What Makes Online Communities 'Better'? Measuring Values, Consensus, and Conflict across Thousands of Subreddits

Galen Weld, Amy X. Zhang, Tim Althoff

Paul G. Allen School of Computer Science & Engineering, University of Washington
{gweld, axz, althoff}@cs.washington.edu

Abstract

Making online social communities 'better' is a challenging undertaking, as online communities are extraordinarily varied in their size, topical focus, and governance. As such, what is valued by one community may not be valued by another. However, community values are challenging to measure as they are rarely explicitly stated. In this work, we measure community values through the first large-scale survey of community values, including 2,769 reddit users in 2,151 unique subreddits. Through a combination of survey responses and a quantitative analysis of public reddit data, we characterize how these values vary within and across communities.

Amongst other findings, we show that community members disagree about how safe their communities are, that long-standing communities place 30.1% more importance on trustworthiness than newer communities, and that community moderators want their communities to be 56.7% less democratic than non-moderator community members. These findings have important implications, including suggesting that care must be taken to protect vulnerable community members, and that participatory governance strategies may be difficult to implement. Accurate and scalable modeling of community values enables research and governance which is tuned to each community's different values. To this end, we demonstrate that a small number of automatically quantifiable features capture a significant yet limited amount of the variation in values between communities with a ROC AUC of 0.667 on a binary classification task. However, substantial variation remains, and modeling community values remains an important topic for future work. We make our models and data public to inform community design and governance.

1 Introduction

Online social communities are extraordinarily varied and capture almost every aspect of our society. Every day, people use millions of online communities to get the news (Weld, Glenski, and Althoff 2021; Geiger 2019; Volkova et al. 2017), for support (Sharma et al. 2020a,b; Wadden et al. 2021), for entertainment (Ling et al. 2021; Centivany and Glushko 2016), to discuss with others (Tan et al. 2016; Zhang et al. 2018; Chang, Cheng, and Danescu-Niculescu-Mizil 2020), to compete with others (Shameli et al. 2017;

Althoff, Jindal, and Leskovec 2017), and many other purposes. Some aspects of communities have been tied to specific societal harms, including distribution of misinformation (Jahanbakhsh et al. 2021; Tran et al. 2020; Anagnostopoulos et al. 2014; Zollo and Quattrociocchi 2018), harassment and bullying (Lenhart et al. 2016; Jhaver et al. 2018; Matias 2019b; Pater et al. 2016; Burke Winkelman et al. 2015; Ybarra and Mitchell 2008; van Laer 2014; Bretschneider, Wöhner, and Peters 2014; Matias, Simko, and Reddan 2020), and increasing polarization (Bessi and Ferrara 2016; Shao et al. 2018; Howard et al. 2018; Grinberg et al. 2019; Bossetta 2018; Bovet and Makse 2019).

However, there are many important aspects of community health beyond these harms, such as the quality of content and the diversity of the community. Given the immense diversity of online communities, it follows that there is no 'one size fits all' approach to making communities 'better' (Weld, Zhang, and Althoff 2021). What is strongly valued by members of one community may not be valued by another, and furthermore, members within a community may disagree with one another about what values are most important.

It is challenging to measure community values across many communities, as they are infrequently formalized or explicitly enumerated. Some work has attempted to study values implicitly by examining communities' rules (Fiesler et al. 2018) or removed content (Chandrasekharan et al. 2018); however, these approaches only capture values as implemented by moderators (Matias 2019a), and are unable to measure the degree to which community members disagree.

In this work, we contribute the first large-scale survey of community members' values to date. Specifically, we survey, analyze, and model community values on reddit. Using a taxonomy of nine different values (§3.1) previously developed from qualitative user studies (Weld, Zhang, and Althoff 2021), we ask community members about (1) which of these values are most and least important to their community, (2) the current state of each value in the community, and (3) how they would like the community to change with regards to each value (§3.3). We recruit survey respondents from a diverse set of reddit users, ranging from very new reddit users to moderators with 10 years of experience. 2,769 members of 2,151 different subreddits completed our survey, making this survey an order of magnitude larger than previous small-scale surveys (Weld, Zhang, and Althoff 2021).

