Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators

ESHWAR CHANDRASEKHARAN, Georgia Institute of Technology, USA
CHAITRALI GANDHI, University of Michigan, USA
MATTHEW WORTLEY MUSTELIER, University of Michigan, USA
ERIC GILBERT, University of Michigan, USA

In this paper, we introduce a novel sociotechnical moderation system for Reddit called *Crossmod*. Through formative interviews with 11 active moderators from 10 different subreddits, we learned about the limitations of currently available automated tools, and how a new system could extend their capabilities. Developed out of these interviews, Crossmod makes its decisions based on *cross-community learning*—an approach that leverages a large corpus of previous moderator decisions via an ensemble of classifiers. Finally, we deployed Crossmod in a controlled environment, simulating real-time conversations from two large subreddits with over 10M subscribers each. To evaluate Crossmod's moderation recommendations, 4 moderators reviewed comments scored by Crossmod that had been drawn randomly from existing threads. Crossmod achieved an overall accuracy of 86% when detecting comments that would be removed by moderators, with high recall (over 87.5%). Additionally, moderators reported that they would have removed 95.3% of the comments flagged by Crossmod; however, 98.3% of these comments were still online at the time of this writing (i.e., not removed by the current moderation system). To the best of our knowledge, Crossmod is the first open source, AI-backed sociotechnical moderation system to be designed using participatory methods.

CCS Concepts: • **Human-centered computing → Collaborative and social computing systems and tools**.

Additional Key Words and Phrases: sociotechnical systems; moderation; AI; machine learning; mixed initiative; participatory design; community norms; online communities; online governance; open source.

ACM Reference Format:
Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. *Crossmod*: A Cross-Community Learning-based System to Assist Reddit Moderators. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 174 (November 2019), 30 pages. https://doi.org/10.1145/3359276

1 INTRODUCTION

Recently, the moderators (or "mods") of a large gaming community on Reddit, r/Games, released screenshots[1] of bigoted, transphobic, racist, misogynistic, pedophilic, and otherwise hateful comments that they had moderated [46]. In their statement, the mods wrote:

[1]Content warning (see above): https://imgur.com/a/umrdBYF

Authors' addresses: Eshwar Chandrasekharan, Georgia Institute of Technology, School of Interactive Computing, USA, eshwar3@gatech.edu; Chaitrali Gandhi, University of Michigan, School of Information, USA, ckgandhi@umich.edu; Matthew Wortley Mustelier, University of Michigan, School of Information, USA, mustelie@umich.edu; Eric Gilbert, University of Michigan, School of Information, USA, eegg@umich.edu.

2573-0142/2019/11-ART174 $15.00
https://doi.org/10.1145/3359276

> Unfortunately, this inflammatory content is not infrequent ... These are some of the more awful comments we see regarding transphobia, homophobia, islamophobia, racism, misogyny, pro-pedophilia/pro-rape, and vitriolic personal attacks against other users. These kinds of comments occur on a daily basis. We've compiled an entire album of examples of the horrible things people say on this subreddit. From bigotry to vitriol, this album merely scratches the surface of the magnitude of the problem.[2]

While most mainstream platforms prohibit obviously racist, homophobic, and hateful content, platforms still intake vast amounts of it [18, 67]. As the r/Games mods tried to make visible, moderators are on the front lines of the battle to keep such content out of their online communities and off platforms [28]. Platforms and their moderators use a variety of different approaches to regulate behavior in online communities, and subsequently limit the damage that bad actors cause [37]. On most sites, those techniques take two primary forms: human moderation, and human moderation augmented by automated techniques.

1.1 Regulating platforms through human moderation

Most social platforms employ the services of moderators (either paid or unpaid) who regulate content generated within the platform. Human moderation typically takes two forms: centralized and distributed approaches. In the centralized approach, teams of human moderators such as externally contracted workers, and/or a small number of power users—manually go through posts, and scrub the site of content with racist, homophobic, or misogynist language or imagery [60]. In the distributed approach, a social platform's users triage inappropriate submissions via voting or reporting mechanisms—after which the site can take action (often, moderators take these actions).

1.1.1 Challenges faced by human moderation. Human moderation approaches require moderators to perform a great deal of manual labor, and these suffer from drawbacks when deployed within large-scale platforms [55, 68]. In the centralized approach, groups of paid or volunteer moderators constantly regulate all of the content generated within platforms. This constant exposure to disturbing content negatively and substantially affects the mental health of moderators [20, 59]. In the distributed approach, platforms require users to report inappropriate content before taking action—the exact type of content platforms wish their users did not have to encounter in the first place. In addition, human moderation struggles to keep up with the immense volume of content generated within large-scale platforms—plenty of content that violates site guidelines remains online for years [28].

1.2 Using AI to triage content for human moderators

To keep up with the volume of content created by users, social platforms—like Facebook [6], YouTube [30], and Twitter [54]—are known to train machine learning algorithms by compiling large datasets of past moderation decisions on the platform. Deploying these algorithms without any human oversight can be detrimental; for example, Tumblr "caused chaos" recently when it launched a new, unsupervised anti-porn algorithm on the site [39]. Nonetheless, machine learning approaches can be especially helpful for algorithmically *triaging* comments for human moderators to review. However machine learning-based approaches face drawbacks that prevent them from being easily deployed—scarcity of labeled ground truth data, and the contextual nature of moderation.

1.2.1 Scarcity of labeled ground truth data. First, machine learning-based approaches require vast amounts of labeled ground truth data for training effective models. These data are difficult to obtain since platforms do not share moderation decisions publicly due to privacy and public relations

[2] https://www.reddit.com/r/Games/comments/b7ubwm/rgames_is_closed_for_april_fools_find_out_why_in/

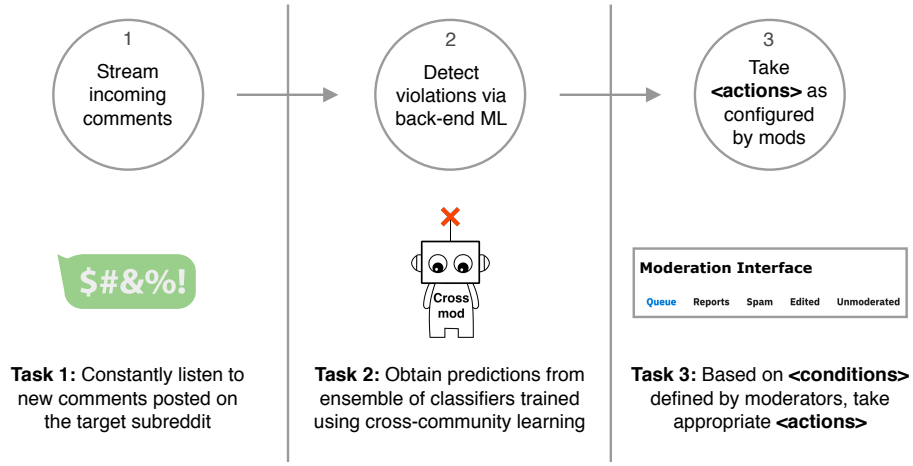

Fig. 1. Broad, illustrative overview of how Crossmod works.

concerns. Brand new or small online communities, by definition, have little to no training data at all. New techniques such as *cross-community learning* can side-step the need for site-specific data and classifiers for moderation [19]. In cross-community learning, data obtained from one or more source communities is used to detect violations within a completely different target community.

1.2.2 Contextual nature of moderation. Second, moderation is a highly contextual task. Moderation decisions about what is considered acceptable or undesirable are guided by an online community's norms [17, 43]. But norms vary widely across communities; even behavior considered undesirable in one community may be valuable in another [5, 50]. A common failure mode for moderation algorithms is failing to understand the community norms where they are being deployed.

1.3 Formative interviews, system, and evaluation

In this paper we build a new, open-source, AI-based moderation system for assisting moderators of communities on Reddit. We call this system the CrossModerator or *Crossmod*. Crossmod aims to overcome the problems above by embracing a sociotechnical partnership with mods, who understand their community's norms. Specifically, we adopt a *mixed-initiative* [31] approach in Crossmod, allowing moderators of subreddits to augment the automatic predictions obtained from cross-community learning with human decisions and oversight.

1.3.1 Formative interviews. We conduct a formative interview study with 11 mods from 10 different subreddits to understand the current state of automated moderation tools on Reddit, as well as opportunities for extending those tools. We also work closely and iteratively with these moderators through all stages of building Crossmod.

1.3.2 System. Next, we introduce the primary contribution of this work: Crossmod. Developed with iterative, participatory methods, Crossmod is a machine-learning based moderation system that is freely available and open source. The machine learning back-end for Crossmod leverages *cross-community learning* [18, 19]; specifically, it uses classifiers trained on the moderation decisions from 100 other communities over roughly a year. For example, Crossmod's ML-backend provides counterfactual estimates about what 100 communities would do with new content, as well as

whether that content resembles racism, homophobia, or other types of abuse. Driven by our formative interviews, Crossmod wraps this backend in a sociotechnical architecture that fits into existing moderator workflows and practices. Figure 1 depicts an overview of how Crossmod works.

1.3.3 Summative evaluation. Finally, we deploy Crossmod in a controlled environment, simulating real-time conversations from two large subreddits with over 10M subscribers each—r/science and r/Futurology. Two moderators from each subreddit evaluated Crossmod's moderation recommendations by manually reviewing comments scored by Crossmod that are drawn randomly from existing threads in their own subreddit. Moderators reported that they would have removed 648 (95.3%) of the 680 comments surfaced by Crossmod; however, 637 (98.3%) of these comments were still online at the time of this writing. In other words, moderators reported that those comments should have been removed, but that the current sociotechnical moderation architecture failed to help them do so.

1.3.4 Contributions and implications. The contributions of our work are two-fold. Firstly, we make a systems contribution where we develop a novel sociotechnical system for moderation on Reddit. To the best of our knowledge, Crossmod is the first open source, AI-backed sociotechnical moderation system to be developed and released publicly.[3] Prior work on online governance tend to focus on the algorithmic *detection-side* of moderation. Crossmod extends this line of research by exploring ways to go beyond detection towards enforcement. Second, we develop Crossmod by upholding the principles of participatory design—workers who are involved in the work system should be given a voice in the design process to determine how the new system could improve the quality of their work [38, 48]. In this paper, we present a grounded approach to include the voice of the moderators (i.e., the workers) in the design process to determine how Crossmod (i.e., the work system) could improve the quality of moderation (i.e., their work). Through our work, we hope to inform the design of similar AI-backed sociotechnical systems in the future, and ameliorate the *social-technical gap* in similar CSCW applications [2].

2 BACKGROUND

Most online platforms today curate content generated by their users in different ways—presenting content in an ordered form to increase user engagement (e.g., Facebook NewsFeed [56]), promoting sponsored content (e.g., ads [44]), and moderating (or screening) undesirable content [33, 41, 63]. Different social platforms adopt different approaches to moderating content. But these approaches can be broadly categorized as two types—human moderation (centralized and distributed), and automated moderation. Next we describe commonly deployed approaches from each category in detail, highlighting how they are currently used in practice.

2.1 Human approaches to content moderation

Human moderation take two primary forms—centralized and distributed approaches. In the central moderation approach, most sites employ the services of commercial content moderation workers or moderators who regulate content generated within the platform (e.g., Facebook, Twitter, YouTube, Reddit) [13, 29, 52]. Moderators either screen all of the content before it gets posted (*proactive*), or deal with it after the content is either reported by other users or triaged by an AI-agent (*reactive*). Content found to violate site guidelines, community norms, or even the law are typically removed off-site, with users being sanctioned and even banned from the platform in severe cases [59]. In the distributed approach, users up-vote desirable content making it more visible, and down-vote undesirable content, sometimes even reporting content [24, 41, 57, 58].

