Mitigating Harassment in Online Communities with Human-Bot Moderation: Insights from Reddit Communities

An Nguyen[[1]](#footnote-20)

Arun Rai[[2]](#footnote-21)

Likoebe Maruping[[3]](#footnote-22)

This paper sets out to explore the duality of human-bot moderation in mitigating harassment. We examine communities that use a block-list type of bot to prevent harassment from the source of harassment. We expect that the employment of the bot alongside human moderation will create a *shielding effect*- a declining trend of harassment towards community members, followed by an *emboldening effect* - an uptick in harassment by community members towards their perceived outgroup members, and, finally, a *spillover effect* - an increase in harassment in neighbor communities who share the same topic of discussion but not the same moderators. We use Detoxify, a BERT-based model, to determine the probability that a comment represents harassment. . We then use Bayesian Structure Time Series to examine the three types of effects of human-bot moderation in the communities under study

# Introduction

The need to converse emerged since the dawn of human civilization. In this Internet age, that need for face-to-face conversation has evolved into the need for online conversation, which ranges from decade-old Internet Relay Chat and forums to today’s newsgroup and social networks. Not all conversations are civilized conversations. In online settings, because of online anonymity, people may engage in even more harmful activities Duggan (2021b), such as harassment and cyberbullying. Harassment does not only negatively affects individuals who participate in online communities but also harms said communities as a whole.

In a recent survey by Pew Research (Vogels 2021) , 41% of American states that they have personally experienced some forms of harassment, while 25% of those surveyed stated that they even received threats, stalking, sexual and sustained harassment. Although the consequences of online harassment on individuals are more prominent in the short term, a few people, especially women who experienced it, reported that they suffered negative long-term impact (Duggan 2021a). More than that, harassment also poses great threats for the community as a whole, ranging from the quality erosion of content to exclusion of certain groups of contributors. A prominent example of this is 8chan, a loosely moderated image-board site where users can post images annonymously about anime, popular culture, to politics and sports (Wikipedia n.d.). From its inception in 2013, the site was gradually dominated by a group of users with extreme ideologies, who drove away other genuninely interested contributors along with a variety of other topics. The site then turned into a home for antisemitism, misogyny, and anti-immigration ideologies. It was associated with a series of mass shooting in 2019: the New Zealand mass shooting at two mosques (Regan and Sidhu 2019), El Paso mass shooting at Walmart (Mezzofiore and OŚullivan 2021), and Dayton, Ohio mass shooting (Paul P. Murphy and Levenson 2019). Consequently, the site was shut down by its network infrastructure provider (Cloudfare 2019), web service (Robertson 2019). Such incidents along with other related concerns pressure platforms to mainly focus on addressing harassment.

Most platforms used human moderation as the immediate solution when they first encountered online toxicity. However, moderation by just human ran into several issues for platforms owner. *First,* as platform expands, governing with just human moderation as traditionally done is costly to scale up, considering how vast platforms have become over the last decades. Gillespie (2018). *Second,* a number of research and reports show that human moderators, no matter paid or volunteer, experience burnout and emotional distress Snyder (2020). Along with that, human moderators cannot be solely relied on in case of emergency. During the Covid-19 pandemic, platforms sent their human moderators home while starting to automate the moderating process Lapowsky (2020). . *Last but not least,* although human judgement is always used as a standard, it known to have biases, especially towards delicate matter such as political ideology Diakopoulos and Naaman (2011). Thus, the future of moderation cannot depend on human actions alone.

Preparing themselves for the future of moderation, platform, and community owners shifted from using mainly humans to relying mostly on machines for moderation tasks, reasoning that machine is faster at scale Gorwa, Binns, and Katzenbach (2020).

For example, Wikipedia implemented a wide range of bots to automate tasks on each Wiki page, one of which is the ClueBot NG bot. This bot claims to detect whether an edit is an act of vandalism (Wikipedia 2021a). Facebook also used AI intensively in its online moderation ranging from hate speech to misinformation detection (Schroepfer 2021).

