Best practices for moderation

A comprehensive write-up of existing peer reviewed research on moderation and communities for node staff. This covers important context for how digital communities form and operate; various styles of moderation; likely sources of user and moderator strife; and overall best practices for moderation on the Website League.

- Noteworthy community dynamics
- Approaches to moderation
- Shortcomings and difficulties of moderation
- Solutions, conclusions, and prescriptions
- References

Noteworthy community dynamics

- Seering et al. propose there are three simultaneous processes which define moderation and its relation to online communities. These are:
  - (1) Over the course of weeks or months, new moderators are chosen, learn through daily interactions, and develop a moderation philosophy.
  - (2) Moderators interact on a daily basis with users and make individual short-term decisions about specific incidents, ranging from warnings to light penalties and eventually to bans if necessary.
  - (3) Finally, throughout the life cycle of the community, moderators make important decisions about policies that impact how the community evolves, usually in reaction to problems that emerge. [Seering et al. 2019]
- Moderation can be thought of as an act of public work--defined by Boyte and Kari as an activity of cooperative citizenship that creates social as well as material culture. danah boyd likewise considers volunteer moderation as creating, maintaining, and defining “networked publics,” imagined collective spaces that “allow people to gather for social, cultural, and civic purposes”. [Matias 2019a]
- Online communities tend to follow a pattern wherein a “group of early members consolidate and exercise a monopoly of power within the organization as their interests diverge from the collective’s.” [Matias 2019a]
- When people feel like they belong in a space, they are more influenced by the posted norms of a space; they do likewise when behavior is less private, more monitored, and norms are more explicitly enforced. [Matias 2019b]

Approaches to moderation

Top-down

- **Top-down (or centralized) approaches to moderation** are predominant at the platform level and include "individual-level sanctions (such as removing content posted by users and banning users), community-level sanctions (in which entire communities are isolated or banned), algorithmic downranking, and some forms of fact-checking and labeling controversial contentment." [Gilbert 2023]
- Top-down approaches to moderation typically seek to position sites as neutral, displaying standardized guidelines that “perform, and therefore reveal in oblique ways, how platforms see themselves as public arbiters of cultural value” [Schoenebeck, Haimson, and Nakamura 2021]
- Top-down approaches are effective at immediate harm reduction and reducing hate speech. Community-level sanctions are also effective at reducing overall toxicity on a platform. [Gilbert 2023]
- However, top-down approaches have several drawbacks, including:
  1. Content removal is ineffective without transparency and explicit rules; as Gilbert says, "when users are unaware of why content was removed, the community never develops beneficial social norms." But transparency can also enable harassment. [Gilbert 2023]
  2. Bans are less effective in communities where it is easy to make new accounts (especially decentralized communities). Banning may also disproportionately impact users who have been historically marginalized and/or make their living online. Civility moderation algorithms have been found to perpetuate racism. [Gilbert 2023]
  3. Community-level sanctions may lead to radicalization of the sanctioned community; when applied wrongfully to a community, such sanctions can also lead to massive loss of information. [Gilbert 2023]

Bottom-up

- **Bottom-up (decentralized) approaches to moderation** are frequently allowed at the individual level by platforms, but they are not typically framed as moderation. The most common forms are distributed moderation (in which "users make judgements about the quality of content") and individual-level controls such as blocks and filters. [Gilbert 2023]
- Bottom-up approaches often involve more people in moderation decisions than top-down models and are, in a sense, democratic. Because ordinary users can be active participants, such models may receive stronger buy-in than top-down approaches. [Gilbert 2023]