[3]The code for Crossmod is publicly available at https://github.com/ceshwar/crossmod.

Moderators are critical to platforms that rely heavily on human moderation, acting as digital gatekeepers who decide what content gets posted on the platform and what content gets taken down [29]. By curating content generated by users, moderators help guard against serious infractions that might do harm to a social platform's digital presence [28]. More importantly, moderators keep users from (unintentionally) viewing racist, homophobic, violent, misogynist, etc., content. [18, 60].

2.1.1 Challenges faced by human moderation approaches. Human moderation approaches suffer from drawbacks when deployed at scale, especially the need for a great deal of human labor [55, 68]. In the centralized approach, the labor falls on a small number of paid workers or volunteers who must work tirelessly to maintain the community [59]. Despite their significant role in shaping online discourse and improving user experience on social platforms, moderation workers typically receive low wages and work long hours. Moderators on most platforms are dispersed globally, typically hired from firms located overseas [20]. Prior work on commercial content moderation observed that:

> The work is almost always done in secret for low wages by relatively low-status workers, who review content day in and day out, digital content that may be pornographic, violent, disturbing, or disgusting. [60]

Recent reports have found that continued exposure to such violent and disturbing content can lead to detrimental effects on the mental well-being of moderators [4, 20]. The tasks performed by moderation workers range from repetitive and pedestrian normative violations like posting spoilers about a TV show, to exposure to images and material that can be violent, disturbing and, at worst, psychologically damaging [18, 61]. The workers are often further isolated because the work they do is carried out in secret, because their employers consider their work to be a threat to the brand [59].

In addition, the rapid pace and scale at which content is constantly generated within large-scale platforms restricts the effectiveness of *proactive* review processes that rely on human moderation alone—where a moderator examines every new piece of content before it appears on the site. As a result, most platforms employ human moderation in a *reactive* manner. In other words, even horrific content may exist on the site for some period of time where other users view and experience it, until the content gets reported, and subsequently removed off-site by a human moderator [60]. Due to the sheer challenge of regulating content within such large-scale platforms, plenty of content that violates site guidelines remain online for days, and sometimes even years [28]. Moderation is hard, and not all types of content are simple or easy for moderation workers to spot or to adjudicate, either. Some of these tasks involve complicated matters of judgment that involve familiarity with community norms and context information. Moreover, when the work of moderating platforms is outsourced to other parts of the world (e.g., Philippines [69]), it creates an additional hurdle to be familiar with social norms: workers must become steeped in the racist, homophobic, and misogynist tropes and language of another culture for which the content is destined [29].

2.2 Automated approaches to content moderation

In an effort to keep up with ever-growing content, technical approaches are reported to exist within social platforms communities [1, 27, 33]. These range from simple word and source-ban lists, to more sophisticated AI-based techniques that can flag inappropriate content. Word and source-ban lists take automated actions by filtering content based on either the use of black-listed words [12], or posting from blacklisted IP addresses [66]. Social platforms are also known to train machine

learning algorithms by compiling large datasets of example posts that have been moderated off-site [6, 30, 54]. Machine learning-based approaches can be especially helpful for algorithmically *triaging* comments for a much smaller number of (perhaps paid) human moderators.

2.2.1 Challenges faced by automated approaches to content moderation. Automated approaches to triage undesirable content have the potential to allow human moderators to review content more effectively and target their manual labor towards reviewing content that are likely to be violations. But current automated moderation approaches face some key drawbacks.

Static methods are crude by modern standards. Static word and source-ban lists have been observed to perform poorly [64]. They need to be constantly updated to keep up with the evolving nature of online behaviors to remain effective, and they are also prone to "false positives" as they typically do not account for context-specific information [11, 15].

Scarcity of labeled ground truth data. Machine learning-based approaches are generally more effective than static word-ban lists, but they require vast amounts of labeled ground truth data for developing reliable models. Moreover, new and emerging platforms suffer from the "cold start problem"—they lack enough data from their respective users. Such platforms lack the data and resources required to develop reliable automated moderation systems for triaging undesirable content [65]. In this paper, we use *cross-community learning* to address the scarcity of labeled ground truth. The core idea behind cross-community learning is to learn from data obtained from different source communities to detect violations within a completely different target community [19]. Recent research has shown that *cross-community learning* can be useful to side-step the need for site-specific data and classifiers for moderation [18, 19]. We extend this line of work.

Online moderation is contextual. Online moderation is a highly contextual task, and community norms guide moderation decisions by defining what is acceptable, or undesirable within an online community [5, 43, 50]. Community norms can vary widely across communities, and sometimes behavior considered undesirable by most may even be promoted in certain places (e.g., r/fatpeoplehate [17], Something Awful Forums [50], and r/RoastMe [36]). Such nuances are important to take into account, as platforms and researchers are doubling down on automated approaches towards moderation.

2.3 Recent advances in online moderation

Prior work on ML-based approaches to detect online misbehavior (e.g., abuse [19, 49], toxicity [70], hate speech [17], violations [51]) focus mainly on the detection-side of online moderation as if it is the ultimate step. Despite being a critical part of how norms are enforced and communities are regulated, the enforcement-side of online moderation remains relatively unexplored [17, 35]. As a first step towards this goal, we work closely with Reddit moderators to develop a new AI-based moderation system that can be easily customized to detect content that violate a target community's norms and enforce a range of moderation actions. Though moderation systems have been developed in the past, they are either completely proprietary, therefore unavailable for public use and study (e.g., internal tools at Facebook [6], The New York Times[4], YouTube [30]), or they are not supported by AI (e.g., Twitter Blocklists [27], The Coral Project[5], HeartMob [8], and SquadBox [47]). Crossmod is the first open source, AI-backed moderation system to be released publicly.

3 FORMATIVE INTERVIEW STUDY

Like many social platforms, Reddit relies on human moderation to regulate content. Moderators on Reddit are groups of users with special privileges who regulate content generated within subreddits

[4]https://www.nytimes.com/2017/06/13/insider/have-a-comment-leave-a-comment.html
[5]https://coralproject.net

Subreddit Name	Participant Name	Age	Country
r/AutoModerator	Chad Birch (P0)	35	Canada
r/photoshopbattles, r/blackpeopletwitter	P1	18	USA
r/science	P2	44	USA
r/news, r/funny, r/todayilearned	P3	60	UK
r/science	P4	30	USA
r/relationships	P5	31	USA
r/Sakartvelo	P6	39	UK
r/femalefashionadvice	P7	29	Australia
r/homeimprovement, r/homeautomation	P8	36	USA
r/itsaunixsystem, r/jailbreak	P9	23	Canada
r/computers	P10	39	UK

Table 1. The list of moderators we interviewed. P0 preferred to be identified by his real name in this paper.

on a voluntary basis—they are not paid for their labor. Moderators enforce rules that are subreddit-specific [25], in addition to site-wide (content[6] and anti-harassment[7]) policies. Reddit has existing architecture in place to support human moderation on the platform, as well as automated moderation tools to assist moderators. However recent reports have found that Reddit is systematically failing to limit the damage caused by bad actors, and moderators are struggling due to the large volumes of abuse constantly directed at them—resulting in moderator burnout, and even mental health risks [53]. We begin with a formative interview study to learn about the current state of automated moderation tools on Reddit, and explore opportunities for extending these tools[8]. The goal of this study is to understand how moderators use automated tools on different parts of Reddit (either to *proactively* remove content, or to triage content for human review). Through these interviews, we examine challenges faced by existing automated tools and identify unmet moderator needs.

3.1 Methodology

3.1.1 Recruitment. We conducted in-depth, formative interviews with 11 individuals who moderate Reddit communities on a regular basis. Interviewees were recruited by sending private messages through Reddit.[9] In order to maintain diversity in the topics and sizes of the communities represented in our interviews, we targeted five mainstream subreddits with over 1M subscribers each, five mid-level subreddits with 100k to 1M subscribers each, and five niche subreddits with less than 10k subscribers. Through this procedure, we reached out to moderator groups from 15 subreddits, and eventually recruited 10 active moderators who regulate content across 11 unique subreddits. In addition, we also interviewed Chad Birch, a Reddit moderator who created AutoModerator (or Automod)[10]—an automated moderation tool that is widely used on Reddit. Overall, we interviewed 11 participants based on a variety of English-speaking countries. Further details about our interview participants are shown in Table 1. All the interviewees were compensated with a $20 Amazon gift card for their time; further participation in the study was completely voluntary.

[6] https://www.redditinc.com/policies/content-policy
[7] https://redditblog.com/2015/05/14/promote-ideas-protect-people/
[8] This study was approved by the IRB at the authors' institution.
[9] Reddit allowed us to contact each subreddits' moderators through a group *private message*, following which interested moderators responded to our recruitment message individually.
[10] https://www.reddit.com/wiki/automoderator/full-documentation

3.1.2 Interview goals. We began the interviews by asking moderators about their general experiences regulating content on Reddit, and then dove into specific moderation practices currently employed within their subreddits. Next, we explored the breadth of existing automated moderation tools for Reddit, learning about commonalities and differences between how subreddits employ these tools. Finally, the moderators explained the advantages and drawbacks of existing automated tools. Through this process, we were able to understand the needs of moderators and the challenges faced by existing tools. In Chad's interview, we also asked about his experiences creating Automod for a handful of subreddits, and it subsequently getting adopted by Reddit as an internal moderation tool. All 11 interviews were conducted remotely over Skype or Hangouts, and lasted between 40-60 minutes. Interviews were recorded and then transcribed using a paid transcription service.[11] Finally, the first and second author read through the transcripts, coded them using thematic analysis.

3.2 Current state of automated moderation tools on Reddit

Reddit has existing infrastructure built for supporting moderators in the process of manually curating content within subreddits. Different subreddits use Reddit's infrastructure differently, but the underlying moderation interface and internal tools remain the same across subreddits.

3.2.1 Existing moderation interface. Reddit's current moderation interface is shown in Figure 2. Each subreddit has a dedicated moderation queue or "mod queue," which is a central listing of all the content generated within the community that needs to be reviewed by moderators. This includes all of the posts and comments reported by users, and content marked as spam by Reddit's site-wide spam filter. In addition to the mod queue, moderators also use two other tabs to review specific types of content. First is the "Reports" tab, which only shows content flagged through user-reporting, and the second is the "Spam" tab, which only lists removed content (mods said that these might be used by Reddit for refining the spam filter). In addition to Reddit's moderation interface, 9 out of the 11 moderators also use third-party browser extensions like Toolbox and Reddit Enhancement Suite (RES). P4 said, "*RES and toolbox are the standard ones (i.e., third party tools) that make the site usable and (provide) extra analytics.*" These browser extensions offer better user-interface and additional features that help mods sift through content manually.

Given the sheer amount of content generated due to Reddit's popularity, reviewing all content manually is infeasible. P2 said, "*Ultimately, for smaller websites you can do human based moderation. But right now, r/science has like 20 million subscribers.*" In order to keep up with the large volumes of user-generated content within subreddits, 10 out of the 11 moderators we interviewed rely heavily on automated tools to triage comments for manual review.

