The Procedural Justice Framework for Tech Professionals

A Practical Guide for Building and Maintaining Healthy Online Environments

THE JUSTICE COLLABORATORY
Yale Law School

SOCIAL MEDIA GOVERNANCE INITIATIVE

Table of Contents

Executive Summary

OVERVIEW

Social media plays an increasingly central role in the information landscape.[1] [2] In the United States, platforms host a substantial portion of national and political discourse, and their regulatory approaches have come under growing scrutiny.[3] This report uses the procedural justice theory to provide a framework for building effective content moderation strategies.

THE STATUS QUO

Currently, online platforms primarily rely on a deterrence approach, using punishment to discourage unwanted behavior. Content in violation of applicable rules is taken down, and the platform may sanction an individual with an escalating sequence of punishments from suspension to a permanent ban. The underlying logic is that users follow rules to avoid punishment. This method is not novel and is reflective of our offline criminal legal systems.

THE PROCEDURAL JUSTICE FRAMEWORK

Decades of research from the criminal legal setting, along with more recent research in the online space, suggests that a deterrence approach may not be the most optimal. Studies indicate that platforms can build trust and improve rule adherence by adopting a procedural justice approach. As its name suggests, procedural justice is concerned with the fairness of the **process** through which decisions are made and communicated.

The procedural justice framework is founded upon four pillars:

1. Treating the individual with dignity and respect
2. Giving the individual a voice
3. Maintaining neutrality and transparency
4. Acting with trustworthy motives

These principles may be reflected in the way that policies are designed and enforced. For example, providing explanations for post removals can help individuals understand the platform's decision-making process, making them more likely to respect the outcome even if they did not initially agree with it. When users who violate a platform's rules feel that they have been treated fairly by the platform, **they are less likely to violate these rules in the future.**

There is no one-size-fits-all approach to content moderation. However, if used as part of a multifaceted strategy, **procedural justice can help build trust with users** while stemming the tide of violative content on the platform.

About SMGI

This report was created as part of the **Justice Collaboratory's Social Media Governance Initiative** (SMGI). The SMGI aims to create an online environment that is healthy for society. We envision a place in which communities can freely exchange information and engage in socially beneficial civil discourse. Our approach is distinctive because of our engagement with industry practitioners. We collaborate with several online platforms that are central to our initiative and share ideas across the boundaries of the academy and corporation. We believe it is critical to work as independent scholars, but also to engage with the companies that benefit from applying scientific research.

This report is a reflection of some of our learnings partnering with different platforms to incorporate procedural justice into their products over the past five years. This project was made possible through the generosity of an Stavros Niarchos Foundation grant by the Yale Law School.

AUTHORS

Vivian Zhao is a junior at Yale pursuing a major in economics and a Certificate of Advanced Language Study in Chinese. In addition to serving as a research assistant for the Social Media Governance Initiative, Vivian has worked with the U.S. Department of State, American Bar Association, and Yale Jackson School for Global Affairs. Her interests lie at the intersection of economics, policy, and law.

Jackson Akselrad is a second-year student at Yale Law School, where he served as a research assistant for the Social Media Governance Initiative at the Justice Collaboratory. Jackson has also participated in the Law School's Criminal Justice Clinic and Veterans Legal Services Clinic. Prior to attending Yale, Jackson served in the U.S. Intelligence Community.

Matt Katsaros is the Director of the Social Media Governance Initiative at the Justice Collaboratory at Yale Law School. Before this, he spent the past decade working in the tech industry (two years at Twitter and seven years at Facebook). During that time, he worked as a Researcher and Advisor supporting product teams on various online governance issues from developing machine learning, automation of detection of offensive content, and reshaping systems to better incorporate principles of procedural justice theory after users violate platform rules. Outside of his research interests, Matt is deeply engaged in his art practice working in textiles, natural dyes, and printmaking.

DESIGN

Nicole Lavelle is a designer, researcher, and creative strategist who contributes to teams solving complex, human-centered problems. She works with design consultancies and large technology corporations in the San Francisco Bay Area. She is also an artist who makes work about place and identity.

THANK YOU

Thank you to Beth Parker, Michael Swenson, and Sudhir Venkatesh for providing feedback on drafts of this guide.

What is Procedural Justice?

It has been repeatedly demonstrated that individuals are more likely to obey the law and comply with orders — and less likely to commit future offenses — if they view the legal processes they were subjected to as fair.[4] The central idea of procedural justice is that **people often care more about the process than the outcome**. When an individual views an interaction as just, they are more likely to believe that the respective authority figure is legitimate. This perception of legitimacy in turn drives one's willingness to voluntarily abide by the rules put in place by that authority.

But what makes a fair process? Research points to four main principles that contribute to people's understanding of this concept: treating the individual with dignity and respect; giving the individual a voice; maintaining neutrality and transparency; and acting with trustworthy motives. In this section, we dive into each of these in more detail.