With our participants' consent, we gather their reddit post and comment history, along with metadata and six months of content from each subreddit in our dataset (§3.4). Using these data, we answer four research questions:

RQ1 What are communities' values, and how do they vary across communities? (§4)

RQ2 Within communities, where is there disagreement over values? (§5)

RQ3 How do moderators differ in their values from non-moderator community members? (§6)

RQ4 To what degree can community values be predicted based on automatically measurable features? (§7)

We find that there is substantial variation in values both within and across communities, especially with regards to safety, for which there is 47.4% more disagreement within communities than other values. We leverage theories of group bonds from sociology (Prentice, Miller, and Lightdale 1994; Grabowicz et al. 2013; Ren, Kraut, and Kiesler 2007) that suggest that communities built around interpersonal connection place greater emphasis on safety, engagement, and inclusion than communities built around shared interests. We find communities for specific groups of people (*e.g.,*/r/teenagers) place 1.21 points (out of 8) more importance on inclusion than communities for the sharing of pictures and video (§4). We examine differences between newcomers and senior community members in the context of literature on the challenges of managing community growth and new members (Lin et al. 2017; Danescu-Niculescu-Mizil et al. 2013) and find that, on reddit, new members are more positive in their perception of the current states of their communities than more senior members (§5). Given that governance on reddit is often characterized by divisions between moderators and non-moderators (Matias 2019a), we measure differences in values between moderators and non-moderators. We find that moderators perceive their communities as 14.7% less democratic, think they should be 56.7% less democratic, and that democracy is 23.6% less important, relative to the average non-moderator in each community. This has important implications for the implementation of participatory governance practices in online communities (Zhang, Hugh, and Bernstein 2020) (§6).

Given the large amount of variation between communities, we suggest that researchers and community leaders consider the specific values and needs of each community when making decisions about how to change those communities. As measuring community values with survey responses is time-consuming and expensive, the ability to accurately model community values with automatically quantifiable features would be of great value. Through a binary classification task which seeks to differentiate between above- and below-average communities, we show that such features are able to predict a substantial amount of the variation between communities' values with a ROC AUC of 0.667 (§7). However, much variation remains challenging to predict, and additional research is needed on modeling and measuring community values. We make our models and anonymized responses public to support further research[1].

2   Related Work

Content Moderation, Rules, and Norms. A community's formal rules can offer insight into that community's values. On reddit, rules have been studied by Fiesler et al. (2018), who produced a taxonomy of 24 different types of rules in use across 1,000 subreddits. These rules are enforced by volunteer moderators (Matias 2019a), and in some cases, content removed by moderators for violating the rules can be recovered and used to characterize community norms (Chandrasekharan et al. 2018). However, one significant drawback of these approaches is that rules are both set and enforced by moderators, in almost all cases without any input from non-moderator community members (Zhang, Hugh, and Bernstein 2020). As such, such analyses may fail to represent the interests of the non-moderator majority of the subreddit. Further evidence for this can be found in studies of user reactions to moderator actions, which find that there is often conflict and disagreement between moderators and non-moderators (Srinivasan et al. 2019; Jhaver et al. 2019). In contrast, our method of explicitly surveying both moderators and non-moderators enables us to directly measure the differences between these two groups (§6).

Community Governance. Nearly all social media communities (*e.g.,* Subreddits, Facebook Groups, Twitter) adopt a strictly hierarchical governance model, where each community is managed by a small group of privileged moderators (sometimes also called admins) who have the authority to set rules and enforce them (Zhang, Hugh, and Bernstein 2020; Matias 2019a). On reddit, while moderators are beholden to platform administrators (Jhaver, Frey, and Zhang 2021), they typically have wide latitude to set and enforce policies as they see fit, with no requirement for community input. Social media communities stand in contrast to many peer-production communities such as Wikipedia, which operates primarily on a consensus model (Halfaker et al. 2013), or StackExchange, which holds formal elections. While some systems to incorporate democracy into reddit have been developed (Zhang, Hugh, and Bernstein 2020), such systems have not been widely adopted, and moderators often face conflict and accusations of corruption (Matias 2019a). In this work, we ask community members about their perceptions of democracy, and examine how moderators' and non-mods' responses differ (§6).