    - In contrast to fixed, top-down approaches, bottom-up approaches essentially allow for users with similar values to come together and engage in networked, collective sensemaking [Jhaver et al. 2018]
- Distributed moderation in particular provides feedback on *what* a community likes and dislikes, or what is acceptable or unacceptable to post. [Gilbert 2023] For moderators and admins, block and filter data can also provide useful metrics. Says Derek Powazek: "when a user blocks another user, that's an important signal. If you watch to see who the most-blocked users are, that's a goldmine of insight into how your users are behaving, and you should absolutely monitor who is getting blocked a lot today." [Powazek 2024]
- However, bottom-up approaches have several drawbacks:
  1. Bottom-up approaches generally demand uncompensated (and often significant) labor from volunteers; frequently this labor must also be supplemented by some degree of automation to be scalable. [Schoenebeck, Haimson, and Nakamura 2021]
  2. Distributed moderation is capable of false-positives, especially when automation is involved. This can have the effect of censoring innocent users and leaving them with little recourse, particularly when forking of lists and tools occurs; it can also be dangerous for falsely-added users, who may be branded harassers, abusers, or other negative traits. [Jhaver et al. 2018]
  3. Distributed moderation is prone to punishing or burying contrarian views and minority communities. It frequently creates echo chambers by promoting dominant viewpoints and silencing people who are marginalized. [Gilbert 2023]
     - Public opinion can also reinforce existing social prejudice, rendering it unacceptable for involving the broader community in decision-making. [Matias 2019a]
  4. Distributed moderation can be a vector for misinformation if users are not subject matter experts. [Gilbert 2023]
  5. Blocking and filtering are time consuming, ongoing tasks that demand continual user labor. Both actions are also individual; they merely hide problematic actors in a space and do not remove them from it. [Gilbert 2023]

Proactive

- **Proactive moderation** is both a strategy and approach to moderation; as implied by the name, it seeks to handle potentially harmful content before any harm can be done. Proactive moderation is frequently top-down, but can also be bottom-up or its own style. [Gilbert 2023]
- Proactive strategies at the platform level seem largely successful--on Reddit, proactive moderation identified problematic communities earlier and with less human effort. Proactive moderation also seems to build high(er)-quality communities. Facebook Groups in which moderators approved posts had fewer but higher-quality posts, and more active community members that reported fewer posts. [Gilbert 2023]
- Strategies that make users aware of community norms (explicitly posted rules, stickying moderator comments, explaining post removals, etc.) reduce initial rule violations, reduce future post removals, and increase perceptions of fairness. Visible moderator involvement

in the community also reinforces community norms. [Gilbert 2023]
    - Private communication of community norms can be particularly helpful in non-urgent or sensitive circumstances; these ensure that external users cannot cause digressions, and protect users because interventions are not made public. [Gilbert 2023]
    - Moderation visibility is particularly important for marginalized users; transparency allows them to see action being taken but also to highlight inconsistencies in moderator action. [Thach et al. 2022]
- However, Gilbert notes that "proactive models can be difficult to implement because they require prediction and human labor, and risk backlash." They can also replicate harms found in top-down approaches. [Gilbert 2023]

Justice model

- **Justice model moderation** is based on the premise that people and their relationships are what must be mediated online, rather than individual pieces of harmful content. [Salehi 2020]
- Justice models therefore focus on behavior rather than content. As observed by Salehi: "Once the problem of online harm is framed as content moderation, it is already a lost cause for victims." [Salehi 2020]. They also account for the context of behavior--for example, length of time in a community--when interventions become necessary. [Schoenebeck and Blackwell 2021]
- Adopting a behavioral focus allows moderators to distinguish genuine bad faith or harmful actors from other users, and target punitive sanctions against them when necessary. [Schoenebeck and Blackwell 2021]
- Justice models place emphasis on "accountability and reparation to victims of online harms" with the goal of "foster[ing] education, rehabilitation, and forgiveness" [Gilbert 2023]. Any sanctions levied in a justice based model would be proportionate to the violation. [Schoenebeck and Blackwell 2021]

Shortcomings and difficulties of moderation

The moderation process itself

- Moderation often serves to reinforce existing personal (power upheld within relationships), community (power developed within cultural contexts and upheld by groups), and systemic (power upheld by social institutions) power structures. [Gilbert 2023]
- A subset of people in an ostensibly structureless group can come to occupy informal positions of power, creating an unacknowledged structure. [Gilbert 2023]
- Friendships can make moderation work more difficult due to power relationships that developed between members of the moderation team. [Gilbert 2023]
- All moderators, to some degree, manage relationships and navigate power structures in the course of maintaining their communities; this work is typically unacknowledged but an additional source of emotional toll on moderators. [Gilbert 2023]
- When moderators quit or burnout, the most important and most likely reasons are “struggles with other moderators in the group” and “too little available time”. [Schöpke-Gonzalez et al. 2022]
- Moderation processes may exhibit herding or an "information cascade" effect, in which previous decisions oblige moderators to make similar decisions. [Lampe et al. 2014]