3.2.2 Moderation bots. Moderation bots are a popular class of automated tools that currently support Reddit moderators [45]. Bot development is facilitated through the openly-available Reddit API, and the associated Python Reddit API Wrapper (PRAW), both of which offer a range of scripted functionality [9]. A well-known bot that performs moderation tasks is */u/Botwatchman*, which detects and removes other *blacklisted* bots (based on a predefined list of bots that are not allowed on the subreddit). Additionally, moderators also employ Reddit bots specifically written to perform certain tasks, based on a set of predefined conditions. P1 said, "*r/photoshopbattles only allows submission of images with reasonable quality, and employs a bot to automate checking image quality.*"

3.2.3 AutoModerator. From our interviews, we found that AutoModerator or *Automod* is the most commonly used automated tool on Reddit, and some subreddits rely entirely on Automod to regulate content. P9 said, "*AutoModerator does bulk of the moderation for our subreddit.*" Automod is a customizable moderation tool that was created by Chad Birch (P0) for a handful of subreddits in

[11] www.rev.com

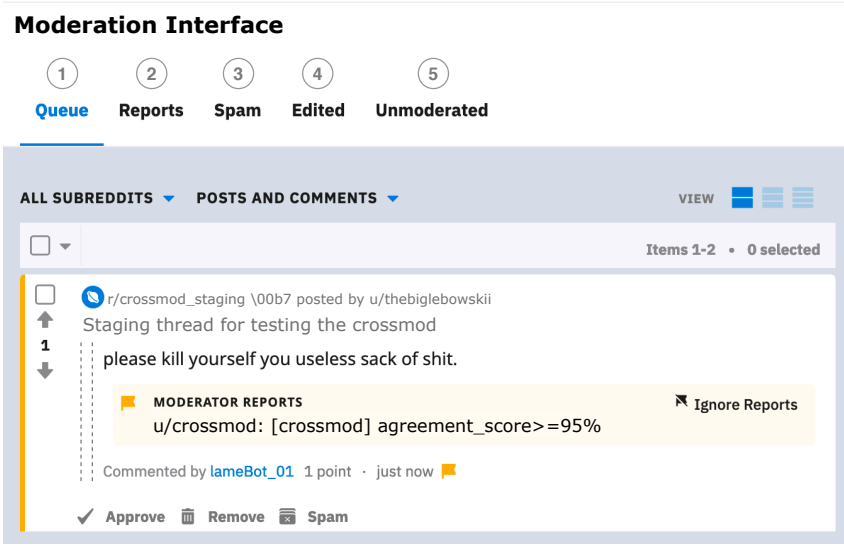

Fig. 2. Reddit's moderation interface for moderators to curate content manually. Five different tabs exist in this interface: *Queue* (or mod queue), *Reports*, *Spam*, *Edited*, and *Unmoderated*. When automated moderation tools are configured to "triage" content violating community norms, posts and comments are sent to the Mod queue (labeled as tab #1 in the Moderation Interface) for manual review by moderators. All posts and comments in the Moderation *Queue* are reviewed by moderators, after which they make moderation decisions. If a comment does not violate community norms, they can "approve" allowing it to remain on the subreddit. If they feel that a comment violates community norms, then they can either "remove" the comment, or mark it as "spam", taking the content down (i.e., off-site).

its initial stages. Automod was subsequently adopted by Reddit, and released as an internal tool to assist moderators on the platform. All of the moderators we interviewed had used Automod for moderating their subreddits at some point, and 9 moderators continue to actively use Automod. As a result, we focused our interviews on understanding how Automod is employed within subreddits, and its strengths and weaknesses in the opinions of moderators.

3.3 Current uses of Automod

Automod employs a regular expression (*regex*)-based filtering approach to scan, and proactively remove or triage[12] content within a subreddit. Automod rules are written and maintained by moderators (within a subreddit-specific config file). P9 mentioned that the documentation for Automod is well-maintained and easy to follow. A full description[13] of all the capabilities of Automod can be found on the subreddit called *r/AutoModerator*. 9 out of 11 moderators used Automod predominantly as a *triaging tool*—sending content violating hard-coded rules to the moderation queue for human review. P9 said, "*AutoModerator can* filter *comments: remove from public thread, and push to moderation queue for human review. Apparently, this functionality isn't available to any-one/thing other than AutoModerator.*"

[12]Triaging comments and posts for further review by moderators.
[13]https://www.reddit.com/wiki/automoderator/full-documentation

3.3.1 Automod can detect violations based on simple, hard-coded rules. Automod works based on simple, hard-coded rules, and is effective at taking care of repetitious and mundane tasks. P0 said, "*What Automod did was get rid of all the really tedious, easy moderation work and allow the moderators to pay attention to more subjective things.*" For example, P1 uses Automod to enforce formatting guidelines within their subreddit:

> r/photoshopbattles has a submission rule that states, "All titles must begin with *PsBattle*:". This rule helps distinguish r/photoshopbattles images from other pictures when they appear on the Reddit front page along with images from other subreddits. To ensure that users comply with this rule, we have hard-coded an Automod rule that detects and removes all submissions that do not begin with "PsBattle:"

In addition to enforcing formatting restrictions, Automod is also used to ban problematic users, prohibit URLs linking to objectionable websites (P6, P9, P10), and detect comments using phrases that are known to commonly occur in undesirable content within their subreddit (P1, P2, P4).

Chad, creator of Automod, had created a library of common rules available to everyone for reference. Shown below is an example rule that is used to configure AutoModerator to remove comments that consist of only capital letters:[14]

> **type:** comment
> **body (case-sensitive, regex, full-text):** "([A-Z0-9]|\\W)+"
> **action:** remove
> **action_reason:** CAPS ONLY

Many subreddits also share AutoModerator config files among themselves, allowing them to re-use commonly employed rules. P3 said, "*I just copy paste them from a preexisting AutoModerator in another subreddit? You know, the knowledge, you don't need to know how to create it from scratch.*"

3.4 Challenges in using Automod

3.4.1 Moderators noted that hard-coded rules and regexes are prone to mistakes. Automod's filtering mechanism, based on hard-coded rules and regexes, is effective at tedious and straightforward tasks. But Automod is prone to make mistakes as tasks get harder and need to take contextual information into account. P1 and P5 said that Automod consistently misses content that are violations (*false negatives*). Additionally, Automod filters "innocuous" content by mistake (*false positives*), and moderators stopped using Automod due to this reason. P6 said that, "*A lot of good comments get flagged by the AutoModerator and there was a lot of backslash from the community which is why I stopped using Automoderator.*" This aligns with findings from prior work examining AutoModerator's challenges [32]. Moderation is a highly contextual task, that needs to take the community's norms into consideration before taking down content. Encoding all of the norms of a community in the form of simple regular expressions is not feasible.

3.4.2 Moderators find it hard to configure Automod. Given that regexes are written by the moderators themselves, configuring Automod is non-trivial. Despite the presence of thorough documentation, a lot of moderators have a hard time configuring Automod by themselves, including long-time moderators like P3 with over 8 years of experience (and currently moderates 10 different subreddits). P3 said, "*I personally have problems with configuring Automod because I don't understand the process. It gets quite complex because it uses filters, mark down rules and so on which are beyond me. They seem to require a level of coding knowledge.*" Moreover, the current design of Automod's configuration file makes it restrictive for a lot of moderators to use. Along these lines, P2 said, "*You have to be very comfortable with what feels a lot like a command line interface in a lot of ways.*

[14] https://www.reddit.com/r/AutoModerator/comments/2l4e2j/can_automod_remove_or_report_postscomments_that/

And that restricts the number of people who are able to interact with AutoModerator, to the ones who have the skillset and the inclination, which is probably the biggest restriction." Additionally, Automod configuration files for most subreddits quickly grow into massive lines of hard-coded rules in order to keep up different types of misbehavior. P5 said that this makes it hard to maintain: "*When configuration files for AutoModerator become huge, it can cause errors on Reddit due to large amounts of processing time required by the rules.*"

3.4.3 Moderators need to manually come up with new rules and constantly update Automod. In order to keep up with the dynamic nature of abuse on the platform, 5 out of 11 moderators said that they have to constantly update Automod rules (e.g., come up with new lists of undesirable phrases or domain names for URLs). P2 said,

> "The AutoModerator is what it is, right? It is text string recognition. It doesn't learn, it doesn't adapt. If you're not constantly on top of it, it will quickly lose its relevancy. So what I had to do was constantly read bad comment strings and constantly update it with the latest dumb jokes, and that's an entirely manual process."

This static nature of Automod rules is a major drawback, despite its effectiveness in enforcing simpler formatting rules.

3.5 Summary of findings from formative interviews

Via formative interviews with 11 active Reddit moderators, we learned about currently available automated moderation tools on Reddit. Current automated tools like Automod fit into a deeply sociotechnical system of human- and machine-moderation. We identified three major areas of improvement for extending the capabilities of these tools.

- Current automated tools like Automod, though effective at performing simpler tasks, but are prone to making mistakes when tasks get harder.
- Moderators find that configuring Automod is hard and unintuitive, and the lack of an easy way to configure the tool is a major drawback.
- In order to keep up with the evolving nature of misbehaviors on the platform, moderators maintain long lists of rules that need to be manually updated on a regular basis.

4 CROSSMOD

Through our interviews, we found that mods need tools that adapt and learn. As P4 said:

> "I just need a smarter Automod. Automod is great because it can act on regular expressions. It can ban (spam) bots and report problems. It's a very strong tool, but it's a very simple tool. A machine learning model that can learn from past mod actions and remove content would be powerful, especially if it can do what a properly socialized and culturalized moderator can."

Next, we introduce *Crossmod*, a sociotechnical moderation system built to extend the capabilities of moderators while also fitting into their existing workflows reported in Section 3. Figure 3 demonstrates this through an example walkthrough. Scaling the idea of machine learning-based sociotechnical interventions up to groups interacting via an online community presents challenges, because of the social norms that emerge within a particular group [26]. In order to address this challenge, we construct Crossmod by working with the selected partner subreddits in Table 1, incorporating key principles of mixed-initiative user interfaces [31]. The moderators (or "mods") of those groups know those norms best, and expend considerable effort enforcing them. At each stage of our system design process, we worked with mods in a participatory framework to inform