Growing Pains and Internal Conflict. Community growth and differences between new and senior members are a frequent source of conflict within communities that have been studied on a range of platforms (Robert E Kraut et al. 2012; Halfaker, Kittur, and Riedl 2011). Research investigating this tension on reddit has taken either a high-level approach which relies on implicit signals of conflict such as linguistic change and negative sentiment (Danescu-Niculescu-Mizil et al. 2013; Lin et al. 2017), or a qualitative interview with a small number of participants, focusing only on a single subreddit (Kiene, Monroy-Hernández, and Hill 2016; Cho and

---

[1]https://behavioral-data.github.io/reddit_values_surveys_public/

Quality	Quality of the content
Variety	Variety in/of the content
Diversity	Diversity of the people
Trust	Trustworthiness of the people and information
Engagement	Members' engagement with one another
Inclusion	Members' inclusion and ability to contribute
Size	Size of the community
Democracy	Community input into moderator decisions
Safety	Absence of offensive or harassing behavior

Table 1: We leverage the taxonomy of widely-held values on reddit by Weld, Zhang, and Althoff (2021), which was developed through user studies and iterative categorization.

Wash 2021). In contrast, our approach allows us to include over 2,000 communities while still gathering granular information about values through explicit survey questions.

Community Bonds and Membership. We draw upon the Theory of Common Identity and Common Bonds (Prentice, Miller, and Lightdale 1994), which suggests that some communities form due to common identities (*i.e.,* shared interests) while others form due to common bonds (*i.e.,* social relationships). Some previous work has examined this theory in the context of online communities (Ren, Kraut, and Kiesler 2007; Grabowicz et al. 2013); in contrast, our work explicitly surveys community members on their values.

3 Methods

3.1 Measuring Community Values

Central to this work is the set of nine values around which we design our survey instrument (§3.3). The set of values we use is grounded in sociology literature on different dimensions of social relations (Deri et al. 2018; Bao et al. 2021; Choi et al. 2020) and drawn directly from the taxonomy developed by Weld, Zhang, and Althoff (2021) via iterative categorization of unstructured survey responses from redditors. The complete Weld, Zhang, and Althoff (2021) taxonomy consists of 29 different values in nine major categories. As it is impractical to ask about 29 different values, we use the nine major categories with minor modifications; we include both Variety of Content and Diversity of People, we break Offensive, Abusive, and Harassing Content or Behaviors into a separate category called Safety, and we drop the Technical Features category as items within this category are outside the scope of control of community members and moderators. As such, the nine values we consider in this work are listed in Table 1. For each of these values, we ask community members about three dimensions: (1) the overall importance of the value to their community, (2) their perception of the value's current state in their community, and (3) their desire to change their community with regards to the value.

3.2 Reddit Background

Reddit is the fifth most popular social media site in the United States (Statista 2021), and is an ideal platform for researching the values of online communities as reddit is explicitly divided into many thousands of discrete communities, known as 'subreddits.' Each subreddit has its own topic, moderators, rules, and community norms. Within a subreddit, a user may post a link to another website (a linkpost), some text (a selfpost), or may comment on an existing post. Almost all content on reddit is publicly available (Baumgartner et al. 2020), and reddit has been widely studied (Medvedev, Lambiotte, and Delvenne 2019).

3.3 Data Collection: Online Survey

Responses were gathered through an online survey hosted on the Qualtrics platform. We summarize the survey here, a complete copy is online[2]. The survey consists of three parts: (1) informed consent, (2) general reddit questions, and (3) subreddit-specific questions. Before any other questions are asked, the participant is shown a brief summary of the survey, study, and IRB information (§3.5) and asked for their consent. After this point, all questions are optional.

The general reddit questions ask the participant about their usage of the platform across all subreddits. First, the participant is optionally asked to provide their reddit username, which is used to query the reddit API for their post/comment history. Then, the participant is asked about how often and how much time they spend on reddit, how frequently they 'lurk' vs. posting or commenting, how often they browse content aggregated from multiple subreddits (*e.g.,* on their front page), and their mobile vs. PC usage of reddit. At the end of this section, the participant is asked to select up to three subreddits that they consider themselves a member of. For reddit users who choose to provide their username, subreddits from their recent post and comment history are automatically suggested.

