Procedural Justice and Self Governance on Twitter: Unpacking the Experience of Rule Breaking on Twitter

Matthew Katsaros, Tom Tyler, Jisu Kim, and Tracey Meares

Abstract. Online platforms are increasingly being held to account for the content that their users post. Regulation of content has long been a secondary concern of platforms, but more recently as platforms focus on their content governance, they have typically drawn their regulatory model from offline legal frameworks built around sanctioning and punishment of rule violators. This study approaches these problems using an alternative approach, also drawn from legal scholarship, that is based upon motivating voluntary rule following by emphasizing the fairness of platform rules and the justice of the processes used to communicate content moderation decisions. Using a survey (n=10,487) sent to rule violators on Twitter paired with an analysis of participants' platform behaviors, this study looks at the relationship between people's judgments of the procedural justice of an enforcement action and the participants' likelihood of reoffending in the future. We find that those who felt more fairly treated during their enforcement were less likely to recidivate (beta = -.05, $p < .001$). This, along with the study's other findings, indicates an opportunity for platforms to put a stronger focus on people's experience with enforcement systems as a potential pathway for reducing recidivism.

1 Introduction

Over the past decade, the adoption of social media has rapidly increased around the world (Pew Research Center 2022). While the number of people on these platforms has dramatically increased, so has the amount of communication and activity that takes place on them. One of the most significant challenges for operators of these platforms in the face of this increased activity is how to appropriately govern the content and behaviors of its users (Tyler et al. 2021). While some of these emergent problems confronting social media platforms may be unique to these online social spaces, many of the issues these platforms seek to govern mirror long-standing issues in the offline world. These problems also range dramatically—in both frequency and severity—from seeing repeated and irrelevant content that might occur frequently but be only a slight annoyance to platform users, to issues like harassment or doxxing, which, while less frequently occurring than spam, can result in severe harms to the individual targeted

doi:10.54501/jots.v1i3.38

(Hinduja and Patchin 2013; Hoff and Mitchell 2009). Many platforms have created transparency reports providing the prevalence and other details of the various issues they work to govern (Singh and Doty 2021). Although many of these transparency reports claim that any given issue makes up only a fraction of a percent of all content on the platform, in absolute terms, this can mean millions of pieces of violative content a month; aggregated over time, this can end up affecting a large portion of the platform's user base. A 2020 Pew survey found that 41% of US adults reported personally experiencing some form of online harassment, a figure that has been increasing since 2014 (Auxier and Anderson 2021).

Many platforms have looked to the deterrence model common in legal settings as an initial framework through which to regulate content (Massanari 2017; Pater et al. 2016). Platforms write rules and create technical and operational mechanisms to evaluate user content against those rules (Meta, Inc. 2022a; Twitter, Inc. 2022d; Nextdoor 2022; Reddit, Inc. 2022). Those who violate rules by posting violative content get sanctioned in some way, typically with a graduated series of punishments and/or restrictions: users' posts are removed, their accounts might be suspended for some period of time, and/or users might even be banned from a platform. In adopting this approach, online platforms have inherited both the strengths and weaknesses of traditional law.

This set of technical and operational systems to govern content and behavior on the platforms comes at a significant and ongoing investment from platform operators. In this research, we seek to better understand these governance systems through a survey paired with a behavioral analysis of some of the individuals who pass through these systems, specifically users on Twitter.

This study examined whether the procedural justice model, which has been widely shown to influence behavior in real-world settings, can be applied to content moderation in an online platform setting. The specific question addressed is whether managing a content moderation decision in ways that users view as procedurally just leads users to follow platform rules more in their later online activity.

This research was conducted in collaboration with Twitter to better understand people's encounters with Twitter's content moderation system. A survey was administered to accounts who had recently violated one of Twitter's rules, asking them about this rule-breaking experience. This survey finds that 10% of those who break rules on the platform are unaware that the platform even has rules, while many others are aware of the existence of rules but have never read them. When asking participants why they posted the rule-breaking Tweet, it becomes clear that there are a wide variety of circumstances that lead to rule-breaking behavior. Nearly a third of participants answered "I thought it was appropriate to post," while 18.5% answered "I was defending myself or someone else." One in ten participants answered "I lost my temper," while very few participants (1.6%) posted the rule-breaking Tweet because "I wanted to harm someone." Survey responses are then paired with logged platform data on rule violations in the six months before and 90 days after the survey in order to understand the relationship between people's evaluation of their experience having a Tweet removed and their likelihood to break Twitter's rules in the future.

In support of Procedural Justice theory, when controlling for past rule violations we see that those who felt more fairly treated during the enforcement experience were less likely to break rules in the future. This survey also allowed us to understand which was more strongly related to people's overall evaluations of the fairness of the enforcement—people's agreement with the platform's decision or people's assessments of the fairness of various procedural elements of the enforcement. Counter to the idea that people's assessments of fairness are driven primarily by their agreement with outcomes, we

see that people's judgments of the procedural elements of the enforcement experience (transparency, voice, being treated with dignity, etc.) were more strongly associated with overall fairness. Lastly, we looked at how people's orientation toward values of free speech and safety online played a role in these models. We saw that, as anticipated, those who more strongly valued free speech were more likely to recidivate, while those who more strongly valued safety were less likely to recidivate. These findings point to limitations that platforms must confront when designing content moderation systems.

2 Related Work

2.1 Content Moderation

How to effectively manage and govern a wide spectrum of content online has long been an area of concern for platform operators and researchers alike. Grimmelmann (2015) puts forth a helpful taxonomy or "grammar" of moderation outlining a set of techniques, distinctions, and community characteristics that try to describe various forms that moderation online can take. Grimmelmann (2015) provides helpful distinctions, for example, between moderation done with transparency vs secrecy or moderation done manually vs. automatically. Bradford et al. (2019) describe governance structures as being "top down"—where officials implement a relatively detailed set of rules over a given community—or "bottom up"—where users define their own norms and actively intervene in their enforcement—or some combination of the two.

There are endless combinations and forms that online content moderation can take, but one of the predominant mechanisms used to govern content online is through removal after content has been posted and later deemed to violate some rule. These content removals are most frequently done by the operators of the platform, but similar moderation actions are taken by volunteer moderators in online spaces that use community governance structures (Seering et al. 2019). While other enforcement actions like ranking and sorting content are likely to be more prevalent than content removals, these content removals continue to be the focus of newsworthy controversy and public debate.

Content removals are primarily a result of the content being found in violation of norms or rules set forth by the community and/or by the platform operators. As an example, Twitter maintains what it calls "The Twitter Rules" (Twitter, Inc. 2022d), whereas Facebook maintains a similar set of rules they call "Community Standards" (Meta, Inc. 2022a). While the issues these rules seek to regulate are often quite similar, they can be defined very differently. Pater et al. (2016) looked at the way 15 different online platforms defined harassment and found there was little consistency in the definition or even range of behaviors described by the various platforms (Pater et al. 2016). In online spaces with community governance models, users are responsible for creating their own rules and, as such, definitions of similar behaviors become even more inconsistent. Fiesler et al. (2018) conducted an in-depth qualitative and quantitative analysis of the varying rules created by 100,000 different communities on Reddit, which found a wide range of rules set forth by community moderators (Fiesler et al. 2018). The impact of removals or sanctions imposed on the content creator varies depending on the platform as well as the specific rule violated, ranging from the content no longer being accessible to the community to permanent removal of the authoring account from the platform.

