Measuring User-Moderator Alignment on r/ChangeMyView

VINAY KOSHY, University of Illinois, Urbana-Champaign, USA
TANVI BAJPAI, University of Illinois, Urbana-Champaign, USA
ESHWAR CHANDRASEKHARAN, University of Illinois, Urbana-Champaign, USA
HARI SUNDARAM, University of Illinois, Urbana-Champaign, USA
KARRIE KARAHALIOS, University of Illinois, Urbana-Champaign, USA

Social media sites like Reddit, Discord, and Clubhouse utilize a community-reliant approach to content moderation. Under this model, volunteer moderators are tasked with setting and enforcing content rules within the platforms' sub-communities. However, few mechanisms exist to ensure that the rules set by moderators reflect the values of their community. Misalignments between users and moderators can be detrimental to community health. Yet little quantitative work has been done to evaluate the prevalence or nature of user-moderator misalignment. Through a survey of 798 users on `r/ChangeMyView`, we evaluate user-moderator alignment at the level of policy-awareness (does users know what the rules are?), practice-awareness (do users know how the rules are applied?) and policy-/practice-support (do users agree with the rules and how they are applied?). We find that policy-support is high, while practice-support is low – using a hierarchical Bayesian model we estimate the correlation between community opinion and moderator decisions to range from .14 to .45 across subreddit rules. Surprisingly, these correlations were only slightly higher when users were asked to predict moderator actions, demonstrating low awareness of moderation practices. Our findings demonstrate the need for careful analysis of user-moderator alignment at multiple levels. We argue that future work should focus on building tools to empower communities to conduct these analyses themselves.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing**; **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: content moderation; content regulation; user-moderator alignment

ACM Reference Format:
Vinay Koshy, Tanvi Bajpai, Eshwar Chandrasekharan, Hari Sundaram, and Karrie Karahalios. 2023. Measuring User-Moderator Alignment on r/ChangeMyView. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 286 (October 2023), 36 pages. https://doi.org/10.1145/3610077

1 INTRODUCTION

Reddit is one of several social media platforms today that utilize a *community-reliant* model of moderation [38]. On Reddit, volunteer community moderators create and enforce rules to govern permissible behavior within their respective communities (or subreddits). However, no built-in channels exist to ensure that the rules moderators create reflect the needs of their community

Authors' addresses: Vinay Koshy, vkoshy2@illinois.edu, University of Illinois, Urbana-Champaign, Urbana, Illinois, USA; Tanvi Bajpai, tbajpai2@illinois.edu, University of Illinois, Urbana-Champaign, Urbana, Illinois, USA; Eshwar Chandrasekharan, hs1@illinois.edu, University of Illinois, Urbana-Champaign, Urbana, Illinois, USA; Hari Sundaram, hs1@illinois.edu, University of Illinois, Urbana-Champaign, Urbana, Illinois, USA; Karrie Karahalios, kkarahal@illinois.edu, University of Illinois, Urbana-Champaign, Urbana, Illinois, USA.

2573-0142/2023/10-ART286 $15.00
https://doi.org/10.1145/3610077

members. Past work has surfaced some of the harms that can arise when a community's rules are misaligned with its participants' values. For example, community members may become frustrated when well-intended contributions are taken down by moderators [20]. And moderators, who rely on user-reports to catch rule violations at scale, may have to manage an increased workload to compensate for low quality user reports [4].

CSCW researchers have proposed a number of systems to improve different aspects of user-moderator alignment: user-jury systems for reviewing moderation decisions [7, 18, 35], voting processes to create new subreddit rules [44], and machine learning techniques for automating moderation in the presence of disagreement [15, 16]. While these systems are promising, we still lack an understanding of the precise nature and degree of user-moderator misalignments. This makes it difficult to predict which solutions would be most effective in a real community.

In this paper, we conduct an analysis of user-moderator alignment on `r/ChangeMyView` (henceforth `CMV`), a large, discussion-based subreddit. `CMV` was chosen for two reasons. First, it is an established and active community that enforces a diverse set of community rules. Second, it is one of the few subreddits for which community moderators were willing to give us access to the fine-grained moderation data necessary to conduct our study.

1.1 Research Questions

We focus our analysis on measuring alignment at the level of *policy-awareness* (do users know what the rules are?), *practice-awareness* (do users know how the rules are applied?) and *policy-/practice-support* (do users agree with the rules and how they are applied?).

We focus on the above dimensions because of their role in determining which interventions would be most effective for improving user-moderator alignment. If, for example, the rules themselves are unpopular amongst community members (low policy-support), rule co-creation could be an important potential solution. On the other hand, a subreddit's rules could be broadly popular, but spark disagreements when applied (low practice-support). In this case giving users a voice in specific moderation decisions, perhaps in the form of a jury system, might be a more effective intervention. And finally, it's possible users are unaware of a subreddit's moderation policies and practices (low awareness), but find them acceptable once presented with them. This could suggest the need for greater transparency in the moderation process.

To target each form of alignment, we ask the following research questions:

- **RQ1a**: To what extent are users aware of the subreddit rules set by moderators? *(Policy-Awareness)*
- **RQ1b** To what extent are users aware of how those rules are enforced? *(Practice-Awareness)*
- **RQ2a**: To what extent do `CMV` users support the existing subreddit rules set by moderators? *(Policy-Support)*
- **RQ2b**: To what extent do `CMV` users support the way in which moderators enforce the existing set of rules? *(Practice-Support)*

To answer these questions we conducted a survey of `CMV` community members (N=798). To assess policy support and awareness, participants were asked questions to gauge their knowledge of and agreement with the existing subreddit rules. To assess practice support and awareness, participants were asked questions about specific comments that were previously reported for violating subreddit rules. These comments (1876 total) were randomly assigned to users, and were drawn from a large-scale dataset of `CMV` report- and comment- data created by the research team (containing 441,000 comments). A pilot survey was conducted with the subreddit moderators to ensure that the amount of context provided for each sampled comment was enough to allow for meaningful survey ratings.

To analyze the data, we t a series of hierarchical Bayesian models. Our ndings paint a nuanced picture of user-moderator alignment within CMV. On the one hand, we nd that policy-support is relatively high for 4 of the 5 subreddit rules, more than 70% of participants rated their support at least a four on a ve point Likert scale. On the other hand, practice-support was limited. We estimate that correlation between community opinion and the real-world moderator actions ranges from .15 to .45 across rules. Practice-awareness was also relatively low, with users signi cantly underestimating the rate at which moderators removed comments. Crucially, these results held even after using multilevel regression and poststrati cation (MRP) to adjust for biases in the survey sample [10].

We conclude the paper by discussing how our ndings inform the design of previously proposed tools for improving user-moderator alignment. In addition to providing evidence for these guidelines, our analysis demonstrates a novel, systematic approach to evaluating alignment between users and moderators in an online community. Although our study is limited in scope to one community, it may serve as a guide to future studies within this area.

We emphasize that although the goals of this study are oriented around measuring user-moderator alignment, it is only one amongst several measures of community health. Communities may deem low alignment to be acceptable for a number of reasons. For example, moderators of a subreddit experiencing rapid growth may choose to prioritize the needs of long-time subreddit members over those of newcomers. Rather than suggesting that all communities adopt the recommendations in our paper, we argue future work should focus on building tools to allow online communities to monitor alignment themselves. This would make transparent the trade-o s being made between alignment and other values within a community. By lowering the barrier to conducting internal opinion-polling, we could empower communities to adopt the appropriate intervention themselves, and obviate the need for researchers to intervene directly.

2 RELATED WORK

2.1 Community-reliant Moderation

Community-reliant moderation systems are those in which content moderation policies vary between di erent communities within the same platform [17]. This is in contrast with centralized-moderation systems, in which a single set of rules is enforced platform-wide (e.g. as on Twitter). For many community-reliant systems, community rules are created and enforced by volunteer users rather than paid employees of the platform. Examples of such volunteer-based systems include content moderation within Reddit subreddits, Facebook groups, or Discord channels [28]. In practice, most community-reliant moderation systems exist alongside centralized ones [22, 38]. On Reddit for example, a global set of rules is enforced by the platform, while local rule sets are enforced by volunteers within each subreddit. Platform-wide rules tend to be oriented around preventing extreme or illegal forms of antisocial behavior, such as doxxing or incitements of violence. In contrast, community-speci c rules tend to focus on keeping content inline with the community's goals.

Prior work has explored some of the strengths and weaknesses of community-based moderation approaches, both theoretically and empirically. From a theoretical lens, Grimmelmann [17] argues that community-based moderation systems allow users to agree to disagree over moderation disputes by sorting themselves into the communities that match their preferences. Empirically, Chandrasekharan et al. [5] and Fiesler et al. [8] both study the variation in norms across Reddit communities, nding a wide range of rules employed across subreddits. Chandrasekharan et al. [5] identify a hierarchical structure to these rules, categorizing rules as either macro-norms, meso-norms, or micro-norms, depending on how common they were across communities. Jiang et al

[24] also find that moderation preferences vary by country, highlighting one source of the rule variation surfaced by Fiesler et al. [8] and Chandrasekharan et al. [5].

Although the ability to enforce context-sensitive rules is valuable, recent work has surfaced the immense burden volunteer-based community moderation systems place on their moderators [4, 12, 14, 29, 31]. Specifically, moderators are tasked with addressing issues of governance (identifying which rules best serve the communities goals) [12, 14, 31] and of scale (enforcing these rules across potentially thousands of comments each day) [5, 29]. Li et al. [29] estimate the monetary value of this labor to be high, at roughly 3.4 million USD per year on Reddit alone. Further, several studies highlight the emotional dimension to this labor: moderators may self-describe their role as thankless [43] and can be subject to abuse when enforcing policies their communities disagree with [6, 12].

2.2 Misalignments Between Users and Community Moderators

Given their role in community governance, community moderators can find themselves in conflict with users when expectations around community rules are mismatched. Such tensions have been touched on in past studies, both quantitative and qualitative. However, with few exceptions, their exact degree and nature have not been systematically evaluated. Given our interest in evaluating alignment along the dimensions of practice/policy and support/awareness, we here discuss prior studies that address each combination of dimensions, as well as studies that do not fit neatly into these categories.