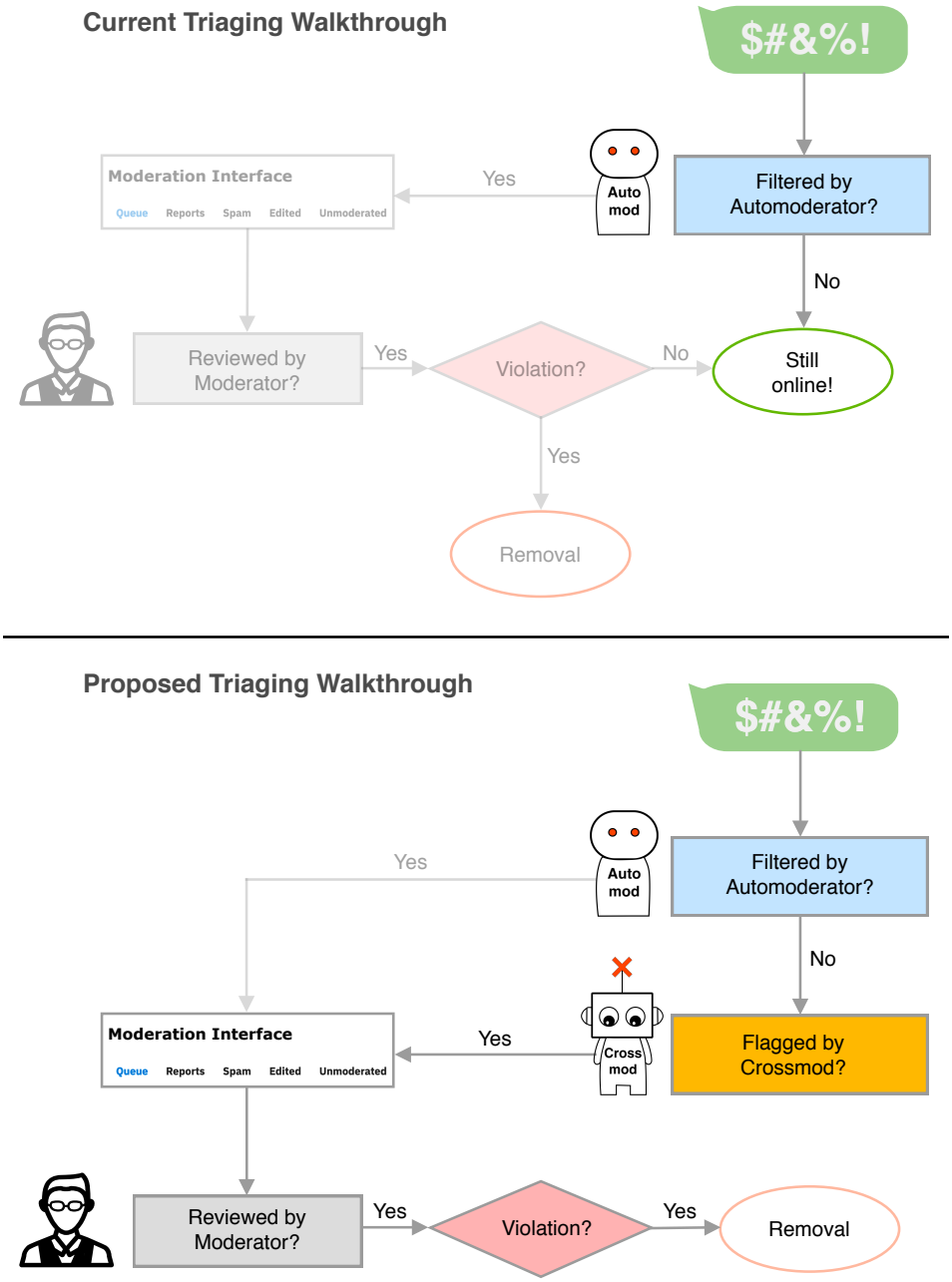

Fig. 3. Comparison between the current triaging workflow a comment undergoes when posted to a subreddit, and how it changes with Crossmod. Note that this illustration represents subreddits where moderators only review content flagged by automated tools (i.e., the complete workflow is more complex than depicted here).

key features: as a consequence, Crossmod permits great deal of moderator control over internal machine-generated predictions.

4.1 Crossmod: System design

Crossmod provides moderation recommendations for incoming comments in an automated manner, and makes its decisions based on *cross-community learning* (described in more detail below)—an approach that leverages a large corpus of previous moderation decisions via an ensemble of classifiers. Crossmod is designed so that it can be easily integrated into each subreddit's dedicated moderation interface.

4.1.1 How Crossmod is integrated into Reddit's moderation interface. Figure 2 illustrates how Crossmod is integrated into Reddit's interface to support different moderation actions. For example, if Crossmod is configured to report comments that violate community norms, the comments flagged by the system will be added to the *Reports* and *Mod Queue* tab in the Moderation interface, and this can be further reviewed by human moderators (since Crossmod has been granted *moderator-permissions* on the subreddit shown in Figure 2, all comments reported by Crossmod are also added to the central Mod Queue). In this illustration, both of the comments in the Mod Queue are automatically reported by Crossmod. All comments in the Mod Queue remain online until they are reviewed by human moderators. A moderator can perform one of three types of mod actions based on whether they agree with Crossmod's moderation recommendations or not. If they agree with Crossmod's moderation recommendation, then they just click "remove" on the interface, or mark the comment as "spam". If instead they disagree with Crossmod's report, and feel that the comment does not actually violate community norms, they can just "approve" the comment, and take it out of the Mod Queue.

Fig. 4. An illustration of the core idea behind *cross-community learning*. Using an ensemble of classifiers, we provide counterfactual estimates about what a set of source communities would do with new content from a completely different target community. In other words, "What would r/science do if this comment was posted there?" In our work, Crossmod's ML-backend provides counterfactual estimates about what 100 subreddits would do with new content, as well as whether that content resembles racism, homophobia, and so on.

In addition to reporting comments, Crossmod is designed to support a range of other functionalities: to send alerts to moderators in case of particularly sensitive topics (in the form of *modmails*), or to proactively remove comments that garner very high *abuse* scores computed by the pre-trained machine learning models described in Section 4.3. We detail the different moderation actions that

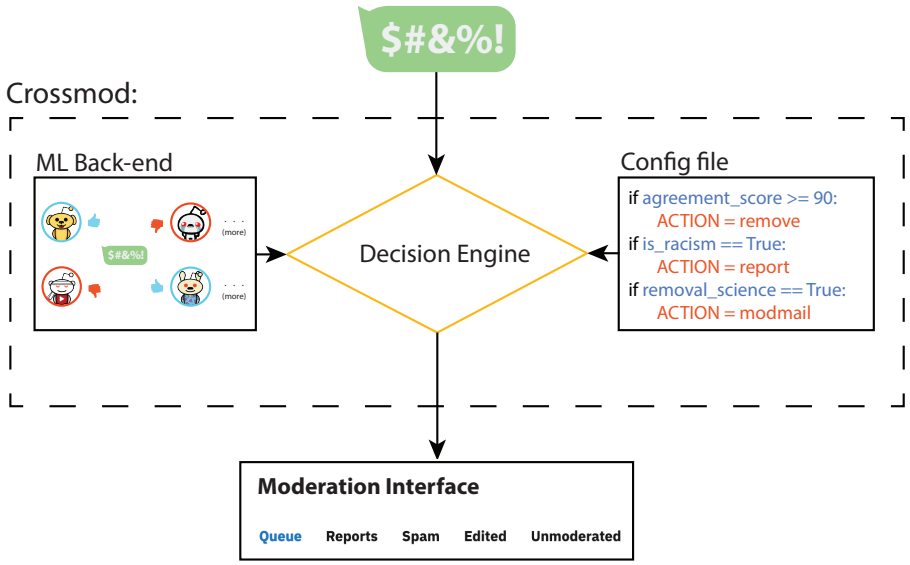

Fig. 5. Flowchart depicting Crossmod's system pipeline. Crossmod makes its moderation decisions by obtaining predictions from an ensemble of *cross-community learning*-based classifiers. Crossmod wraps this back-end in a sociotechnical architecture that fits into Reddit's existing *Moderation Interface*. Our system design allows moderators to easily configure Crossmod using simple conditional statements, and tailor its actions to suit community-specific needs.

can be supported by the proposed tool in further sections (an overview is shown in Table 3). The system pipeline is shown in Figure 5, and we describe each component in detail next.

4.2 System pipeline for Crossmod

Crossmod is a Reddit bot written in Python, and works as a subreddit-level moderator tool that can automatically detect comments violating community norms. Crossmod performs three main tasks:

Task 1: Constantly listening to the stream of incoming comments posted on a particular target subreddit. Every new comment that is posted on the subreddit will be ingested by Crossmod, and sent to the appropriate tab in the moderation interface, if necessary.

Task 2: Querying the ML back-end to obtain scores (or predictions) for each ingested comment. Crossmod's ML back-end is developed using *cross-community learning*, leveraging a large corpus of previous moderator decisions via an ensemble of classifiers. The outputs obtained from the ensemble of classifiers are sent further down the pipeline. Currently Crossmod queries the back-end classifiers through a request to a remote server.

Task 3: Finally, the outputs from the ensemble of classifiers in the ML back-end are aggregated to compute *scores*. We describe the different types of scores that can be computed in

further sections. Based on these aggregated scores, Crossmod detects comments that violate community norms. Depending on how Crossmod has been configured by the subreddit's moderators, Crossmod takes the appropriate moderation *<actions>* that were triggered based on the *<conditions>* specified in the *configuration file*.

4.3 ML back-end based on cross-community learning

Crossmod's ML back-end comprises an ensemble of cross-community learning-based classifiers that are of two types: subreddit classifiers, and norm violation classifiers. The different classifiers we use in this work were generated as part of our prior work on Reddit moderation [18]. Detailed information about the constructions of these classifiers are provided in the *Appendix A*.

Subreddit classifiers: First, we obtained 100 classifiers trained to detect whether a given comment will be removed by moderators of 100 popular subreddits. These were developed as part of our prior work, and the names of all 100 subreddits we trained classifiers for can be found on GitHub [15]. These classifiers were trained using FastText, a state-of-the-art library [10, 34].

Macro norms classifiers: Next, we obtained 8 classifiers trained to detect comments that violate macro norms on Reddit. In our prior work, we found the existence of "macro norms" that are known to be enforced across most parts of Reddit [18]. For example, posting hate speech in the form of homophobic and racist slurs, using misogynistic slurs, graphic verbal attacks, and distributing pornographic material are removed by moderators of most subreddits. As a follow-up to the study, we publicly released a labeled dataset of Reddit comments labeled violating 8 different types of macro norms [16]. Using this labeled dataset of macro norm violations,[16] we trained FastText classifiers to identify comments violating the different types of *macro norms* shared across Reddit [18].

This ensemble of 108 pre-trained, cross-community learning-based classifiers—100 subreddit classifiers, and 8 (macro) norm violation classifiers—constitute Crossmod's ML back-end. The main function of the ML back-end is to make predictions about a new query comment using the ensemble of classifiers. The final output from the ML back-end is the list of predictions obtained from the ensemble of classifiers. This is depicted in Figure 4. Through Crossmod, we bring the idea of *cross-community learning* to inform online moderation into production. Crossmod is the first open source, AI-backed moderation system to be released publicly, and the system can be easily adopted by new and emerging online communities *off-the-shelf*. The goal of providing the ensemble of classifiers in Crossmod is to provide moderators with the following choice—which subreddits would they like to emulate. It may be possible to use the frame of macro norms to derive normative guidelines for new and emerging online communities. In other words, the ensemble of classifiers used in Crossmod's back-end may serve as sensible defaults for a new online community. Additionally, using predictions obtained from an ensemble of 108 classifiers, instead of relying on a purely in-domain classifier trained on comments that are inherent to the target subreddit only, brings in more diversity and robustness into the decision-making process.

4.4 Configuring the Crossmod: *If This Then That* (IFTTT) format

In our sociotechnical intervention, a machine learning system will intervene in a normally unmediated process. In other words, Reddit would usually immediately post those comments. Issues of agency naturally arise in mixed-initiative systems like the one we build. Our approach is to empower the mods in Crossmod. Here, this means that moderators can review and ultimately reject moderation recommendations made by Crossmod's back-end ML (the models can even learn from

[15] List of 100 study subreddits: https://github.com/ceshwar/reddit-norm-violations/blob/master/data/study-subreddits.csv
[16] A detailed description of our dataset can be found on Github: https://github.com/ceshwar/reddit-norm-violations

```
if agreement_score >= 95:
    ACTION = remove
if agreement_score >= 80:
    ACTION = report
if is_racism == True:
    ACTION = modmail
if is_misogyny == True and is_homophobic == True:
    ACTION = report
if removal_science == True:
    ACTION = report
if agreement_score >= 90 and removal_The_Donald == False:
    ACTION = remove
```

Fig. 6. Example configuration file for Crossmod. In this config file, the mod is auto-removing comments with very high agreement scores, and reporting those with moderate scores. In addition to using specific macro norm and subreddit scores, on the last line the mod has exempted r/The_Donald from the agreement score.

those corrections). Beyond that, we develop rich sets of options moderators can use to control and configure Crossmod to meet their subreddit-specific needs.

