Social media governance:

Can companies motivate voluntary rule following behavior among their users

Tom Tyler
Yale Law School

Matt Katsaros
Facebook

Tracey Meares
Yale Law School

Sudhir Venkatesh
Columbia University

Introduction

One view of social media communication is that users are free to say whatever they want, and social media companies act as pass-through agents without evaluating or restricting the content of communications. From this perspective, posts are like letters, and we would no more expect a private company to evaluate posts than we would expect the post office to read and censor a letter.

A contrasting view of social media communication is that companies should remove posts deemed to contain inappropriate content such as nudity, bullying, and hate speech. Many private companies have accepted the argument that they should take responsibility for content and have developed systems for reporting and reviewing posts.

One set of topics not addressed here are normative and are concerned with what the rules ought to be. Any rulemaking system is generally structured around a set of affirmative rights and duties, which often must be balanced so that the enjoyment of one right does not result in the deprivation of another. For instance, in the civil sphere, one's right of free speech must be balanced against their duty not to cause harm to others. In legal settings, this balance is defined by laws (Waldron, 2014) or courts' interpretation of those laws. In the social media world, this balance is defined through the standards and policies that social media companies create.

We set aside the normative issue on what the rules ought to be; instead, the issue we address is how social media companies can effectively enforce the rules that they create for their sites. In particular, whether they can motivate their users to voluntarily adhere what they post to the standards.

The question of how to effectively enforce rules on social media mirrors the general question of how to enforce laws in society. One model for both governments and private companies is incapacitation—preventing people from taking particular actions. Using this approach, private companies can exercise control over their sites by removing content that violates their standards and placing restrictions on the account. This parallels the governmental process of removing people who break the law from society through incarceration. The problem in both settings is that people seek ways to get around such controls, trying to hide their actions. It is better if people willingly follow the rules, something referred to as self-regulation. Research with legal authority makes clear that such self-regulation is possible in the case of public legal authority and is linked to the legitimacy of that authority (Tyler, 2006). Our question is whether private companies can have similar legitimacy and can thereby motivate their users to voluntarily follow content rules.

The social media setting

In one respect, social media companies operate at an advantage in comparison to many of the situations facing governments. They have virtually complete control over their platforms and can simply remove content or even block access to their sites. They have less difficulty observing the types of behavior that their rules proscribe and are not as susceptible to people attempting to hide their actions. They do not have to use incarceration to incapacitate violators and thereby avoid the attendant procedures that governments must follow. Further, when they remove content or block a user, they have control over the limits of that action.

On the other hand, private companies may lack legitimacy for engaging in rulemaking and enforcement. For example, in the US many users believe that they are entitled to free expression in their social media behavior and are upset when their content is removed, even though the right to free speech enshrined in the First Amendment prohibits intrusions by the government, not private entities. This impacts sites in several ways. First, unless users buy into rule enforcement decisions, they can seek to subvert them. One common approach is to open multiple user accounts. Another is to operate in smaller or more private spaces to avoid notice by the companies. Further, companies may find that their approach of removing problematic content has the effect of alienating users and increasing future rule-breaking behavior. Hence, the advantages of social media companies in the arena of rulemaking and enforcement are not unlimited. The same question arise in both private and public arenas: can authorities enforce rules in ways that promote user/citizen acceptance and enhance their willingness to self-regulate in the future?

This study examines whether social media authorities can shape user behavior following a particular instance of posting "inappropriate" content. This is a classic issue of seeking to influence possible recidivism and implicates the legitimacy of the authority that finds a violation and imposes the appropriate consequence. If the articulated standards are viewed by users as fairly created and enforced, studies of public authority hypothesize that this should lead to a higher level of future rule acceptance. In particular, studies of law enforcement and courts suggest that evaluations of the procedural justice of the actions of the authority will influence later behavior (Tyler, 2006; Tyler, Goff & MacCoun, 2015; Tyler & Huo, 2002). While this extension might appear to be straightforward, there are two reasons for viewing it as potentially uncertain:

First, private companies may be viewed as inappropriate authorities for making decisions about what users are entitled to do. They may be viewed as lacking the training and expertise associated with authorities like judges, and, aside from their control over user accounts, the nature of their authority to make such decisions is sometimes disputed by users. In other words, users may not view them as legitimate authorities.

