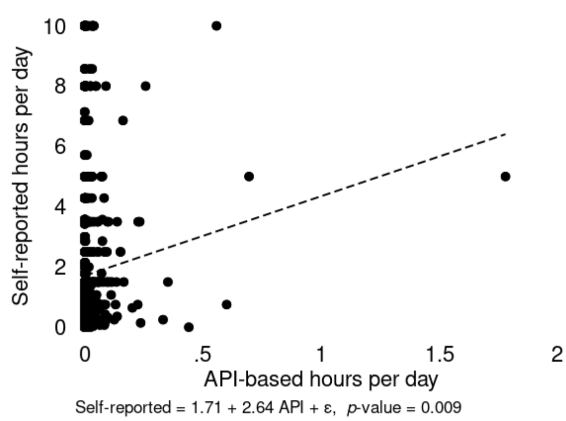


Figure C.8: Self-reported and API-based time spent on Twitter

Notes: This figure presents a comparison between the self-reported hours that participants spend on Twitter with the hours implied by their statuses and likes per day obtained through Twitter's API. The dashed line comes from a linear regression of self-reported hours on API-based hours.

Table C.7: Effects on sanctions among Tweets with replied and attacked users

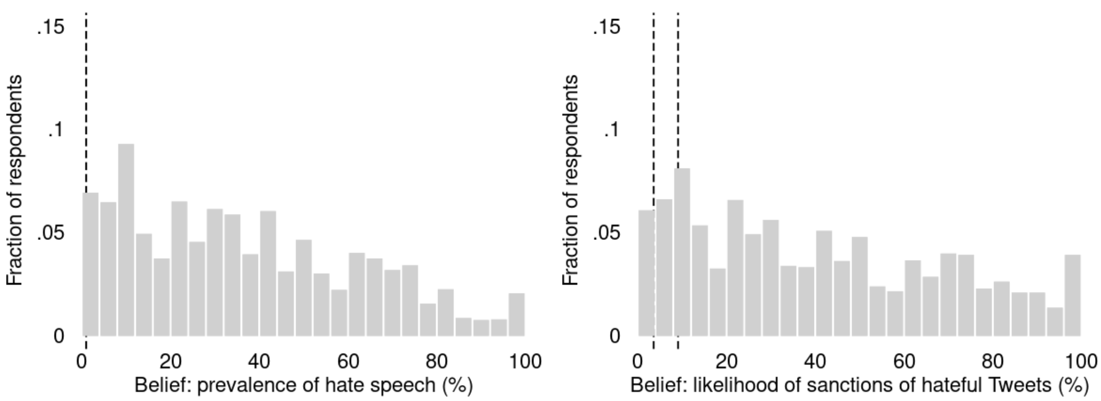
	Tweet removals			Suspensions			Shadow-bans		
Panel A: Sample of replies									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment	0.008*	0.008*	0.008*	0.003	0.002	0.002	-0.003	-0.001	-0.004
	(0.004)	(0.004)	(0.004)	(0.007)	(0.006)	(0.006)	(0.013)	(0.013)	(0.013)
y Mean	0.02	0.02	0.02	0.05	0.05	0.05	0.24	0.24	0.24
y SD	0.15	0.15	0.15	0.22	0.22	0.22	0.43	0.43	0.43
R^2	0.00	0.08	0.08	0.00	0.03	0.03	0.00	0.02	0.08
Obs.	4,752	4,734	4,734	4,752	4,734	4,734	4,404	4,388	4,388
Panel B: Sample of attacks									
	Tweet removals			Suspensions			Shadow-bans		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment	0.006	0.004	0.004	0.002	0.001	0.001	-0.010	-0.008	-0.010
	(0.005)	(0.005)	(0.005)	(0.007)	(0.007)	(0.007)	(0.013)	(0.013)	(0.013)
y Mean	0.02	0.02	0.02	0.05	0.05	0.05	0.22	0.22	0.22
y SD	0.15	0.15	0.15	0.22	0.22	0.22	0.42	0.42	0.42
R^2	0.00	0.04	0.04	0.00	0.02	0.02	0.00	0.02	0.06
Obs.	4,165	4,149	4,149	4,165	4,149	4,149	3,860	3,845	3,845
Strata FE	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes	No	No	Yes