Consistency is also a good virtue that machine brings to moderation tasks. In the case of Covid-19 pandemic, technology companies boasted their innitial results of switching to machine moderation. In a report, Facebook stated its independence of human moderators stating that 95% of the hate speech they have taken down was performed by Artificial Intelligence (Schroepfer 2021) despite several criticisms for dismissing its human moderators Stokel-Walker (2020). Despite its rising popularity in recent years, moderation relying solely on machines ignites a whole new level of concern given the inherently complicated landscape of moderation. *First*, as with many other tasks being automated, there is a burning question about whether machines can totally replace humans in such delicate matter. In this vein of discussion, Gillespie (2020) argues that although machine-based moderation is inevitable, humans must remain in the loop. The bias could come from the training datasets as proven in Binns et al. (2017). Gorwa, Binns, and Katzenbach (2020) also raised the same concern by arguing that algorithmic moderation could create injustice in large-scale socio-technological systems. *Secondly*, in a similar vein with Gorwa, Binns, and Katzenbach (2020), Gillespie (2020) concerns about transparency and accountability should there be no human in the loop. *Moreover*, the platform’s point of view about the trade between free speech and safety, Gorwa, Binns, and Katzenbach (2020) pointed out that purely algorithmic moderation would undermine the political nature of speech. *Last but not least,* Mark Zuckerberg admitted himself that machine is not sensitive to “nuances” in languages or the intent behind the comment yet, which inevitably can lead to misclassification (Canales).

Past discussions on platforms’ governance suggest that neither relying on solely humans nor machines works effectively for online moderation against harassment. Thus, human-machine moderation is a viable solution for the future of platform governance. More recent discussions, research included, on online moderation shifted their attention to human-machine moderation (Chandrasekharan, Gandhi, Mustelier, & Gilbert, 2019; Jhaver, Birman, Gilbert, & Bruckman, 2019; Kiene & Hill, 2020; Kiene, Jiang, & Hill, 2019). Through-out these studies, Reddit, Discord, and Twitch stand out as the most studied platforms for human-machine moderation. These platforms allow customized moderation at the community level: each community is run by a team of moderators, human and machine included. As a result, this decentralized governance invites a variety of governing modes across different communities. Kiene, Jiang, and Hill (2019) and Kiene and Hill (2020) studied the successful use of bot moderators when human moderators faced an exploded amount of content on Discord and Reddit. Seering et al. (2018) discovered that the moderation bot also played a social role in facilitating discussion on Twitch. Not only does the literature explore the overall effect of human-bot moderation, but it also dives into specific bots’ effects. Chandrasekharan et al. (2019) claimed to study the first “open source, AI-backed socio-technological moderation systems” - the Crossmod. Although the study confirmed the bot’s superior performance judging from human moderators’ positive feedback, in reality, it is not used as much as other bots on Reddit. As of July 2022, the bot seems to cease operation, judging from the two communities it is monitoring. The mixed results of human-machine’s moderation in research could be due to the fact that not all human-bot moderation is executed the exact same way.

Generally, there are two mechanisms on which these moderation machines are built on. Consequently, there are two ways in which human and machine collaborate in moderating task. *The first mechanism*, also the more news-dominating one, is *based on Artifical Intelligence* (AI). The moderation machine that Facebook is using right now is a great example for the AI-based moderation machine (Schroepfer 2021). Another example is Wikipedia’s ClueBot NG bot, which claims to autonomously detect whether an edit is an act of vandalism (Wikipedia 2021a). Because of the autonomous property of the technology, AI-based bot may be associated with *human-out-of-the-loop* approach, which does not require human interaction in the process, rather than *human-in-the-loop* (HITL) approach. *Another mechanism* on which moderation machine is built is based on rules. Human will define rules and adjust the parameters according to their needs. Opposed to the AI-based machine, rule-based machines involves human’s interaction to adjust the parameter as human see fit. Often, these rule-based machines are embodied in bots. While getting little to no attention on the news headlines, rule-based bots are surprisingly common among platforms like Reddit and Discords.

As the literature suggests (Gillespie 2020), *human-in-the-loop* moderation is a better fit for moderation at community level as opposed to *human-out-of-the-loop moderation*. One prominent example of the two mechanism’s application is the Crossmod bot (Chandrasekharan et al. 2019) and the AutoModerator bot (Jhaver et al. 2019) on Reddit. Automod was developed independently in 2012, then was officially adopted as Reddit’s official tool in 2015. From its grassroots popularity, Automod rose to become the only platform-incorporated machine moderator as well as the most adopted bot across all communities. The authors highly believe that the stark contrast between the performance of Crossmod and Automod is due to the configuration of human-machine collaboration. While Auto¬mod includes human in many of its actions, human-in-the-loop mechanism, Crossmod only involves human’s judgement at the beginning and the end of the process, human-out-of-the-loop mechanism.