Increasingly, this content removal approach relies on the use of algorithms and machine learning to identify or help sort and prioritize content for human review (Chandrasekharan et al. 2017; Wulczyn, Thain, and Dixon 2017; Yin et al. 2009). As just one example,

Facebook removed 31.5 million pieces of content for violating its policies on hate speech, of which 97.6% of that content was proactively identified using machine learning (Meta, Inc. 2022b). While this can be a tremendously useful approach to manage the massive scale at which many platforms are operating, the use of algorithms for moderation undoubtedly comes with its own risks and issues, including false positives (Hosseini et al. 2017) and many opportunities for algorithmic bias (Vaidya, Mai, and Ning 2020; Binns et al. 2017).

Prior research regarding the effect of content moderation on user behavior has shown that a platform's content moderation policy has the potential to discourage users to engage in repeated violative behaviors(Tyler et al. 2021; Jhaver, Bruckman, and Gilbert 2019). However, many social media users who have their content removed by a platform frequently express confusion about what exactly triggered a moderation action (Saltz, Leibowicz, and Wardle 2021; Suzor et al. 2019; Myers West 2018). Myers West (2018) explored the folk theories that people have developed surrounding these content moderation systems and how those folk theories shape people's relationship to the platform. As one example, Myers West (2018) found that many users "expressed confusion and frustration about the [content moderation] process and suggested that content moderation systems appeared to be designed to escalate these emotional responses." Saltz, Leibowicz, and Wardle (2021) conducted interviews and found that some users perceived the warning labels added on visual misinformation related to COVID-19 as politically biased acts of censorship from the platforms. Based on survey responses submitted by users, Suzor et al. (2019) also found that moderated users were uncertain about what content triggered a moderation decision and did not receive sufficient information to understand why a moderation decision was made. Thus, it is clear there is an opportunity for social media platforms to orient moderation systems toward education, rather than punishment, by providing them an understanding of why their content was removed and increasing the transparency and clarity of their moderation processes (Myers West 2018; Suzor et al. 2019).

2.2 Procedural Justice

An alternative, or complementary, approach to managing content by the traditional approach of sanctioning users for inappropriate content is to motivate users to take personal responsibility for following rules (Tyler et al. 2021). This approach has been shown to be effective in legal settings; people have been found to be more willing to follow rules in their everyday lives, to adhere to court orders and mediated agreements, and to voluntarily follow police directives when they feel that authorities are legitimate and, therefore, entitled to be obeyed (Tyler 2003). This has led to a research literature examining what factors shape the legitimacy of legal authorities (Hinds and Murphy 2007; Tyler and Huo 2002; Lind and Tyler 1988; Tyler and Wakslak 2004; Wolfe et al. 2016; Jackson et al. 2012). That literature suggests that evaluations of the justice of the procedures authorities use to create and implement rules are a central factor shaping the legitimacy of those rules and the authorities who implement them. This legitimacy, in turn, influences willing deference to rules and decisions. These findings have led to a literature on the meaning of procedural justice that now includes studies of a variety of types of authorities, including legal authorities, managers, teachers, and others (Trinkner and Cohn 2014).

Procedural justice is the evaluation of how authorities manage their authority (Tyler 2004). This includes how they make decisions and how they treat those with whom they deal. Procedural justice is distinct from the actual decisions made by an authority. The core research finding is that people evaluate procedures separately from outcomes and that they accept outcomes that go against their desires more readily when they believe

that the procedures used are fair.

Studies of the meaning of procedural justice generally suggest that people make evaluations of procedural justice using four core criteria: voice, neutrality, respect, and trust (Tyler 2007). The fist two of these criteria concern the way a decision is made. Voice is the opportunity to participate and express one's views or explain one's situation during the creation and implementation of rules. Neutrality is evidence that the rules are created and applied in a consistent and impartial way, without bias or preferential treatment. This includes transparency and explanation during rule creation and implementation. A separate set of criteria concern quality of treatment; these are sometimes referred to as relational issues (Tyler 2003; Blader and Tyler 2015). The first is the respect and courtesy that people experience when dealing with others. The second is whether people infer that those others are seen as trustworthy—benevolent and sincerely concerned about their needs and issues.

These criteria shape evaluations of the procedural justice of some procedures, and they can be applied to any procedure, including bilateral interactions, third-party procedures, and interactions with institutions (Lind and Tyler 1988). This study examines their extension to an online context involving a social media platform's content moderation system (Tyler et al. 2021). It asks whether the fairness of that system is correlated to its success in influencing user acceptance and compliance of the platform rules.

When seeking a way to manage online content moderation, it makes sense to draw upon models of regulation that have been effective in law. At the same time, it is important to recognize that legal authorities have a type of inherent legitimacy as state actors, which may not be replicable with private vendors. Further, that legitimacy is connected to shared identity, and online communities may not have the same type or degree of shared identity (Tyler et al. 2021). Hence, the degree to which legitimacy motivates self-regulation in online settings and the question of whether procedural justice works in this space are open issues for empirical study.

To understand why the extension of procedural justice to online settings is an open question, it is important to consider the psychological mechanisms underlying procedural justice effects. People care about experiencing procedural fairness because treatment with fairness both implies inclusion in a social group, within a scope of justice, and signals status in a relationship, group, or community (Tyler and Huo 2002). For this reason, experiencing fairness builds self-esteem and enhances feelings of self-worth (Smith et al. 1998). However, for these effects to occur, people have to have a shared social connection with others. Treatment by strangers, who are outside one's community and scope of justice, does not have identity-relevant implications. When dealing with outsiders, people are more outcome focused. Hence, a key issue is whether online users feel that they are members of a community. This can include a community of specific users and/or the community defined by the platform itself. A motivation can be to follow the informal consensus of users: to not violate rules of civility or to adhere to the rules as defined and managed by the platform.

Some researchers have looked at the role of transparency in the content moderation experience. Much of this work suggests that providing more transparency throughout the moderation experience can increase future rule compliance (Tyler et al. 2021; Jhaver et al. 2019; Jhaver, Bruckman, and Gilbert 2019). Jhaver, Bruckman, and Gilbert (2019) found that users who had posts removed from Reddit were less likely to have future posts removed when provided explanations. Tyler et al. (2021) similarly found that users who were provided education about platform rules in the week following their post removal were less likely to have their future posts removed. Matias (2019) found that posting the rules to the r/Science Reddit discussion board increased rule compliance among new

members. Jhaver et al. (2019) found that, among users on Reddit who had recently had a post removed, those who had previously read the Subreddit's rules were more likely to perceive their post removal as fair.