2.2.1 Policy Awareness and Support.

Few studies directly measure popular support for and awareness of community rules. Still, prior work has highlighted the importance of establishing explicit guidelines and policies for platforms and online communities [18, 27, 40]. Jhaver et al. [21] found that users who are aware of community rules are more likely to view moderator removals of their own comments as fair. Matias [32] found that providing community members with reminders of community guidelines led to more rule compliance for newcomers. Despite these potential benefits Juneja et al. [25] note that moderators themselves have mixed feelings about transparency. In interviews with moderators, they found that while some moderators saw transparency as a virtue, others worried it would only lead to further anti-social behavior within their communities.

2.2.2 Practice Awareness.

Jhaver et al. [20]'s study of user reactions to content removals on Reddit provides indirect support for the importance of practice awareness. Specifically, in a survey of authors of recently removed posts, they found that 73% of respondents disagreed when asked if they suspected their post would be removed. This is in spite of the fact that, just under half of all participants claimed to have read the rules before posting. This emphasizes the gap between policy-awareness and practice-awareness: even when users read the community rules, they may be unclear on how they're applied.

2.2.3 Practice Support.

Lampe and Resnick conducted one of the first quantitative studies on stakeholder disagreements around specific moderation decisions. In analyzing SlashDot's built-in "meta-moderation" system, they found relatively high rates of agreement when moderators were asked to evaluate other moderators' decisions: 92% of all post-hoc reviews being deemed "fair" by another moderator. Atreja et al. [1] conduct a similar study of agreement rates amongst crowdworkers for a misinformation detection task, though they find lower rates of consensus.

Although these studies provide insight into agreement rates for various moderation tasks, none look for systematic disagreements between different groups of stakeholders. Jhaver et al. [20]'s study on user reactions to content removals does get at practice-support indirectly. In their survey of users who recently had a post removed from Reddit, they also asked participants to self-evaluate whether

the moderator action taken against them was fair. A majority (59.9%) of surveyed users found the moderation decision to be unfair. This gives us some signal into the presence of misalignment in practice-support, though it is unclear whether users may be biased when evaluating their own posts. Finally, Resnick et al. [37] conduct another study of moderation preferences across stakeholders. They contrast the ratings of crowdworkers against those of journalists for a misinformation detection task. They find that the ratings of 8 crowdworkers served as a reasonable approximation of the rating of a single journalist. Although insightful, this study is divorced from any particular community context, motivating our work.

2.2.4 Other dimensions. Although we believe the above dimensions are essential to understanding user-moderator alignment, we highlight a few dimensions investigated in other studies. Weld et al. [42] conduct a large-scale evaluation of value-alignment between Reddit users and moderator. They find that moderators tend to find community values like diversity more important than users and values like democracy less important. Unlike our study, their work captures alignment across a wide range of subreddits. However, their methodology is not grounded in the specifics of the communities they study, making it unclear how their findings affect day-to-day operations. Qualitatively, Gilbert [12] observe that differences in demographics can also be a source of tension between users and moderators on Reddit.

2.3 Tools and Interventions for Promoting User-Moderator Alignment

Although modern social media platforms lack built-in systems for promoting user-moderator alignment, many communities have adopted their own ad-hoc practices to achieve this goal. In interviews with moderators, Seering et al. [39] found moderators occasionally used informal channels, like polls and discussion threads, to solicit community feedback on rule changes.

Increased transparency into the moderation process has also been surfaced in prior work as a potentially effective means for improving both policy- and practice-awareness. Compared to other researcher-proposed interventions, transparency measures are also relatively simple for moderators to adopt. Jhaver et al. [21] find that content removal explanations, one of the few built-in tools on Reddit for improving alignment, were associated with a small decrease in the probability of future rule violations. Matias [32] similarly found that providing users reminders of subreddit rules was associated with a 10% increase in rule compliance rate among first time commentors, as well as a 70% increase in the participation rate of first-time commentors.

In an effort to improve practice-alignment, researchers have also proposed utilizing juries of users to evaluate moderation decisions [7, 18, 35]. However, the efficacy of jury systems for content moderation [23], as well as the details of their implementation [7, 35], remain up for debate. Fan and Zhang [7] outline a model for designing digital juries, but highlight several open design dimensions, such as jury composition and deliberation processes. After implementing two examples of digital jury designs, they find that users perceived the decisions made with juries as more just than those made by an algorithm [7]. However, Pan et al. [35] found that decisions made using expert panels were considered more legitimate than those made by juries. Gordon et al. [15] integrates the jury model with ML-based approaches, developing a system to automate moderation decisions in a manner that highlights disagreement within the set of raters used to generate training data. Gordon et al. [16] also develop machine learning techniques to adjust model performance against human-rater disagreements.

Researchers have also built systems beyond juries for broadening participation in community governance. Zhang et al. [44] developed PolicyKit, a software infrastructure that allows online community members to develop, implement, and amend community guidelines and governance processes. By design, PolicyKit is able to handle varying balances of user-moderator participation.

This is a strength of the system, as the right balance between user and moderator control is likely context dependent. However, determining the right balance of control in different contexts remains an open question, motivating our work as a potential first step. Matias and Mou [33] developed an experimentation infrastructure, CivilServant, to enable community members and moderators to evaluate policies. After conducting two case studies using CivilServant, they found that experiments conducted using CivilServant ultimately informed moderator practices as well as community guidelines [33].

Each of the described tools aims to improve user-mod alignment in at least one of the forms we conceptualize in this paper (i.e. practice-/policy- support and awareness). However, what is missing is a systematic framework to measure where misalignments occur within a community. Our work is motivated by the need for such a framework, as it allows us to identify which solutions are most promising for improving alignment, and to evaluate whether they actually work once they have been deployed.

3 DATA COLLECTION

An example of an r/ChangeMyView post:

An example of a user awarding a delta (in the second comment):

Fig. 1. A successful viewpoint change from r/ChangeMyView. Usernames are redacted.

In the following subsections, we will describe the design of our study and the factors that motivated it. We begin by discussing our approach to recruiting subreddits and provide an overview

of CMV, the subreddit we ultimately study. Next, we describe the data collection procedures of our study. This includes a longitudinal scrape of moderation data (Section 3.2.1), a pilot survey conducted with CMV moderators (Section 3.2.2), and a large-scale survey of CMV users (Section 3.3). Ethical considerations made for each procedure will be discussed within their respective subsection. A discussion of broader ethical considerations can be found in Section 3.4. All described data collection procedures were approved by our university's Institutional Review Board (IRB).

3.1 Selection of CMV

3.1.1 Recruitment Procedure. Initially, we reached out to the mods of 20 subreddits about participating in our study. We targeted subreddits that were large (at least 100k subscribers) and contained distinct, subreddit-specific rules (i.e. micro-norms as per Chandrasekharan et al. [5]'s terminology). For recruitment purposes, a preliminary assessment of a subreddit's rules was done by a member of the research team. Rules were subjectively categorized as micro-, meso-, or macro-norms if they were similar in scope to the example norms surfaced in [5].

Understandably, most of the subreddits' moderators expressed discomfort with giving an external party access to private mod-queue and -log data. Unfortunately, this data was necessary to conduct an audit of their moderation practices (see Section 3.2). As a result, CMV was the only subreddit to agree to all the study procedures. Although the inclusion of more subreddits would have been useful, we believe CMV has several properties which make it an interesting case study. We describe these below.

3.1.2 What is CMV? CMV describes itself as "A place to post an opinion you accept may be flawed, in an effort to understand other perspectives on the issue." The subreddit is relatively well-established, and has existed since 2013. CMV posters submit an opinion they hold that they are open to changing, and commenters present them with arguments in an attempt to persuade them to change their view. A unique feature of the subreddit is its deltapoint system. When a commenter successfully changes the viewpoint of a poster, they can be awarded a delta. Accrued deltas are displayed next to a commenter's username, and are tracked on subreddit leaderboards. Figure 1 demonstrates an example of a typical post in CMV, as well as an example of a delta being awarded. A list of CMV commenting rules can be found in Table 1.

3.1.3 Why study CMV? Beyond meeting our inclusion criterion, CMV is noteworthy because its moderators have already taken steps to increase alignment within their community, in line with recommendations from prior work [20, 21, 25]. Thus, our access to CMV allows us to gain insights into what user-moderator alignment looks like in a community where moderators have made reasonable effort to improve it. Of course, this limits our study's generalizability, since many subreddits do not go to these lengths [25].

3.2 Preliminary Dataset Construction

Fig. 2. Diagram of the order in which relevant study datasets were constructed. We first collected 3 months of comment, user report, and moderator action data. This data was then used to generate a pilot survey sent to CMV moderators. Based on feedback from moderators, we redesigned the survey format, and sent the main survey to CMV users.

Table 1. List of rules on CMV, as well as an example of a comment removed for violating each. Example comments were pulled from our scrape of the subreddit's moderation data.

Rule Name	Rule Description	Example Violation
Rule 1: Doesn't Challenge OP (top-level only)	Direct responses to a CMV post must challenge at least one aspect of OP's stated view (however minor), unless they are asking a clarifying question.	I don't really have anything to disagree with you on. I just wanted to say that I've felt this way for a while [...]
Rule 2: Rude/Hostile Comment	Don't be rude or hostile to other users. Your comment will be removed even if the rest of it is solid. They started it is not an excuse. You should report it, not respond to it.	You're kind of a tool. [...]
Rule 3: Bad Faith Accusation	Refrain from accusing OP or anyone else of being unwilling to change their view, or of arguing in bad faith. If you are unsure whether someone is genuine, ask clarifying questions (see: socratic method). If you think they are still exhibiting ill behaviour, please message us.	Why post this in this subreddit? You're clearly not interested in hearing the other side.
Rule 4: Delta Abuse/Misuse or Should Award Delta	Award a delta if you've acknowledged a change in your view. Do not use deltas for any other purpose. You must include an explanation of the change along with the delta so we know it's genuine. Delta abuse includes sarcastic deltas, joke deltas, super-upvote deltas, etc.	I'm giving you a !delta because at least you understand that['s] why I feel what I feel.
Rule 5: Doesn't Contribute Meaningfully	Comments must contribute meaningfully to the conversation. Comments that are only links, jokes, or written upvotes will be removed. Humor and a rmations of agreement can be contained within more substantial comments.	Big word brain hurt. Me smash.