The material being moderated

- Online environments have a platonic ideal of contribution: too few contributors and shared interpersonal interactions or experiences become difficult, but too many contributors and information overload results. When a space undergoes information overload, the sheer amount of information being created leads to an inability to make a decision or stay informed; users begin to participate more simply or withdraw from doing so at all. [Lampe et al. 2014]
  - Public discussions in online spaces can be overloaded--intentionally or unintentionally--through uncivil discussions, flaming, trolling, or even messages that are just off topic. [Lampe et al. 2014]
  - Users "express a greater intent to comment in conversation environments that include continuous monitoring and enforcement of moderation policies." [Matias 2019b]
- Because moderation is governance, its perception as legitimate is tied to community acceptance; moderator decisions with negligible community buy-in are problematic.

[Matias 2019a]

- A number of unique factors--persistence, searchability, replicability, and invisible audiences--make online harassment uniquely harmful for its targets. [Jhaver et al. 2018]
  - Because online contributions can exist indefinitely and algorithmic can resurface previous traumatic content, the possibility of online re-traumatization is substantial. [Scott et al. 2023]
  - Online harassment has a chilling effect: after significant incidents of harassment, many users will begin to censor themselves for fear of being harassed for what they say. [Scott et al. 2023]
- Targets of offensive or harmful content are often not brought into or allowed to be visible in the moderation process; this forecloses any potential for restorative justice or reparation of harm. Schoenebeck et al. note that “processes optimized solely for stopping harassment are unlikely to address the larger impact of the harassment on the targeted user.” [Schoenebeck, Haimson, and Nakamura 2021]. See also Salehi's observation that "Once the problem of online harm is framed as content moderation, it is already a lost cause for victims." [Salehi 2020].

User and moderator considerations to account for

- Users and moderators may have (or develop) varying forms of trauma that should be considered; conversely, these may be avenues for harassers to exploit. Per Scott et al. these include but are not limited to:
  1. **Individual trauma**, wherein users or moderators experience or have experienced harassment through direct messages or public venues;
  2. **Interpersonal trauma**, which is frequently caused through unwanted, persistent, reoccurring, and/or hateful messages. This form of trauma is frequent in circumstances of abuse or intimate partner violence;
  3. **Secondary or vicarious trauma**, wherein users and moderators exposed to harassment of others become traumatized by the experience themselves. This particular form of trauma can lead to burnout;
  4. **Developmental trauma**, wherein users are exposed to age-inappropriate and traumatic content. This form of trauma is particularly frequent in the context of attempted or successful grooming;
  5. **Group or collective trauma**, wherein an entire group or identity are experienced to harmful content. This is frequently experienced during networked harassment or harassment campaigns, and can also occur during violent or impactful events ranging from pandemics to terrorism;
  6. **Racial and cultural trauma**, wherein collective experiences of racial or cultural harm (anti-Blackness, anti-Indigenous racism, antisemitism, etc.) are aggravated or reaggravated through online [Scott et al. 2023]
- Users and moderators may feel trauma through the loss of access to social media, especially if they are marginalized and rely on social media for social connections. [Scott

et al. 2023]

- Retraumatization becomes likely when one or more of the following conditions are met: users and moderators are obliged to continually tell their story; are treated as numbers; are seen as labels and not people; feel unseen and unheard; and are uninvolved with moderation processes. [Scott et al. 2023]

Solutions, conclusions, and prescriptions

Cultural norms

- Content warnings are of debatable effectiveness and, when poorly implemented, can be more harmful than not. They are most valuable in a broader, trauma-informed approach. [Scott et al. 2023]