3 Hypotheses and Research Questions

3.1 The awareness and comprehension of rules on the platform.

Other work has demonstrated the existence of a connection between the awareness and transparency of an online community's rules and rule-violating behavior within that community (Jhaver et al. 2019; Matias 2019). In hopes of contextualizing our respondents' experience with Twitter's content enforcement system, we wanted to ask two basic questions about their awareness and familiarity with Twitter's rules:

- *RQ1: Are those who break rules on Twitter aware of the Twitter Rules?*
- *RQ2: For those aware of the Twitter Rules, have they read the rules, and how familiar do they self-report being with the Twitter Rules?*

3.2 The context surrounding the rule-breaking event.

When trying to understand someone's experience with Twitter's content enforcement system, it can be particularly helpful to understand the context surrounding their alleged violation of the Twitter Rules. The myth of the online troll—that individuals knowingly and willfully violate platform rules seeking to cause havoc and harm to others—is pervasive in the media coverage of online abuse and harassment. While research has demonstrated that some individuals engage in online harassment with intentions to cause harm to others (Goodboy and Martin 2015; Lopes and Yu 2017), there is little empirical work demonstrating what proportion of online hate speech is motivated by malice. What little research that has been done to understand motivations behind the creation of offensive online content more broadly has shown there is actually a number of circumstances and motivations behind these content creators (Blackwell et al. 2018). As such, to help contextualize the experiences that people have with Twitter's content enforcement system, we additionally wanted to ask:

- *RQ3: What are the circumstances leading to rule breaking on Twitter?*

3.3 Evaluations of procedural justice and future rule breaking.

The primary focus of content enforcement efforts of online platforms has been on defining rules for appropriate content and evaluating whether individual pieces of content violate these rules. This focus on decision outcomes has diverted resources and attention away from understanding how people going through these content enforcement systems are actually experiencing this process. This neglect has led platforms to default to a sanction-based model of enforcement, which does not require attention to what users are experiencing because it is based on the presumption that platforms can control user behavior exclusively by graduated systems of access to the platform. This study approaches these problems using an alternative model of governance, a model based on voluntary rule acceptance by using rule enforcement processes that users perceive as fair and just. The core hypotheses we seek to test in this study are understanding the relationship between people's perceptions of the current enforcement process and their propensity to self-govern moving forward.

- *H1: People who feel more fairly treated by the platform during decisions regarding their content removal are less likely to recidivate.*
- *H2: Perceptions of the fairness of the enforcement process will have a stronger correlation with recidivism than people's agreement with the decision to remove their Tweet.*

3.4 The underlying makeup of procedural justice.

Beyond understanding the relationship between people's perceptions of the fairness of rule enforcement process and their future violations, we want to understand what specific aspects contribute to perceptions of fairness of the enforcement process. Leveraging decades of research on procedural justice in many other contexts, we developed multiple survey items to ask about individual elements of procedural justice—transparency, consistency, voice, and respect—during the enforcement process. Using these items, we pose the following research question:

- *RQ4: How are individual elements of procedural justice correlated to people's overall assessment of the fairness of their enforcement process?*

3.5 Values of free speech and safety online and future rule breaking.

The enforcement of rules on social media platforms is a constant and delicate balance between providing a platform to express oneself freely and providing a platform where people feel safe to do so. While a platform has control over changing its enforcement experience to improve people's perceptions of procedural justice in the enforcement (for example, by introducing more clarity and transparency regarding a rule violation), there may be limitations in using this approach alone. Individual users of a platform also bring with them their own orientation toward these values of safety and free speech. Individuals who deeply value free speech might see any attempt to infringe on this value as illegitimate and be more likely to violate platform rules. By contrast, those who value feeling safe on the platform may be more likely to internalize a platform's rules following their own violation. In our survey, we posed two questions side by side to participants, asking them to indicate the importance for them to speak freely and to feel safe on the platform. From this, we pose the following hypotheses:

- *H3: Those who more highly value the ability to speak freely on Twitter will be more likely to recidivate.*
- *H4: Those who more highly value the ability to feel safe on Twitter will be less likely to recidivate.*

4 Study Design

To understand these research questions and test these hypotheses, in collaboration with Twitter, Inc., we developed a survey that was paired with an analysis of user behavior data provided by Twitter. Participants were recruited for this survey through a message shown on their Twitter timeline asking if they would like to provide their feedback through a survey. Those who clicked on this survey recruitment message were taken off of Twitter to a survey platform, where an online questionnaire was administered in English (full questionnaire text in Appendix D).

The recruitment for this survey was not sent to all platform users. To be eligible for this survey recruitment message, a participant had to have met both of the following two

inclusion criteria:

1. A participant had to be using Twitter from a primarily English-speaking country.
2. A participant had to have created a Tweet that was determined by Twitter within the last 30 days to have violated one of three of Twitter's rules: Hateful Conduct Policy (Twitter, Inc. 2022b), Abusive Behavior Policy (Twitter, Inc. 2022a), or Promoting Suicide and Self-harm Policy (Twitter, Inc. 2022c). These three rules are a subset of the full Twitter Rules (Twitter, Inc. 2022d) but account for 64% of all accounts actioned, 25% of all accounts suspended, and 74% of all content removed during the second half of 2020 (Twitter, Inc. 2022e).

The total number of accounts meeting this inclusion criteria was 235,626. Of those eligible, 79% (n=186,405) accounts were shown the survey recruitment message. Reasons an account may not have been shown the message can include having not logged on to the platform during the fielding of the survey or an eligible account no longer being active (voluntarily or removed as a result of further/more severe rule violations) during the fielding of the survey. Of those 186,405 accounts that were shown a recruitment message, 5.6% (n=10,487) started the survey, of which, 61% (n=6,385) participants completed the entire survey. The analysis utilized both complete and partial responses.

Anonymized survey responses were combined with platform activity. Specifically, data on rule violations during the six months prior to the survey and three months following the survey was appended to anonymized survey responses. This anonymized violation data was limited to the date of the violation and rule violated during that nine month period, with no other personally identifiable information about the violation appended to survey responses. The percent of respondents who recidivated and violated a rule on Twitter in the 90 days following the survey was 22%; 16% had one violation during this period, 4% had two violations, and 2% had three or more. Table 5 in Appendix A contains summary statistics of the variables in the dataset.

5 Results

5.1 RQ1: Rule Awareness

Prior research on people who violate rules on social media platforms has shown that a lack of rule awareness can help explain these rule violations, especially among first-time rule violators. To answer RQ1, in our questionnaire we asked participants "Does Twitter have any rules about appropriate behavior on the platform?" with answer options "Yes," "No," and "I'm not sure." It is worth clarifying that this survey was administered after a rule violation had occurred; all participants of the survey had a rule violation within the last 30 days, during which they were notified of their violation by Twitter. Unfortunately, these findings do not provide us insight into people's awareness of rules preceding their violation; nonetheless, the findings of people's rule awareness following their rule violation still provide important insights. 74.4% of respondents answered "Yes," while 9.9% answered "No" and another 15.7% answered "I'm not sure." Given this finding that a quarter of respondents answered "No" or "I'm not sure" to this very basic question, "Does Twitter have any rules about appropriate behavior on the platform?," it becomes clear that there is an opportunity to better socialize users to the existence of rules during the enforcement process.