During the interviews, we learned that configuring AutoModerator (through regexes in a YAML file) was a challenge for most communities. We found that one of the needs of moderators was just to design a moderator tool that simplifies the way it can be configured. P2 said:

> "Well, it also needs to be straight up more user friendly to just even do what we're currently doing. Like even if it was just easier to type in the text without having to get all of the semantics of the Automod configuration properly done, that would be helpful, just in and of itself."

We adopt an *If This Then That* (IFTTT) format where moderators just create chains of simple conditional statements to trigger moderation actions. In other words, the configuration file of Crossmod contains a list of IFTTT commands written in the following format:

if **<condition>**: **<action>**

Through our initial rounds of feedback obtained from moderators, we found the use of such simple conditional statements makes configuring Crossmod intuitive and straightforward to moderators without (any) coding experience. Next we present the different types of **<conditions>** that are supported by Crossmod's ML back-end, and then review the different types of **<actions>** that Crossmod can be configured to perform. An example Crossmod configuration file containing some example conditional statements is shown in Figure 6.

4.5 <conditions> supported by Crossmod's ML back-end

Crossmod makes its decisions based on predictions about a *new query comment* obtained from the ML back-end. Using the output returned by the ML back-end, the decision engine of Crossmod can be configured to evaluate the three types of **<conditions>** shown in Table 2. We describe each of them in detail below.

<condition-type>	**Example conditional statements**
removal agreement score	if (agreement_score >95%):
macro norm violation	if (is_racism = True):
specific subreddits would remove	if (removal_nfl = True):

Table 2. Different types of conditional statements that can be used to configure the Crossmod's decision engine. The values for agreement_score, is_racism, removal_science are computed using the predictions returned by the different classifiers present in the ML back-end.

4.5.1 ***<condition>****: agreement_score.* First, we use the subreddit-specific classifiers built for detecting comments that would be removed by moderators of 100 popular subreddits. Using the predictions obtained from all of the 100 source subreddits, for a given unseen comment from the target community, we obtain an *agreement_score*—the percentage of source subreddits that consider the particular comment to be a norm violation (i.e., "If the comment were hypothetically posted on the subreddit, would moderators remove it?"). In other words, if the *agreement_score* computed for a comment is high (say 99.99%), it denotes a majority agreement to remove the comment, among most of the subreddit classifiers present in the ensemble. An example conditional statement of this type would be: if(agreement_score) > 0.95. Given this **<condition>**, Crossmod will only detect comments that at least 95% of subreddit classifiers in the ensemble predict to remove.

When interviewing moderators for design feedback on our system prototype, we observed that moderators liked the idea of using Crossmod to make moderation recommendations within their subreddit by simulating moderation decisions by other moderators from 100 popular subreddits. P4 said that conditional statements based on *agreement_score* would definitely be useful for the subreddits they moderate (i.e., r/news, and r/todayilearned): "*Most useful for me would be the agreement_scores returned by the 100 subreddit classifiers*". Along similar lines, P8 said: "*Agreement is super helpful where you can straight up remove stuff without intervention or flagging it.*"

4.5.2 ***<condition>****:* $removal\_\{subreddit_i\}$. Next, we use the predictions obtained from individual source subreddits or a group of specific source subreddits for a given unseen comment from the target subreddit. Moderators can configure Crossmod to detect target community comments based on whether a subset of source communities would consider a given comment as a norm violation (if it were hypothetically posted on the source subreddit, a moderator would remove it). An example conditional statement of this type would be: if(removal_nfl) = $True$, where we are checking if the classifiers trained for a specific subreddit, r/nfl, predicts to remove the given comment.

In addition to the *agreement_score* described previously, moderators asked for this ability to configure Crossmod to make moderation recommendations based on only certain subsets of classifiers (instead of all of them). P0, P1, and P4 said that such conditional statements help smaller, and topically-similar subreddits. In particular, P1 said that,

> "That seems useful because let's say, sports subreddits, for example, r/soccer. If you wants to make a subreddit for a local football team that's not so broad, then learning from what r/soccer does would be useful for moderating."

4.5.3 ***<condition>****:* $is\_\{violation_{norm}\}$. The last type of **<conditions>** supported by Crossmod is based on predictions obtained from the 8 classifiers trained to detect specific types of macro norm violations (described in Table 5). The goal of each of these classifiers is to identify whether an unseen comment from the target subreddit is a macro-norm violation, or not.

An example conditional statement of this type would be: if(is_racism) = $True$, where we are checking if the classifier trained to detect macro norm violations predicts the given comment to be

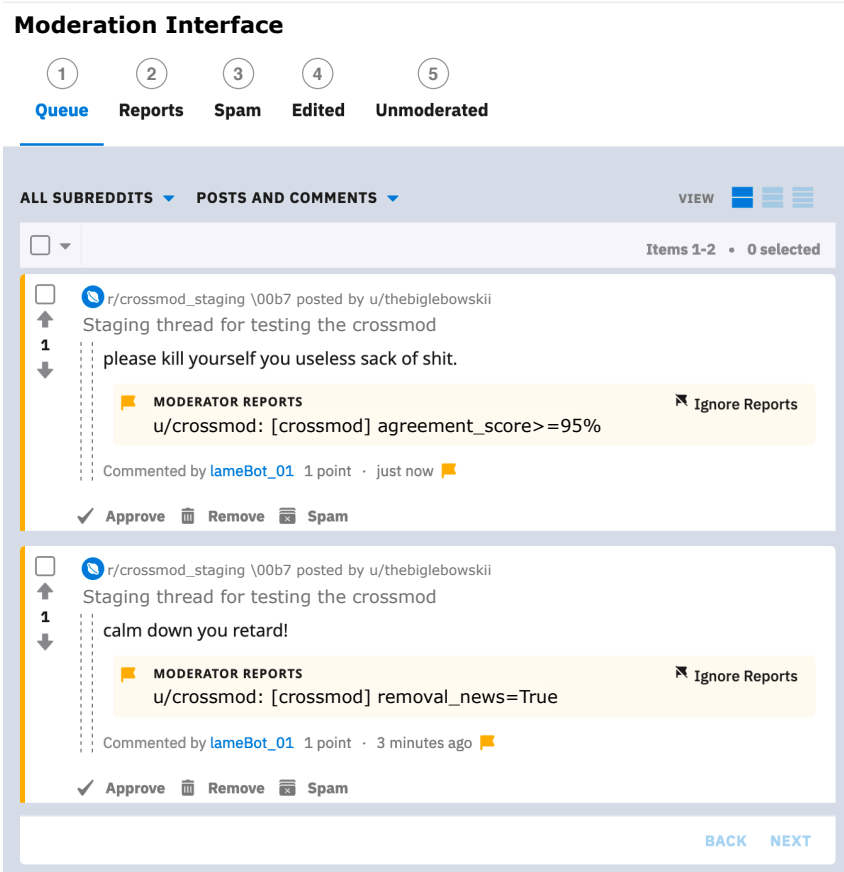

Fig. 7. How Crossmod is integrated into Reddit's existing moderation interface to support *triaging*. When Crossmod is configured to triage (as opposed to outright remove), comments flagged by Crossmod will be *reported*, and sent for further review by human moderators. In this example, the first comment in the moderation queue is reported for obtaining an *agreement_score* over 95%, while the second comment in the queue is reported for because the classifier trained for r/news predicts removal (i.e., $removal\_news = True$). All comments pushed to the moderation queue are reviewed by moderators, and the moderator can perform three different actions based on whether they agree/disagree with Crossmod's report. If they disagree with the report, and feel that the comment does not violate community norms, they can "approve" the comment. Instead, if they agree that the comment does indeed violate community norms, then they can either "remove" the comment, or mark it as "spam".

racist in nature. P3 (r/news) and P4 (r/science) said that Crossmod's ability to automatically detect such harmful content would be very useful for the subreddits they moderate. In particular P4 said,

> "(Detecting) macro norm violations would be very helpful for my subreddit (r/science). Automod can capture obvious use of hate speech, but a tool that can recognize less obvious speech is a very useful tool."

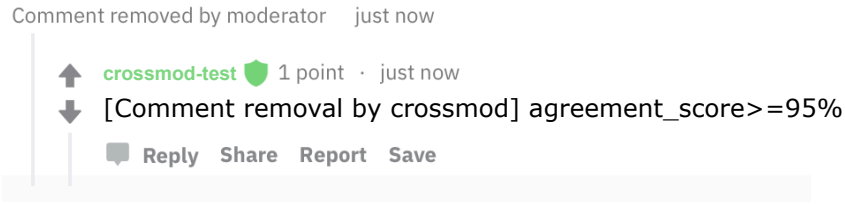

Fig. 8. A comment found to violate community norms is proactively removed by Crossmod, and leaves a *tombstone* in place of the comment. All comments that are proactively removed by Crossmod will be sent to a *Spam* tab in the Moderation Interface (denoted by tab #3 in Figure 2).

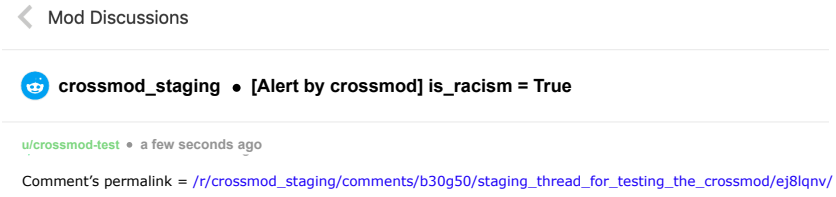

Fig. 9. In this example, the comment triggered the *modmail* <action> supported by Crossmod. As a result, a modmail was sent to alert the moderators, along with the corresponding <condition> violated by this comment. Moderators can review this comment by clicking on the URL (i.e., *permalink*) pointed to in the modmail from Crossmod. The mods can then choose what to do next.

4.6 <actions> that Crossmod can perform

Currently, Crossmod's design allows the system to perform three types of **<actions>** with varying levels of autonomy, shown in Table 3. Depending on the severity of removal **<conditions>**, and the target subreddit's preferences, moderators can configure Crossmod to take specific types <actions>. The goal here is to provide moderators with the flexibility in choosing the **<conditions>** and **<actions>** based on their specific requirements, or level of autonomy they would like to grant Crossmod. Though there were a few other specific moderation **<actions>** that were requested by individual moderators (e.g., locking the post, adding a spoiler, or marking as not safe for work), we focus on three types of **<actions>** in Crossmod's current design. We found that almost all of the moderators we spoke to agreed on the need for these three **<actions>**.

4.6.1 ***<action>****: remove.* As illustrated in Table 3, the first type of moderation *action* that Crossmod can perform is *direct removal*. Once a comment violates the **<condition>** which triggers the direct removal **<action>**, Crossmod automatically removes the comment from the subreddit. As shown in Figure 8, the comment is replaced with a *tombstone* indicating that it was removed by a moderator, in this case the crossmod.