Notes: This table reports estimates from OLS regressions on treatment assignment. Robust standard errors are parenthesized. Controls are selected using the double-lasso method of Belloni et al. (2014) recommended by Urminsky et al. (2016). *,** , and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table C.8: Harassment and moderation experience by subsample

	Means			Difference t-stat.
	Survey	Minority	Not minority	Min.-Not min.
Observations	3,027	1,440	1,587	
Has been harassed	25.2	28.8	20.8	4.07
Prevalence of hate in feed	18.1	20.5	15.1	5.52
Has reported content	32.2	35.7	27.8	3.62
Has been sanctioned or reported	18.5	19.9	16.6	1.98
Tweet removal	9.6	10.4	8.8	1.33
Suspension	5.0	6.0	3.7	2.65
Shadow-ban	6.3	6.2	6.4	-0.16
Account locked	9.8	10.9	8.5	1.86
Has been reported	9.0	9.5	8.3	1.01

Notes: This table presents mean values of variables across different subsamples. It also presents t-statistics from tests of difference in means between minorities and not minorities. Observations are weighted to match representative Twitter users. Minority status Minority status depends on religion, sexual preference, gender, and race.



(a) Prevalence **(b)** Likelihood of moderation

Figure C.9: Beliefs about prevalence and moderation of hate speech

Notes: These figures present histograms of beliefs about prevalence and moderation of hate speech among survey respondents. Prevalence is the fraction of Tweets that are classified as hate speech. Likelihood of moderation is the fraction of hate speech Tweets or users that get removed or de-platformed after 1 month of posting. The dashed lines indicate the observed values of prevalence and moderation in my sample of Tweets. One line in panel (b) corresponds to the majority rule and one to the consensus rule for classifying hate speech.

Table C.9: Effects of information on other measures of socia-media valuation

	WTA uniform distribution			WTA upper endpoint		
Panel A	(1)	(2)	(3)	(4)	(5)	(6)
High moderation	-0.150	0.024	0.024	-0.123	0.066	0.030
	(1.802)	(1.778)	(1.778)	(1.879)	(1.853)	(1.853)
y Mean	33.59	33.59	33.59	38.36	38.36	38.36
y SD	36.33	36.33	36.33	37.91	37.91	37.91
R^2	0.00	0.02	0.02	0.00	0.02	0.02
Obs.	2,998	2,998	2,998	2,998	2,998	2,998

	WTA heuristic			TIOLI		
Panel B	(1)	(2)	(3)	(4)	(5)	(6)
High moderation	0.276	0.630	0.382	0.011	0.009	0.010
	(2.902)	(2.855)	(2.857)	(0.020)	(0.020)	(0.020)
y Mean	36.40	36.40	36.40	0.78	0.78	0.78
y SD	58.81	58.81	58.81	0.42	0.42	0.42
R^2	0.00	0.02	0.03	0.00	0.01	0.02
Obs.	2,998	2,998	2,998	2,998	2,998	2,998
Strata FE	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

Notes: This table reports estimates from OLS regressions on treatment assignment. Robust standard errors are parenthesized. Controls are selected using the double-lasso method of Belloni et al. (2014) recommended by Urminsky et al. (2016). *,**, and *** denote significance at the 10%, 5%, and 1% levels, respectively. Observations are reweighted to match Twitter users from the ATP on observables.

Table C.10: Effects of information on other measures of activity

Panel A: Tweets and Likes

	Tweets			Likes		
	(1)	(2)	(3)	(4)	(5)	(6)
High moderation	3.488	3.359	1.567	13.480*	13.752*	8.927
	(3.733)	(3.621)	(2.623)	(7.636)	(7.356)	(6.140)
y Mean	9.64	9.64	9.64	27.97	27.97	27.97
y SD	70.91	70.91	70.91	140.25	140.25	140.25
R^2	0.00	0.02	0.67	0.00	0.04	0.40
Obs.	1,427	1,427	1,427	1,427	1,427	1,427