Dignity and Respect

Voice

Neutrality and transparency

Trustworthy motives

Dignity and Respect

Individuals should feel that they are treated respectfully when interacting with the rule-making or rule-implementing authority. In the criminal justice context, this might include aspects of interpersonal communication, like the politeness of a police officer, prosecutor, or judge. Online, this could involve the tone of correspondence from the platform to the user or whether a moderator acts in a manner that comes off as accusatory, dismissive, or hostile.

IN ACTION

An important consideration in translating dignity to online spaces is the way in which automation is used. People who pass through platform enforcement systems often feel like it is a black hole of automation where their cases are merely being screened by computers with no regards to their individuality. Many say that this impersonal treatment leaves them feeling disrespected. While in reality, many platforms are expending tremendous resources to include humans in the review process, users on the other end rarely understand this. Consider ways to individualize and personalize communication so that people know their cases are being carefully considered. When automation is used, be clear about how it is used; conversely, when real people are involved, it can be beneficial to make that known to those on the receiving end.

Voice

People want to be able to express their side of the story. Try to provide opportunities for users to give feedback, share their experience, and be prepared to demonstrate that you are actively listening to them. The pillar of voice has been demonstrated to be valuable to people's conceptualization of a fair process even when it doesn't change the outcome. In other words, being presented with the chance to explain oneself is still very valuable in making a process procedurally just, regardless of the result. While it can be daunting to consider how to incorporate this at the large scale that many platforms operate under, these principles are guiding values and not "all or none." Consider ways to incrementally include feedback and voice in your products.

IN ACTION

Offering users a way to appeal moderation decisions (something that is increasingly required by new regulations) is one method of integrating voice into your process.[5] It's important to note that too often we see platforms build appeal systems that focus entirely on outcomes or trying to "fix" perceived enforcement mistakes. Certainly, an appeal system should be aimed at error correction. However, from a procedural justice perspective, it is more important to ensure people are provided with an opportunity to share their opinion. A system which only offers a button to submit an appeal doesn't actually offer individuals a voice. Instead, providing a text box for a written response, or a series of questions to answer can be a way to integrate voice to a greater extent in the appeals process.

An appeal process is designed for the period after a decision has been made, but it's also important to incorporate voice further upstream. People are more likely to see the implementation of a policy as procedurally just if they had an opportunity to shape it through some sort of participatory process. One way to do this is through a "notice and comment" period on new policies; by publicly announcing potential areas of change, you can allow members to provide input on the development of these rules.

Neutrality and Transparency

Unbiasedness in the process can be achieved by acting with transparency and consistency. If individuals feel that they understand how a decision was reached, such as who participated in the decision-making and what factors were weighed, they are more likely to consider the process fair and abide by the result, even if the outcome was unfavorable.

Early on, many platforms sought to hide or obfuscate their rules, thinking that malicious users would use this information to circumvent them. Of course, it's hard to imagine how individuals are expected to follow rules if they don't know what they are; more damaging is that such an approach undermines the legitimacy of a platform and erodes trust with its users. At a minimum, platforms need to be transparent about what the rules are and actively seek to familiarize community members with them. In an online environment, individuals may view a process as more transparent if they are aware that a platform has rules, understand what they are, and know how those rules specifically apply to their case.

When thinking about consistency, it can be helpful to consider if the process delivers predictably similar outcomes when the inputs are also relatively similar. Individuals may believe that a platform's enforcement is inconsistent, and thus unfair, if certain content is removed some of the time and left undisturbed at others. Likewise, they may feel wronged if their post is taken down but apparently interchangeable posts from accounts with larger followings or different political orientations are not.

IN ACTION

Procedural justice favors transparency that speaks directly to an individual and their circumstances. Often, while communicating with users, platforms will indicate that a rule was broken and point them to the full set of its rules. In these cases, it can be hard for people to connect their post to the guideline it violated; it's best to be specific when possible. It can also be helpful to provide details about how the decision was made, such as clarifying whether it was decided through automation or if a human reviewed the post.

Trustworthy Motives

It's important that people understand the motivations of the authority making decisions and believe that they are trustworthy arbiters. This includes knowing not only what the rules are, but also the rationale for why they exist. At the core of many online platform policies is a desire to create safe spaces and protect people from harms. When communicating your rules, it can be helpful to explain reasons behind their existence.

In a study surveying recent rule violators on Twitter, over half of the respondents indicated that they believed it is extremely or very important to feel safe on the platform. Given this finding, an appeal towards people's desire to keep their online spaces safe will likely strongly resonate with them.