4.6.2 ***<action>****: report.* The next type of moderation action that Crossmod can perform is reporting a comment, and thereby sending a comment for further review by moderators. Through the report **<action>**, Crossmod can serve as a *triaging* tool for subreddits, automatically detecting comments that are likely to be undesirable and sending them for further review by moderators. This is illustrated in Figure 7. All reported comments continue to remain online until a moderators decides

<action>	Description
remove	Remove from the subreddit, and move to *Spam* tab in moderator interface
report	Send to mod queue for review by human moderator
modmail	Alert human moderators through a modmail or message

Table 3. Different types of moderation actions statements that are supported by Crossmod.

to eventually remove it. Instead of taking moderation decisions autonomously, Crossmod can help augment automated predictions with human judgment. P5 and P9 said that,

> "If there's a rule you're not sure about, then don't remove the comment directly. Instead, triage it through the bot so that the human mod can remove/approve it after review."

4.6.3 ***<action>****: modmail.* The final type of **<action>** that Crossmod can perform is alerting the moderators about the presence of new undesirable content on their subreddit. Crossmod can be configured to send alerts over modmail, or in the form of direct messages to specific moderators. An example of Crossmod sending an alert over modmail is shown in Figure 9. P0 said,

> "Sending a Modmail would be a good one for sure. That's a different way of reporting. So it to be, oh, I detected a comment that appears like I'm 90% confident is hostile, you should take it and look at it, here's the link. That can be a more detailed way of reporting things, since it's difficult to put much information in a report."

For instance, though most moderators liked the idea of alerting through *modmail*, P1 said they preferred the *remove* and *report* **<actions>** over *modmail*:

> "For r/photoshopbattles, sending comments that violate community norms to Mod queue would be more helpful than sending it to the Modmail directly."

4.7 Design elements to track Crossmod <actions>

Under **<action>**: remove, Crossmod has full autonomy over its moderation decision and it can directly alter conversations within a target subreddit by removing comments (without the need for human review). As a result, we included design elements in the system that can help moderators easily track moderation actions, and reverse Crossmod removals if necessary.

Send removals to Spam tab in Moderation Interface. All direct removals made by Crossmod are automatically sent to the *Spam* tab in the moderation interface. This allows moderators the flexibility to audit direct removals made by Crossmod, and even reverse Crossmod removals by approving the comment, in case of false positives. This also enables a post-hoc analysis of Crossmod's moderation decisions in the future, and monitor false positive rates (i.e., how many times did human moderators overturn Crossmod's moderation actions?).

Provide removal reasons. We designed Crossmod so that it can be configured to provide a *removal reason* to moderators and/or the user (i.e., author of the comment), in the form of a *reply* and/or a *modmail*. P3, P6 and P9 mentioned that providing removal reasons allows the users and other moderators to understand why a certain comment was removed by the system:

> "You'd have to spend time either figuring out or guess and approve, remove it based on your own quick judgment. So you should have a removal reason, this has been removed because X, Y, Z." (P3)

Alert moderators after removing content. Finally, Crossmod can be configured to send out a *modmail* informing the moderators about a direct removal made by the system. An example of this is shown in Figure 10. Additionally, we are also brainstorming design ideas with moderators on

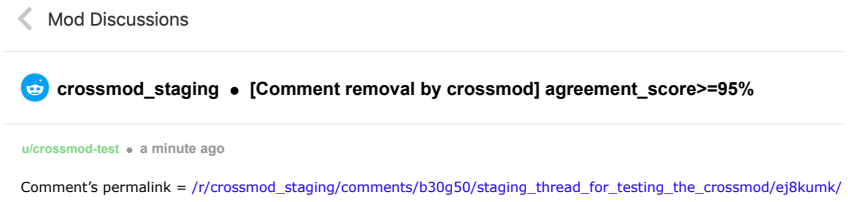

Fig. 10. A comment was directly removed by Crossmod, and a modmail was sent to alert the moderators about this action. Moderators can review this automated decision by clicking on the URL pointing to the removed comment. If they disagree with Crossmod, they can reverse the decision by "approving" the comment.

ways to use community feedback to reverse Crossmod removals when necessary. P10 suggested an idea to use community-feedback on the Crossmod's reply containing the removal reason to identify "false positives", and reverse an automated removal. For example, if Crossmod's reply receives over 5 downvotes (indicating that at least 5 users disagree with Crossmod's decision that the comment is a violation), Crossmod could be configured to reverse its **<action>** by *approving* the removed comment. This functionality is not supported currently, but we plan to explore this in future iterations of Crossmod.

5 DEPLOYING CROSSMOD IN A CONTROLLED ENVIRONMENT

As a proof-of-concept summative evaluation, we deployed Crossmod in a controlled environment, simulating real-time conversations from two large test subreddits with over 10M subscribers each. We created a staging instance for each test subreddit, and simulated actual conversations that took place within the test subreddit, in an unobtrusive manner. When we interviewed Chad (P0), the creator of Automod, he suggested a similar deployment study:

> I think even if you could just come up with something that like hypothetically shows, "hey the bot is watching all the comments in our subreddit, here's the one who's it would have acted on." And if they (moderators) can see that and just see, oh yeah, that actually catches a ton of stuff we're doing manually right now, then that would be a huge thing to make them more confident in implementing it. And that gives them an upfront way to see what things it would do instead of saying, "here, put this in and only then we can see what it does."

5.1 Test subreddits: r/science and r/Futurology

We deployed Crossmod in a controlled environment simulating real-time conversations from r/science and r/Futurology. r/science is one of the largest communities on Reddit with over 20 million subscribers, and it is a "place to share and discuss new scientific research." r/Futurology is another large community with over 13 million subscribers, and it is "devoted to the field of Future(s) Studies and speculation about the development of humanity, technology, and civilization."

5.1.1 Creating staging instances to mirror test subreddits. First, we created a staging instance for each of our test subreddits. The staging instance is a controlled environment created to mimic conversations within test subreddits, and we obtained test subreddit comments in an unobtrusive manner. For example, the staging instance created for r/science was a subreddit mirroring r/science, and therefore contained a copy of all comments posted to r/science. We created the staging instances by streaming all test subreddit comments posted during a 4-month period, from September 2018 to

December 2018. We used Google BigQuery to obtain these comments, and then (re-)posted all of them on the staging instances using the Reddit API.

5.1.2 Deploying Crossmod in the staging instances. We deployed Crossmod in the staging instances so that it would monitor all comments streamed from r/science and r/Futurology. Our goal was to simulate the real-time conditions under which Crossmod is intended to be used by moderators. For the purpose of this evaluation, Crossmod was configured to "report" comments that obtained an *agreement_score* >= 85% from the system's ML back-end. Under real-world circumstances, mods would configure the Crossmod with their own custom config file; here, we tested a baseline scenario to see how well Crossmod could perform relative to existing tools. (See Future Work for longitudinal deployment plans.)

5.2 Evaluating reported comments with the help of moderators

We recorded the moderation recommendations made by Crossmod when deployed in this controlled environment. In particular, we stored the *id* (i.e., unique identifier for the comment) and *body* (i.e., actual text in the comment) of all comments reported by Crossmod, along with the *agreement_score*'s computed for each comment. In order to evaluate Crossmod's moderation recommendations, we asked 2 moderators from each test subreddit to decide whether they would allow the comment on their subreddit or not. We conducted human review of comments in 2 phases.

5.2.1 Phase 1. In Phase 1, two moderators from r/science independently rated a set of 100 comments collected from their subreddit. Out of the 100 comments shown to both moderators, 50 of these comments received an agreement_score >= 85 (i.e., high scoring), while 50 of these comments received an agreement_score <= 1 (i.e., low scoring). Moderators were not told about the distribution of high scoring to low scoring comments in order to avoid any cognitive biases in the decision-making process. Similar to the above process, two moderators from r/Futurology also independently rated a set of 100 comments collected from their subreddit.

5.2.2 Phase 2. In Phase 2, the four moderators were asked to rate a larger set of comments obtained from their respective subreddits. In this phase, we asked moderators to only review a random sample of 650 comments scored highly by Crossmod (i.e., agreement_score >= 85). In Phase 1, we found that Crossmod achieved a low false negative rate (less than 0.125). As a result, we chose to have the moderators only focus on the comments that were flagged by Crossmod (i.e., high scoring) in Phase 2, as these resemble the type of comments that would be reported by Crossmod upon real-time deployment. In order to prevent bias from the manual review process, we did not disclose the classifier scores (or sampling strategy) to the moderators. As a result, the moderators were only told that the comments were obtained from their respective subreddits, and they did not know whether a comment was actually reported by Crossmod or not.

We would like to note that moderators on Reddit perform large amounts of human labor on a voluntary basis, manually regulating content in order to help maintain healthy conversations on the platform. As mentioned earlier, moderators are struggling to keep up with the vast amounts of content generated within their communities [52]. In addition to this manual labor, the task of moderating content also involves emotional labor as well, with moderators having to view gruesome and disturbing content regularly [53]. Out of respect for the moderators' time and to keep the amount of effort required to review content manually, we asked each moderator to review a random sample of 170 comments scored highly by Crossmod in Phase 2.

	Accuracy	Precision	Recall
r/science	0.92	0.98	0.875
r/Futurology	0.86	0.72	1.0

Table 4. Evaluation of Crossmod in Phase 1 where we compare labels provided by 4 moderators and predictions from Crossmod on an equal distribution of high-scoring and low-scoring comments.

5.3 Results

5.3.1 Phase 1. In Phase 1, a total of 200 comments obtained from 2 large-scale subreddits were labeled by 4 moderators from these subreddits, ensuring that an equal distribution of high-scoring and low-scoring comments were reviewed by each moderator. We found high inter-rater agreement among the two moderators from each subreddit when deciding whether they would allow the set of 100 comments on their subreddit or not—r/science moderators disagreed on 3 comments, and r/Futurology moderators disagreed on just 1 comment. In the case of disagreements, we asked moderators to discuss and provide us with a consensus label for the comments (i.e., remove or not). Using the labels assigned by moderators upon review as ground truth, we evaluate Crossmod's performance and the results are shown in Table 4. In Phase 1, Crossmod achieved high recall of over 87.5% for both r/science and r/Futurology, with an overall accuracy of over 86%. Moderators told us that they preferred that Crossmod achieve higher recall over precision. This is because moderators' intend to use Crossmod as a reporting tool to triage norm violating comments. Therefore a system that is able to detect as many violations as possible (i.e., higher recall at the cost of precision) is desirable because every reported comment will be eventually be going through human review—moderators can correct false positives during the review.

Error analysis: Next, we examined the comments that were misclassified by Crossmod in Phase 1, and found that *false positives* were comments that contained URLs (e.g., links to Wikipedia, or Imgur), and comments that excessively used swear words for exclamation. While *false negatives* included comments that were off-topic and anecdotal, and prior work has found that these are considered to be micro norm violations within scientific communities like r/science [18]. One approach to account for community-specific norms would be to use Crossmod along-side a purely in-domain classifier trained on moderated comments from a target community (e.g., r/science). This would allow moderators to make decisions based on moderation recommendations obtained from an ensemble of classifiers trained to encode *Reddit-wide* norm enforcements (e.g., macro norms) along with *community-specific* recommendations (e.g., micro norms). We plan to explore this line of research in our future work.

5.3.2 Phase 2. In Phase 2, the 4 moderators reviewed a total of 680 unique comments (each mod reviewed 170 unique comments found within their subreddit) reported through Crossmod. Given the high inter-rater agreement observed in Phase 1, we decided to show a non-overlapping set of unique comments to each moderator from r/science (and r/Futurology) in Phase 2. This allowed us to evaluate Crossmod's performance by reviewing a larger sample of comments. Overall, moderators decided that they would have removed 648 (95.3%) of the comments detected by Crossmod. The r/science moderators decided that 338 out of the 340 comments detected by Crossmod would have been removed from r/science. The r/Futurology moderators decided that 310 out of the 340 comments detected by Crossmod would have been removed from r/Futurology.