Panel B: other activity measures

	Winsorized time			Fraction of active days		
	(1)	(2)	(3)	(4)	(5)	(6)
High moderation	0.022	0.022	0.014	0.017	0.015	0.014
	(0.017)	(0.016)	(0.014)	(0.027)	(0.026)	(0.023)
y Mean	0.07	0.07	0.07	0.28	0.28	0.28
y SD	0.27	0.27	0.27	0.37	0.37	0.37
R^2	0.00	0.05	0.34	0.00	0.04	0.21
Obs.	1,427	1,427	1,427	1,427	1,427	1,427
Strata FE	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

Notes: This table reports estimates from OLS regressions on treatment assignment. Robust standard errors are parenthesized. Controls are selected using the double-lasso method of Belloni et al. (2014) recommended by Urminsky et al. (2016). *,**, and *** denote significance at the 10%, 5%, and 1% levels, respectively. Observations are reweighted to match Twitter users from the ATP on observables.

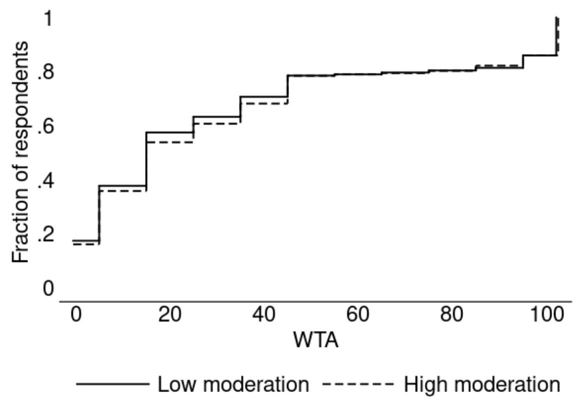


Figure C.10: CDF of the WTA to stop using social media

Notes: This figure displays the CDF of the WTA to stop using social media during one week, by treatment arm. Observations are reweighted to match Twitter users from the ATP on observables.

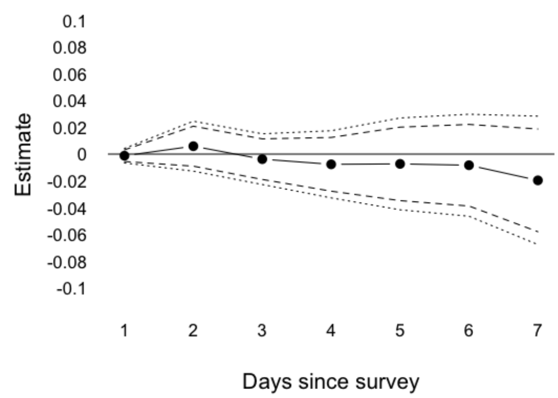


Figure C.11: Cumulative dynamic treatment effects on hours spent on Twitter

Notes: This figure presents dynamic treatment effects of hours spent one week after the survey, pointwise confidence intervals (dashed), and sup-t simultaneous confidence bands (dotted).

Table C.11: Effects of information on inattention and attrition

Panel A: Tweets and Likes	Inattention: |recollection-info.|			Attrition		
	(1)	(2)	(3)	(4)	(5)	(6)
High moderation	1.252	1.013	1.127	0.004	0.005	0.005
	(0.898)	(0.856)	(0.820)	(0.004)	(0.004)	(0.004)
y Mean	8.90	8.90	8.90	0.01	0.01	0.01
y SD	19.82	19.82	19.82	0.10	0.10	0.10
R^2	0.00	0.12	0.24	0.00	0.01	0.01
Obs.	2,997	2,997	2,997	3,027	3,027	3,027
Strata FE	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

Notes: This table reports estimates from OLS regressions on treatment assignment. Attrition indicates whether participants who finished the prescreening questions did not finish the survey. Robust standard errors are parenthesized. Controls are selected using the double-lasso method of Belloni et al. (2014) recommended by Urminsky et al. (2016). *,**, and *** denote significance at the 10%, 5%, and 1% levels, respectively. Observations are reweighted to match Twitter users from the ATP on observables.