Consistent with the examples and what have been theorized in the literature, we argue that human-in-the-loop mechanism works best for collaboration, particularly in the case of community moderation. Adding on to this growing human-bot moderation literature, we aim to explore one type of bot that has grown in popularity on such platforms. Specifically, we studied a block-list anti-harassment bot with human-in-the-loop mechanism. We expect that this human-in-the-loop mechanism will help ease the collaboration between bot and human, which eventually strengthens the good sides of solely human moderation. However, the machine’s block-list feature would work so effectively that the bot will lead the community to radicalism, with little to no space for civilized discussions.

More than that, we also examine the effect of the human-bot moderation on other communties. Since most platforms host a series of communities (e.g., Facebook, Reddit, Discord, etc.), at least some communities within a platform may discuss the same issues. As the literature on moderation provide a strong patterns of spillover effect when moderation measurements are in place Jhaver et al. (2021), we suggest that the introduction of bot to one community’s moderator team may affect other communities with similar topics.

Thus, we propose the two research questions below to investigate this mode of cooperation between human and machine in content moderation. We expect that this collaboration will have effect on both the *focal community*, the community which is under the human-bot moderation, and the *neighbor community*, whose topic of interest is similar to that of the focal commnunity but lack of human-bot moderation team. Specifically, we ask: - How does human-bot moderation impact the level of harassment within the focal community? - How does community-level human-bot moderation impact the level of harassment in other similar communities?

# Theoretical Background

## Toxicity and Harassmnet

Although the topic of online toxicity attracts scholars from various fields, policymakers, and platforms, little has been agreed on about the formal definition of toxicity (Jhaver et al., 2021). In this paper, we define toxicity similarly to the general belief, in which toxicity is an umbrella term for any daily offensive language, harassment, hate speech, etc. Under this umbrella term, we identify *harassment* and *non-harassment*. *Harassment* is any type of language that causes a reasonable person to feel unease or uncomfortable. The necessary condition for a harassing comment is that it must be targeted at either a person or a group. Otherwise, it is non-harassment.

Researchers have yet to agree on any unified classification in harassment. In this paper, we are interested in two specific definitions that fall into harassment: identity attack and insult. Identity attack is a ”negative or hateful comment targeting someone because of their identity”, while insult is ”inflammatory, or negative comment towards a person or a group of people” (PerspectiveAPI, n.d.). Below is an example of the two types of harassment in the community r/femaledatingstrategy:

* Insult: “I diagnose you with retardation”; “youre honestly a disgusting person and its sad you exist. maybe one day you’ll stop hating yourself and mature, but probably not. have a good day you cancer-to-society. I can only hope you all realize how insane and toxic you are”; “this subreddit is fucking evil - a man.”
* Identity Attack: “Shut up feminist”; “If women are so smart and capable compared to men how did they get subdued for all of human history so easily?Like, really easy. Like, pervasively and unanimously easily.You’re like the right wing nutjobs that screech about how jews have infiltrated all of high society and are masterminds yet somehow are brainless, genetically inferior slugs that deserve to be wiped out”;“F\*ck these n\*ggas.”

As in the two examples above, not all comments use profane language to be classify as harassment. This is due to the fact that language can be used in a seemingly civilized way but not to construct a civilized conversation. Furthermore, as we explore further in the below sections, language use in a community will evolve due to the community’s characteristics as well as evade moderation.

## Social Catergorization Lense

Much has been researched about the language of harassment. In this paper, we aim to investigate the source of harassment and its underlying mechanism under the condition of human-bot moderation. We use the Social Categorization theory lens(Kawakami, Amodio, and Hugenberg 2017) to understand why harassment exists among different group of people. Kawakami, Amodio, and Hugenberg (2017)’s synthesis proposed that human generally put other humans into groups to navigate the complex world while minimizing their cognitive efforts. This effect intensifies under the condition of uncertainty and excessive information, which characterizes the Internet and online communities well. The most effortless and basic categorization is simplifying other people into two groups: *ingroup* versus *outgroup*, or *us* vs. *them*. This mentality plays an important role in motivating humans to conduct acts of harassment.

### Source, Target, and Acts of Harassment

Because of social categorization, an individual may have positive feelings towards the in-group members and negative feelings towards the outgroup members (Dovidio, Kawakami, and Gaertner 2002). This bias serves as a source of hostility, derogation, and intent to harm to the outgroup members (Pettigrew and Meertens 1995). We will refer these groups with intent to harm outgeroup members as the *source of harassment*. The target for the ingroup’s derogation is the perceived outgroup, which will be called the *target of harassment* in the study.