The subreddit-specific section of the survey asks questions specific to the subreddits the participant listed themselves as a member of. For each subreddit, the participant is asked separately about nine different community values (§3.1). For each value, the participant is asked about their perception of the *current state* of the subreddit on an 11-point rating scale (*e.g.,* Safety: 'How much offensive or harassing behavior is there in /r/science?' with scale ends 'Lots of offensive behavior' and 'No offensive behavior') and their *desired change* for the subreddit on a 3-point rating scale (*e.g.,* Safety: 'Would you change the safety of /r/science?' with options of 'The community should be less focused on safety,' 'the focus on safety is about right,' or 'The community should be more focused on safety.') Last, the participant is asked to rank all nine values in order from most important to least important to their experience in the subreddit.

The survey was piloted with 13 participants from a variety of departments at two large American universities. All pilot participants reported no difficulties completing the survey.

Participant Recruiting and Incentives. Survey participants were recruited through multiple channels, including reddit advertisements, private messages, and distribution on /r/SampleSize, a subreddit for the recruiting of survey participants. Community moderators were additionally recruited via reddit moderator mail. Responses were collected from

---

[2]https://behavioral-data.github.io/reddit_values_surveys_public/

May-July 2021, with a total of 2,769 people participating. Additional details on recruiting, participation, and compensation are included in Appendix C.

Quantifying Community Values and Disagreement. We compare values at the subreddit level instead of the survey response level, in order to avoid biasing our findings towards particularly popular communities that may receive a larger number of responses. To measure the degree of (dis)agreement on values at the subreddit level, we compute the mean average deviation (MAD) from the subreddit mean by computing the mean difference between each response for a subreddit and the average response for that subreddit.

Ensuring Response Validity. Generally only a subset of community members will respond to our survey. To ensure reasonably representative results when extrapolating from survey responses, we exclude subreddits with fewer than 15 responses from our analyses. This threshold was selected through an empirical power analysis (Appendix B) leveraging subreddits with a high number of responses, which indicated that subreddit averages have stabilized at this number of responses.

3.4 Data Collection: User & Subreddit Information

To augment survey responses from participants, we additionally compute user and subreddit features from publicly available reddit data. We source data from two locations: for each participant in the survey, we download their entire public reddit history and metadata such as account age from the reddit API. To more comprehensively characterize entire subreddits, we extract the most recent six months (January-June 2021) of posts and comments for each subreddit that participants in the survey are members of, using the Pushshift reddit corpus (Baumgartner et al. 2020).

User features are computed from the participant's entire public reddit history, and include the age of their account, their total number of posts, linkposts, selfposts, comments, as well as the mean length (# of characters) of each of the previous, along with their ratio of posts:comments, ratio of selfposts:posts, the mean and cumulative scores (# upvotes-downvotes) of their posts and comments, and the mean number of comments received for each of their posts. Then, for each subreddit the user is a member of, we extract the same set of features while considering only content from that subreddit. Finally, we compute the fraction of a user's total posts across all of reddit that are in the subreddit(s) they indicated they were a member of. For example, if a person answers survey questions for /r/science, we compute their number of posts in all subreddits, their number of posts in /r/science only, and the fraction of their posts (in any subreddit) which are in /r/science.

Subreddit features are computed from the most recent six months of posts and comments (January-June 2021) in that subreddit. These features include the age of the subreddit, the number of posts, linkposts, selfposts, and comments, as well as the number of removed (by moderators) and deleted (by their author) posts and comments, and the number of distinct users and subscribers each subreddit has. We also compute the mean score of posts and comments in each subreddit, the number of posts/comments per distinct user, and the number of rules declared by the community moderators.

Categorizing Subreddit Topics. For the 122 largest communities, we additionally hand-label the community topic.[3] For more details on this taxonomy, see Appendix A. The six topic categories we use are: **Hobby** communities *e.g.,* /r/nba, /r/bicycling, **Discussion** communities *e.g.,* /r/AskReddit, /r/relationship_advice, **Media-sharing** communities *e.g.,* /r/pics, /r/CrappyDesign, **News** communities *e.g.,* /r/worldnews, /r/science, **Meme** communities *e.g.,* /r/dankmemes, /r/me_irl, and **Identity-based** communities *e.g.,* /r/india, /r/teenagers.