Actions before the moderation process

- A moderation team's structure should be explicitly delineated or informally acknowledged; moderators should also generally oppose formal and informal power structures. [Gilbert 2023]
- It is important to acknowledge that part of moderation work is navigating and mediating power structures; most moderators quit for interpersonal reasons, so good working relationships within a mod team are vital to prevent attrition. [Gilbert 2023; Schöpke-Gonzalez et al. 2022]
- Proactive strategies (and clear setting of contribution expectations) tend to reduce user rule violations, reduce future user recidivism, and increase user perceptions of fairness [Gilbert 2023]. Although it may lead to fewer contributions, proactive strategies will generally lead to higher-quality contributions; incentivize a more diverse set of contributors; and lessen the chances of information overload [Lampe et al. 2014]. Users also "express a greater intent to comment in conversation environments that include continuous monitoring and enforcement of moderation policies." [Matias 2019b]
  - Transparency and visibility of moderator action is particularly important to minority communities; this allows them to see action being taken but also to highlight inconsistencies in moderator action. [Thach et al. 2022]
  - Private communication of community norms is valuable both in non-sensitive and sensitive circumstances; external users cannot cause unnecessary digressions, and private contact lessens the feeling of persecution because interventions are not made public. [Gilbert 2023]
- First-time participators are unlikely to know expectations and are more likely to violate community policies than experienced community members. Best practice to proactively integrate new users is to welcome them; describe expectations and consequences for rule violations; and explain how enforcement is done, who does enforcement, and the level of enforcement. Such an intervention leads to stronger compliance and higher levels of

participation. [Matias 2019b]
    - Platforms should teach users how use the anti-abuse and distributed moderation tools available to them. [Jhaver et al. 2018]
- Moderation and rule enforcement receives the strongest buy-in from users when they feel like they belong; buy-in also occurs when spaces are less private, clearly monitored, and expectations are consistently enforced when posted. [Matias 2019b]
- Metrics can be valuable for identifying community norms, patterns of problematic behavior, and trouble users or communities. Distributed moderation allows direct community feedback on acceptable and unacceptable behavior [Gilbert 2023]; block and filter data allows for moderators and administrators to both see the macro- and microdynamics of their communities, and where interventions are necessary. [Powazek 2024]
- Universal community surveys, community check-ins, and self-reported data on potentially triggering, harmful, sensitive, or traumatic content (e.g., check boxes for all that apply) can be valuable for establishing particular areas of community consensus or need; when asked universally, such questions can be inclusive and prevent users with a history of trauma from being singled out. [Scott et al. 2023]
- Empowering users with tools to curate and manage their own experience can provide a solution to last-mile moderation problems (circumstances where moderator intervention may not be possible, but users still feel action is warranted).
    - Granular and flexible visibility options are valuable, particularly for minority communities. Allowing users (or specific posts by users) to be selectively visible can create regulated spaces which minimize disruptive participation and protect users involved in sensitive conversations [Gilbert 2023]. More broadly, visibility options give users a degree of privacy and significant control over when, how, and in what ways they are perceived.
    - Audience governing tools (settings that manage who can interact with a user) are also helpful, especially for marginalized users at risk of harassment or users with specific curation needs. [Scott et al. 2023]

Actions during and after the moderation process

- It is beneficial for moderators to assume their users, more likely than not, have a history of traumatic experiences that should be factored into removals, interventions, restorative procedures, or sanctions. [Scott et al. 2023]
- Graduated sanctions (sanctions that increase with severity for each offense) increase the perceived legitimacy and effectiveness of sanctions overall. [Jhaver et al. 2018]
- Different moderator actions have different outcomes, may be desirable in different circumstances, may be supported or not supported by differing groups of users, and can be reconciled more or less easily with certain approaches to moderation.
    - Top-down strategies are generally effective at immediate harm reduction and reducing hate speech. Effectively-applied community-level sanctions reduce overall