In online communities, individuals interact with one another using text-based and image-based exchange. Since the communities lack of the real world’s physicality, the *act of harassment* can be seen through these exchanges: text and images. In this study, we emphasize on the language side of the exchange. Based on the language used in the communities, we identify whether the language is harassment or not; thus, whether there is an *act of harassment* or not. (**fig1-source-target-act?**) presents the relationship among the three components we presented: the source, the target, and the act of harassment.

|  |
| --- |
| Figure 1: Source, Target, and Acts of Harassment. The circle depicts online communities |

### Focal Community and Neighbor Community

*Good. I wonder if we might flip the order of the argument in this paragraph. Instead of opening with definitions of focal versus neighbor community, we could open with the idea that we want a broader accounting of the effects of human-bot moderation. Such approaches could be locally beneficial but broadly deleterious in their effects. To explore this idea, we distinguish between the focal community—where human-bot moderation is being enacted—and neighbor community—which are adjacent (content/topic-wise) but do not use such moderation.*

Based on the social categorization lens, acts of harassment are best manifested between the two contrasted groups based on their characteristics such as gender, race, age, etc. In this study, we focus on one of those characteristics: gender. Recent years have seen an explosion of attention to manosphere, an umbrella term for online communities with misogyny and anti-feminism propaganda (Chandrasekharan, Jhaver, Bruckman, & Gilbert, 2020). These communities are dangerous because of their linkage to real life mass shootings (Chemaly, 2015). Some most popular and prominent manosphere online communities include r/TheRedPill, r/mengoingtheirway, r/pickupartists on Reddit, 4chan’s “ROBOT 90001” , (**4chan-4chan-incels?**). The quick growth and spread of these communities draw a stark contrast between male and female’s value in online communities. In response, many ultra-feminist groups were established as a safe space for women to share and discuss misandry. These groups are relatively younger than misogynistic groups; thus, they accumulate enough platform experience to live among other healthy communities, specifically the use of bots against any acts of harassment.

Similar to misogynistic communities, misandrist communities often exist in clusters, meaning that there exists more than just one community (e.g., r/femaledating trategy, r/wgtw, r/WitchesVSPatriarchy, etc.). Some of these communities utilize bots as a defense mechanism against the source of harassment while others do not. This fact enables us to study beneficial effect of human-bot moderation locally and its deleterious effect at large. Specifically, we distinguish between the *focal community* and *neighbor community*, with *focal community* being the commnunity the bot was implemented while the *neighbor community* being the community with similar topic of discussion but not the human-bot moderation.

The local and adjacent effect on focal and neighbor communities are common in the online community stream of literature, especially when the focus is on evaluating the effects of moderation strategy. For example, in studying deplatforming, removing one person from the platform, (Jhaver et al. 2021) concludes that deplatforming helps reduce the level of offensiveness and toxicity in that platform However, (Ali et al. 2021), (Rauchfleisch and Kaiser 2021) discovers that deplatforming pushes extreme users to more toxic venues outside of the platform. Thus, we argue that to study the whole picture of human-bot moderation, one must examine its focal effects and its spillover, which, in this case, the effects on focal communities and neighbor communities.

## Human Machine Moderation

*Do we need a broad conceptualization of bots somewhere? Either here or somewhere in the introduction?* In our research, we focus on the rule-based bot rather than AI-based bot because of the two following reasons. *First*, although AI-powered moderation has been significantly advanced in these years and proudly used by platform owners Schroepfer (2021), community moderators still opt for rule-based bot moderators. Prominently, the rule-based Automod rose to popularity and recommended by Reddit without any AI-backed algorithm. Moreover, on Reddit and Discord almost all of the most adopted moderation bots are also rule-based rather than AI-powered bots www.top.gg (2022). One can argue that these human moderators are not paid or trained to fine-tune moderation bots, thus the low adoption rate. Another possible explanation for rule-based bot’s popularity is that these unpaid, volunteer moderators do the work out of their passion for the community. Thus, they require the bot to be more controllable. In other words, the more transparency the bot is, the more collaborative it is to these human moderators. *Second*, no matter under which mechanism the bot is based on, the decision of moderation is of high sensitivity and often finally made by human. The main difference among human-machine moderation strategies is not what type of bot is implemented but how much and at which stage human’s action is involved. As we ground our research on online communities rather than platform, we focus on the rule-based bots rather than the AI-based ones.