Table 1: Responses to "Have you ever read Twitter's rules about appropriate behavior on the platform?"

Response	Percent Response	Number of Responses
Yes, I've read them completely	29.9%	1,931
Yes, I've read them partially	47.1%	3,131
No, I haven't read them	21.7%	1,444
I'm not sure	2.2%	147

Table 2: Responses to "How familiar are you with Twitter's rules on appropriate behavior on the platform?"

Response	Percent Response	Number of Responses
Extremely familiar	17.8%	1,180
Very familiar	27.0%	1,793
Somewhat familiar	38.6%	2,563
A little bit familiar	11.6%	769
Not at all familiar	5.0%	333

5.2 RQ2: Reading Rules and Rule Familiarity

For those that answered "Yes" to this rule-awareness question, follow-up questions about their familiarity with these rules were asked to address RQ2. Participants were asked "Have you ever read Twitter's rules about appropriate behavior on the platform?" as well as "How familiar are you with Twitter's rules on appropriate behavior on the platform?" Results from these questions are in Table 1 and Table 2. We see that most of those aware of Twitter's rules self-report having read them completely (29%) or partially (47.1%); however, 21.7% self-report never having read them despite being aware of their existence.

5.3 RQ3: Circumstances Surrounding Rule Breaking Tweets

To answer RQ3, we asked participants the following question about their Tweet that was deemed to have violated the Twitter Rules: "Which of the following describe why you posted this Tweet? Please select all that apply." A list of options was shown to participants, with checkboxes allowing them to select as many options as they wanted along with an "Other" option that allowed participants to write in an answer. The responses to this question are shown in Figure 1. From these answers, we are able to better understand some of the motivations and circumstances surrounding rule violations on Twitter.

As we saw from RQ1 and RQ2, many who violate rules on Twitter are unaware of rules. Unsurprisingly, we see that 32.8% selected "I thought it was appropriate to post" to our question. This result adds to a growing understanding that a significant proportion of those who break rules on platforms are simply unaware of the norms of appropriate behavior and conduct that platforms seek to promote. Similarly, 12.8% selected "I saw other people posting similar things." Behavioral research across many contexts points to people learning about norms and appropriate behavior by observing others (Seering, Kraut, and Dabbish 2017). In this study, we see that many respondents observed some behavior on the platform and inadvertently ran afoul of the Twitter Rules when they

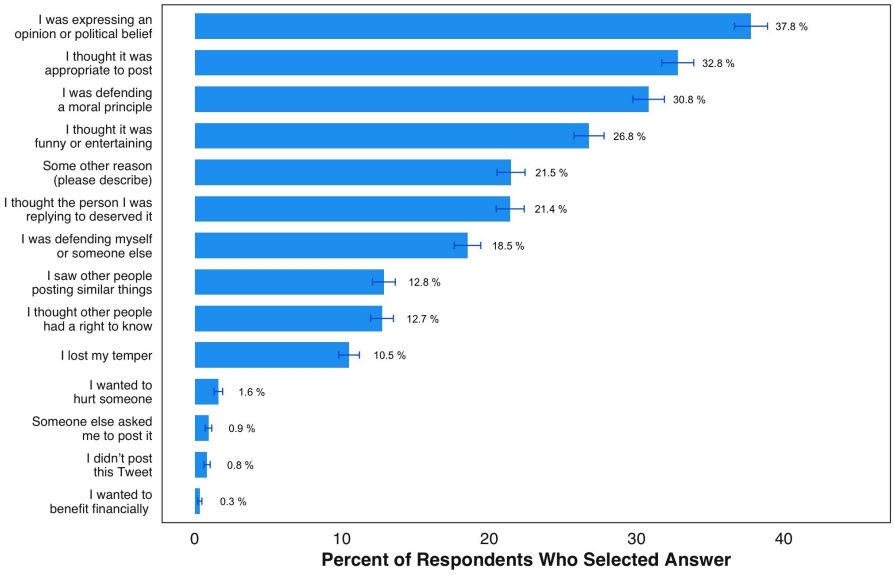

Figure 1: Participant responses to the multiple-choice question "Which of the following describe why you posted [the rule-violating] Tweet? Please select all that apply."

exhibited similar behavior.

5.4 H1 and H2: Procedural Justice and Future Rule Violations

For the participants in this study, when an account was found in violation of Twitter's rules, rather than Twitter removing the content, the account's owner was asked to acknowledge the rule violation and was forced to remove the content themselves before being able to return to using Twitter. To test our first two hypotheses, we asked participants questions about their experience being asked by Twitter to remove their Tweet.

One survey question about the overall fairness of the process asked participants to agree or disagree with the statement "The process Twitter used to ask me to remove my Tweet was fair" (overall fairness). Seven survey items that asked about individual elements of procedural justice, shown in Table 3, were combined into one variable (procedural justice elements). Another question asked participants to consider the decision Twitter made about their Tweet: "To what extent do you agree with Twitter's decision to ask you to remove your Tweet?" (agreement with decision). A variable was created counting the number of rule violations in the six months prior to the survey (prior offending) with another variable (recidivism) indicating whether or not someone had any rule violations in the 90 days following the survey.

A structural equation model using SPSS AMOS 26 software was used to look at the relationship the three independent variables (procedural justice elements, agreement with decision, and prior offending) have with recidivism. Overall fairness was used in the model as a mediating variable. Results from this structural equation modeling are shown in Figure 2. The absence of arrows from two of the independent variables (procedural justice elements and agreement with decision) directly to recidivism indicates that there is no statistically significant direct relationship to recidivism. In other words, these factors only correlate to recidivism to the degree that they correlate to the mediating variable (overall fairness).

As shown in Figure 2, assessments of overall fairness were correlated to recidivism (beta

Table 3: Survey items of underlying procedural justice elements. Participants were asked to agree or disagree with the statements and given a five-point agree-disagree scale.

Item Name	Questionnaire Text
Considered my POV	Twitter considered my point of view when asking me to remove my Tweet.
Twitter Explained their Decision	Twitter clearly explained why my Tweet wasn't allowed.
Opportunity to Provide my POV	Twitter gave me an opportunity to provide my point of view.
Treats Everyone the Same	Twitter treats everyone the same when it asks people to remove their Tweets.
Considers my Feedback	Twitter considers my feedback about what the rules are and how to enforce them.
I Understand Why	I understand why Twitter asked me to remove my Tweet.
Treated with Respect	Twitter treated me with respect.

= -.05, $p < .001$), with those who felt fairly treated less likely to recidivate. The presence or absence of procedural justice elements and the extent to which the participant agreed with Twitter's removal decision were both correlated to overall fairness (beta = 0.52, $p < .001$ and beta = 0.35, $p < .001$ respectively). Finally, prior violation history was the only of the three independent variables directly correlated to recidivism (beta = 0.13, $p < .001$). This model had a CFI = 0.94 and an RMSEA = .081.

