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]

Revision #1
Created 1 December 2024 02:56:23 by Alyaza Birze
Updated 1 December 2024 02:57:11 by Alyaza Birze