IN ACTION

Take the common case of a parent who shared a photo of their unclothed child jumping through sprinklers on a hot summer day. The platform asks them to remove the post. Often in these situations, the user was simply unaware that a rule applying to this content exists. Most parents, when provided a respectful explanation that photos of unclothed children are removed to protect them from online predators, are happy to comply. In fact, it's easy to see why they may actually feel a higher level of trust in the platform knowing that it is helping to keep children safe. That person is in turn less likely to post similar pictures in the future. Conversely, a process which doesn't properly communicate the rationale and motives behind such a post removal may result in that parent trying to post the same or similar images again.

→ When these conditions are met—that is, **when people feel their experience with a rule-making or rule-implementing authority was procedurally just—individuals tend to regard the authority as more legitimate.** This legitimacy, in turn, results in a higher level of self governance and rule following.

Research Applying Procedural Justice Principles to Online Governance

Several recent studies have explored whether a procedural justice approach translates to online spaces. The results have shown strong support for the theory, demonstrating that procedural justice can be used to build trust and increase voluntary rule-following. Below, we summarize a few of these studies and point to ways these learnings can be used to improve trust and legitimacy across online platforms.

- Procedural Justice and Rule Breaking on Twitter
- Procedural Justice and Repeat Offenses on Facebook
- Procedural Justice and Repeat Offenses on Reddit

Procedural Justice and Rule Breaking on Twitter

A recent study of Twitter users who had a Tweet removed by the platform found that users who felt more fairly treated during the enforcement process were less likely to violate the rules in the future.[6]

STUDY DESIGN

People who had violated one of Twitter's rules in the last 30 days were invited to take a survey on their rule-breaking experience. Participants were asked whether they thought Twitter clearly explained its decision, whether Twitter gave them an opportunity to provide their point of view, whether Twitter treats all users the same when it asks them to remove their Tweets, and whether Twitter treated them with respect. The survey data was paired with participants' rule-breaking actions in the six months before and three months after the survey.

INSIGHTS

- **10% of participants were not aware that Twitter even had rules** (despite recently violating one of them) while another 15% expressed that they were not sure. Of the remaining 75% who were aware of Twitter's rules, most were not very familiar with and had not read them.
- **Participants who felt that Twitter's enforcement process was fairer were less likely to break rules in the future.** Interestingly, a question which asked whether participants agreed with Twitter's decision to remove their post (their perception of the outcome) was not significantly correlated to future violations.

HOW CAN WE USE THIS?

- **There are many opportunities for platforms to better familiarize users to their rules.** In the absence of sufficient understanding, individuals will develop their own folk theories (which are likely to be untrue). People should not be discovering that a platform has rules only after someone has broken one. Introducing rules to users as they join your platform is a great way to socialize your safety efforts from the jump.
- When it comes to actual rule enforcement, this work shows that **incorporating transparency, voice, respect, and other elements of procedural justice can translate to reductions in violations**.

Procedural Justice and Repeat Offenses on Facebook

A similar study of Facebook users who violated the platform's Community Standards found that providing more transparency following a post removal resulted in higher rule-following and a reduction in appeals submitted.[7]

STUDY DESIGN

A survey was sent to recent rule-breakers on Facebook asking about their experience having a post removed. Participants' answers were then matched against rule-breaking behavior before and after the survey. Then, a second study was performed in which the authors used a randomized controlled test to educate a group of users about the platform's rules in the week following their post removal.

INSIGHTS

- The results were consistent with the Twitter study: when controlling for prior rule-breaking behavior, **participants in the survey who felt more fairly treated by Facebook during the removal process were less likely to break rules in the future**. This finding was consistent across different types of policy violations (nudity, harassment, etc.).
- The experiment, which provided the treatment group with an educational unit about the platform's rules in the week following the violation, found that this **additional transparency resulted in a small but significant increase in rule-following** and a significant decrease in appeal submissions. This provides empirical support for the procedural justice approach and aligns with the results of a study which introduced rules to new users joining a subreddit.[8]

HOW CAN WE USE THIS?

- **Education and transparency can have a meaningful impact on rule-following.** The design of enforcement flows is often ephemeral. To return to the platform, people move through informational screens quickly yet lack a way of revisiting these resources. Consider users' interactions with your rules as an ongoing journey with different touchpoints as opposed to a one-time flow. It is likely that you are trying to communicate a lot through a small set of screens. Reminding people of the rules or allowing them to access this information at a time of their choosing may be more effective ways to distribute this information.

Procedural Justice and Repeat Offenses on Reddit

Using a sample of 32 million Reddit posts, this study sought to determine the relationship between removal explanations and future behavior on Reddit.[9]

STUDY DESIGN

Researchers scraped millions of Reddit posts, looking for instances in which posts were taken down by volunteer moderators. Analyzing the removals, they identified whether explanations were provided and if so, the figure through which this occurred (e.g. a bot, human, etc.). The authors then examined if the individual who had their post removed went on to have future posts removed.