Second, the effectiveness of procedural justice depends upon people identifying with the group, organization or community the authority represents (Schultz, 2006). It is unclear whether social media users feel any type of identification with the proprietors of such sites. Social media sites are communities of a type, and they may be what communities look like in the modern world (Gruzd, Jacobson, Wellman & Mai, 2016; Kraut & Resnick, 2012). The question is whether they support the type of value-based self-regulation found with legal authorities.

While the extension of the ideas of procedural justice and legitimacy to rule compliance on social media platforms might seem obvious, private entities are not state authorities and social media users are not citizens. Hence, the question is whether a similar model of the dynamics of authority applies to the enforcement decisions of private companies in social media situations. As issues of inappropriate and harmful content delivered via social media have become important in society, it has become increasingly important to ask how social media content can and should be managed by companies. The concern of this study is with one issue within this general question; can social media companies address issues of potentially inappropriate content in ways that motivate future rule following behavior? And, are the mechanisms facilitating deference similar to those found to be effective in public settings?

Study One

This study examines enforcement in the framework used by one social media site: Facebook. Facebook has a set of rules called "Community Standards" that are available to users on the Facebook site (https://www.facebook.com/communitystandards/). Facebook enforces those standards by removing content that violates those standards. After content is removed, users are notified of their violation and may have their account blocked or, in extreme cases, closed. In most cases, users can appeal the removal of a particular post.

This study uses a large sample to consider several types of violating content: nudity; hate speech; bullying and other violations. Peoples' views are sought after they have violated one of these standards. They have the opportunity to complete a questionnaire between one and three weeks following the notification of a precipitating violation. Our sample included 40,482 users whose precipitating violation was for nudity; 7,310 for hate speech; 2,699 for bullying and 4,115 for other violations. The questionnaire asked respondents about the fairness of their content violation procedure. Their future compliance with standards was then tracked for up to 45 days.

Methods

Users in the study completed the questionnaire on Facebook. The questions asked, and response categories provided are included in Appendix A.

Perceived procedural justice

Respondents were asked six questions about the procedural fairness through which Facebook handled their post, which we are referring to herein as perceived procedural justice. The scales included in this study were: "How fair was the procedure used when Facebook removed your post?"; "How supported did you feel by Facebook?"; "How well did you understand why your post was removed?"; How clearly did Facebook explain to you why your post was removed?"; "Facebook has the necessary information to make the decision?"; and "How well did Facebook understand your perspective when removing your post?". These items were found to be highly intercorrelated and were averaged to form a single scale of procedural justice (alpha = 0.91). A high score reflects high perceived justice.

Impact upon behavior.

This study focused upon four types of violating content: nudity; hate speech; bullying and other violations. In each case a user had content removed for violating Facebook's Community Standards. Seven days after a removal, users were asked to complete an online questionnaire about the process. The study includes responses received within 14 days of being eligible for the survey. Responses were connected to a user's history, including number of content removals prior to completion of the survey, as well as number of content removals in the 45 days after completion of the survey.

In the case of nudity, 9% of respondents had content removed for nudity after the survey. With hate speech, 2% of respondents had content removed again after the survey. With bullying, 1% of respondents had content removed again after they completed the questionnaire. With other violations, 2% of respondents had content removed after the survey. These lower numbers do not mean that the rate of behavior has necessarily been lowered, because the post-survey data is for a limited period of time. It does demonstrate that the base rates for content removals are very low.

Table 1 shows the influence of perceived fairness on future violating postings. In all four cases and controlling for violating posting prior to the precipitating removal, the perceived fairness of the procedures used by Facebook significantly influenced post removal behavior.

Table 1. Study 1. The impact of perceived procedural justice on future violating postings.