These results indicate support for H1: People who feel more fairly treated by the platform during decisions regarding their content removal are less likely to recidivate. This finding is consistent with the literature in contexts outside of social media and the limited literature that looks at procedural justice within a social media context. People's perceptions of how fairly they are treated when having rules enforced by a platform is negatively correlated with future rule-breaking behavior. Our model also shows that the procedural justice elements had a stronger correlation to overall fairness (beta = 0.52, $p < .001$) relative to the agreement with decision (beta = 0.35, $p < .001$). This indicates strong support for H2: Perceptions of the fairness of the enforcement process will have a stronger correlation with recidivism than people's agreement with the decision to remove their Tweet. Both of these findings take into account an individual's prior rule-breaking history, which, unsurprisingly, as seen in Figure 2 has a strong and direct correlation to future rule-breaking behavior.

5.5 RQ4: Individual Elements of Procedural Justice and Overall Fairness

Significant research on procedural justice in other contexts such as criminal justice, policing, and doctor-patient relationships indicate key factors that make up people's overall impression of procedural justice. Those include transparency (explaining the rules and decision-making process), consistency (treating people the same), providing voice (allowing people to express themselves and their side of the story), and treating people with dignity and respect. Seven survey items (listed in Table 3) on a five-point agree-disagree scale were presented to participants in a matrix that asked about these underlying elements of procedural justice. We estimated the correlation of these seven items with overall procedural fairness judgments and recidivism in order to understand RQ4: How are individual elements of procedural justice correlated to people's overall

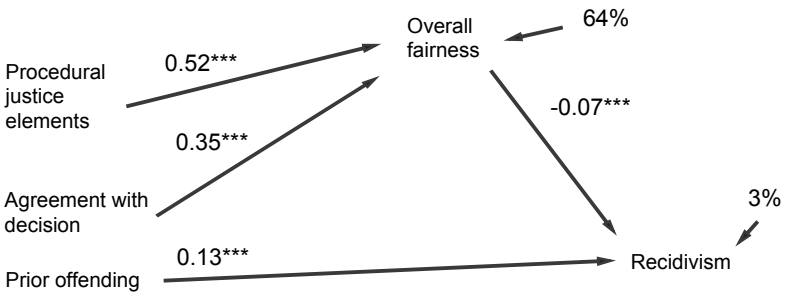

Figure 2: Structural equation modeling using procedural justice elements, agreement with decision, prior offending as independent variables, recidivism as the dependent variable, and overall fairness as the mediating variable. The 64% and 3% indicate the proportion of variance in Overall Fairness and Recidivism (respectively) explained by the independent variables. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

assessment of the fairness of their enforcement process? The results of this statistical model are shown in Table 4.

Table 4: Correlation coefficients of individual elements of procedural justice with overall fairness perceptions and recidivism during the 90 days following the survey.

PJ Element	Overall Fairness	Recidivism
Considered my POV	0.55***	-0.04**
Twitter Explained their Decision	0.52***	-0.06***
Opportunity for Provide my POV	0.42***	-0.03**
Treats Everyone the Same	0.49***	-0.07***
Considers my Feedback	0.58***	-0.08***
I Understand Why	0.61***	-0.06***
Treated with Respect	0.64***	-0.09***

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

These results indicate that all seven items are significantly correlated (p < .001) to overall fairness perceptions. However, some of these elements are more strongly correlated to an individual's overall perceptions of fairness. Specifically, people's sense of being treated with dignity and respect appears to have the strongest correlation to overall fairness (beta = 0.64, p < 0.01). This is a particularly surprising finding, as much of the prior procedural justice literature on dignity and respect is in relation to interpersonal communication—for example, the way police officers interact with citizens. Given that, one might suspect that dignity and respect would be unimportant as there is no direct communication from platform representatives with these individuals during the rule enforcement process. On the other hand, perhaps this finding indicates that the lack of any direct communication from the platform leaves individuals feeling disrespected. Understanding why the violative Tweet was removed was a similarly important element in overall fairness evaluations. Taken with the results from RQ1 and RQ2, bridging the gap in the awareness and comprehension of rules on the platform might be an approach to building more positive perceptions of enforcement fairness.

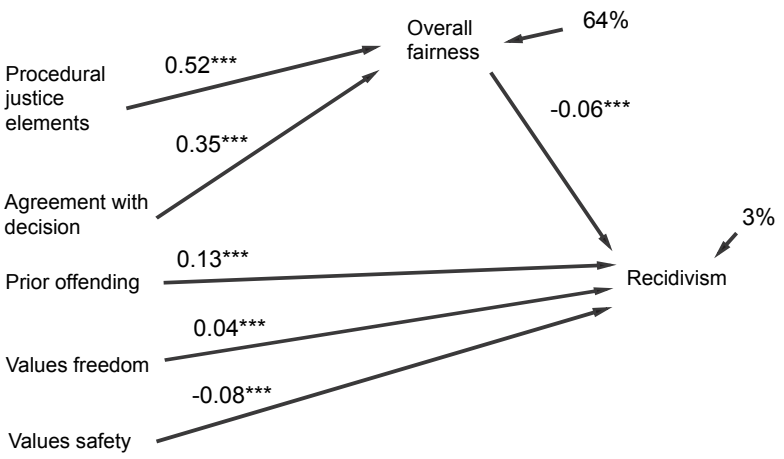

Figure 3: Structural equation modeling using procedural justice elements, agreement with decision, prior offending, values safety, and values freedom as independent variables, recidivism as the dependent variable, and overall fairness as the mediating variable. The 64% and 3% indicate the proportion of variance in Overall Fairness and Recidivism (respectively) explained by the independent variables. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

5.6 H3 and H4: Valuing Free Speech and Safety and Future Rule Violations

The analyses reported thus far have focused on the process of removing content, inherently limiting someone's ability to speak freely on the platform. This limitation of free speech is balanced with ensuring everyone on the platform feels safe enough to exercise that freedom. But perspectives towards these two values—speaking freely and feeling safe on the platform—are important individual differences, and respondents do not all see them as equally important. This analysis also considers the importance respondents place on these two values. Participants were asked two questions, shown in Table 3, to indicate the importance of these two values. We used a similar structural equation model used in Figure 2, but this time added these two survey questions as two more independent variables (values safety and values freedom) in order to test H3 and H4; the results of this structural equation model expanding on the model shown in Figure 2 are presented in Figure 3.

6 Discussion and Conclusions

Platform operators have invested significant efforts to curb unwanted behavior from its users. But how to allocate those resources and the ways to most effectively manage unwanted behavior are still open questions. Significant research in the criminal justice and policing context points to procedural justice as a way to build legitimacy and promote an internalized approach to rule following. This study seeks to understand how this procedural justice theory translates to a social media context, specifically when it comes to enforcing rules on Twitter.