INSIGHTS

- The provision of an explanation was associated with a reduction in future post removal rates. **The researchers estimated that requiring explanations for all removals could reduce the likelihood of subsequent post removals by 20 percent**. This higher degree of transparency appeared to improve users' and bystanders' understanding of community guidelines and social norms, thereby reducing tendencies to violate them in the future.
- **The way in which an explanation was provided sometimes had an effect on future post removals.** Specifically, users who were given explanations through more substantial comments as opposed to short tags ("flairs") were less likely to have their content taken down in the future. Encouragingly for platforms operating at large scales, the researchers did not see any difference in future post removals when an explanation was provided by a bot as opposed to a human.

HOW CAN WE USE THIS?

- **People can more quickly understand rules and social norms when they are provided clear explanations in the face of violations.** Avoid brief explanations in favor of more substantial communications drawing on the principles of procedural justice to determine what types of information you'd want to include.

Additional Questions to Consider

Procedural justice is not a one-size-fits-all solution. Instead, it is a theory that provides a helpful framework for building trust. It's important to tailor these principles to fit your specific context. Below are just a few additional considerations as you seek to apply them.

Varying needs.

One study found that users' preferences for content moderation approaches vary based on racial identity, sexual orientation, and the platform they are on. **It's critical to develop an understanding of your user base as well as its needs to create an appropriate moderation tool.**

Legitimacy of the authority.

Research suggests that an authority's position in its community – specifically, whether users view them as an "outsider" – affects people's perceptions of their legitimacy. Finding opportunities to involve community members may result in a higher sense of legitimacy. For example, many platforms, such as Reddit and Discord, use volunteer moderators (often seen as trustworthy leaders within their communities) to perform a significant amount of the moderation.

While there is room for more research in this area, the existing research predicts that the same decisions would be seen as more legitimate if they came from trusted individuals rather than the platform itself. **Platforms may therefore find it valuable to consider how integrated their moderators are within their respective communities.**

Additional Resources

This guide is meant to serve as an introduction and primer to the procedural justice theory while providing ideas for how it can be applied to online environments. If you are looking to dive deeper into this and related research, we have compiled all of the referenced studies into the Endnotes on the following page. We also invite you to visit our website, The Justice Collaboratory's Social Media Governance Initiative, and check out some of the other projects we have recently worked on:

***Sudhir Breaks the Internet* Podcast**

***Reimagining the Internet* Podcast**
Episode 68: Justice That We Can Trust with Tracey Meares and Tom Tyler

Facebook's Data Transparency Advisory Group Report

Special issue of the *Yale Journal of Law and Technology*:
In a New Light: Social Media Governance Reconsidered.

Op Ed: "Spotify must be more transparent about its rules of the road" (*Tech Crunch*)

Presentation of research collaboration with Twitter
(Trust & Safety Research Conference, 2022)

Presentation of research collaboration with Nextdoor
(Trust & Safety Research Conference, 2022)

The Social Media Governance Initiative at the Justice Collaboratory is always eager to talk to industry professionals working on these issues.

Please feel free to get in touch if you'd like to be made aware of future research or convenings, collaborate on a project, or just have questions about this work. Contact: **smgi@yale.edu**.

Endnotes

1. Shearer, Elisa. "More than Eight-in-Ten Americans Get News from Digital Devices." *Pew Research Center*, Pew Research Center, 12 Jan. 2021.

2. Suciu, Peter. "Americans Spent On Average More Than 1,300 Hours On Social Media Last Year." *Forbes*, Forbes Media LLC, 24 June 2021.

3. Mitchell, Amy, et al. "Americans Who Mainly Get Their News on Social Media Are Less Engaged, Less Knowledgeable." *Pew Research Center*, Pew Research Center, 30 July 2020.

4. Tom Tyler's "Why People Obey the Law" is one of the most highly cited works on procedural justice making it an excellent resource for those looking for a starting place to dive more deeply into this topic.

5. Meosky, Paul. "Europe's Digital Services Package: What It Means for Online Services and Big Tech." *Electronic Privacy Information Center*, EPIC, 23 Aug. 2022.

6. Katsaros, Matthew, et al. "Procedural Justice and Self Governance on Twitter." *Journal of Online Trust and Safety*, vol. 1, no. 3, 31 Aug. 2022.

7. Tyler, Tom, et al. "Social media governance: can social media companies motivate voluntary rule following behavior among their users?" *Journal of Experimental Criminology*, vol. 17, no. 1, 27 Dec. 2019.

8. Matias, J. Nathan. "Preventing harassment and increasing group participation through social norms in 2,190 online science discussions." *Proceedings of the National Academy of Sciences*, vol. 116, no. 20, 29 Apr. 2019.

9. Jhaver, Shagun, et al. "Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit." *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, Nov. 2019.