### Human-out-of-the-loop versus Human-in-the-loop Moderation

Human-in-the-loop (HITL) and human-out-of-the-loop (HOOTL) are the two different mechanisms in which humans involve differently in the human-bot collaboration. In both cases, human is the locus of decision. However, we define HITL moderation as the type of human-machine collaboration in which human provides iterative inputs to fine-tune the parameters of the machine and jointly produce the decision. Meanwhile, HOOTL mechanism only requires humans to provide the input only. Once the machine produces the results, human will rely on that result and make final decision. Since HITL involves continuous feedback through various steps of data processing (Wu et al., 2022), it is more practical for delicate matters like online moderation against ha¬rassment. In this argument, we propose a further implication: regardless of the nature of machine, HITL is useful for online moderation. Figure 2 illustrates the two concepts. It does not only dictate how much the bot is used, but it also implies how effective the bot is.

|  |
| --- |
| Figure 2: Human-out-of-the-loop (below) and human-in-the-loop (above) illustration. Adapted from Wu et al. (2022) |

### AI-based Moderation

In detecting harassment, AI-based moderation often relies on the performance of natural language processing (NLP) models. Traditionally, the NLP model will predict whether a text instance is classified as harassment or not. In Crossmod example, the bot scans comments, then detects harmful content via back-end, cross-community Machine Learning algorithm, and finally informs human moderators if a comment is suspected to be harmful (Chandrasekharan et al. 2019). This mechanism fits with HOOTL definition in that human does not interfere with any fine-tuning. As we already cover, this type of moderation falls out of favor because it is not practical for volunteer, unpaid, untrained community moderators. Trujillo et al. (2021) discovered that communities’ specialized language does not stay static. While communities naturally develop their sets of specialized language, the bot is fed on a cross-community data. Inevitably, the bot cannot detect harassment in a specific community’s language as it has never learned it before. Moreover, harmful users may evade ban by using new ways of insulting without explicit profanity that goes undetectable by machine. To resolve this, newly developed framework use Adversarial Attacks (Dinan et al. 2019), which includes human in the loop. The adversarial attack starts with build it process where a language model is applied to detect harassment. Next, crowd-sourced workers, humans, are asked to write harassing comments that go unde-tected by machine to break the model. Then, during the fix it phase, the whole model is retrained with newly collected data. The whole process repeats at build it. This inclusion of human in the model promised a better detection, especially when context-based harassment evolves to avoid machine detection.

### Rule-based Moderation

Although most of the conversations around human-in-the-loop focuses on AI-type of machine, we extend the definition beyond that. Human-in-the-loop also appears in other less sophisticated machine like rule-based bots. Platforms that rely on community level mod-erators like Reddit and Discord made customization accessible to community moderators. For Discord, once a server (community) employed a bot, human moderator only needs a single line command to ask the bot to perform a task or the human mod will see intuitive interface as depicted in Figure 4. Reddit’s open data coupled with its readily available Application Programming Interface (API) and pushshift.io Reddit API allows independent developers to scan contents and perform basic moderation tasks such as remove duplicated posts, detect spams, welcome new members, answer basic conversion questions, etc. To communicate with human moderators, most of the bots allow human moderator to compose syntactic rules in YAML format - human-friendly data language. Figure 3 presents an example of YAML syntax by the saferbot.

|  |
| --- |
| Figure 3: A snapshot of interface between human and rule-based bot on Discord. |

Because the bot’s functions are straight-forward and its interface is intuitive, human can interact with the bot to change the parameter of moderation anytime. For example, human may first set the definition of spammers as any users who post 5 comments per minute. However, they soon learn that normal users can contribute 5 comments in one minute. In that case, human moderators can easily and quickly change the parameters to ban users who post 10 comment per minute. We define this immediate and customized feature as human-in-the-loop mechanism for rule-based bots.

|  |
| --- |
| Figure 4: A snapshot of interface between human and rule-based bot on Reddit. |

### Moderation Strategy

Most commonly, moderation is characterized by its scope of impact: content moderation, user moderation, and community moderation. Content moderation is the lowest level of mod¬eration, which focuses on the specific information exchanged by individuals. By default, virtually all platforms hosting online conversation has rules and regulations to moderate content. For example, Wikipedia’s editor can add, revert, delete, and flag information deemed inappropriate for the platform (Wikipedia 2021b). Twitter also employs a variety of content moderation such as labeling, limiting visibility, and removing Tweets for its users (Twitter 2020). Because of its ubiquity, much of past and current research effort has been spent at content level of moderation.