Given the relatively small amount of literature that exists on the experience of individuals who break rules on social media, the first set of research questions was aimed at shedding light on these experiences. We find that there exists a gap in rule awareness and comprehension; about one-third of users who had broken rules on the platform were unaware that rules existed, and for those that were aware, many had never read or were unfamiliar with the rules. Relatedly, when participants were asked why they

posted the Tweet that was later removed, responses reveal a similar trend—people simply thought their Tweet was appropriate or they had seen others doing something similar. All of these findings together indicate that a significant amount of rule-breaking behavior is explained quite simply by a lack of awareness rather than premeditated malicious activity. Many running afoul of the rules had no idea they were doing so—either because they didn't know about rules or saw others doing something similar. These point to an opportunity to focus more on education, awareness, and comprehension of rules. This study replicates prior findings that the procedural justice of a platform's rule enforcement process is correlated to rule-following behavior. Specifically, those that feel more fairly treated during the enforcement process are more likely to self-govern and avoid rule breaking in the future. This replication is important because it involves a different platform with a distinct user base, adding support for the robustness of the procedural justice effect across social media platforms. Much of the discussion in the media coverage of online content moderation is focused on specific outcomes of these decisions—often simply whether posts and Tweets are taken down or left up. Left out of these discussions is a focus on the process in which these decisions are made. The results from this study are quite clear: when it comes to the people who had their content removed, whether or not they agreed with Twitter's decision to remove their Tweet had a relatively smaller relationship on their future rule-following behavior. By contrast, their assessment of how fairly their case was handled by Twitter was correlated with future recidivism. These findings indicate a major opportunity for platforms. If platforms seek to have individuals self-govern, there is a clear pathway to this end through focusing on how they treat people when enforcing rules. A focus on ensuring people feel treated with dignity and respect alongside a comprehensive explanation of how their Tweet violated Twitter's rules are some of the more critical elements to address to improve perceptions of fairness. Further qualitative research could be pursued to unpack these points and better understand how these insights translate into specific designs to build on those opportunities.

The last set of hypotheses indicate limitations to this procedural justice approach. While platforms have relatively significant control over the enforcement system they build—how much transparency or voice is afforded to those who go through it—other important elements contributing to rule-breaking behavior may lay further outside a platform's control. The two elements we focused on with this study were the importance participants placed on speaking freely on Twitter and feeling safe on Twitter. We see that both of these are significantly correlated with recidivism. Those who more strongly value free speech are more likely to break rules and, by contrast, those who more strongly value feeling safe are less likely to break rules. While interesting, the application of these findings are limited; it is hard to imagine in what ways platforms could successfully seek to influence people's attitudes on these values. Furthermore, the fact that the ties these values have to recidivism are of relatively similar strength to perceptions of fairness suggest that a platform's focus on procedural justice has limitations.

One limitation of this study is its cross-sectional design; because of this, we are only able to identify associations and cannot draw causal conclusions as other studies have done (Tyler et al. 2021). As more research is conducted in this arena, we should seek to produce studies that allow causal-inference, by, for example, designing experiments where users are provided enforcement procedures that vary in their procedural justice to understand how these changes might affect recidivism and people's evaluations of fairness.

As noted, there are reasons to believe that the models of legal authority that work in real-world settings may not be effective with online platforms. Legal authorities have legitimacy as state actors, which may not be replicable with private platform operators.

Further, that legitimacy is connected to shared identity, and online communities may not have the same type or degree of shared identity (Tyler et al. 2021). Hence, the degree to which legitimacy motivates self-regulation in online settings as well as the question of whether procedural justice works in this space are open issues for empirical study. We are not able to directly compare the utility of procedural justice in online and real-world settings. However, our study does make clear that procedural justice is effective in online platform environments.

This study demonstrates empirically that there is an association between online platform user's experiences with a content moderation action and their later rule-violating behavior. If users experience the moderation action as involving more just procedures, they are more likely to adhere to platform rules in the future.

References

Auxier, Brooke, and Monica Anderson. 2021. "Social Media Use in 2021." *Pew Research Center.*

Binns, Reuben, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. "Like trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation." In *International Conference on Social Informatics,* 405–15. Springer.

Blackwell, Lindsay, Mark Handel, Sarah T. Roberts, Amy Bruckman, and Kimberly Voll. 2018. "Understanding "Bad Actors" Online." In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems,* 1–7.

Blader, Steven L., and Tom R. Tyler. 2015. "Relational Models of Procedural Justice." *The Oxford Handbook of Justice in the Workplace* 351:370.

Bradford, Ben, Florian Grisel, Tracey L. Meares, Emily Owens, Baron L. Pineda, Jacob N. Shapiro, Tom R. Tyler, and Danieli Evans Peterman. 2019. "Report of the Facebook Data Transparency Advisory Group." *Yale Justice Collaboratory.*

Chandrasekharan, Eshwar, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. "The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems,* 3175–87.

Fiesler, Casey, Joshua McCann, Kyle Frye, Jed R. Brubaker, et al. 2018. "Reddit Rules! Characterizing an Ecosystem of Governance." In *Twelfth International AAAI Conference on Web and Social Media.*

Goodboy, Alan K., and Matthew M. Martin. 2015. "The Personality Profile of a Cyberbully: Examining the Dark Triad." *Computers in Human Behavior* 49:1–4.

Grimmelmann, James. 2015. "The Virtues of Moderation." *Yale JL & Tech.* 17:42.

Hinds, Lyn, and Kristina Murphy. 2007. "Public Satisfaction with Police: Using Procedural Justice to Improve Police Legitimacy." *Australian & New Zealand Journal of Criminology* 40 (1): 27–42.

Hinduja, Sameer, and Justin W. Patchin. 2013. "Social Influences on Cyberbullying Behaviors among Middle and High School Students." *Journal of Youth and Adolescence* 42 (5): 711–22.

Hoff, Dianne L., and Sidney N. Mitchell. 2009. "Cyberbullying: Causes, Effects, and Remedies." *Journal of Educational Administration.*

Hosseini, Hossein, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. "Deceiving Google's Perspective API Built for Detecting Toxic Comments." *arXiv preprint arXiv:1702.08138.*

Jackson, Jonathan, Ben Bradford, Mike Hough, Andy Myhill, Paul Quinton, and Tom R. Tyler. 2012. "Why Do People Comply with the Law? Legitimacy and the Influence of Legal Institutions." *British Journal of Criminology* 52 (6): 1051–71.

Jhaver, Shagun, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "'Did You Suspect the Post Would be Removed?' Understanding User Reactions to Content Removals on Reddit." *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW): 1–33.

Jhaver, Shagun, Amy Bruckman, and Eric Gilbert. 2019. "Does Transparency in Moderation Really Matter? User Behavior after Content Removal Explanations on Reddit." *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW): 1–27.

Lind, E. Allan, and Tom R. Tyler. 1988. *The Social Psychology of Procedural Justice.* Springer Science & Business Media.

Lopes, Barbara, and Hui Yu. 2017. "Who Do You Troll and Why: An Investigation into the Relationship between the Dark Triad Personalities and Online Trolling Behaviors Towards Popular and Less Popular Facebook Profiles." *Computers in Human Behavior* 77:69–76.

Massanari, Adrienne. 2017. "# Gamergate and The Fappening: How Reddit's Algorithm, Governance, and Culture Support Toxic Technocultures." *New Media & Society* 19 (3): 329–46.