	Future nudity postings	**Future hate speech postings**	**Future bullying posts**	**Future other violating posts**
Pre-removal history	0.080(.005)***	0.126(.003)***	0.074(.003)***	.035(.004)***
Perceived procedural justice of FB	-.026(.010)**	-.018(.002)***	-.005(.001)***	.005(.001)***
Age	0.001(.001)	0.001(.000)***	0.000(.000)	.000(.000)
Gender	0.048(.011)*	0.011(.003)***	0.001(.001)	.009(.003)***
Adjusted R-sq.	1%***	4%***	2%***	1%***
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*				

These results demonstrate that the perceived procedural justice through which Facebook manages decisions about whether to remove violating content has an influence on later rates of rule violation in the cases of nudity; hate speech; bullying and other content. However, the strength of the impact of perceived procedural fairness on later behavior was not the same across all content violation types. In the case of nudity, 23% of those in the study had a prior nudity removal; with hate speech 4% had prior removals and with bullying 2% had prior removals. Hence, the removal experience was more unique with hate speech and bullying. This is reflected in the results. The impact of Facebook actions was weaker with nudity. In the case of other posts, the event was infrequent, but the impact of removal was weak, although significant.

What do these effects mean in terms of user behavior? It is possible to divide perceived procedural justice into four quadrants of approximately equal size and ranging from very fair to very unfair. In the case of bullying, the number of users involved in future removals among these subgroups was: highly fair (0.4%); fair (0.6%); unfair (0.8%); and highly unfair (1.5%). Comparing the groups suggests an approximately 75% drop in the rate of recidivism between the highly fair and the highly unfair quartiles. With hate speech, the number of users involved in future removals was: highly fair (1.0%); fair (1.4%); unfair (2.8%) and highly unfair (5.1%). Comparing the groups suggests an approximately 80% drop in the rate of recidivism between the highly fair and the highly unfair quartiles for hate speech. Finally, with other types of posts the numbers were highly fair (1.2%); fair (1.5%); unfair (1.5%) and highly unfair (1.9%) suggesting a modest drop in repeat violations of about 40%. In all of these cases, it is important to note that the base rates are low; these numbers might best be viewed as a general suggestion that violating posts generally declined after a content removal experience and those declines were greater if users indicated that Facebook procedures were fair.

In the case of nudity, the rate of repeat behavior was higher because people are generally more likely to violate rules about nudity than they are to violate rules about other issues like hate speech. In the case of nudity, the influence of the perceived fairness of the removal process upon later posting of violating content was weaker, although violating posting behavior was still significantly lower when users felt fairly treated. The average number of repeat violations was 0.21 of those who felt highly fairly treated; 0.21 of those who felt fairly treated; 0.25 of those who felt unfairly treated; and 0.28 of those who felt highly fairly treated. In other words, fair treatment decreased the frequency of repeat violations approximately 4%.

Users were also asked whether they were less likely to post similar material in the future and were more likely to say that they would not if they felt that procedures were fair ($r = -.04$, $p < .001$). Similarly, they were more likely to have removed the post voluntarily ($r = -.04$, $p < .001$).

Discussion: study one

The focus of study 1 is upon addressing a broad range of removed violating content and focusing upon repeat behavior. Even with a large sample of respondents (n = 64,042) it is difficult to examine changes in behavior because the frequency of behavior is low. However, the study results suggest that in all three cases there is a clear impact of the perceived justice of Facebook actions. If people feel that they were fairly treated by Facebook, their future rate of posting content which is then removed for violating Facebook's rules declines. The experience of having content removed generally led to lower levels of inappropriate content in all cases. In each case variations in perceived justice significantly shaped later behavior even when controlling for prior history.

These findings support the suggestion that it is important to consider the fairness that users feel they receive from Facebook when it decides whether to remove a violating post. The user experience matters. As noted, Facebook is a private company and people may not view it as a legitimate arbiter of content. Further, the psychological dynamics underlying procedural justice may be weaker in a social media community context than within court systems. These limitations aside, it is clear that perceived justice mattered and that it shaped future behavior, including future hate speech and bullying.

The magnitude of the perceived justice effects was low. One possible reason is that many users did not rate their experience as fair. When asked how fairly their content removal was handled, 48% said unfairly. And 57% said it was unlikely that Facebook understood their perspective. Given these generally negative views, it is impressive that the levels of fairness that people did feel motivated their behavior.

This is a test of the extension of procedural justice models to a new arena involving private authorities and voluntary users. These findings validate the importance of considering the user experience. They also suggest that perceptions of procedural justice can have similar effects in both the private and public context.