3.5 Ethical Considerations

We strongly believe that this work will have a positive broader impact by informing the design of online communities in a manner which is aligned with the values of their members. The most serious potential negative impact of this work is the potential for deanonymization of responses. We take this possibility seriously and have taken numerous steps to mitigate this risk. To ensure the anonymity of our participants, we do not publish their usernames nor any of their reddit usage data, and remove responses from subreddits whose names or small size could enable deanonymization of individual contributors to our public dataset. All participants were informed of the goals of the study and how we would use and share their data before consenting to participate. In a separate step of the survey, we collect specific additional consent to access users' public reddit histories and use them for research (Fiesler and Proferes 2018), which we do not publish. This study was approved by the University of Washington IRB under ID number STUDY00011457.

4 RQ1: What Are Communities' Values, and How Do They Vary across Communities?

Understanding what communities' values are in general, and how these values vary from community to community, are key questions with implications for community design that also provide context for further analyses in this paper. In this section, we begin by quantifying what values are most important to communities, the current state of these values, and the level of variability across communities. Then, we explore how these values vary across communities according to community topic, age, and size of community. Informed by Common Identity and Common Bond Theory (§2), we hypothesize that communities with relatively strong interpersonal relationships, such as Identity-based communities, smaller communities, and older communities, will place greater emphasis on values related to interaction with community members, such as Inclusion, Engagement, and Safety. On the other hand, we hypothesize that larger, younger, and more content-consumption focused communities based on shared interests will place greater emphasis on

---

[3]We additionally experimented with pre-computed subreddit embeddings (Kumar et al. 2018; Martin 2017; Waller and Anderson 2019), but these did not explain significant variation in values.
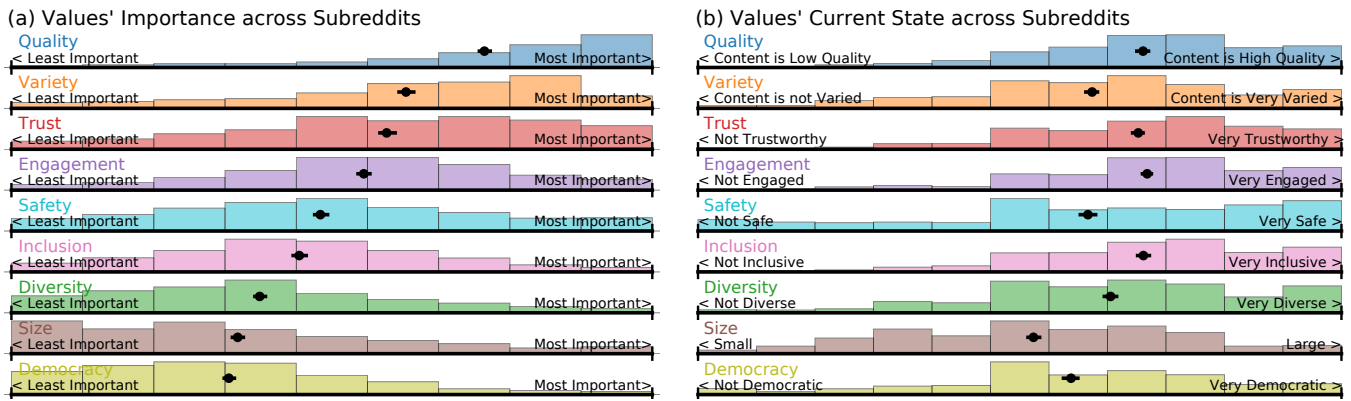
Figure 1: To understand what communities' values are, we average all responses for each community. (a) shows the distribution of the relative importance of each value across communities. Quality of Content is most frequently considered the most important value, while Size and Democracy are generally considered to be the least important. (b) shows the distribution of communities' perception of their current state. Black points indicate the average community. In this and all figures, bars indicate 95% bootstrapped confidence intervals.

Quality and Variety of Content and Size.