toxicity on a platform. [Gilbert 2023]
    - **Banning actions** [defined by Schoenebeck et al. as “banning the person from the site”] are broadly popular but retributive; these are most associated with top-down moderation and may disproportionately affect minority communities if poorly applied. [Schoenebeck, Haimson, and Nakamura 2021]
    - **Apology actions** [defined by Schoenebeck et al. as “requiring a public apology from the person”] that express responsibility and remorse are considered highly fair, just, and desirable by most users; these are a cornerstone of justice model moderation and enable a gradated sanction before banning a user. Care should be taken, however: some vulnerable groups may suffer disproportionate harm from non-genuine apologies, and in some cases this may lead to further harassment. [Schoenebeck, Haimson, and Nakamura 2021]
    - **Educating actions** [defined by Schoenebeck et al. as “educating the person about your identities and experiences”] are desirable to racial minorities and queer people, who frequently feel obligated to explain and justify their identities. [Schoenebeck, Haimson, and Nakamura 2021]
    - **Exposure actions** [“allowing you to have more exposure to a large audience on the site”] are considered broadly just, fair, and desirable to previous sufferers of harassment; however, groups such as women and queer people are largely unfavorable to this. [Schoenebeck, Haimson, and Nakamura 2021]
    - **Listings of online offenders** [defined by Schoenebeck et al. as “adding the person to an online public list of offenders”] are considered just, fair, and desirable by most users; however, these are problematic. Shaming behaviors often lead to disproportionate harm or punishment in online spaces, and they frequently dehumanize and degrade the persons being shamed. To the extent that this is effective, it is primarily effectively if it shames a violation, not the person who did the violation. [Schoenebeck, Haimson, and Nakamura 2021]
    - **Payment actions** [“paying you and your supporters”] are not considered just, fair, or desirable even by minority groups. [Schoenebeck, Haimson, and Nakamura 2021]
- Users often draw comfort from sharing their own experiences and seeing support from others over these experiences. [Jhaver et al. 2018]
- Both moderators and users benefit from continuous discussions about acceptable behaviors; these discussions contribute to community growth and development. [Seering et al. 2019]

References

1. Sarah Gilbert. 2023. *Towards Intersectional Moderation: An Alternative Model of Moderation Built on Care and Power*.
2. Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. *Online Harassment and Content Moderation: The Case of Blocklists*. https://dl.acm.org/doi/10.1145/3185593
3. Cliff Lampe, Paul Zube, Jusil Lee, Chul Hyun Park, and Erik Johnston. 2014. *Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums*. https://linkinghub.elsevier.com/retrieve/pii/S0740624X14000021
4. J. Nathan Matias. 2019a. *The Civic Labor of Volunteer Moderators Online*. https://journals.sagepub.com/doi/10.1177/2056305119836778
5. J. Nathan Matias. 2019b. *Preventing harassment and increasing group participation through social norms in 2,190 online science discussions*. https://www.pnas.org/doi/full/10.1073/pnas.1813486116
6. Derek Powazek. 2024. On Blocking. *Powazek.com*. https://powazek.com/posts/3591
7. Niloufar Salehi. 2024. Do no harm. *Logic Magazine*. https://logicmag.io/care/do-no-harm/
8. Sarita Schoenebeck, Oliver L Haimson and Lisa Nakamura. 2021. *Drawing from justice theories to support targets of online harassment*. https://journals.sagepub.com/doi/10.1177/1461444820913122
9. Sarita Schoenebeck and Lindsay Blackwell. 2021. *Reimagining social media governance: harm, accountability, and repair*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3895779
10. Angela M. Schöpke-Gonzalez, Shubham Atreja, Han Na Shin, Najmin Ahmed, and Libby Hemphill. 2022. *Why do volunteer content moderators quit? Burnout, conflict, and harmful behaviors*. https://journals.sagepub.com/doi/10.1177/14614448221138529
11. Carol Scott, Gabriela Marcu, Riana Elyse Anderson, Mark Newman, and Sarita Schoenebeck. 2023. *Trauma-Informed Social Media: Towards Solutions for Reducing and Healing Online Harm*. https://dl.acm.org/doi/10.1145/3544548.3581512
12. Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. *Moderator engagement and community development in the age of algorithms*. https://journals.sagepub.com/doi/10.1177/1461444818821316
13. Hibby Thach, Samuel Mayworm, Daniel Delmonaco, and Oliver Haimson. 2022. *(In) visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit*. https://journals.sagepub.com/doi/full/10.1177/14614448221109804