Matias, J. Nathan. 2019. "Preventing Harassment and Increasing Group Participation through Social Norms in 2,190 Online Science Discussions." *Proceedings of the National Academy of Sciences* 116 (20): 9785–89.

Meta, Inc. 2022a. "Facebook Community Standards." Accessed August 24, 2022. https://www.facebook.com/communitystandards/.

———. 2022b. "Facebook Community Standards Enforcement Report: Hate Speech." Accessed August 24, 2022. https://transparency.fb.com/data/community-standards-enforcement/hate-speech/facebook/.

Myers West, Sarah. 2018. "Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms." *New Media & Society* 20 (11): 4366–83.

Nextdoor. 2022. "Nextdoor Community Guidelines." Accessed August 24, 2022. https://help.nextdoor.com/s/article/community-guidelines.

Pater, Jessica A., Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. 2016. "Characterizations of Online Harassment: Comparing Policies across Social Media Platforms." In *Proceedings of the 19th International Conference on Supporting Group Work,* 369–74.

Pew Research Center. 2022. "Social Media Fact Sheet." Accessed August 24, 2022. https://www.pewresearch.org/internet/fact-sheet/social-media/.

Reddit, Inc. 2022. "Reddit Content Policy." Accessed August 24, 2022. https://www.redditinc.com/policies/content-policy.

Saltz, Emily, Claire R. Leibowicz, and Claire Wardle. 2021. "Encounters with Visual Misinformation and Labels across Platforms: An Interview and Diary Study to Inform Ecosystem Approaches to Misinformation Interventions." In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems,* 1–6.

Seering, Joseph, Robert Kraut, and Laura Dabbish. 2017. "Shaping Pro and Anti-Social Behavior on Twitch through Moderation and Example-Setting." In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing,* 111–25.

Seering, Joseph, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. "Moderator Engagement and Community Development in the Age of Algorithms." *New Media & Society* 21 (7): 1417–43.

Singh, Spandana, and Leila Doty. 2021. "The Transparency Report Tracking Tool: How Internet Platforms Are Reporting on the Enforcement of Their Content Rules." New America. Accessed August 24, 2022. https://www.newamerica.org/oti/reports/transparency-report-tracking-tool/.

Smith, Heather J., Tom R. Tyler, Yuen J. Huo, Daniel J. Ortiz, and E. Allan Lind. 1998. "The Self-Relevant Implications of the Group-Value Model: Group Membership, Self-Worth, and Treatment Quality." *Journal of Experimental Social Psychology* 34 (5): 470–93.

Suzor, Nicolas P., Sarah Myers West, Andrew Quodling, and Jillian York. 2019. "What Do We Mean When We Talk about Transparency? Toward Meaningful Transparency in Commercial Content Moderation." *International Journal of Communication* 13:18.

Trinkner, Rick, and Ellen S. Cohn. 2014. "Putting the "Social" Back in Legal Socialization: Procedural Justice, Legitimacy, and Cynicism in Legal and Nonlegal Authorities." *Law and Human Behavior* 38 (6): 602.

Twitter, Inc. 2022a. "Abusive Behavior." Accessed August 24, 2022. https://help.twitter.com/en/rules-and-policies/abusive-behavior.

———. 2022b. "Hateful Conduct Policy." Accessed August 24, 2022. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

———. 2022c. "Suicide and Self-Harm Policy." Accessed August 24, 2022. https://help.twitter.com/en/rules-and-policies/glorifying-self-harm.

———. 2022d. "The Twitter Rules." Accessed August 24, 2022. https://help.twitter.com/en/rules-and-policies/twitter-rules.

———. 2022e. "Transparency: Rules Enforcement." Accessed August 24, 2022. https://transparency.twitter.com/en/reports/rules-enforcement.html%5C#2020-jul-dec.

Tyler, Tom, Matt Katsaros, Tracey Meares, and Sudhir Venkatesh. 2021. "Social Media Governance: Can Social Media Companies Motivate Voluntary Rule Following Behavior among their Users?" *Journal of Experimental Criminology* 17 (1): 109–27.

Tyler, Tom R. 2003. "Procedural Justice, Legitimacy, and the Effective Rule of Law." *Crime and justice* 30:283–357.

———. 2004. "Enhancing Police Legitimacy." *The Annals of the American Academy of Political and Social Science* 593 (1): 84–99.

———. 2007. "Court Review: volume 44, issue 1/2-Procedural Justice and the Courts." *Court Review: The Journal of the American Judges Association,* 217.

Tyler, Tom R., and Yuen J. Huo. 2002. *Trust in the Law: Encouraging Public Cooperation with the Police and Courts.* Russell Sage Foundation.

Tyler, Tom R., and Cheryl J. Wakslak. 2004. "Profiling and Police Legitimacy: Procedural Justice, Attributions of Motive, and Acceptance of Police Authority." *Criminology* 42 (2): 253–82.

Vaidya, Ameya, Feng Mai, and Yue Ning. 2020. "Empirical Analysis of Multi-Task Learning for Reducing Identity Bias in Toxic Comment Detection." In *Proceedings of the International AAAI Conference on Web and Social Media,* 14:683–93.

Wolfe, Scott E., Justin Nix, Robert Kaminski, and Jeff Rojek. 2016. "Is the Effect of Procedural Justice on Police Legitimacy Invariant? Testing the Generality of Procedural Justice and Competing Antecedents of Legitimacy." *Journal of Quantitative Criminology* 32 (2): 253–82.

Wulczyn, Ellery, Nithum Thain, and Lucas Dixon. 2017. "Ex Machina: Personal Attacks Seen at Scale." In *Proceedings of the 26th International Conference on World Wide Web,* 1391–99.

Yin, Dawei, Zhenzhen Xue, Liangjie Hong, Brian D. Davison, April Kontostathis, and Lynne Edwards. 2009. "Detection of Harassment on Web 2.0." *Proceedings of the Content Analysis in the WEB* 2:1–7.

Authors

Matthew Katsaros is the Director of the Social Media Governance Initiative at the Justice Collaboratory at Yale Law School (matthew.katsaros@yale.com). **Tom Tyler** is the Macklin Fleming Professor of Law at Yale University and a Professor in the Yale Psychology Department. **Jisu Kim** is an assistant professor at the Singapore Institute of Technology where she teaches in the Digital Communications and Interactive Media program. **Tracey Meares** is the Walton Hale Hamilton Professor of Law and Founding Director of The Justice Collaboratory at Yale Law School.

Acknowledgements

We'd like to thank Aruna Balakrishnan, Lindsay Blackwell, Lauren Fratamico, and Sarah Anoke for their contributions and feedback on early drafts of this work.

Data Availability Statement

The data for this study is not available publicly for replication or other uses. This survey included standard disclosure language ensuring individual responses would not be shared externally. In accordance with this disclosure made to study participants, we are unable to release the underlying individual response data for replication.

Funding Statement

This study was funded in large part by Twitter, Inc. Two of the study's authors (Matt Katsaros and Tracey Meares) were, at the time of this study, under contract as part-time researcher advisors to Twitter Inc.