In this section, we detail the datasets constructed prior to the main survey of CMV users. Recall that to evaluate practice-support and awareness, we needed to have community members review past moderation decisions. As such, we began by scraping data from CMV's moderation logs to generate a pool of moderator actions for CMV users to audit in our main survey. We also conducted a small scale pilot survey with CMV moderators to help nalize the design of our main user survey. All primary measures of interest (policy/practice-awareness and policy/practice-support) were collected in the user survey. We describe the moderator action and pilot survey datasets below, and describe the user survey dataset in the next subsection.

3.2.1 Constructing the CMV Moderator Action Dataset. Data was collected over a 3 month period between July 28 and Oct 19, 2021. We used the Reddit API to scrape every comment posted to CMV during this period (441,000 total), to the best of our ability. Due to technical issues, our data collection was down for a two-day period (August 25-26). For each comment, we collected basic information. This includes, but is not limited to: the text of the comment, all user reports issued against the comment, and all moderator actions taken on the comment. User reports are anonymous, and typically include a subreddit-speci c rule the comment was accused of violating. Less frequently, comments were reported for violating platform-wide rules or were given a custom report reason. Moderator actions were always either a removal, meaning the comment gets taken down, or an approval, meaning the comment would be allowed to stay up. A full list of variables stored for each comment can be found in the supplementary materials. These variables were chosen to ensure that we would be able to provide enough context for comments when including them in the user survey. It is important to note that user report and moderator action data were sourced from CMV's mod-queue and mod-log respectively. These are not publicly accessible through the Reddit API. To make this data available, the moderators of CMV had to grant access to the research team's Reddit account.

Prior to data collection, moderators were briefed on the purpose of our study and given an overview of the data we wished to collect. Upon request, we also gave the moderators a chance to

inspect our data collection scripts for the sake of transparency.[1] Moderators were also given the option to view the raw data collected, though none requested this.

After the data collection period, we used comment IDs to match reply comments to their parents. Because CMV moderators typically share public removal explanations after taking down comments we also matched each removed comment to its stated removal reason where possible. To help preserve the privacy of commenters, we dropped all comment IDs from the dataset. In some cases, comments in our dataset were deleted by their original poster after being ingested into our database. To respect the posters' privacy, these comments were dropped from our dataset, and not included in subsequent surveys. This accounted for less than 1% of all collected data.

3.2.2 The Moderator Survey Dataset. Once the CMV moderation action dataset was constructed, we used it to generate a preliminary set of comments to show moderators in a pilot survey. Two goals motivated our decision to conduct the pilot survey. First, we wanted to understand how much context was necessary to identify rule violating comments. Second, we wanted to get a qualitative sense of how often moderators disagreed on how to handle a comment. Although our primary goal is to understand user-moderator alignment, getting a sense of the alignment within the mod team provides useful contextualizing information for our user-moderator findings.

We had 4 of the 19 CMV moderators go through previously posted comments and label each with rule violations where appropriate. Moderators reviewed 134 comments in total, with each comment receiving 2 moderator labels. Unless otherwise specified, comments were assigned to moderators to ensure that the moderator had not removed, approved, or reported the comment before in the past. The survey took moderators an average of 1 hour to complete, and participating moderators were compensated with a $20 Amazon gift card. Because of our uncertainty around how much context to provide for each comment, we felt it was important to ensure that comments with varying trajectories through the moderation pipeline were represented. Specifically we chose to include:

- 20 comments that were never reported by users or removed by moderators
- For each rule A:
  - 5 comments that were reported and removed for violating A
  - 5 comments that were not reported, but were removed for violating A
  - 5 comments that were reported for violating a different rule, but removed for violating A
  - 5 comments that were reported for violating A but were not removed
- For each mod:
  - 5 comments in the mod queue that were previously reviewed by the same moderator under consideration

For Rule 4, which was fairly uncommon and rarely resulted in removal, it was not always possible to find comments for each of the five sub-categories.

For each comment, moderators were provided with the text of the comment, the text of the immediate parent of the comment (if applicable), and the title and body of the post associated with the comment. Moderators were asked to select which rule they believe the comment violated out of a list, if any, and to provide any additional information they felt like they would need to see in order to make a decision.

We found only a few cases where moderators needed additional information before making their decisions (details can be found in the supplementary materials). Hence, we decided our survey format provided an adequate amount of context. However, moderators found the survey's representation of comments and surrounding context confusing. As a result, we modified the survey

[1]These scripts can be viewed here: https://github.com/vinyoshy2/public-moderation-data-collection

representation of comments to more directly mirror the Reddit interface. The updated comment representation can be found in g. 5, while the original can be found in Appendix B.

3.3 The User Survey Dataset

Fig. 3. Diagram of user survey flow. Users completed a series of tasks designed to measure their participation rates on Reddit, and evaluate their alignment with the moderators

After the survey design was nalized, we distributed it to CMV users. The distribution and design of the survey will be described in this section. The survey consisted of ve major sections (see Figure 3). In the rst section we collected participation metrics for each user to conduct response bias correction. In the second section, users were shown a series of previously submitted comments, and were asked questions to assess practice support and practice awareness. In the third section, users completed a rule recognition task to assess policy awareness. In the fourth section, we assessed policy support by asking users to provide a Likert rating of their support for each subreddit rule. Finally, having been shown the subreddit rules, users were asked to relabel comments from the second section based on which rules they believed would apply as written. Although not directly associated with any of our research questions, this section was included to assess whether our practice awareness stemmed from a genuine lack of knowledge of how the rules were applied, or simply a high degree of subjectivity in how the rules are worded.

3.3.1 Survey Distribution.

We distributed survey links through Reddit direct messages. Links were sent out over a one month period between March 17th and April 19th, 2022. Users who commented on CMV were sent a survey link with a 20% probability each time they commented, and survey links were never sent to the same user twice. A total of 8376 links were sent out over the 34 day period, yielding 798 complete survey responses. One out of every 100 participants received a $50 Amazon gift card. Only users who supplied a username were eligible to receive this, as we did not collect any other personal information that would allow us to contact users if they won. Survey participation was restricted to participants over 18 and under 65 years of age, and to participants who resided in the United States.

3.3.2 Participation Metrics Collection.

Given that some degree of response bias is expected in any survey, we wanted to collect background metrics for each sampled participant to compare against the subreddit population. Here, we chose metrics that we believed would a ect both whether a participant would respond to our survey, and how they would respond. We identi ed a few factors that met these criteria: amount of experience with Reddit, activity level (in r/ChangeMyView and across subreddits), and prior interactions with moderators (on CMV and across subreddits). Speci cally we collected:

(1) Account age
(2) Number of comments posted on Reddit in the last month

(3) Number of comments posted on CMV in the last month
(4) Number of comments removed on Reddit in the last month
(5) Number of comments removed on CMV in the last month
(6) Whether or not the user was a moderator of any subreddits

Fig. 4. Comparison of survey respondents to survey link recipients (both respondents and non-respondents) along various measures. Notice the survey over-samples users with lower commenting rates, older accounts, and fewer comment removals.

Users were given two means to provide this information. First, they were given the option to provide us with their Reddit usernames, allowing us to scrape these metrics from their pro le through the Reddit API. They were also given the option to self report this data rather than provide their username. Self-reported data was collected in interval format, rather than in precise numbers (e.g. "1-2 years" for account age). 103 of the 798 participants chose to self report their information. If users opted to self report their information, their usernames were not linked to their survey response at any point in our dataset. If a username was supplied, it would be linked to their survey response until we scraped the relevant metrics from their pro le, at which point it was deleted from our dataset. To conduct response bias adjustment, we also needed to estimate the joint distribution of the collected variables in the subreddit as a whole. To do so, we scraped these same metrics for 1270 randomly selected user accounts who were sent a survey link, regardless of whether they responded or not.

Two issues arose when scraping usernames collected through the survey. First, a substantial number of respondents provided us with usernames that did not correspond to real Reddit accounts. In some cases, these appeared to be typos (e.g. users used underscores instead of hyphens). In other cases these usernames seemed to be deliberately false. Additionally, some accounts appeared to have been deleted or suspended before we were able to scrape them. Of the 695 usernames provided, we collected participation data from 518 successfully. The MRP analysis used for response bias correction (see Section 4) was conducted using these 518 responses and the 103 responses containing self-reported participation rates. All unadjusted results are presented using the full 798 responses. The validity of our MRP analysis rests on the assumption that the relationship between the adjustment variables and the survey measures does not di er between accounts that were and were not successfully scraped. This assumption should be taken into account when interpreting our results.

We found a handful of biases in our data when comparing the 518 scraped participation metrics against the subreddit population (see Figure 4). We nd a skew towards users with lower commenting rates across Reddit (median of 87.5 comments in the sample vs 131 in the population) and lower commenting rates on CMV (median of 4 vs 9). This skew ran counter to expectations, as we predicted highly active users would be more likely to participate. Survey participants were also a little less likely to have comments removed on Reddit (median of 0 vs 1), and tended to have older accounts (median of 5.78 years compared to 3.94). Survey participants were only slightly less likely to be the moderator of any subreddits.

All biases along these variables should be taken into account when interpreting any unadjusted results presented in later sections. For adjusted results, one should still be mindful of potential biasing factors we were unable to adjust for.

3.3.3 Practice Awareness and Support. In the second section of the survey, participants were shown limited re-creations of comments previously posted to the subreddit. Figure 5 contains an example of a comment shown to survey takers. Users were shown ve comments, and each comment was shown to two users. Each time a new comment needed to be included in the survey, we rst randomly selected a subreddit rule, and then randomly selected a comment that had been reported by a user for violating it. We opted to focus on reported comments because they are guaranteed to have a moderator decision associated with them, allowing for direct comparison between user and moderator opinion. This comes with a tradeo . A reported comment is more likely to be controversial, since at least one user thought it was acceptable to post (the author), and at least one user thought it was unacceptable (the reporter). Although reported comments re ect the bulk of moderators' day-to-day workload, our analysis may not re ect the degree to which users are happy with the end results of CMV's moderation system.

Fig. 5. Example of the interface used to show comments to users. The text of the relevant comment is highlighted. Where applicable, users are also shown the immediate parent of the comment, as well as the post associated with the comment

Comments used in the pilot moderator survey were also included as part of a separate pool to allow for a small, but direct comparison between users and moderators in the survey environment. In total, 1876 comments were rated. For each comment, participants were asked how they would handle the comment if they were a moderator (practice support), and how they believed the current moderators would handle the comment (practice awareness). Response options corresponded to removal, approval, unsure, and a write-in to specify some other action.

