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]

Revision #1
Created 2024-12-01 02:55:30 UTC by Alyaza Birze
Updated 2024-12-01 02:56:13 UTC by Alyaza Birze