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]

Revision #1
Created 2024-12-01 02:54:02 UTC by Alyaza Birze
Updated 2024-12-01 02:55:20 UTC by Alyaza Birze