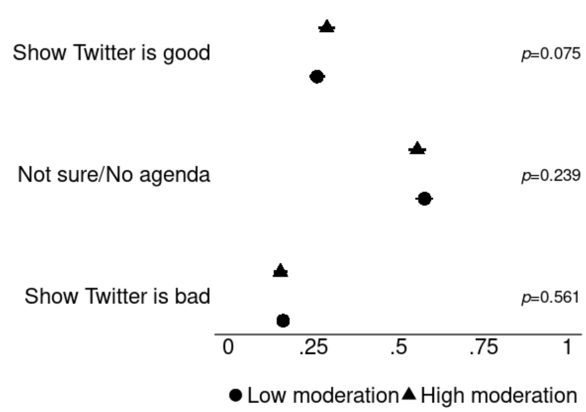
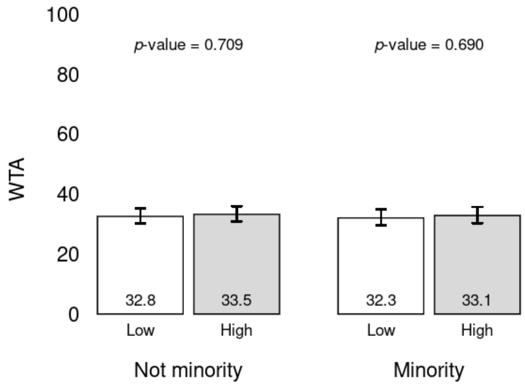
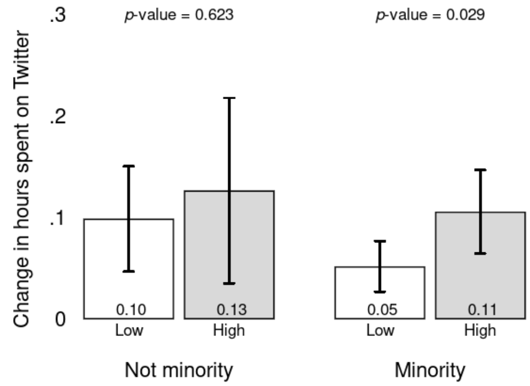


Figure C.12: Treatment effect on perceived experimenter's agenda

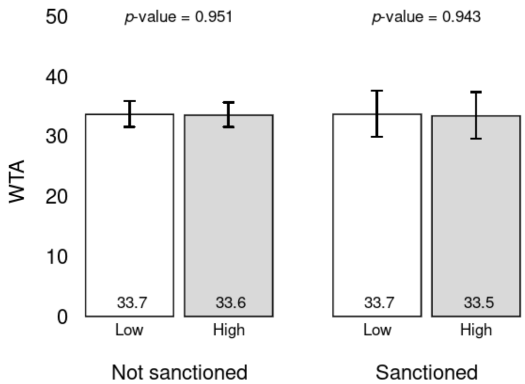
Notes: This figure presents means and 95% confidence intervals by treatment arm. The dependent variables are answers to the question "Do you think that the researchers in this study had an agenda?". The p-values come from independent OLS regressions.