Furthermore, for all of the reported comments that were decided to be violations by moderators, we queried the Reddit API (by their unique *id*). Our goal was to examine whether any of these comments were still online on r/science and r/Futurology. We found that out of the 338 comments

decided to be violations by r/science moderators, only 10 comments were actually removed from r/science (i.e., taken off the site). In other words, Crossmod was able to detect 328 comments that moderators would have removed, but were previously *missed* by existing moderation tools like AutoModerator. Similarly, we found that 309 out of the 310 comments decided to be violations by r/Futurology moderators were still present online, but were detected by Crossmod. As mentioned earlier, human moderation struggles to keep up with the immense volume of content generated within large-scale platforms—plenty of content that violates site guidelines remains online for days, sometimes even years [28]. Additionally, we found that current automated tools on Reddit are not robust, and prone to mistakes through our interview study (see Section 3.4). In particular, it is hard to quantify the rate of *false negatives* (i.e., how many actual norm violations are being missed by existing tools) due to the reasons mentioned above. Our findings provide some clarity on this issue. The results from Phase 2 of our evaluation quantifies the gap in currently deployed moderation approaches (i.e., what are current approaches missing out on), and how Crossmod can help address this gap by extending their capabilities.

5.3.3 Summary of findings. We deployed Crossmod in a controlled environment, simulating real-time conversations from two large subreddits with over 10M subscribers each—r/science and r/Futurology. Two moderators from each subreddit evaluated Crossmod's moderation recommendations by manually reviewing comments scored by Crossmod that had been drawn randomly from existing threads. In Phase 1, Crossmod achieved an overall accuracy of 86% when detecting comments that would be removed by moderators, with high recall (over 87.5%). In Phase 2, moderators decided that they would have removed 648 of the 680 (95.3%) comments detected by Crossmod; however, 637 of 648 (98.3%) were still online at the time of this writing (i.e., not removed by current moderation tools). While necessarily incomplete and proof-of-concept, these results indicate that Crossmod will significantly extend the capabilities of moderators when deployed in the wild.

6 DISCUSSION

Through our formative interviews, we found that the majority of Reddit moderators use automated tools to make the task of curating a large-scale platform like Reddit managable. However, there are very few automated tools that can assist moderators, and existing tools are limited in their scope. We developed *Crossmod*, a sociotechnical moderation system built to extend the capabilities of moderators, while also fitting into their existing workflows. We would like to emphasize that Crossmod is not intended to replace human moderation or any of the existing automated tools used by moderators. Instead, Crossmod aims to extend the current capabilities provided by existing social and automated tools. Though limited in scope, our summative evaluation shows that Crossmod can improve moderation by detecting undesirable content which is currently undetected by the existing sociotechnical infrastructure.

6.1 Theoretical implications

6.1.1 Online moderation as a sociotechnical phenomenon. Sociotechnical research is premised on the interdependent and inextricably linked relationships among the features of any technological object or system and the social norms, rules of use, and participation by a broad range of human stakeholders [62]. This mutual constitution of social and technological is the basis of the term sociotechnical. Mutual constitution directs scholars to consider a phenomenon without making a priori judgments regarding the relative importance or significance of social or technological aspects (e.g., [7, 42]). In our work, we examined currently deployed approaches to content moderation on Reddit, shedding light on the sociotechnical nature of online governance. Prior work on online moderation has focused on just the social system (e.g., [41]), or the technological system (e.g., [40]),

or even the two side by side (e.g., [33, 47]). Our work extends this line of work by investigating the phenomena that emerges when the two interact, and aims to ameliorate the *socio-technical* gap [2] in moderation systems. Instead of focusing on the behavioral or the technological, we study the emergent sociotechnical phenomena that sets our work apart. We hope to see more theoretical work exploring online moderation along these lines in the future.

6.1.2 Going beyond detection towards enforcement in online moderation. Current research on automated approaches is focused on the detection-side of online moderation, but there is a gap in the study of the enforcement-side, and the considerations to be made during this process [35]. Depending on the platform and the specific community under consideration, enforcement strategies vary significantly in intent and execution [18], and these nuances should inform detection strategies in the future. Our work aims to bridge this gap, and explores how to take the next step by developing a new AI-based moderation system that can be easily customized to detect content that violate a target community's norms and enforce a range of moderation actions.

6.2 Design implications

6.2.1 Participatory methods to inform the design of socio-algorithmic governance. Machine learning algorithms have become core design elements in modern social computing systems. Most often, they are designed and implemented by corporations who control access to an algorithm's core data and code. Here, we have followed emerging research in exploring a different path: working in dialog with users and key stakeholders to design socio-algorithmic systems [3, 72] that govern large social spaces. Crossmod is one data point in the space of possible socio-algorithmic governance systems that could result with participatory input.

6.2.2 Agency and configurability in AI-backed sociotechnical interventions. Crossmod turns over control and oversight to empowered mods who can direct the underlying algorithms as they see fit. In sociotechnical interventions like Crossmod, an algorithm (e.g., machine learning) intervenes in a normally unmediated process. Issues of agency naturally arise in mixed-initiative systems like the one presented here. We turned to building configurability and oversight into Crossmod as a solution, by developing rich sets of options mods can use to control and configure Crossmod. As a consequence, Crossmod permits a great deal of mod control over internal machine-generated predictions, and can be tailored to meet subreddit-specific needs. Moderators can easily configure Crossmod using simple conditional statements, and also track moderation actions taken by Crossmod. In cases where moderators disagree with Crossmod, they can overrule its judgments.

6.3 Next steps for Crossmod

We plan to extend Crossmod through the following ways:

6.3.1 Graphic UI to help mods configure the system. Moderators found configuring Crossmod using simple conditional statements straightforward. In addition to this, we plan on introducing a graphical user-interface that simplifies configuring Crossmod even further.

6.3.2 Incorporate community feedback into Crossmod. We are currently brainstorming design ideas with moderators on ways to use community feedback in Crossmod. One idea is to reverse Crossmod removals when necessary. For example, P10 suggested an idea to use community-feedback on the Crossmod's reply containing the removal reason to identify "false positives", and reverse an automated removal. For example, if Crossmod's reply receives over 5 downvotes (indicating that at least 5 users disagree with Crossmod's decision that the comment is a violation), Crossmod could be configured to reverse its **<action>** by *approving* the removed comment. This functionality is not supported currently, but we plan to explore this in future iterations of Crossmod.

6.3.3 Ensuring fairness and transparency in moderation decisions. We plan to conduct a systematic analysis of any algorithmic biases [8, 23] inherent in Crossmod's machine learning back-end, and incorporate design changes that ensure fairness, accountability and transparency in the moderation system [21, 22, 71]. As a first step, we added "removal_reasons" to the system's design, facilitating better sense-making of Crossmod's actions. Additionally, using an ensemble of classifiers to make moderation decisions introduces diversity into Crossmod's decision-making process, potentially reducing biases that are likely.

6.3.4 Publicly release Crossmod with API keys. We plan to release Crossmod publicly through an API, and this would allow us to evaluate the entire system, in addition to the core algorithm that was already evaluated in this work. Prior work has identified that norms are shared across different Reddit communities (e.g., macro norms are found to be universal to most parts of Reddit) [18]. Though we evaluated Crossmod only on a couple of subreddits in this paper, we believe that the different **<conditions>** supported by the system would allow Crossmod to identify norm violations within other subreddits as well. But moderators asked us to take caution before releasing the data and code for Crossmod publicly since bad actors may easily find ways to circumvent the system. In order to prevent such situation, we will be using API keys to control who has access to Crossmod.

6.3.5 Default configurations for Crossmod's back-end. In future iterations of Crossmod, we plan to include default configurations for the ensemble based on prior configs used by previous moderators, or automatic tuning based on historical moderated data obtained from the target subreddit under consideration (i.e., automatically selecting which classifiers to use, in addition to asking the moderators to choose manually).

6.3.6 Deployment on Reddit. Our goal is to finally push Crossmod into production across Reddit. As Chad said, "*A lot of moderators are quite disappointed in how few moderators tools there are. So when something new comes out, they're pretty quick to adopt that.*" We are currently in conversations with moderators from several subreddits, including r/Futurology, about deploying our system real-time on Reddit. As the next step, we plan to deploy Crossmod as a real-time *reporting* tool to triage norm violating comments for further moderator review. We hope that by releasing Crossmod publicly, Crossmod can be adopted by moderators and researchers going forward.

7 CONCLUSION

In this paper, we build a new AI-based moderation system for assisting Reddit moderators called *Crossmod*. To the best of our knowledge, Crossmod is the first open source, AI-backed sociotechnical moderation system to be designed using participatory methods. We adopted a *mixed-initiative* approach in Crossmod, allowing moderators of subreddits to augment the automatic predictions obtained from an ensemble of classifiers (trained using *cross-community learning*) with human decisions and oversight. First, we conducted formative interviews to examine challenges faced by existing automated tools, and then developed Crossmod to address unmet moderator needs. Finally, we deployed Crossmod in a controlled environment, simulating real-time conversations from two large subreddits with over 10M subscribers each—r/science and r/Futurology. Results from the two-phase evaluation of Crossmod's moderation recommendations through manual review indicate that Crossmod would significantly extend the capabilities of moderators when deployed on Reddit.

8 APPENDIX A: CROSS-COMMUNITY LEARNING-BASED CLASSIFIERS

This appendix provides additional information and detail about the cross-community learning classifiers that act as Crossmod's machine learning back end.

8.1 Counterfactual moderation classifiers

Chandrasekharan et al. (2018) trained classifiers to detect whether a given comment will be removed by moderators of 100 popular subreddits [18]. For each subreddit S_k, we built a classifier clf_{S_k} trained on comments removed by moderators of S_k, along with an equal number of randomly sampled comments from S_k that were not removed, at the time of data collection (i.e., *unremoved* comments). Each in-domain classifier was built entirely using removed and unremoved comments from a single subreddit is referred to as a "subreddit classifier". The mean 10-fold cross-validation F1 score for the 100 study subreddits was 71.4%. This was comparable to the performance achieved in prior work on building purely in-domain classifiers to identify moderated comments within an online community [14, 19]. In our current work, we use all of the 100 subreddit classifiers we trained previously, one for each study subreddit.

8.2 Classifiers for macro norm violations

Chandrasekharan et al. (2018) also factored out "macro norms" that are enforced across most parts of Reddit [18]. For example, posting hate speech in the form of homophobic and racist slurs, using misogynistic slurs, graphic verbal attacks, and distributing pornographic material are removed by moderators of almost all subreddits. The list of all 8 types of macro norms we identified in our prior work are shown in Table 5.

Macro norm violation
Using misogynistic slurs
Opposing political views around Donald Trump (depends on originating subreddit)
Hate speech that is racist or homophobic
Verbal attacks on Reddit or specific subreddits
Posting pornographic links
Personal attacks
Abusing and criticizing moderators
Name-calling or claiming the other person is too sensitive

Table 5. List of all the *macro norm* violations that can be detected by the system's back-end ML models. These are detected using classifiers trained on the macro norm violations identified in our prior work [18].