Ethical Standards

The study was reviewed and approved by the Yale University IRB. The approval was given May 20, 2021. The IRB approval number is 2000030435.

Keywords

Social Media; Procedural Justice; Content Moderation

Appendices

Appendix A: Summary Statistics

Summary statistics for variables included in this study. Included in the table is the variable name, the interpretation of higher value, the range of values, number of respondents, and the mean & standard deviation.

Table 5

Variable	Higher Value	Range	# Respondents	Mean(SD)
Overall Fairness	More Fair	1-5	7,079	2.16(1.04)
Considered my POV	Stronger Agreement	1-5	6,635	1.63(1.09)
Treated with Respect	Stronger Agreement	1-5	6,632	2.51(1.40)
Opportunity to Provide my POV	Stronger Agreement	1-5	6,635	1.88(1.31)
Considers my Feedback	Stronger Agreement	1-5	6,621	2.02(1.25)
Twitter Explained their Decision	Stronger Agreement	1-5	6,650	2.53(1.51))
I Understand Why	Stronger Agreement	1-5	6,643	2.57(1.52)
Treats Everyone the Same	Stronger Agreement	1-5	6,640	1.87(1.27)
Combined PJ Elements	Stronger Agreement	1-5	7,079	2.16(1.04)
Agreement with Decision	Stronger Agreement	1-5	7,092	1.90(1.28)
Values Safety	More Important	1-4	9,407	2.67(1.12)
Values Freedom	More Important	1-4	9,410	3.41(0.80)
6-Months Prior Violations	More Prior Violations	1-37	10,487	1.66(1.33)
90-Day Recidivism	Did Recidivate	0-1	10,487	0.22(0.42)
Education	Higher Education	1-7	6,207	4.04(1.72)
Age Range	Older	1-6	8,545	3.61(1.60)
Race/Ethnicity	0=Non-White; 1=White	0-1	10,487	28%
Gender	0=Non-Male; 1=Male	0-1	10,487	32%

Appendix B: Regression Table for Recidivism

A regression of the variable for recidivism 90 days after the survey. The entries are the standardized regression coefficient and the unstandardized regression coefficient and the standard deviation. The adjusted R-sq. reflects the proportion of the dependent variable explained by all factors adjusted for the number of independent variables.

$^{****}p < 0.001$, $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 6

Variable	Interpretation of Higher Value	Standardized	Unstandardized
Procedural Justice Elements	More Fair	-0.07	-.028(0.008)****
Agreement with Decision	Stronger Agreement	0.02	0.006(0.005)
Prior Offending	More Prior Violations	0.13	0.039(0.004)****
Values Safety	More Important	-0.07	-.028(0.005)****
Values Freedom	More Important	0.04	0.021(0.007)**
Age	Older	-0.03	-.008(0.004)**
Education	Higher Education	-0.02	-.005(0.004)
Gender	1=Male; 0=Non-Male	0.04	0.034(0.014)***
Race/Ethnicity	1=White; 0=Non-White	-0.03	-.03(0.01)**
Constant	—	—	0.26(0.04)
Adjusted R.-sq.	3.5%		
ANOVA	F(9,5234)=21.0****		

Appendix C: Regression Table for Overall Fairness

A regression of the Overall Fairness question "The process Twitter used to ask me to remove my Tweet was fair." The entries are the standardized regression coefficient in one column with the unstandardized regression coefficient and the standard deviation in the other column. The adjusted R-sq. reflects the proportion of the dependent variable explained by all factors adjusted for the number of independent variables.

$^{****}p < 0.001$, $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 7

Variable	Interpretation of Higher Value	Standardized	Unstandardized
Procedural Justice Elements	More Fair	0.60	0.78(0.01)****
Agreement with Decision	Stronger Agreement	0.29	0.31(0.01)****
Prior Offending	More Prior Violations	-.01	-.01(0.008)
Values Safety	More Important	-.01	-.02(0.01)
Values Freedom	More Important	-.01	-.02(0.01)*
Age	Older	-.02	-.02(0.01)**
Education	Higher Education	0.00	0.00(0.01)
Gender	1=Male; 0=Non-Male	-.01	-.04(0.03)*
Race/Ethnicity	1=White; 0=Non-White	0.05	0.15(0.03)**
Constant	—	—	-.01(0.07)
Adjusted R.-sq.	69%****		
ANOVA	F(9,5243)= 1327.40****		

Appendix D: Questionnaire Text

NOTE: Content in [brackets] is intended to help the reader understand logic that is programmed into the survey.

How important is it to feel safe on Twitter?

- Extremely important
- Very important
- Somewhat important
- Not at all important

How important is it for people to be able to speak their minds freely on Twitter?

- Extremely important
- Very important
- Somewhat important
- Not at all important

[Page Break]

Does Twitter have any rules about appropriate behavior on the platform?

- Yes
- No
- I'm not sure

[If *"Yes"* or *"I'm not sure"*, continue to next question. If *"No"*, skip to reporting awareness question]

[Page Break]

Have you ever read Twitter's rules about appropriate behavior on the platform?

- Yes, I've read them completely
- Yes, I've read them partially
- No, I haven't read them
- I'm not sure

How familiar are you with Twitter's rules on appropriate behavior on the platform?

- Extremely familiar
- Very familiar
- Somewhat familiar
- A little bit familiar
- Not at all familiar

[Page Break]

In as much detail as possible, please describe the most recent time Twitter asked you to remove something you tweeted. Please include information about the Tweet that you posted, why you posted it, and any other relevant information. [open-end]

[Page Break]

All of the remaining questions are about this most recent time that Twitter asked you to remove something that you tweeted.

[Page Break]

Which of the following describe why you posted this Tweet? Please select all that apply.

- I was defending myself or someone else
- I was defending a moral principle
- I was expressing an opinion or political belief
- I thought it was appropriate to post
- I thought it was funny or entertaining
- I thought the person I was replying to deserved it
- I lost my temper
- I thought other people had a right to know
- I saw other people posting similar things
- I wanted to hurt someone
- I wanted to benefit financially
- Someone else asked me to post it
- I didn't post this Tweet
- Some other reason (please describe)

[Page Break]

To what extent do you agree with Twitter's decision to ask you to remove your Tweet?

- Strongly agreee
- Somewhat agreee
- Neither agree nor disagreee
- Somewhat disagreee
- Strongly disagree

[Page Break]

Please agree or disagree with the following statement: The process Twitter used to ask me to remove my Tweet was fair.

- Strongly agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Strongly disagree

[Page Break]

[matrix-style question with *Statements* as rows and *Answers* as columns] Thinking about this most recent time Twitter asked you to remove your Tweet, please agree or disagree with the following:

Statements:

- Twitter considered my point of view when asking me to remove my Tweet.
- Twitter treated me with respect.
- Twitter gave me an opportunity to provide my point of view.
- Twitter considers my feedback about what the rules are and how to enforce them.
- Twitter clearly explained why my Tweet wasn't allowed.
- I understand why Twitter asked me to remove my Tweet.
- Twitter treats everyone the same when it asks people to remove their Tweets.

Answers:

- Strongly agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Strongly disagree