To determine appropriate sample size for this study, we followed recommendations by Gelman and Carlin[9] and conducted extensive analysis on simulated data prior to deploying the survey. We determined how precisely we could capture practice-support and -awareness with varying numbers of sampled comments and ratings per comment. The results of these simulations are included in the supplementary materials. We also note that because our analysis is Bayesian, our model results have built-in uncertainty quanti cation. As long as there is enough data to ensure the model is identi able, credible intervals generated from our model's posterior should capture the degree of certainty contained in the data. We provide evidence of model t in Appendix A.4.

Due to the potentially sensitive/o ensive nature of some previously removed comments (e.g. comments containing references to suicide or hate speech), users were provided with a short content warning prior to being shown this section and given the option to skip it entirely.[1] Only a single user who completed the survey selected this option.

To avoid priming participants, this comment rating task was completed prior to the tasks in which participants were shown the actual subreddit rules.

3.3.4 Policy Awareness. To assess policy awareness, we set up a rule recognition task in which users would be asked to identify CMV rules out of a list containing the ve actual CMV rules (Table 1),

[1]The text of the content warning was: On the next few pages, you will see comments that were previously posted to r/ChangeMyView, some of which were removed by the moderators. The research team conducting this study feels that some o these comments may contain text that is potentially o ensive (e.g. references to suicide, hate speech, body-shaming, etc). If you do not feel comfortable viewing such text, please indicate so below, and you will skip this portion of the study.

Table 2. A list of decoy rules used in the rule recognition task

Rule Name	Subreddit	Rule Description
Invalidation	r/AskWomen	No invalidation of others' experiences. Invalidation includes but is not limited to: Stating or implying that a user's personal experiences or opinions are wrong or otherwise invalid, or debating someone's personal experiences or opinions
Information Quality	r/Coronavirus	Keep information quality high. There are many places online to discuss conspiracies and speculate, we ask you not to do so here
Replies to Removal	r/CanadaPolitics	Replies to removed comments or removal notices will be removed without notice, at the discretion of the moderators.
Medical Advice	r/AskScience	Medical advice is strictly prohibited on CMV. Asking for or giving medical advice are both against the rules.
Cite Sources	r/NeutralPolitics	If you're claiming something to be true, you need to back it up with a quali ed source. There is no 'common knowledge' exception, and anecdotal evidence is not allowed.

and ve additional decoy rules (Table 2). Decoy rules were selected from other subreddits oriented around discussion of sensitive or political topics. In selecting decoy rules, the research team tried to ensure that they would still make sense in the context of CMV, while being su ciently di erent from the existing CMV rules. Each rule's text description was taken from the sidebar of its source subreddit. Although we acknowledge that the selection of decoy rules introduces subjectivity into the experiment design, it provides some signal of rule awareness while minimizing cognitive load for participants. Further, it is easy to make this subjectivity transparent since we provide the chosen decoy rules above. Since measuring policy awareness relies on participants not having seen the subreddit rules, we included it before the policy support Likert rating question.

3.3.5 Policy Support. Participants were next shown a list of the real subreddit rules, and were asked to rate their support for each of the subreddit rules on a 1 to 5 point Likert scale. Likert scales are commonly used in political science and psychology research, and provide a low-cost way to measure positive or negative sentiment towards a policy [30].

3.3.6 Rule Application Task. Finally, participants asked to read through the comments from the practice-support and practice-awareness tasks a second time. For each comment, they were shown the list of CMV rules and asked to select any they believed applied to the comment. This section was included to add context to the practice awareness task. Speci cally, it enables us to distinguish a genuine lack of knowledge of how a rule is enforced from subjectivity that may be inherent to the application of a rule.

3.4 Ethical Considerations

In designing this study, care was taken to respect the autonomy and privacy of CMV moderators and community members. In general, we followed guidelines from Bruckman's work on ethically studying online communities [2]. This included seeking permission from community moderators before distributing survey links, and orienting our study around producing information that can directly inform community policy. Additional measures were taken with non-publicly available data collected in our study. At no point during collection were moderator usernames associated with speci c actions. Instead, moderator usernames were replaced with salted cryptographic hashes to prevent moderator identity from being revealed. Further, we remove potentially identifying

information from comments and survey responses in our datasets before releasing them to a wider audience (e.g. text, author, exact time of posting, etc).

4 DATA ANALYSIS

We applied three primary forms of analysis to our data. For the policy awareness and policy support questions, we compute simple summary statistics to characterize our in-sample data. For the practice-support and practice-awareness questions, we rely on a more sophisticated hierarchical Bayesian model to disentangle the relationship between user- and moderator-supplied labels across comments and rules. Finally, to understand how our results generalize to the population of non-survey-respondents, we apply MRP analysis to all four of our primary alignment measures (policy/practice- support and awareness). We begin by providing an overview of Bayesian modeling, and motivate its use in this study. We then describe the model used to evaluate practice-support and -awareness, and conclude by describing our use of MRP.

4.1 Bayesian Modeling

Hierarchical Bayesian models were used throughout our analysis. This decision is motivated by several factors. First, our dataset is naturally hierarchical, as we are interested in estimating values across subreddit rules (e.g. On average, what proportion of users think that a comment reported for violating Rule X should be removed?). Partial pooling allows us to share information across rules, improving estimation. Second, Bayesian modeling allows us to incorporate weak prior information into our analysis, biasing our estimation towards reasonable values [34]. This is essential in small studies where variance is high. Additionally, because we use maximum-entropy priors, our results are maximally conservative given the assumptions our priors encode [34]. Kay et al [26] provide a detailed discussion of why Bayesian modeling is a good fit for HCI research.

All models are fit using the Numpyro implementation of the No U-Turn Sampler [36]. Full mathematical descriptions of each model are included in Appendix A.2. All analysis code (and the requisiste data) is accessible through publically available Google Colab notebooks.[2] To assess model fit, we follow the recommendation of Gelman et al [11] and use visual posterior predictive checks. Because Bayesian models are generative in nature, a posterior predictive check entails simulating data according to the fitted model parameters and then comparing these simulations to the observed data. To assess fit, we can check whether the model has learned to replicate the patterns of interest from the dataset. Posterior predictive checks for the analysis of the comment-rating model are included in Appendix A.4.

4.2 Modeling Practice Support and Practice Awareness

Recall that in our user survey, participants were asked to provide three ratings for each comment shown. In particular, participants were asked to predict what the moderators would do with the comment, to state what they themselves would do if they were a moderator, and to state whether a subreddit rule was applicable to the comment. We are interested in understanding whether user-supplied labels correlate with real life actions taken by moderators. For the first two comment rating tasks presented to the users (eliciting their personal opinion and their prediction of how the mods would act), additional steps had to be taken to compare the provided labels against real life moderator actions. First, responses were treated as binary, with *This comment should be removed* denoting an express preference for removal and *No action is necessary* and *Unsure/Need more context to tell* denoting a lack of immediate preference for removal.

[2] https://drive.google.com/drive/folders/1EqZ1RBWFZokdoHGT-pB-7pKPbHBM2IJT?usp=sharing

Second, special considerations had to be made for the write-in option associated with the Some other action should be taken answers. This option was selected in 2% of all ratings for the rst practice-awareness task, and in 6% of all ratings for the practice-support task. The write-in options for these questions were manually parsed to determine if participants indicated support for immediate removal or not, and were mapped to the corresponding answer choice. Speci c details on how these responses were parsed are included in the supplementary materials.

In general, we will refer to a user-provided label as positive when a user states that a comment should be removed, predicts a comment will be removed, or states that a rule applies to a comment. We will refer to a user-provided label as negative when they do not express a preference for removal, predict a removal, or state that a rule applies to the comment. Mod-provided labels (either in a survey or real life) are positive when they correspond to a comment being removed, and negative when they do not.

We compute two measures of alignment between real life and survey labels for these questions. First, we consider whether the rate of positive labels are similar. Second, we consider whether real life labels are correlated with probability of observing a positive survey label. In other words, are comments that receive more positive labels in the survey likelier to have been removed in real life?

Let $2_{8,A}$ correspond to the 8^{th} comment reported for violating rule A. Our model says that underlying each comment is:

A latent tendency $j_{A,D,8}$ that user D would remove the comment if they were a moderator. The label observed is denoted by the variable $G_{A,D,8}$

A latent tendency $k_{A,D,8}$ that user D thinks a moderator will remove the comment. The label observed is denoted by the variable $~_{A,D,8}$

A latent tendency $l_{A,D,8}$ that a user D thinks a rule applies to the comment. The label observed is denoted by the variable $I_{A,D,8}$

A latent tendency $[_{A,8}$ that the comment will be removed in real life. The label observed is denoted by the variable $<_{A,8}$

A latent tendency $\_{A,E,8}$ that a moderator E will think a comment should be removed within the survey environment. Recall that the comments included in the moderator survey were also included in the user survey, though these only account for only a small percentage of comments shown to users. The label observed is denoted by the variable $B_{A,E,8}$

We want to allow these latent probabilities to co-vary, and learn the degree to which they co-vary from the data. We model this as a generative process. The latent continuous tendencies above are drawn from a multivariate-Normal distribution for each comment and are transformed into positive label probabilities via logistic transformation. Thus, for each rule A we learn a set of means $\grave{}_A()$ and covariance matrices $(_A)$ that govern the distribution of labels across all comments reported for violating that rule. Intuitively, the learned means can tell us whether the rates of positive labels are similar in di erent contexts. The learned covariance matrices can tell us whether the occurrences of positive labels are correlated across contexts. This is described in the set of equations below:

$$
\begin{aligned}
G_{A,D,8} &\sim \text{Bernoulli}(\text{Logistic}(j_{A,8})) \\
~_{A,D,8} &\sim \text{Bernoulli}(\text{Logistic}(k_{A,8})) \\
I_{A,D,8} &\sim \text{Bernoulli}(\text{Logistic}(l_{A,8}, l_{D}^{00})) \\
<_{A,8} &\sim \text{Bernoulli}(\text{Logistic}([_{A,8})) \\
B_{A,E,8} &\sim \text{Bernoulli}(\text{Logistic}(\_{A,8})) \\
\langle j_{A,8}, k_{A,8}, l_{A,8}, [_{A,8}, \_{A,8} \rangle &\sim \text{MvNormal}(\grave{}_A, (_A)
\end{aligned}
\tag{1}
$$

We set priors on model parameters $\grave{}_A$ and $_A$ to be weakly informative. We allow for any value of the parameters to be possible, but bias them somewhat against extreme values. Where possible, we incorporate partial pooling into the priors for $\grave{}_A$ across rating contexts.