We used the dataset of all comments labeled to be macro norm violations on Reddit.[17] Using this dataset of macro norm violations, we train FastText classifiers to identify comments violating the different types of *macro norm* violations shared across Reddit. Specifically, we train classifiers to detect 8 types of macro norm violations which are mentioned in Table 5. For training the classifiers to detect each type of macro norm violation, we use a combination of 5,000 comments labeled to be macro norm violations, and a random sample of 5,000 unmoderated Reddit comments as training examples. The average F1 score obtained from 10-fold cross-validation was over 95% for each classifier. The goal of each of these classifiers is to identify whether an unseen comment is a macro norm violation, or not.

8.2.1 Pre-trained classifiers serve as the ML Back-end. All of these pre-trained classifiers are loaded into a back-end for further querying. The two main functions of the back-end classifiers are to make the following predictions about a *new comment*:

- Would the comment be removed by moderators if it were posted on each of the 100 subreddits?
- Is the comment violating a *macro norm*?

[17] A detailed description of this dataset can be found on Github: https://github.com/ceshwar/reddit-norm-violations

The final output from the back-end is the list of predictions obtained from each of subreddit classifiers, and (macro) norm violation classifiers. This takes the form a list of 108 values, and is passed up the stack to higher parts of Crossmod's architecture.

9 ACKNOWLEDGMENTS

We thank Amy Bruckman, Jane Im, Kathryn Cunningham, Cliff Lampe, Sarita Schoenebeck, and Srividhya Chandrasekharan for their valuable inputs that improved this work. We also thank all of the Reddit moderators who took the time to chat with us when we asked, and manually reviewed comments to help evaluate Crossmod. Chandrasekharan and Gilbert were supported by the National Science Foundation under grant IIS-1553376.

REFERENCES

[1] [n. d.]. Perspective API. *https://conversationai.github.io/* ([n. d.]).
[2] Mark S Ackerman. 2000. The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human–Computer Interaction* 15, 2-3 (2000), 179–203.
[3] Zahra Ashktorab and Jessica Vitak. 2016. Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 3895–3905.
[4] Jennifer Beckett. 2018. We need to talk about the mental health of content moderators, September 2018. *http://theconversation.com/we-need-to-talk-about-the-mental-health-of-content-moderators-103830* (2018).
[5] Michael S Bernstein, Andrés Monroy-Hernández, Drew Harry, Paul André, Katrina Panovich, and Gregory G Vargas. 2011. 4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online Community.. In *ICWSM*. 50–57.
[6] Monika Bickert. 2018. Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process, Apr. 2018. *https://newsroom.fb.com/news/2018/04/comprehensive-community-standards/* (2018).
[7] Wiebe E Bijker. 1987. The social construction of Bakelite: Toward a theory of invention. *The social construction of technological systems: New directions in the sociology and history of technology* (1987), 159–187.
[8] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 24.
[9] Bryce Boe. 2016. Python Reddit API Wrapper (PRAW).
[10] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* (2016).
[11] Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 562–570.
[12] Brooks Buffington. April 4, 2015. Personal communication.
[13] Catherine Buni and Soraya Chemaly. 2016. The Secret Rules of the Internet, Apr. 2016. *http://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech* (2016).
[14] Stevie Chancellor, Zhiyuan Jerry Lin, and Munmun De Choudhury. 2016. "This Post Will Just Get Taken Down": Characterizing Removed Pro-Eating Disorder Social Media Content. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 1157–1162.
[15] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. # thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 1201–1213.
[16] Eshwar Chandrasekharan and Eric Gilbert. 2019. Hybrid Approaches to Detect Comments Violating Macro Norms on Reddit. *arXiv preprint arXiv:1904.03596* (2019).
[17] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 31 (Dec. 2017), 22 pages. https://doi.org/10.1145/3134666
[18] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 32.
[19] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM.

[20] Adrian Chen. 2014. The Laborers Who Keep Dick Pics And Beheadings Out Of Your Facebook Feed, October 2014. *https://www.wired.com/2014/10/content-moderation/* (2014).
[21] Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*. IEEE, 598–617.
[22] Nicholas Diakopoulos. 2015. Algorithmic accountability: Journalistic investigation of computational power structures. *Digital journalism* 3, 3 (2015), 398–415.
[23] Motahhare Eslami. 2017. Understanding and Designing Around Users' Interaction with Hidden Algorithms in Sociotechnical Systems. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 57–60.
[24] Randy Farmer and Bryce Glass. 2010. *Building web reputation systems*. " O'Reilly Media, Inc.". 243–276 pages.
[25] Casey Fiesler, Joshua McCann, Kyle Frye, Jed R Brubaker, et al. 2018. Reddit rules! characterizing an ecosystem of governance. In *Twelfth International AAAI Conference on Web and Social Media*.
[26] Laura Garton, Caroline Haythornthwaite, and Barry Wellman. 1997. Studying online social networks. *Journal of computer-mediated communication* 3, 1 (1997), JCMC313.
[27] R Stuart Geiger. 2016. Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society* 19, 6 (2016), 787–803.
[28] Tarleton Gillespie. 2017. Governance of and by platforms. *Sage handbook of social media. London: Sage* (2017).
[29] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
[30] Google. 2018. YouTube Community Guidelines enforcement in Google's Tranparency Report for 2018. *https://transparencyreport.google.com/youtube-policy/removals* (2018).
[31] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 159–166.
[32] Shagun Jhaver, 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 (2019), 31.
[33] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 2 (2018), 12.
[34] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Vol. 2. 427–431.
[35] David Jurgens, Eshwar Chandrasekharan, and Libby Hemphill. 2019. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. *arXiv preprint arXiv:1906.01738* (2019).
[36] Anna Kasunic and Geoff Kaufman. 2018. " At Least the Pizzas You Make Are Hot": Norms, Values, and Abrasive Humor on the Subreddit r/RoastMe. In *Twelfth International AAAI Conference on Web and Social Media*.
[37] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design. MIT Press, Cambridge, MA* (2012), 125–178.
[38] Rob Kling, Roberta Lamb, et al. 2000. IT and organizational change in digital economies: A sociotechnical approach. *Understanding the Digital Economy. Data, Tools, and Research. The MIT Press, Cambridge, MA* (2000).
[39] Rachael Krishna. 2018. Tumblr Launched An Algorithm To Flag Porn And So Far It's Just Caused Chaos, Dec 2018. *https://www.buzzfeednews.com/article/krishrach/tumblr-porn-algorithm-ban* (2018).
[40] Saebom Kwon, Puhe Liang, Sonali Tandon, Jacob Berman, Pai-ju Chang, and Eric Gilbert. 2018. Tweety Holmes: A Browser Extension for Abusive Twitter Profile Detection. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 17–20.
[41] Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 543–550.
[42] John Law and John Hassard. 1999. Actor network theory and after. (1999).
[43] Lawrence Lessig. 1999. *Code and other laws of cyberspace*. Vol. 3. Basic books New York.
[44] Cheng Li, Yue Lu, Qiaozhu Mei, Dong Wang, and Sandeep Pandey. 2015. Click-through prediction for advertising in twitter timeline. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1959–1968.
[45] Kiel Long, John Vines, Selina Sutton, Phillip Brooker, Tom Feltwell, Ben Kirman, Julie Barnett, and Shaun Lawson. 2017. Could You Define That in Bot Terms?: Requesting, Creating and Using Bots on Reddit. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3488–3500.
[46] Sam Machkovech. 2014. No fooling: Reddit's r/games goes silent for one day to call out hate, bigotry, April 2019. *https://arstechnica.com/gaming/2019/04/no-fooling-reddits-rgames-goes-silent-for-one-day-to-call-out-hate/* (2014).

[47] Kaitlin Mahar, Amy X Zhang, and David Karger. 2018. Squadbox: A tool to combat email harassment using friendsourced moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 586.
[48] Enid Mumford. 2000. Socio-technical design: An unfulfilled promise or a future opportunity? In *Organizational and social perspectives on information technology*. Springer, 33–46.
[49] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 145–153.
[50] Jessica Annette Pater, Yacin Nadji, Elizabeth D Mynatt, and Amy S Bruckman. 2014. Just awful enough: the functional dysfunction of the something awful forums. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2407–2410.
[51] John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deep learning for user comment moderation. *arXiv preprint arXiv:1705.09993* (2017).
[52] Robert Peck. 2019. The Punishing Ecstasy of Being a Reddit Moderator, Mar. 2019. *https://www.wired.com/story/the-punishing-ecstasy-of-being-a-reddit-moderator/* (2019).
[53] Benjamin Plackett. 2018. Unpaid and abused: Moderators speak out against Reddit, Aug 2018. *https://www.engadget.com/2018/08/31/reddit-moderators-speak-out/* (2018).
[54] Twitter Public Policy. 2018. Evolving our Twitter Transparency Report: expanded data and insights, December 2018. *https://blog.twitter.com/official/en_us/topics/company/2018/evolving-our-twitter-transparency-report.html* (2018).
[55] Jenny Preece and Diane Maloney-Krichmar. 2003. Online communities: focusing on sociability and usability. *Handbook of human-computer interaction* (2003), 596–620.
[56] Emilee Rader and Rebecca Gray. 2015. Understanding user beliefs about algorithmic curation in the Facebook news feed. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 173–182.
[57] Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. 2000. Reputation systems. *Commun. ACM* 43, 12 (2000), 45–48.
[58] Paul Resnick and Richard Zeckhauser. 2002. Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system. *The Economics of the Internet and E-commerce* 11, 2 (2002), 23–25.
[59] Sarah T Roberts. 2014. *Behind the screen: The hidden digital labor of commercial content moderation.* Ph.D. Dissertation. University of Illinois at Urbana-Champaign.
[60] Sarah T Roberts. 2016. Commercial content moderation: digital laborers' dirty work. (2016).
[61] Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and Psychological Effects of Hateful Speech in Online College Communities. In *Proceedings of the 11th ACM Conference on Web Science*.
[62] Steven B Sawyer and Mohammad Hossein Jarrahi. 2014. Sociotechnical approaches to the study of Information Systems. In *Computing handbook, third edition: Information systems and information technology*. CRC Press, 5–1.
[63] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* (2019), 1461444818821316.
[64] Sara Sood, Judd Antin, and Elizabeth Churchill. 2012. Profanity use in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1481–1490.
[65] Abhay Sukumaran, Stephanie Vezich, Melanie McHugh, and Clifford Nass. 2011. Normative influences on thoughtful online participation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3401–3410.
[66] HN Moderation Team. July 16, 2015. https://news.ycombinator.com/threads?id=dang. (July 16, 2015).
[67] Nitasha Tiku and Casey Newton. February 4, 2015. Twitter CEO:"We suck at dealing with abuse.". http://www.theverge.com/2015/2/4/7982099/twitter-ceo-sent-memo-taking-personal-responsibility-for-the. *The Verge* (February 4, 2015).
[68] Ruth L Williams and Joseph Cothrel. 2000. Four smart ways to run online communities. *MIT Sloan Management Review* 41, 4 (2000), 81.
[69] Daniel Dylan Wray. 2018. The Companies Cleaning the Deepest, Darkest Parts of Social Media, June 2018. *https://www.vice.com/en_us/article/ywe7gb/the-companies-cleaning-the-deepest-darkest-parts-of-social-media* (2018).
[70] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1391–1399.
[71] Tal Zarsky. 2016. The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values* 41, 1 (2016), 118–132.
[72] Amy X Zhang and Justin Cranshaw. 2018. Making sense of group chat through collaborative tagging and summarization. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 196.

Received April 2019; revised June 2019; accepted August 2019