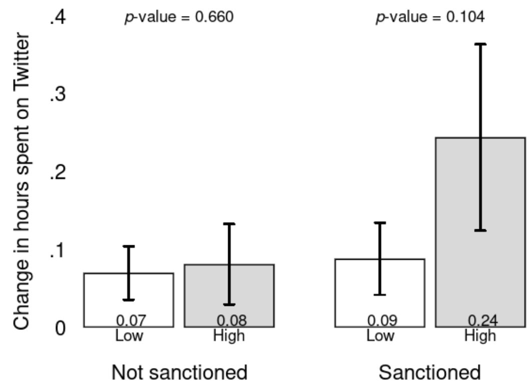
(a) WTA by minority status



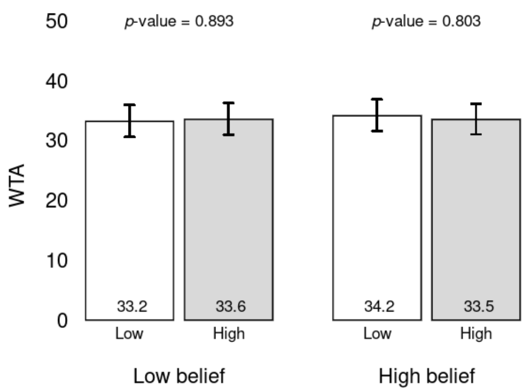
(b) Hours on Twitter by minority status



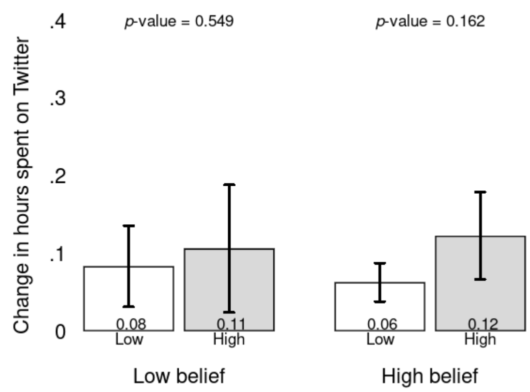
(c) WTA by previous sanctions



(d) Hours on Twitter by previous sanctions



(e) WTA by prior beliefs



(f) Hours on Twitter by prior beliefs

Figure C.13: Heterogeneity of WTA and hours on Twitter by minority status, previous sanctions, and priors

Notes: These figures present means and 95% confidence intervals by treatment arm and minority status. The p-values come from OLS regressions. Observations are reweighted to match Twitter users from the ATP on observables.

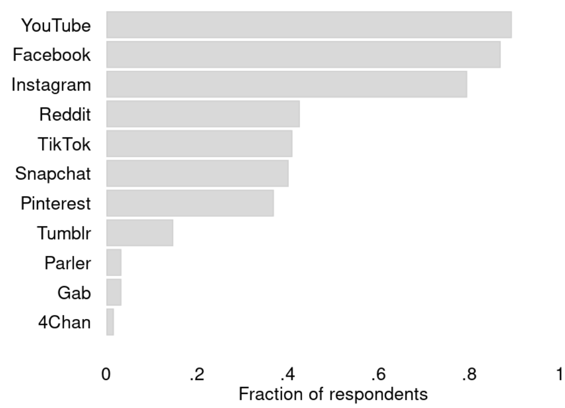


Figure C.14: Other platforms frequented by Twitter users

Notes: This figure presents the fraction of respondents who use other platforms besides Twitter.

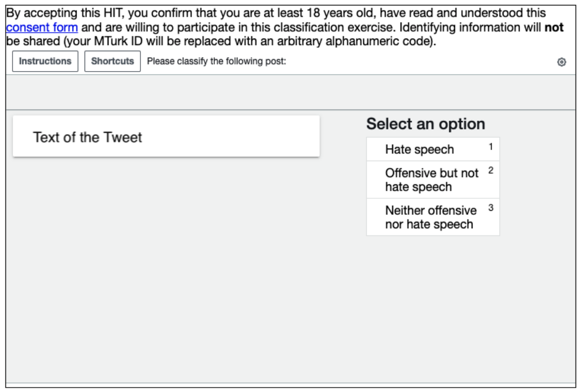
Table C.12: Effects of information on WTA and time spent on Twitter

	WTA			Time spent		
Panel A: Weighted (Twitter ATP)						
	(1)	(2)	(3)	(4)	(5)	(6)
High moderation	-0.3061	0.1035	0.1035	0.0542	0.0576*	0.0576*
	(2.234)	(2.056)	(2.056)	(0.033)	(0.034)	(0.034)
y Mean	33.57	33.57	33.57	0.10	0.10	0.10
y Std. Dev.	36.75	36.75	36.75	0.57	0.57	0.57
R^2	0.00	0.03	0.03	0.00	0.03	0.03
Panel B: Weighted (Social Media ATP)	WTA			Time spent		
	(1)	(2)	(3)	(4)	(5)	(6)
High moderation	-0.0796	-0.0746	-0.0746	0.0400	0.0334	0.0334
	(2.126)	(2.093)	(2.093)	(0.042)	(0.036)	(0.036)
y Mean	34.98	34.98	34.98	0.10	0.10	0.10
y Std. Dev.	37.26	37.26	37.26	0.61	0.61	0.61
R^2	0.00	0.02	0.02	0.00	0.05	0.05
N	2998.00	2998.00	2998.00	1427.00	1427.00	1427.00
Panel C: Unweighted	WTA			Time spent		
	(1)	(2)	(3)	(4)	(5)	(6)
High moderation	0.7230	0.7241	0.7241	0.0456	0.0461	0.0461
	(1.328)	(1.320)	(1.320)	(0.036)	(0.036)	(0.036)
y Mean	32.94	32.94	32.94	0.10	0.10	0.10
y Std. Dev.	36.35	36.35	36.35	0.68	0.68	0.68
R^2	0.00	0.02	0.02	0.00	0.01	0.01
Observations	2,998	2,998	2,998	1,427	1,427	1,427
Strata FE	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