Because each user rates multiple comments in our design, we include user-speci c slopes to account for our repeated-measures design. For a full description of the model, see Appendix A.3

4.3 What is MRP?

Response bias correction for all analysis is done via multilevel regression and post-strati cation (MRP) [10, 41]. In MRP, a linear model is t to predict a participant's response to a survey question based on a set of predictors chosen by the researcher. Ideally, these predictors should represent the factors that a ect both whether a participant will respond, as well as how they respond. To use MRP, the researcher must have access to these variables within their survey sample, as well as in the population they wish to generalize to. The linear model is t using the sample data, and then applied to predict responses in the general population. Crucially, this method assumes that the relationship between the predictors and the outcome are the same for the sample as in the general population. Although model t to the survey sample can be assessed through a number of traditional means (e.g. R^2 values, cross-validation, etc), t to the general population is not able to be assessed without collecting additional data.

Below, we outline how MRP is applied to each alignment method. Additional details for the models can be found in the Appendix A.

4.3.1 MRP for Policy-Awareness.

We t a logistic regression to model relationship between the response bias predictor variables, and the log-odds that a participant will select each rule (real or decoy) out of the provided list. Our model consists of a varying intercept for each rule and a set of slopes for each rule associated with the predictors used for response bias correction.

We need to make special considerations to handle the response bias predictors associated with self-reported participation data. This data is interval-censored. In other words, we observe an interval that contains the true value we're interested in, rather than directly observing the value itself. For example, if a user's account is 18 months old, we only observe the interval 1-2 years. To handle this, we model the true value for each predictor as a latent variable which generates the observed interval. The priors for these latent variables are described in Appendix A.2.

4.3.2 MRP for Policy-Support.

Analysis for this section is similar, but uses an ordinal regression rather than a logistic regression. An ordinal regression is needed to model Likert response data, as it cannot be assumed to be even-interval [34]. In other words, although a response of `2' to more support than a `1', and a `3' corresponds to more support than a `2,' we have no way to claim that the di erence between a `3' and `2' is the same as the di erence between a `2' and `1.'

4.3.3 MRP for Practice-Support and -Awareness.

Applying MRP to our comment rating data requires only a small modi cation to the model described in Section 4.2. We simply treat $\grave{}_A$ as a linear function of the response bias predictor variables. This allows us to learn the relationship between the predictors and the observed outcomes within our survey sample.

5 RESULTS

5.1 Finding 1: Mixed Evidence for Policy Awareness (RQ1a)

First, we describe the results for the rule recognition task. In Figure 6, we observe a clear separation between the real rules (selected an average of 82% of the time) and the fake rules (selected an average of 30% of the time). In Figure 7a, we see that participants were able to identify the correct rules reasonably well, with 52% of participants making two mistakes or fewer. Still, only 13% of

participants achieved a perfect score. Although this provides clear evidence that users have some ability to recognize the subreddit rules, it provides only limited evidence that users are truly aware of the subreddit rules.

Fig. 6. Percentage of respondents who selected each rule as belonging to r/CMV's official list. Decoy rules are indicated in grey. Notice that the real rules are selected at 2-4x the rate of the decoy rules.

Surprisingly, despite the sample skew described in the previous section, we found that the results did not change significantly after conducting MRP. This was the case for all survey sections in which MRP was applied. As such, we save the reporting of MRP-adjusted results for the appendix (Appendix A.1). However, we did find significant associations between participation variables and the response values. Specifically, compared to users who are otherwise similar with respect to the participation variables, users who were more active on r/CMV were significantly more likely to correctly identify most subreddit rules in this section (see the supplementary materials for a table of model coefficients).

5.2 Finding 2: Policy Support is Relatively High (RQ2a)

Evidence for policy support is comparatively much stronger. Figure 7b demonstrates the overwhelmingly positive skew in ratings for all of the subreddit rules. The average ratings for the five rules were 4.5, 4.1, 3.5, 4.5, and 4.1 respectively. Rule 3, which was selected the least often amongst the CMV rules in the recognition task, also received the most negative ratings. This may be because participants view the rule as unnecessary for the subreddit to fulfill its function. This is in stark contrast to Rule 4, which was selected the most often in the rule recognition task, and received the highest share of positive ratings in this section. Rule 4, unlike Rule 3, governs a core element of CMV, the delta system, which may explain its high ratings.

(a) The proportion of users who correctly labeled rules.

(b) Distribution of Likert score ratings for each rule

Fig. 7. On the le (a): the proportion of users who correctly labeled rules. Real CMV rules are considered correctly labeled if the user selected them as belonging to the o icial list of rules. Fake rules are considered correctly labeled if the user did not select them. Almost all scores are above a 5, suggesting that users are at least doing be er than randomly guess. On the right (b): the distribution of Likert scores received by each rule. Support for rules is generally high, with all 5 rules exhibiting a le -skew.

Results from our MRP model indicate that users who were more active on r/ChangeMyView were signi cantly more likely to support all ve subreddit rules compared to users who are otherwise similar with respect to the adjustment variables. On the other hand, users with more prior comment removals (both on CMV and across Reddit) were signi cantly less likely to support all rules (see the supplementary materials).

5.3 Practice Support and Awareness Findings (RQ1b, RQ2b)

We assess practice support using results from our tted Bayesian model. Our model is designed to explain data from all three user comment rating tasks (practice-awareness, practice support, and rule application), the data on real life moderator decisions, and the data on survey moderator decisions. As with prior sections, response-bias adjusted results are included in the appendix. All the inferences discussed below result from inspecting di erent learned components of our unadjusted model. Note that the estimated values presented below are not model parameters themselves. Because we use logistic and ordinal transformations to model the data in this section, our model parameters are not easily interpretable. However, because our model is generative, we can simulate conducting our experiment hundreds of times by generating data according to draws of parameters from the tted model's posterior. We can then compute more interpretable quantities of interest across the simulated experiments, and generate Bayesian credible intervals for them. An important caveat is that these results are only meaningful if the model ts the data well. We present evidence of strong model t in Appendix A.4.

5.3.1 Finding 3: Low to Moderate Practice Support/Awareness. First, we look at whether users and moderators supply positive labels in each context at the same rate. Figure 9a demonstrates that users signi cantly underestimated the rate at which moderators removed comments, suggesting low practice-awareness. Rule 4 was the sole exception to this trend. This is because for Rule 4 reports, CMV moderators tend to provide a warning rst, rather than immediately removing comments. In some cases the underestimation rate was quite large for Rule 1, users predicted moderators would remove comments 39% (95% CI: [34%-44%]) of the time in practice these comments were removed 80% (95% CI: [76% - 84%]) of the time. For other rules, like Rules 2 and 3, this underestimation was smaller. A similar, though slightly more exaggerated trend was present for the practice support questions. In general, survey takers supported comment removal at a slightly lower rate than they predicted.

Although this data suggests that moderators and users have di erent thresholds for removal, it's unclear whether moderator decisions are at least correlated with user opinion. In other words, does removal become more likely as more users think a comment should be removed? Figure 9b shows that such correlations exist, but range from low (.15, 95% CI: [.02-.27] for Rule 1) to modest (.45, 95% CI: [.31-.59] for Rule 4) for both practice support and practice awareness. Interestingly, participants' practice support and practice awareness labels were strongly correlated (.89, 95% CI: [.81-.97]). This could suggest that, beyond their own beliefs, participants had little information to go o of when predicting how moderators would act.

5.3.2 Finding 4: Disagreements Common Between Users. Although we do see a gap between user preference for comment removals and actual moderator actions, it's worth noting that users were not uni ed in their perspectives on this. Figure 8 shows the frequency of disagreements between the pairs of users assigned to each comment in our survey. Disagreements were relatively common, occurring for 30% of all comments rated. Looking at the comments included in the moderator survey gives us a direct way to compare user and moderator disagreement rates. Across these comments, users directly disagreed 28% of the time, while moderators disagreed 14% of the time.[3] Interestingly however, for the moderator survey comments, users and moderators disagreed with the real life moderation decisions at roughly comparable rates (30% of the time compared to 26% of the time). Given the qualitative nature of the pilot survey, it's di cult to evaluate the signi cance of these ndings, though they do suggest that moderator consistency may be an important topic for future work.

5.3.3 Finding 5: Rules Hard to Apply, Even Once Revealed. Recall that in the nal section of our survey, participants were shown a list of subreddit rules and asked to relabel comments with the rules they believed were applicable. Our model reveals that these rule application questions yielded results that di ered from the practice support and practice awareness questions. Figure 9a shows that participants supplied positive labels in this context at a far higher rate than before. Generally, the rate of positive labels here more closely matched the real life comment removal rate, though Rule 4 was again an exception.

Although the di erences in rate were clear, the di erences in correlation were not. Generally, the rule application labels had a slightly stronger correlation with real life removal data than the practice-support and policy-support labels. However, these di erences were not signi cant. Still, the fact that the correlations in the application context were not much higher suggests that the rules (as written in the subreddit's sidebar) may not have a single clear interpretation. This might

[3]Note that the interface presented to the users di ered slightly from the one presented to moderators. Although the same information was provided in both interfaces, this should be considered when comparing these disagreement rates; see the supplementary materials for more information.

Fig. 8. Co-occurrences of practice-support ratings by rule. Notice that for almost all rules, users favor approving comments over removing them. Disagreements are common across rules as well, and were modal for Rules 1 and 2. Co-occurrences of ratings in other contexts are included in the supplementary materials. The subplot in the top le corresponds to the set of comments that were also included in the survey sent to moderators.

explain the apparent contradiction between the high policy support and low-to-moderate practice support observed previously in this section.

No clear trends emerged from the MRP regression conducted for practice-awareness and support data, though a table of coe cients can be found in the supplementary materials.

6 DISCUSSION

In the previous sections we presented an empirical study of user-moderator alignment on CMV. Below, we re ect on what our ndings can tell us about CMV, and distributed moderation more broadly.