Study Two. The impact of general messages on behavior.

This second study is based upon surveys completed by 4,985 users who had content removed for violating Facebook's Community Standards. The first question is whether the perceived justice of treatment shapes later compliance. This is the same issue considered in study one. To examine this question, we again consider a sample of users of one social media platform who have had their account blocked due to the posting of "inappropriate" content. The hypothesis to be tested is that users who feel treated fairly will be more likely to subsequently accept and follow the rules.

The second question is whether messages from the social media company can motivate people to feel more fairly treated. These are general messages indicating that Facebook is concerned about users' needs and tries to explain its policies. The messages are not individualized responses to a particular user about a specific content removal. As such they reflect an approach that is possible even given the large flow of content removal decisions the company must make.

Design

As in study one, responses to questionnaire items were linked to information about future removals to create a single file. The questions asked are shown in Appendix B.

Attitudes

Perceived Procedural justice. Respondents were asked four questions about the fairness of the Facebook content removal process: "How fair was the procedure used when Facebook removed your post?" (67% fair); "How supported did you feel by Facebook during this experience having your post removed?"; "How clearly do you feel Facebook explained to you why your post was removed?"; and "How well did you understand the rules?". These items were combined into an overall scale (alpha = 0.81).

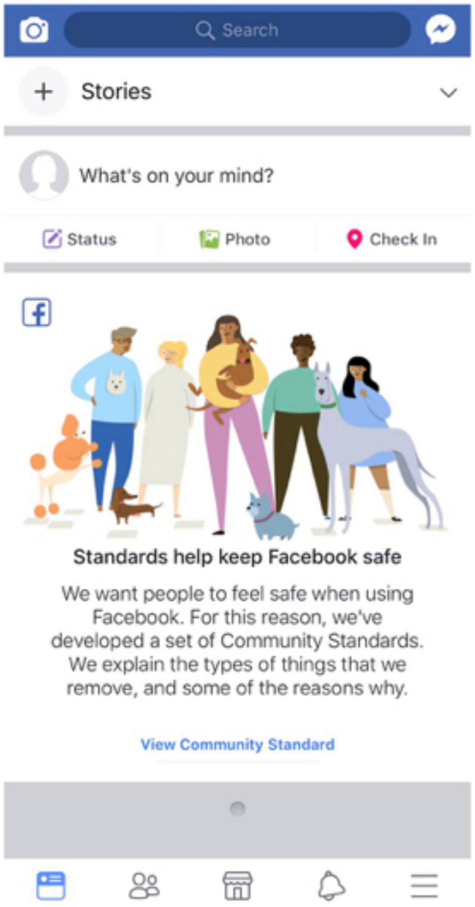
Inconvenience. As a comparison, respondents were asked: "To what extent has being blocked negatively affected you?". Fifty percent of respondents indicated "not at all" and very few indicated any major influence.

Number of restored appeals. This is the success rate of appeals. If Facebook accepts the user's appeal they then restore the original message.

Number of self-initiated take downs. Users have the opportunity to voluntarily remove a post once they have been told that it potentially violates content standards.

Messages

Two messages were composed. One focused on Facebook acting on the basis of safety. It says "Some audiences within our global community may be sensitive to this type of content. It's important that we have one set of standards to keep people safe and welcome on Facebook.". The second message is built around the desire for explanation. It says "We want people to feel safe when using Facebook. For this reason, we've developed a set of community standards. We explain the types of things we remove, and some of the reasons why.". The first message was read by 637 users; the second by 620. The test groups were compared to 1257 control users who received no message.