Srinivasan et al. (2019) found out that *content removal* as a moderation strategy worked best un¬der moderation’s transparency. As platform evolves, so does the challenge for content moderation. Voice-based (Discord) and video-based (Twitch) platform are becoming more relevant today; thus, content moderation departs from merely processing texts to process¬ing voice and images. The challenge was investigated by (Jiang et al. 2019).

An equally popular form of moderation focuses at a higher level: *user moderation*. Twitter has by far more policies on user than most of other platforms. Their user moderation actions include profile editing, placing read-only mode, verifying identity, etc. Reddit, a so¬cial news platform, also impose a number of user-level moderation actions to ensure the health of its communities, namely: banning, muting, shadow banning (Reddit, n.d.). While Twitter’s user ban is a platform decision, Reddit user ban is usually a community level decision as Reddit’s governance is less centralized. Researchers often conduct study on the effectiveness of deplatforming Twitter users with a huge base of followers. Although some studies may show positive effects on the platform after deplatforming vocal, toxic users (Jhaver et al. 2021), others may suggest that deplatforming are not as effective as it is thought to be as users may relocate elsewhere outside of the platform (Ali et al. 2021; Rauchfleisch and Kaiser 2021). Directly related to the current research, block-list type bot is also classified as moderation toward users. (Jhaver, Ghoshal, et al. 2018) presented a well-rounded overview of block-list type of bots on Twitter by inter¬viewing the block source and block target. We are set on a different direction to investigate block-list type bots’ effects on language use rather than community’s sentiments.

The last and the rarest form of moderation level is *community moderation*, which is famously used by Reddit. Different from the majority of social media platforms, Reddit organizes its communities into subreddits, where Reddit users can discuss topics allowed within one community. Similar to deplatforming, moderating community also has its good sides and its bad sides. At first, members of the moderated platform will be less toxic (Chandrasekharan et al. 2017; Trujillo et al. 2021). However, the effects are heterogeneous as some platforms refuse to change their topic (Chandrasekharan et al. 2020) and continue to use toxic language in a way to evade moderation (Trujillo et al. 2021).

To provide our readers with the clearest picture of the type of bot we are pursuing in this study, we include a table of summary type of bots, mode of collaboration, and strategy for moderation in Table 1. While AI-based bots are more popular with platform level moderation strategy, rule-based bots are communities’ favorite tools for bot moderation.

Table 1: Bot Mechanisms and Examples

|  | AI-based Bots | Rule-based Bots |
| --- | --- | --- |
| Human-out-of-the-loop | Facebook’s AI algo­rithms (Schroepfer, 2021) | N/A |
| Human-in-the-loop | Adversarial Attack (Dinan et al., 2019) | Most bots on Reddit and Discord |

## Hypotheses

In the current study, we are interested in block-list bot, a rule-based bot with high degree of human-in-the-loop nature. We would like to examine its effectiveness on communities that are vulnerable to debates (e.g., misandry communities). Specifically, we expect that the

### Shielding Effects of Human-bot Moderation

### Emboldening Effects of Human-bot Moderation

### Spillover Effects of Human-bot Moderation

## Empirical Analysis

### Context

#### Reddit and its moderation policies

#### The bot, *r/saferbot* and the community, *r/femaledatingstrategy*

#### Data Collection, Measurement, and Research Design

### Analysis

## Discussion and Future Research

Ali, Shiza, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. 2021. “Understanding the Effect of Deplatforming on Social Networks.” In *13th ACM Web Science Conference 2021*. ACM. <https://doi.org/10.1145/3447535.3462637>.

August, Tal, Dallas Card, Gary Hsieh, Noah A Smith, and Katharina Reinecke. 2020. “Explain Like i Am a Scientist: The Linguistic Barriers of Entry to r/Science.” In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12.

Binns, Reuben, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. “Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation.” In *International Conference on Social Informatics*, 405–15. Springer.

BotWatch. 2021. “Moderator Bot Growth Statistics Update - December 2021.” <https://www.reddit.com/r/botwatch/comments/rlz7iv/moderator_bot_growth_statistics_update_december/>.

Canales, Katie. “Mark Zuckerberg Said Content Moderation Requires ’Nuances’ That Consider the Intent Behind a Post, but Also Highlighted Facebook’s Reliance on AI to Do That Job.” *Business Insider*. <https://www.businessinsider.com/zuckerberg-nuances-content-moderation-ai-misinformation-hearing-2021-3>.