6.1 Support for Policy, but not Practice

Our results strongly suggest that misalignment between CMV users and moderators occurs at the level of practice rather than policy. Even after making adjustments for potential sources of response bias, we nd that support for the rules themselves is high. Average support was greater than a 4 out of 5 for all rules but Rule 3 (see Table 1). However, when asked to examine speci c comments, users favored removal signi cantly less often than moderators (Figure 9a). And further, correlations between user opinion and moderator decisions were modest at best we estimate these values to

(a) Average probability of a positive label by rule (b) Correlation with simulated decisions by rule

Fig. 9. Figure 9a compares the average probability of observing a positive label in each rating context by rule. Relative to the moderator removal rate, users are typically far less likely to supply a positive label in either the Opinion or Prediction contexts. However, a er seeing the rules, users become far more likely to supply a positive label in the Application context. Rule 4 is the one exception to this trend. Figure 9b contains the estimated correlation between the probability of a positive label and a simulated decision made by a moderator. These correlations generally ranged from low to moderate values. The application labels were most correlated with moderator decisions, though usually not significantly. Rule 4 was again the exception.

range from .15 to .45, across rules (Figure 9b). Taken together, these numbers suggest that although

the CMV community is broadly supportive of the existing rules, there is substantial disagreement over how those rules should be applied.

In spite of this limited practice-support, CMV's rules have remained relatively stable for multiple years. Given the prior work on the harms of user-moderator misalignment [19, 20], this is surprising. One explanation for this stability is that users may make incorrect assumptions about how moderators apply the rules. In our practice awareness section, users significantly underestimated the rate at which moderators removed comments (Figure 9a), and their predictions of moderator actions were only modestly correlated with the real decisions (Figure 9b).

Another possibility is that users and moderators agree on which comments are harmful, but disagree over the appropriate penalty. When given the option to specify an alternative action for moderators to take, users occasionally raised the possibility of graduated penalties. This included written warnings, locking replies, or limiting a comment's visibility without outright removing it. Generally, moderators of CMV do not utilize graduated penalties. However for Rule 4 violations, CMV moderators often provide warnings before escalating to comment removal. This might explain why Rule 4 saw the highest rate of practice support in our survey. Users did not specify such alternate actions very often, but our results do raise this as an area for further investigation.

Finally, it could be that users of CMV do not care about moderation, or that moderation is not as salient to their typical interactions as prior work suggests. Interviews with community members could help tease this apart in future work.

6.2 Awareness of Moderation is Low

Our results also indicate that users have a limited ability to predict how moderators will enforce the rules. Our data is not conclusive about whether these inaccurate predictions are the result of participants not knowing the rules, or not understanding how the rules are applied. We provide limited evidence for rule awareness through our rule recognition task – users selected correct subreddit rules out of a list at 2-4 times the rate of decoy rules (Figure 6). Interestingly, Rule 1, which was selected the second most often in the policy awareness task, scored worst in the practice awareness task. This suggests that there may be a genuine gap between policy and practice awareness, at least for some rules. Still, our policy awareness measure lacks nuance and could be strengthened for future work.

In the past, CSCW researchers have identified transparency as an important tool for promoting policy and practice awareness. Jhaver et al. [20, 21] recommended clear guidelines and removal explanations as a means to improve both kinds of awareness. Matias [32] demonstrated how reminding users of the rules can increase compliance in newcomers. Our results are initially surprising in light of these recommendations. We observe large gaps in practice awareness despite the fact CMV adopts several of the recommended measures – CMV moderators maintain detailed guidelines on how rules are applied[4] and provide public removal explanations for most removed comments. Prior work raises the possibility Reddit's mobile interface, which makes it difficult to find subreddit rule guidelines, may explain why these measures are ineffective [25].

Another possibility is that transparency around what the rules are is inadequate for enabling users to predict moderator behavior. On Reddit, users are unable to see the contents of removed comments, which may make it difficult for them to understand what constitutes a rule violation. Transparency measures that enable users to see examples of rule violating content may allow users to more effectively anticipate moderator behavior.

It is also possible that our results are a byproduct of our decision to audit reported comments. Transparency actions taken by CMV moderators could be effective at preventing most rule-violating

[4] https://www.reddit.com/r/changemyview/wiki/rules/

comments from being posted in the first place, leaving only the borderline cases to be posted and subsequently reported. As such, our study design might produce low practice-awareness rates regardless of how transparent moderators are. If this is the case, our results still have value in providing additional context for Jhaver et al.[20]'s study on user reactions to post removals. Where they found that most users did not expect their own posts to be removed, we find that this extends to others' comments as well. In short, our study confirms that Jhaver et al.[20]'s results are not just a function of users' personal biases.

6.3 Implications for Jury Systems

Our results highlight two potential benefits to incorporating user jury systems into online communities. First, we observed low rates of practice support for most CMV rules. Jury systems give users a direct say in how community rules are enforced, making them a natural solution if improving user-moderator alignment on CMVs desired. Second, we observed non-trivial rates of disagreement between moderators in our moderator survey. Specifically, moderators within the survey disagreed with each other 16% of the time, and disagreed with the real life decision made on comments 26% of the time. Although aggregating multiple moderator opinions could improve decision consistency, it is likely infeasible given that moderation teams are already overburdened. User juries are a promising alternative to this, since they provide a lower-cost channel to get multiple judgements for each reported comment, thereby improving decision consistency. We caution that the small size of the moderator survey makes the results around moderator disagreement rates merely preliminary, and call for future work in this direction.

While the precise details of an ideal jury system work flow are out of scope for this paper, our results also provide additional context for design decisions surfaced in prior work. Hu et al.[18] emphasized the importance of the mechanism used to aggregate juror opinions. They found that deliberative discussion leads to more consistent decision-making when compared to a majority vote, but is more labor intensive. If users agreed about how the rules should be enforced, the two methods would lead to the same outcomes most of the time, making the choice of aggregation method obvious. However, we observed high rates of disagreement between users (Figure 8). Thus, our work reaffirms that either choice of aggregation method comes with real trade-offs.

Past papers have broached the question of who should serve on a jury [15, 23, 35, 37]. Our regression results confirm that the jury selection system is impactful - highly active users were more likely to support the rules, while users with more prior rule violations were less likely to support them. Jury selection procedures that over- or under- sample such users could lead to different outcomes for reported comments.

6.4 All Alignment Measures Vary by Rules

Our hierarchical modeling approach highlights the variation in alignment across rules. Rule 4, governing the subreddit's delta system, stood out as consistently scoring high across all forms of alignment, usually significantly more than the rest of the rules. Other rules scored high in some places, and low in others. Rule 1, for example, scored amongst the highest for both policy-based measures of alignment and lowest for both practice-based measures. However, the source of this variation remains unclear. Intuitively, one might expect that macro-norms would tend to have higher alignment, since many communities have chosen to adopt them. Our results do not show a clear separation in alignment between micro- and macro-norms. Rule 2, the subreddit's only macro-norm, consistently somewhere between the subreddit's micro-norms (Rules 1, 3, and 4) on all measures of alignment. Given the limited precision of our results, and our focus on a single community, we believe future quantitative work is needed to understand the source of variation in alignment.

Qualitatively, insights from the moderators may provide some explanation for why alignment varies by rules. Specifically, we raise the possibility that moderators defer to user opinion when the perceived consequence of rule violation is lower. In our pilot survey, one moderator noted that when enforcing Rule 4 they "prefer to let users award deltas for whatever reason they choose," while another stated that they "see little reason to remove [such comments] for R4 (and claw back the delta awarded)." This makes sense, given that deltas are merely cosmetic and provide no actual privileges. In contrast, after reviewing this study's results, a moderator noted that Rule 3 "was adopted precisely because so many users disagreed [...] that accusations of bad faith were inherently rude." Understanding why moderators choose to enforce controversial rules is an interesting direction for future work. It may also be worth investigating how variation in alignment alters a rule's impact on the community. Enforcing rules with low policy support, for example, might cause users to leave the subreddit. Rules with a higher degree of practice support, on the other hand, may have lower recidivism rates when violated.

6.5 Evaluating User-Mod Alignment on Other Subreddits

Given that our study was limited to a single community, it is natural to wonder how our findings would generalize to other communities. There are a few attributes of CMV to consider when answering this question. First, CMV is relatively large compared to other subreddits. Qualitative work suggests that smaller communities may self-regulate more effectively, without needing to rely upon top-down moderation [19]. Such communities might rely on implicit norms shared by the community rather than explicitly codified ones [3]. Our study, which evaluates alignment around explicit norms, might not capture meaningful information in smaller communities.

Second, CMV is an established community, with a stable set of explicit norms. It's unclear what alignment might look like in a community whose norms are still changing. Qualitative work raises the possibility that in growing communities, norms transition from implicit to explicit precisely when they become controversial amongst the community [39]. If this were the case, we would expect most rules to suffer from poor practice-support. Understanding how subreddit rules are formed and how alignment stabilizes over time is a fruitful area for future research.

Despite these calls for future work, we recognize that the problems we encountered recruiting subreddits make our study difficult to repeat. Instead, we argue that a more fruitful solution may be to build tools that allow communities to audit their own moderation practices along the dimensions surfaced in this study. Such tools could follow a community-driven experimentation model similar to CivilServant [33]. Even without the practical difficulties of recruitment, we think this is a productive idea on its own merits. Social media companies with centralized moderation systems regularly conduct policy research to understand and improve their practices. Yet, community moderators are expected to do the same work of setting and enforcing rules, without the ability to hire researchers to identify the best policies. By lowering the barrier to experimentation and opinion polling, we can empower online communities to adjust their practices over time to suit the needs of their members.

7 ADDITIONAL LIMITATIONS AND FUTURE WORK

Although our results provide nuanced insights into moderation on CMV, we believe there is ample room for future work to improve on our shortcomings. We detail some of these directions below.

First, although in this paper we have focused on measuring user-moderator alignment, we also acknowledge that alignment covers only one amongst many aspects of community health. We do not make the argument that optimizing for high alignment is always the best course of action. Other values, like a commitment to ideological diversity, or a desire to protect the interest of minority groups may take precedence.

We also acknowledge that evaluating whether a comment ought to be removed could require more serious thought than is reasonable to expect in a survey environment. As such, participants may have responded with a gut reaction that does not accurately re ect their true beliefs. Evaluating the consistency and malleability of moderation preferences expressed in the survey context may better contextualize our ndings.