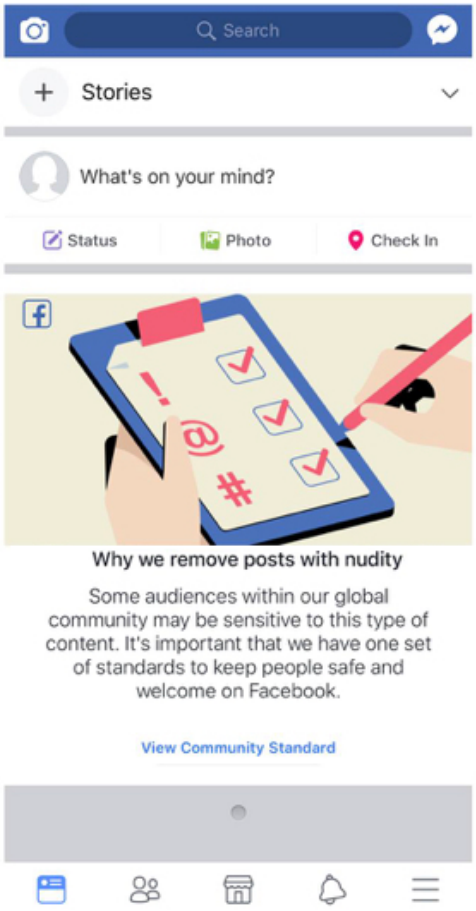


Standards help keep Facebook safe

We want people to feel safe when using Facebook. For this reason, we've developed a set of Community Standards. We explain the types of things that we remove, and some of the reasons why.

View Community Standard

Message 1: "Address user needs"



Why we remove posts with nudity

Some audiences within our global community may be sensitive to this type of content. It's important that we have one set of standards to keep people safe and welcome on Facebook.

View Community Standard

Message 2: "Explain Rules"

Results

The key question is whether the perceived justice of Facebook's content management procedure shaped users' self-regulatory behavior. The first question is whether these individuals expressed regrets. If they felt fairly treated, users were much more likely to express regrets over the post (unstandardized regression coefficient (URC) = 0.996(SE=.047), $p < .001$) and to indicate that they would be less likely to post the same thing again in the future (URC = 0.18(SE=.041), $p < .001$).

An examination of the impact of perceived justice on self-initiated takedowns indicated that more fairness did not lead to more personal deletes (URC = .014(SE=.014), n.s.). Those who felt more fairly treated were, however, less likely to break rules after the survey period (URC = .079(SE=.032), $p < .01$).

This study has a broader focus than the first study. However, it is important to recognize that this study only measures post-exposure content removals, so it is less effective than the behavior measure in study one. Nevertheless, the results replicate those of study one and are extended to other user variables such as regret and intention to do it again.

The influence of fairness can be compared to two outcomes indices. The first is how much the person was impacted by the removal. The second is how many restores they have received from Facebook (a desirable outcome). The results indicate that peoples' rule-oriented behavior is shaped by the impact of the prior removal and by the number of times they have had content restored. It is also distinctly shaped by perceived justice in the case of total takedowns and total appeals.

Table 2. Study 2. Impact of perceived procedural justice.

	Total takedowns	**Total self-deletes**	**Total appeals**
Perceived procedural justice of Facebook actions	-.069(.032)*	0.013(.014)	-.033(.007)***
Number of prior restores (wins)	-1.218(.226)***	0.066(.101)	0.995(.052)***
Impact of removal	-.299(.015)***	0.125(.006)***	0.009(.004)*
	24%	21%	8%
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*			

High scores indicate that Facebook was perceived fair; the user had had prior content removals restored and the impact was high. Totals are high takedowns; self-deletes and appeals.

Messages

Do the "user needs" and "explanation" manipulations influence users? The answer is that both general messages shaped the perceived justice of the experience and altered later rule following behavior.

Table 3. Study 2. Experimental impact of messages.

	Perceived procedural justice	**Total takedowns**
Address user needs vs. control	0.178(.064)***	-.564(.057)***
Explain rules vs. control	0.159(.064)*	-.552(.057)***
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*		

Positive numbers indicate that the message increased perceived justice and negative numbers indicate that it led to fewer Facebook-initiated content removals in the future.

Study Two Discussion

If people feel more fairly treated by Facebook during the content evaluation and removal procedure and possible blocking of their account depending upon their violation history, they are more likely to express regret and to take personal actions to correct the situation. They are also less likely to post violating content in the future.

Study two supports the argument that perceptions of procedural justice can be effective in the social media context. However, this study is limited by the lack of pre-survey behavioral measures and by small sample size. Only posts containing nudity were frequent enough to examine variations in their antecedents.