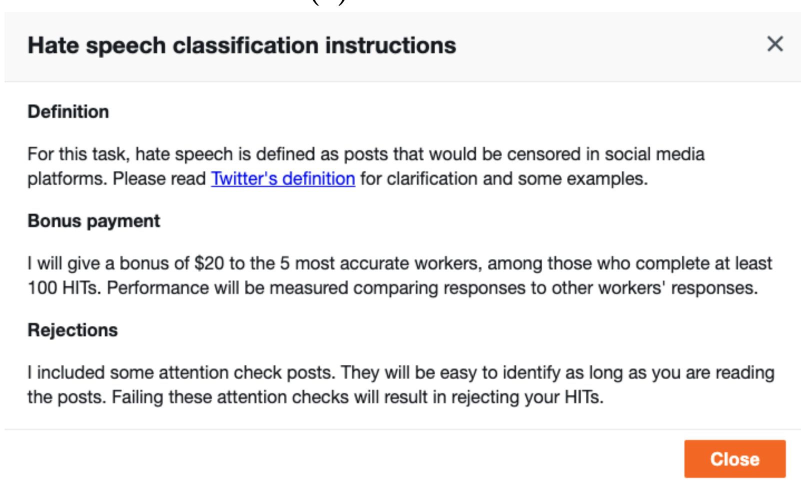
Notes: This table reports estimates from OLS regressions of the WTA and the change in time spent on Twitter on a treatment indicator. Panel A reweights observations to match a representative sample of Twitter users on observables. Panel B reweights observations to match a representative sample of social-media users on observables. Panel C includes unweighted estimates. Robust standard errors are parenthesized. Controls are selected using the double-lasso method of Belloni et al. (2014) recommended by Urminsky et al. (2016).

D Survey Instruments

D.1 Classification of random posts



(a) Task screen



(b) Instructions

Figure D.1: MTurk task to classify posts as hate speech

D.2 Welfare survey

Welcome and thank you for participating in this social media survey!

Which of the following **social media** platforms did you use in the past month? Please select all that apply

- [ ] Facebook
- [ ] Instagram
- [ ] Twitter
- [ ] Snapchat
- [ ] YouTube
- [ ] TikTok
- [ ] Gab
- [ ] Parler
- [ ] Reddit
- [ ] Pinterest
- [ ] Tumblr
- [ ] 4Chan
- [ ] None

The next question is about your interest in sports. In reality, this is an attention check. **If you are reading carefully, please select "A little bit interested" and "Not at all interested"**. How interested are you in sports?

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

After this survey, we will invite some respondents to a **follow-up study**. Below we will describe it in detail

We will also send **$50** Amazon **gift card bonuses** to some participants depending on survey answers and **attention checks**

To proceed with this survey, please confirm if you are willing to **provide a valid email** address. We will use it for the follow-up and bonuses

- Yes
- No: you exit this survey on the next screen

Thank you. Please enter a valid email address:

example@email.com

Now some **demographic** questions

What is the highest level of **education** you have completed?

- Less than high school
- High school graduate
- Some college but no degree
- Associate's degree
- Bachelor's degree or some postgraduate
- Graduate degree (for example: MA, MBA, JD,PhD)

What was your 2020 annual household **income** before taxes?

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

What is your sexual **orientation**?

- Heterosexual
- Homosexual
- Bisexual
- Other
- Prefer not to say

What is your present **religion**, if any?