While we made reasonable e orts to adjust for response bias, we emphasize that the assumptions used in this analysis are untestable with the data at hand. Crucially, our analysis assumes that users who did and did not respond to our survey are comparable, conditional on the predictor variables used [10]. We believe our choice of predictors captures key factors that in uence whether users will respond to our survey and how they will respond. Still, we highlight two potential categories of omitted variables that could impact our results. First, we acknowledge that personality factors, such as social trust, may confound our analysis. Future work could try to use a respondent's participation in other subreddits as a proxy for these variables (e.g. a user active on r/conspiracy may have lower social trust). We caution that informed consent should be sought before performing such analysis to avoid violating users' privacy. Second, we note that it is possible that the decision to participate in our survey was causally in uenced by variables we were interested in measuring (e.g. a user's support for subreddit policy may have caused them to respond). Selection on an outcome variable is a notoriously di cult problem to deal with, and would remain a challenge for future surveys following this methodology [34].

Finally, we note that our survey sample is constrained to users who comment on CMV. This limitation is di cult to work around, as the Reddit API provides no way to identify users who do not actively participate within a community. Still, if disagreements with moderators dissuade users from commenting, we may overestimate the degree of alignment within the community.

8 CONCLUSION

In this paper, we present a systematic approach to evaluate the degree and nature of misalignments between users and moderators in an online community. We conduct one such evaluation on r/ChangeMyView, nding high levels of user-moderator alignment at the level of policy, but lower levels of alignment when these policies are put into practice. Surprisingly, we also surface that users may not always be aware of these misalignments users in our survey signi cantly underestimated how often comments were actually removed.

The boundaries of acceptable speech within online communities will always be contested. An unfortunate consequence of the community-based moderation model is the degree to which volunteer moderators are expected to manage and resolve these con icts. And unlike other measures of community health, user-moderator alignment is not easily observable, making it harder for online communities to have open conversation about their moderation practices. We argue that empowering online communities to conduct similar audits themselves may be one way to reduce this burden, and promote healthier discussion online.

REFERENCES

[1] Shubham Atreja, Libby Hemphill, and Paul Resnick. 2022. What is the Will of the People? Moderation Preferences for Misinformation. arXiv preprint arXiv:2202.00729 (2022).

[2] Amy Bruckman. 2006. Teaching students to study online communities ethically. Journal of Information Ethics 15, 2 (2006), 82.

[3] Gary Burnett and Laurie Bonnici. 2003. Beyond the FAQ: Explicit and implicit norms in Usenet newsgroups. Library & Information Science Research 25, 3 (2003), 333 351.

[4] 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 (nov 2019), 30 pages. https://doi.org/10.1145/3359276

[5] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cli 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), 1 25.
[6] Bryan Dosono and Bryan Semaan. 2019. Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (Glasgow, Scotland Uk) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1 13. https://doi.org/10.1145/3290605.3300372
[7] Jenny Fan and Amy X Zhang. 2020. Digital juries: A civics-oriented approach to platform governance. Proceedings of the 2020 CHI conference on human factors in computing systems 1 14.
[8] Casey Fiesler, Joshua McCann, Kyle Frye, Jed R Brubaker, 2018. Reddit rules! characterizing an ecosystem of governance. In Twelfth International AAAI Conference on Web and Social Media
[9] Andrew Gelman and John Carlin. 2014. Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. Perspectives on Psychological Science 9, 6 (2014), 641 651.
[10] Andrew Gelman and Thomas C Little. 1997. Poststrati cation into many categories using hierarchical logistic regression. (1997).
[11] Andrew Gelman, Xiao-Li Meng, and Hal Stern. 1996. Posterior predictive assessment of model tness via realized discrepancies. Statistica sinica (1996), 733 760.
[12] Sarah A Gilbert. 2020. " I run the world's largest historical outreach project and it's on a cesspool of a website." Moderating a Public Scholarship Site on Reddit: A Case Study of r/AskHistorians. Proceedings of the ACM on Human-Computer Interaction 4, CSCW1 (2020), 1 27.
[13] Lex Gill, Dennis Redeker, and Urs Gasser. 2015. Towards digital constitutionalism? Mapping attempts to craft an internet bill of rights. Mapping Attempts to Craft an Internet Bill of Rights (November 9, 2015). Berkman Center Research Publication 2015-15 (2015).
[14] Tarleton Gillespie. 2018. Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.
[15] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Je Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In CHI Conference on Human Factors in Computing Systems 1 19.
[16] Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems 1 14.
[17] James Grimmelmann. 2015. The Virtues of Moderation. Yale Journal of Law and Technology 17 (2015), 2.
[18] Xinlan Emily Hu, Mark E Whiting, and Michael S Bernstein. 2021. Can Online Juries Make Consistent, Repeatable Decisions?. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems 1 16.
[19] Sohyeon Hwang and Jeremy D Foote. 2021. Why do people participate in small online communities? Proceedings of the ACM on Human-Computer Interaction 5, CSCW2 (2021), 1 25.
[20] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. " Did you suspect the post would be removed?" Understanding user reactions to content removals on Reddit. Proceedings of the ACM on human-computer interaction 3, CSCW (2019), 1 33.
[21] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit. Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 150 (nov 2019), 27 pages. https://doi.org/10.1145/3359252
[22] Shagun Jhaver, Seth Frey, and Amy Zhang. 2021. Designing for Multiple Centers of Power: A Taxonomy of Multi-level Governance in Online Social Platforms. arXiv preprint arXiv:2108.12552 (2021).
[23] Jialun Aaron Jiang, Peipei Nie, Jed R Brubaker, and Casey Fiesler. 2022. A Trade-o -centered Framework of Content Moderation. arXiv preprint arXiv:2206.03450 (2022).
[24] Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R. Brubaker. 2021. Understanding international perceptions of the severity of harmful content online. PLoS ONE 16 (2021).
[25] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. 2020. Through the looking glass: Study of transparency in Reddit's moderation practices. Proceedings of the ACM on Human-Computer Interaction 4, GROUP (2020), 1 35.
[26] Matthew Kay, Gregory L Nelson, and Eric B Hekler. 2016. Researcher-centered design of statistics: Why Bayesian statistics better t the culture and incentives of HCI. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems 4521 4532.
[27] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. Building successful online communities: Evidence-based social design 1 (2012), 4 2.

[28] Cli 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 (Vienna, Austria) (CHI '04). Association for Computing Machinery, New York, NY, USA, 543 550. https://doi.org/10.1145/985692.985761
[29] Hanlin Li, Brent Hecht, and Stevie Chancellor. 2022. Measuring the Monetary Value of Online Volunteer Work. In Proceedings of the International AAAI Conference on Web and Social Media, Vol. 16. 596 606.
[30] Rensis Likert. 1932. A technique for the measurement of attitudes. Archives of psychology (1932).
[31] J. Nathan Matias. 2019. The Civic Labor of Volunteer Moderators Online. Social Media + Society 5, 2 (2019), 2056305119836778. https://doi.org/10.1177/2056305119836778 arXiv:https://doi.org/10.1177/2056305119836778
[32] J Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. Proceedings of the National Academy of Sciences 116, 20 (2019), 9785 9789.
[33] J. Nathan Matias and Merry Mou. 2018. CivilServant: Community-Led Experiments in Platform Governance. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 1 13. https://doi.org/10.1145/3173574.3173583
[34] Richard McElreath. 2020. Statistical rethinking: A Bayesian course with examples in R and Stan. Chapman and Hall/CRC.
[35] Christina A Pan, Sahil Yakhmi, Tara P Iyer, Evan Strasnick, Amy X Zhang, and Michael S Bernstein. 2022. Comparing the Perceived Legitimacy of Content Moderation Processes: Contractors, Algorithms, Expert Panels, and Digital Juries. Proceedings of the ACM on Human-Computer Interaction 6, CSCW1 (2022), 1 31.
[36] Du Phan, Neeraj Pradhan, and Martin Jankowiak. 2019. Composable E ects for Flexible and Accelerated Probabilistic Programming in NumPyro. arXiv preprint arXiv:1912.11554 (2019).
[37] Paul Resnick, Aljohara Alfayez, Jane Im, and Eric Gilbert. 2021. Informed crowds can e ectively identify misinformation. arXiv preprint arXiv:2108.07898 (2021).
[38] Joseph Seering. 2020. Reconsidering community self-moderation: the role of research in supporting community-based models for online content moderation. Proceedings of the ACM on Human-Computer Interaction 4, CSCW2 (2020), 107.
[39] Joseph Seering, Tony Wang, Jina Yoon, and Geo Kaufman. 2019. Moderator engagement and community development in the age of algorithms. New Media & Society 21, 7 (2019), 1417 1443. https://doi.org/10.1177/1461444818821316 arXiv:https://doi.org/10.1177/1461444818821316
[40] Nicolas Suzor, Tess Van Geelen, and Sarah Myers West. 2018. Evaluating the legitimacy of platform governance: A review of research and a shared research agenda. International Communication Gazette 80, 4 (2018), 385 400. https://doi.org/10.1177/1748048518757142
[41] Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. Forecasting elections with non-representative polls. International Journal of Forecasting 31, 3 (2015), 980 991.
[42] Galen Weld, Amy X Zhang, and Tim Altho . 2021. Making Online Communities' Better': A Taxonomy of Community Values on Reddit. arXiv preprint arXiv:2109.05162 (2021).
[43] Donghee Yvette Wohn. 2019. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (Glasgow, Scotland Uk) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1 13. https://doi.org/10.1145/3290605.3300390
[44] Amy X Zhang, Grant Hugh, and Michael S Bernstein. 2020. PolicyKit: building governance in online communities. In Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology. 365 378.

A ADDITIONAL FIGURES, ANALYSES, AND MODEL DESCRIPTIONS

A.1 MRP for Rule Recognition

In this and following Model-speci c subsections, we will be discussing the details of our Bayesian hierarchical models. Our motivation for using such models can be found in Section 4. We assume some familiarity with statistical modeling and Bayesian statistics, but refer the reader to [34] for a more detailed reference of these topics.

A.1.1 Model Description.