General Discussion

The social media universe is unlike society. Society is regulated by government authorities who are empowered to sanction and incapacitate rule violators. Those authorities depend heavily upon the willingness of most members of the community to self-regulate most of their behavior most of the time. The key to peoples' willingness to do so is that they view government authorities as legitimate and they feel connected to others as members of a common community or society.

The results of these studies suggest that social media companies can influence self-regulation. And, they support the model that has been widely supported with public authorities: using procedures perceived as fair leads to greater self-regulation. The effects found are broad in scope and include violations for nudity, hate speech, and bullying. They also include user self-deletes and indications of regret.

These findings support the argument that by focusing on the user experience, and in particular the experience of fairness, social media companies stand to gain because their users are more willing to buy into the rules for content moderation and to take more personal responsibility for following those rules. In the long run, this allows companies to better manage the content appearing on their sites, and in particular, to limit the number of views which occur because the content never appears in the first place. It also suggests a strategy that better manages the user experience. As companies increasingly shift from acquiring new users to maintaining their customer base, the loyalty of users will become more important. Fairness based strategies are particularly desirable from this perspective.

Although the effects are clear, it is also important to note their weakness. No single experience would be expected to strikingly change people's ongoing behavior and these studies suggest that no single experience does. The procedures that people experience have an impact, especially in cases (hate speech; bullying) in which content removal is less of an everyday experience.

One important limit of these studies is that the messages in study two are general messages and are not responses to the individual removal. An individualized response would allow people to appeal in some manner in which they could voice their arguments. And they would receive some type of response indicating that they were being heard. Based upon research in legal settings, we would expect a personalized response to have a stronger impact upon an individual.

While more personalized procedures are less scalable, an important research question for the future is whether, and to what extent, such procedures could offset their costs by eliminating downstream costs associated with managing users' accounts and, perhaps more importantly, promoting user loyalty.

References

Gruzd, A., Jacobson, J., Wellman, B. & Mai, P. (2016). Understanding communities in an age of social media. *Information, Communication & Society,* 19, 1187-1193.

Kraut, R.E. & Resnick, P. (2012). *Building successful online communities*. MIT.

Schultz, M.F. (2006). Fear and norms and rock and roll: What jambands can tell us about persuading people to obey copyright law. *Berkeley Technology Law Journal*, 21, 651-728.

Tyler, T.R. (2006). *Why people obey the law*.

Tyler, T.R., Goff, P.A. & MacCoun, R.J. (2015). The impact of psychological science on policing in the United States. *Psychological Science in the Public Interest*, 16(3), 75-109.

Tyler, T.R. & Huo, Y.J. (2002). Trust in the law. *Russell-Sage.*

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

Appendix

Appendix A. Study 1.

Perceived Procedural Justice (six items, alpha = 0.91).

- How fair was the procedure used when FB removed your post?
    - Very fair
    - Somewhat Fair
    - Somewhat Unfair
    - Very Unfair
- How supported did you feel by FB?
    - Completely
    - Somewhat
    - A little
    - Not at all
- How well did you understand why your post was removed?
    - Completely
    - Somewhat
    - A little
    - Do not understand
- How clearly did FB explain to you why your post was removed?
    - Very clearly
    - Somewhat clearly
    - Slightly clearly
    - Not at all clearly
- FB has the necessary information to make the decision.
    - Agree strongly
    - Agree somewhat
    - neither agree/disagree
    - Disagree somewhat
    - Disagree strongly
- How well did FB understand your perspective when removing your post?
    - Completely
    - Somewhat
    - A little
    - Not at all

Appendix B. Study 2.

Perceived Procedural Justice (four items, alpha = 0.81).

- How fair was the procedure used when FB removed your post?
    - Very fair
    - Somewhat Fair
    - Somewhat Unfair
    - Very Unfair
- How supported did you feel by FB during this experience having your post removed?
    - Completely

    - Somewhat
    - A little
    - Not at all
- How clearly do you feel FB explained to you why your post was removed?
    - Very clearly
    - Somewhat clearly
    - Slightly clearly
    - Not at all clearly
- How well did you understand the rules?
    - Completely
    - Somewhat
    - A little
    - Do not understand

Inconvenience

- To what extent has being blocked negatively affected you?
    - Not at all negatively
    - A little negatively
    - Somewhat negatively
    - Very negatively
    - Extremely negatively