- Christian
- Jewish
- Muslim
- Buddhist
- Hindu
- Atheist
- Agnostic
- Other
- Nothing in particular
- Prefer not to say

Now some questions about your **social media** use

What is your **Twitter handle / user name?** Optional: you may leave it blank

If you choose to share it, we will only use it to get the account creation date and number of tweets and likes. This data improves the quality of the study

Please include the "@"

@username

How many days did you **use Twitter** last week?

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

On an average day that you used Twitter, how much **time** did you spend on it?

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

Have you received one or more of the following **sanctions** on Twitter?

	Yes	No	Don't know
Account suspended permanently / de-platformed	○	○	○
Account locked / suspended temporarily	○	○	○
Tweet removed	○	○	○
Shadow banned / limited visibility	○	○	○

Have **you reported** Tweets or accounts for violating the Twitter Rules?

- Yes
- No

Have you received a notification that someone **reported you** for violating the Twitter Rules?

- Yes
- No
- Don't know

Have you been **harassed, threatened or attacked** online?

- Yes
- No

Out of every 100 Tweets that you read, how many would you say are **hate speech**—not just offensive?

0 10 20 30 40 50 60 70 80 90 100

Slide to select your guess

A crowd-sourced team of annotators helped us identify hate speech using 10,000 random Tweets in August 2020

We call a Tweet "hate speech" if most annotators label its text as hateful—not just offensive

What percentage of Tweets do you think were classified as hate speech?

We will send a **$50** Amazon gift card to one random respondent among those with the **closest guess**

0 10 20 30 40 50 60 70 80 90 100

Slide to select your guess

How **confident** are you about your guess?

- Very confident
- Somewhat confident
- Neutral
- Not too confident
- Not at all confident

The next question is about the **removal, banning, suspension, or de-platforming of hate speech**

We checked if Twitter removed (de-platformed) the hateful Tweets in the sample or the accounts that posted them

What percentage of hateful Tweets or accounts do you think Twitter **removed (de-platformed)** within 1 month?

We will send a **$50** Amazon gift card to one random respondent among those with the **closest guess**

0 10 20 30 40 50 60 70 80 90 100

Slide to select your guess

How **confident** are you about your guess?

- Very confident
- Somewhat confident
- Neutral
- Not too confident
- Not at all confident

Twitter **removed (de-platformed) 3.6%** of hate speech Tweets or the accounts that posted them, within 1 month

Less than 1% of Tweets in the sample were classified as hate speech. Other popular platforms (Youtube, Facebook, and Reddit) have a similar prevalence of hate

We will conduct a small study that **compensates** some participants to **stop using Twitter, Facebook, Instagram, YouTube, Snapchat, TikTok, and Reddit for 1 week**

Similar studies have been conducted in the past (Hunt et al. 2018, Mosquera et al., 2019, or Allcott et al., 2020)

To establish your compensation we ask below if you want to stop using social media **for different Amazon gift card amounts**

A computer will randomly choose some eligible participants whom we will contact for the follow-up

If you are selected, the computer will also choose one of your answers below at random

- If your answer is "yes", we will **ask** you to **stop using social media one week and pay you the offered amount**
- If your answer is "no", we will **not ask** you to **stop using social media**

Each question could be "the one that counts" to determine your compensation

Would you accept a **$50** compensation to stop using Twitter, Facebook, Instagram, YouTube, Snapchat, TikTok, and Reddit for 1 week?

- Yes
- No

Thank you. 4 final questions:

What was the fraction of hate that Twitter **removes (de-platforms)** that **we told you**? Please give the closest number

We will send a **$50** Amazon gift card to one random respondent among those with the **closest guess**

0 10 20 30 40 50 60 70 80 90 100

Slide to select

Do you think the researchers in this study had an **agenda**?

- I don't think they had a particular agenda
- Yes, they wanted to show that Twitter is good
- Yes, they wanted to show that Twitter is bad
- Not sure

What do you think about Twitter's **removal of Tweets or de-platforming of users?**

- Twitter removes too much
- Neutral / Twitter removes enough
- Twitter removes too little

What percentage of hate speech posts or accounts would you guess **Facebook** **removes (de-platforms)** within one month of posting?

0 10 20 30 40 50 60 70 80 90 100

Slide to select