To adjust for response bias, we need to predict whether or not a user will select a rule based on the adjustment variables. We de ne $\sim_{A,D}$ to be equal to one if user D picks rule A and zero otherwise. We use the following variables to model $\sim_{A,D}$:

0_D: the natural log of the number of comments posted anywhere on Reddit in the last month by user D
1_D: the natural log of the number of comments posted on r/ChangeMyView in the last month by user D
3_D: the natural log of the number of comments removed from Reddit in the last month by user D
4_D: the natural log of the number of comments removed from r/ChangeMyView in the last month by user D
5_D: the the age of the account of user D
$<_D$: whether user D is the moderator of any subreddits

We then model $\sim_{A,D}$ as follows:

$$\sim_{A,D} \sim \text{Bernoulli}(\text{Logistic}(Z_A + U_A 0_D + V_A 1_D + X_A 3_D + n_A 4_D + a_A 5_D + h_{A,<_D}[<_D])) \quad (2)$$

Note that because $<_D$ is a binary variable, we use varying intercepts h_A rather than a slope. Thus h_A can be treated as a vector of two values that $<_D$ indexes into. We set priors on the slopes/intercepts as follows

$$\begin{aligned} Z_A &\sim \text{Normal}(Z, \sigma_Z) \\ U_A &\sim \text{Normal}(U, \sigma_U) \\ V_A &\sim \text{Normal}(V, \sigma_V) \\ X_A &\sim \text{Normal}(X, \sigma_X) \\ n_A &\sim \text{Normal}(n, \sigma_n) \\ a_A &\sim \text{Normal}(a, \sigma_a) \\ h_{A,0} &\sim \text{Normal}(h_0, \sigma_h) \\ h_{A,1} &\sim \text{Normal}(h_1, \sigma_h) \end{aligned} \quad (3)$$

Hyperpriors are set such the means for all predictor variables are drawn from $\text{Normal}(0, 1)$.

Recall that some users chose to self report information, rather than allow us to directly measure it from their pro le. In these cases, we observe intervals that contain the true value of the predictor variable (e.g. "1-5 years" or "10+ years"). Taking age as an example, let $\hat{5}_D$ correspond to a vector containing the endpoints of the interval.
If $\hat{5}_D$ contains two endpoints, we say:

$$5_D \quad \text{Uniform}^1 5_D^{\wedge} \cdot 0^1\!/\!4 5_D^{\wedge} \cdot 1^1\!/\!4^o \tag{4}$$

If $5_D^{\wedge}$ contains only a single endpoint, we say:

$$5_D \quad \text{Exponential}^1 1^o \, , \, 5_D^{\wedge} \cdot 0^1\!/\!4 \tag{5}$$

Since all variables are standardized rst (including the endpoint intervals), this corresponds to the assumption that on average, the true value is one standard deviation above the interval's lower bound.

A.1.2 Results.

Fig. 10. A comparison of the policy awareness results before and a er conducting MRP-analysis. In all cases, our observed results fell within a 95% confidence interval.

A.2 MRP For Rule Support

Because rule support data is ordinal (1 to 5 Likert scale) rather than binary, we modify eq. (1) as follows:

$$\sim_{A \bullet D} \quad \text{OrderedLogistic} i\check{Z}_A \, , \, U_A 0_D \, , \, V_A 1_D \, , \, X_A 3_D \, , \, n_A 4_D \, , \, a_A 5_D \, , \, h_A \cdot \prec_D 1\!/\!4 \_A^o \tag{6}$$

Where $\_A$ corresponds to learned cutpoints. Priors for the slopes and intercepts are set in the same way as in the previous section. Priors for $\_A$ are set as follows:

$$\begin{aligned} \_A &= \text{CuSum}^1 !\cdot 68 \text{BC} \cdot 82_A^{oo} \\ ?_A &\quad \text{Dirichlet}^1 \cdot 1 \bullet 1 \bullet 1 \bullet 1 \bullet 1\!/\!4^o \end{aligned} \tag{7}$$

A.2.1 Results.

Fig. 11. A comparison of average policy alignment scores before and a er conducting MRP-analysis. In all cases, our observed results fell within a 95% confidence interval.

A.3 Comment Rating Model

Below we include a full description of the comment rating model, without MRP. Note that because of its alternate construction, we treat the comments included in the mod survey as part of their own rule. This allows us to partially pool it with the other data when tting the model.

A.3.1 Variables.

$G_{A,D}$: whether the user D in our survey thinks the Ath comment reported for violating rule A should be removed
$\sim_{A,D}$: whether the user D in our survey thinks the Ath comment corresponding to rule A would be removed
$I_{A,D}$: whether the user D in our survey thinks one of the subreddit rules applies to the Ath comment corresponding to rule A
$<_A$: whether the Ath comment corresponding to rule A was removed in real life
$B_{A,E}$: whether mod E thinks that the Ath comment corresponding to rule A should be removed in our survey
j_D^0: bias term for user D when deciding whether a comment should be removed
k_D^0: bias term for user D when deciding whether a comment will be removed
l_D^0: bias term for user D when deciding whether a rule applies
$\backslash_E^0$: bias term for mod E when deciding whether a comment will be removed

$_A$: rule-speci c correlation matrix between latent continuous variables that generates $G_{A\bullet D}$ $\sim_{A\bullet D}$ $I_{A\bullet D}$ $<_A$ $B_{A\bullet E}$

A.3.2 Model.

$$\begin{aligned}
G_{A\bullet D} &\sim \text{Bernoulli}(\text{Logistic}(j_{A\bullet 8}, j_D^0)) \\
\sim_{A\bullet D} &\sim \text{Bernoulli}(\text{Logistic}(k_{A\bullet 8}, k_D^0)) \\
I_{A\bullet D} &\sim \text{Bernoulli}(\text{Logistic}(I_{A\bullet 8}, I_D^0)) \\
<_A &\sim \text{Bernoulli}(\text{Logistic}([_{A\bullet 8})) \\
B_{A\bullet E} &\sim \text{Bernoulli}(\text{Logistic}(\backslash_{A\bullet 8}, \backslash_E^0)) \\
j_{A\bullet 8}, k_{A\bullet 8}, I_{A\bullet 8}, [_{A\bullet 8}, \backslash_{A\bullet 8} &\sim \text{MvNormal}(`_A, \ _A)
\end{aligned} \tag{8}$$

$$_A = (f_{j\bullet A} \bullet f_{k\bullet A} \bullet f_{I\bullet A} \bullet f_{[\bullet A} \bullet f_{\backslash\bullet A}) \ (f_{j\bullet A} \bullet f_{k\bullet A} \bullet f_{I\bullet A} \bullet f_{[\bullet A} \bullet f_{\backslash\bullet A}) \ ^T \ ! \ ^2 \tag{9}$$

Model hyperparameters are de ned below:

$$\begin{aligned}
`_A &\sim \text{Normal}(`_A, d_A) \\
f_{j\bullet A} \bullet f_{k\bullet A} \bullet f_{I\bullet A} &\sim \text{HalfNormal}(f_{DB4A}) \\
f_{[\bullet A} \bullet f_{\backslash\bullet A} &\sim \text{HalfNormal}(f_{<>3}) \\
j_D^0 &\sim \text{Normal}(0, g_j) \\
k_D^0 &\sim \text{Normal}(0, g_k) \\
I_D^0 &\sim \text{Normal}(0, g_I) \\
\backslash_E^0 &\sim \text{Normal}(0, g_\backslash)
\end{aligned} \tag{10}$$

And model hyper-hyperparameters are de ned as:

$$\begin{aligned}
`_A &\sim \text{Normal}(0, 1) \\
d_A &\sim \text{HalfNormal}(1) \\
f_{DB4A} &\sim \text{HalfNormal}(1) \\
f_{<>3} &\sim \text{HalfNormal}(1) \\
g_j \bullet g_k \bullet g_I \bullet g_\backslash &\sim \text{HalfNormal}(1)
\end{aligned} \tag{11}$$

All variables are standardized before being used to make predictions

A.3.3 MRP.

To incorporate MRP into this model, we simply modify eq. (8) as follows:

$$\begin{aligned}
G_{A\bullet D} &\sim \text{Bernoulli}(\text{Logistic}(j_{A\bullet 8}, j_D^0, U_{G\bullet A0D}, V_{G\bullet A1D}, \\
&\quad X_{G\bullet A3D}, n_{G\bullet A4D}, a_{G\bullet A5D}, h_{G\bullet A<D}^{\infty})) \\
\sim_{A\bullet D} &\sim \text{Bernoulli}(\text{Logistic}(k_{A\bullet 8}, k_D^0, U_{\sim\bullet A0D}, V_{\sim\bullet A1D}, \\
&\quad X_{\sim\bullet A3D}, n_{\sim\bullet A4D}, a_{\sim\bullet A5D}, h_{\sim\bullet A<D}^{\infty})) \\
I_{A\bullet D} &\sim \text{Bernoulli}(\text{Logistic}(I_{A\bullet 8}, I_D^0, U_{I\bullet A0D}, V_{I\bullet A1D}, \\
&\quad X_{I\bullet A3D}, n_{I\bullet A4D}, a_{I\bullet A5D}, h_{I\bullet A<D}^{\infty}))
\end{aligned} \tag{12}$$

Interval values for the predictors are handled the same way as in the previous models, and priors for the adjustment slopes and intercepts are set similarly as well.

A.3.4 Results.

(a) Adjusted probability of a positive label by rule (b) Adjusted correlation with mod decisions by rule

Fig. 12. Figure 12a compares the adjusted average probability of observing a positive label in each rating context by rule. Results are very similar to the unadjusted ones. Figure 12b contains adjusted estimated correlation values. These correlations are largely unchanged.

A.4 Assessing Practice Support/Awareness Model Fit

We leverage the generative nature of our model to assess model t via a posterior predictive check. In a posterior predictive check, we t the model to the collected data, and then simulate re-running the experiment multiple times. We then plot the number of user removal labels against the moderator decisions made across the simulated comments (see Figure Figure 13). We then compare the outcome of these simulated experiments to the actual data collected. A model that is well t should reproduce key trends in the real data. We opt to use a posterior predictive check over more traditional measures of model t (e.g. R^2 coe cient, cross-validation), because the former is more interpretable and can easily highlight which aspects of the data the model may not be learning. [11]

It should be noted that such posterior predictive checks can serve as an indication of whether our model is under tting, but not necessarily whether our model is over tting. Figures 13 to 15 contain the results of these simulations. One can observe that our model's posterior reproduces trends in the dataset, suggesting that it has appropriately captured both the rate of positive labels in di erent contexts, and their correlation.

Fig. 13. Comparison of the simulated number of comments that receive di erent combinations of practice awareness/moderator labels against the actual number observed. Colored bars denote simulated values, whereas transparent bars denote observed. Error bars correspond to 94% credible intervals around simulated predictions

Fig. 14. Comparison of the simulated number of comments that receive di erent combinations of practice support/moderator labels against the actual number observed.

Fig. 15. Comparison of the simulated number of comments that receive di erent combinations of rule application/moderator labels against the actual number observed.