Chandrasekharan, Eshwar, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. “Crossmod: A Cross-Community Learning-Based System to Assist Reddit Moderators.” *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW): 1–30.

Chandrasekharan, Eshwar, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2020. “Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit.” *arXiv Preprint arXiv:2009.11483*.

Chandrasekharan, Eshwar, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. “You Cant Stay Here.” *Proceedings of the ACM on Human-Computer Interaction* 1 (CSCW): 1–22. <https://doi.org/10.1145/3134666>.

Cloudfare. 2019. “Terminating Service for 8Chan.” <https://blog.cloudflare.com/terminating-service-for-8chan/>.

Diakopoulos, Nicholas, and Mor Naaman. 2011. “Towards Quality Discourse in Online News Comments.” In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, 133–42.

Dinan, Emily, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. “Build It Break It Fix It for Dialogue Safety: Robustness from Adversarial Human Attack.” *arXiv Preprint arXiv:1908.06083*.

Dosono, Bryan, and Bryan Semaan. 2019. “Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit.” In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13.

Dovidio, John F, Kerry Kawakami, and Samuel L Gaertner. 2002. “Implicit and Explicit Prejudice and Interracial Interaction.” *Journal of Personality and Social Psychology* 82 (1): 62.

Duggan, Maeve. 2021a. “Part 4: The Aftermath of Online Harassment.” *Pew Research Center*. <https://www.pewresearch.org/internet/2014/10/22/part-4-the-aftermath-of-online-harassment/>.

———. 2021b. “The Broader Context of Online Harassment.” *Pew Research Center*. <https://www.pewresearch.org/internet/2017/07/11/the-broader-context-of-online-harassment/>.

Dwoskin, Elizabeth, and Nitasha Tiku. 2020. “Facebook Sent Home Thousands of Human Moderators Due to the Coronavirus. Now the Algorithms Are in Charge.” *Washington Post*. <https://www.washingtonpost.com/technology/2020/03/23/facebook-moderators-coronavirus/>.

Geigner, Timothy. 2021. “Facebook AI Moderation Continues to Suck Because Moderation at Scale Is Impossible.” *Tech Dirt*. <https://www.techdirt.com/2021/10/20/facebook-ai-moderation-continues-to-suck-because-moderation-scale-is-impossible/>.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.

———. 2020. “Content Moderation, AI, and the Question of Scale.” *Big Data & Society* 7 (2): 2053951720943234.

Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. “Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance.” *Big Data & Society* 7 (1): 2053951719897945.

Jhaver, Shagun, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2018. “Did You Suspect the Post Would Be Removed?”: User Reactions to Content Removals on Reddit.” *Proceedings of the ACM on Human-Computer Interaction* 2.

Jhaver, Shagun, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. “Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator.” *ACM Transactions on Computer-Human Interaction (TOCHI)* 26 (5): 1–35.

Jhaver, Shagun, Christian Boylston, Diyi Yang, and Amy Bruckman. 2021. “Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter.” *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW2): 1–30. <https://doi.org/10.1145/3479525>.

Jhaver, Shagun, Amy Bruckman, and Eric Gilbert. 2019. “Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit.” *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW): 1–27.

Jhaver, Shagun, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. “Online Harassment and Content Moderation: The Case of Blocklists.” *ACM Trans. Comput.-Hum. Interact.* 25 (2). <https://doi.org/10.1145/3185593>.

Jiang, Jialun Aaron, Charles Kiene, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. 2019. “Moderation Challenges in Voice-Based Online Communities on Discord.” *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW): 1–23. <https://doi.org/10.1145/3359157>.

Kawakami, Kerry, David M Amodio, and Kurt Hugenberg. 2017. “Intergroup Perception and Cognition: An Integrative Framework for Understanding the Causes and Consequences of Social Categorization.” In *Advances in Experimental Social Psychology*, 55:1–80. Elsevier.

Kiene, Charles, and Benjamin Mako Hill. 2020. “Who Uses Bots? A Statistical Analysis of Bot Usage in Moderation Teams.” In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8.

Kiene, Charles, Jialun Aaron Jiang, and Benjamin Mako Hill. 2019. “Technological Frames and User Innovation: Exploring Technological Change in Community Moderation Teams.” *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW): 1–23.

Lapowsky, Isse. 2020. “After Sending Content Moderators Home, YouTube Doubled Its Video Removals.” *Protocol*. <https://www.protocol.com/youtube-content-moderation-covid-19>.

Mezzofiore, Gianluca, and Donie OŚullivan. 2021. “Part 4: The Aftermath of Online Harassment.” *Pew Research Center*. <https://www.cnn.com/2019/08/04/business/el-paso-shooting-8chan-biz>.

Newcomb, Allysa. 2019. “Twitter Says a.i. Is Now Removing over Half of Its Abusive Tweets Before They’re Flagged.” <https://fortune.com/2019/10/24/twitter-abuse-tweets/>.

Paul P. Murphy, Drew Griffin, Konstantin Toropin, and Eric Levenson. 2019. “Dayton Shooter Had an Obsession with Violence and Mass Shootings, Police Say.” <https://www.theguardian.com/technology/2019/aug/04/mass-shootings-el-paso-texas-dayton-ohio-8chan-far-right-website>.

PerspectiveAPI. n.d. “Attributes & Languages.” <https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages>.

Pettigrew, Thomas F, and Roel W Meertens. 1995. “Subtle and Blatant Prejudice in Western Europe.” *European Journal of Social Psychology* 25 (1): 57–75.

Rauchfleisch, Adrian, and Jonas Kaiser. 2021. “Deplatforming the Far-Right: An Analysis of YouTube and BitChute.” *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3867818>.

Reddit. n.d. “Reddit Content Policy.” <https://www.redditinc.com/policies/content-policy>.

Regan, Helen, and Sandi Sidhu. 2019. “49 Killed in Mass Shooting at Two Mosques in Christchurch, New Zealand.” *CNN*. <https://edition.cnn.com/2019/03/14/asia/christchurch-mosque-shooting-intl/index.html>.

Robertson, Adi. 2019. “8chan Goes Dark After Hardware Provider Discontinues Service.” <https://www.theverge.com/2019/8/5/20754943/8chan-epik-offline-voxility-service-cutoff-hate-speech-ban>.

Schroepfer, Mike. 2021. “Update on Our Progress on AI and Hate Speech Detection.” <https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/>.

Seering, Joseph, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. 2018. “The Social Roles of Bots: Evaluating Impact of Bots on Discussions in Online Communities.” *Proceedings of the ACM on Human-Computer Interaction* 2 (CSCW): 1–29.

Snyder, Kristen. 2020. “TikTok Content Moderators Allege Emotional Distress.” <https://dot.la/tiktok-content-moderators-2657593810.html>.

Srinivasan, Kumar Bhargav, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. “Content Removal as a Moderation Strategy: Compliance and Other Outcomes in the Changemyview Community.” *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW): 1–21.

Stokel-Walker, Chris. 2020. “As Humans Go Home, Facebook and YouTube Face a Coronavirus Crisis.” *WIRED*. <https://www.wired.co.uk/article/coronavirus-facts-moderators-facebook-youtube>.

Trujillo, Milo Z, Samuel F Rosenblatt, Guillermo de Anda Jáuregui, Emily Moog, Briane Paul V Samson, Laurent Hébert-Dufresne, and Allison M Roth. 2021. “When the Echo Chamber Shatters: Examining the Use of Community-Specific Language Post-Subreddit Ban.” *arXiv Preprint arXiv:2106.16207*.

Twitter. 2020. “Our Range of Enforcement Options.” <https://help.twitter.com/en/rules-and-policies/enforcement-options>.

Vogels, Emily. 2021. “The State of Online Harassment.” *Pew Research Center*. <https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/>.

Wikipedia. n.d. “8chan,” n.d. <https://en.wikipedia.org/wiki/8chan>.

———. 2021a. “User:ClueBot NG.” *Wikipedia*. <https://en.wikipedia.org/wiki/User:ClueBot_NG>.

———. 2021b. “Wikipedia Is Not a Forum.” <https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_is_not_a_forum>.

www.top.gg. 2022. “List of Discord Moderation Bots.” <https://top.gg/search?q=moderation>.

Yang, Yukun. 2019. “When Power Goes Wild Online: How Did a Voluntary Moderator’s Abuse of Power Affect an Online Community?” *Proceedings of the Association for Information Science and Technology* 56.

1. Georgia State University, anguyen192@gsu.edu [↑](#footnote-ref-20)
2. Georgia State University, arunrai@gsu.edu [↑](#footnote-ref-21)
3. Georgia State University, lmaruping@gsu.edu [↑](#footnote-ref-22)