Method. To analyze how values vary across communities, we group communities based on their topical category (see Appendix A for details on categorization methodology) or by their quartile along a variable of interest, and then average across all communities in each group. When appropriate, we make minor adjustments from true quartile values to improve interpretability. We operationalize community age and size by the time since the community was founded and its number of unique contributors from the reddit API, respectively. We operationalize the degree to which the community is text-based by computing the fraction of text-posts (called selfposts on reddit).

Results. We find that there is substantial variation in both the importance and current state of values from community to community (Fig. 1). On average, Quality of Content is the most important value, with Size and Democracy generally considered the least important (Fig. 1a). Safety is especially varied with regards to both its importance (Fig. 1a) and current state (Fig. 1b), with a standard deviation 7.0% and 20.07% larger than those of any other values' importance and current state, respectively. While the average community rates Safety 5/9 in terms of importance, 171 communities have Safety as their most important value, and 176 have Safety as their least important value.

Our hypothesis that communities with strong interpersonal relationships will place greater emphasis on community-focused values such as Inclusion, Engagement, and Safety is largely upheld by our results. Identity-based Communities place greater than average importance on Diversity and Inclusion (Fig. 2c,d), while Hobby and News Communities place greater importance on Quality (Fig. 2b). Meme and Media-sharing Communities both place higher than average importance on Variety of Content and Size, which includes the amount of content submitted (Fig 2g,h). Identity-based Communities rate Inclusion as 1.21 points (out of 8) more important than Media
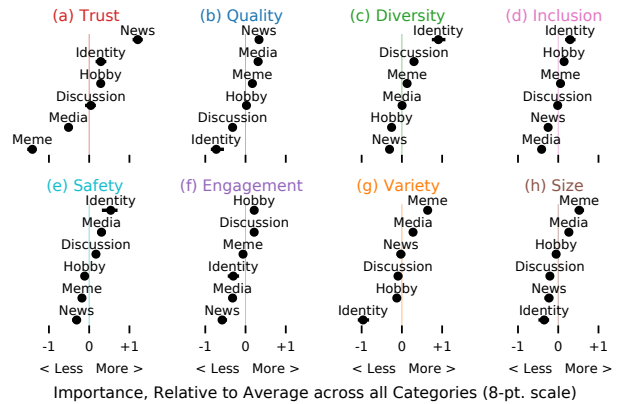

Figure 2: Differences in value importance across communities of different topics. News Communities rate Trust as 2.62 points more important than Meme Communities (out of an 8 point scale). Diversity and Inclusion are especially important to Identity-based Communities. Variety of Content and Size are especially important to Meme and Media-sharing Communities.

Communities (Fig. 2d). It is important to note that Common Identity and Common Bond Theory (Prentice, Miller, and Lightdale 1994) does not explain all observed differences between community categories. News and Meme Communities are both primarily motivated by shared interests, yet News Communities rate Trust as 2.62 (out of 8) points more important than Meme Communities (Fig. 2a).

When examining the differences between new and older communities, and between small and large communities, differences are especially pronounced for Trust, Size, and Safety (Fig. 3). The youngest quartile of communities (established within the past 8 years) have a 41.2% (0.35 vs 0.24) stronger desire to grow than older communities, while older communities have a 30.1% (0.27 vs. 0.20) stronger desire to improve Trust than younger communities (Fig. 3a), which is consistent with our hypothesis that older communi-
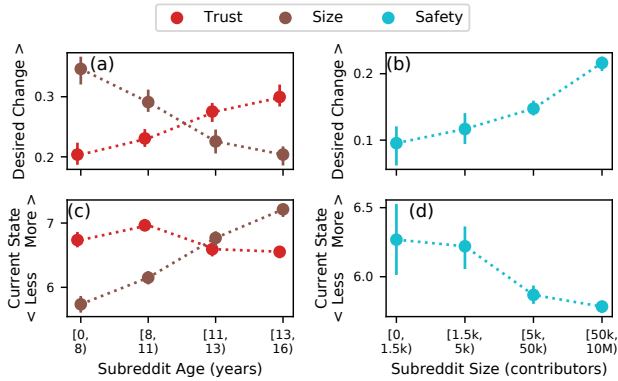
Figure 3: Average importance and desired change across community, binned into approximate quartiles by the age (since founding) and size. (a) Older communities 30.1% more strongly desire increased Trust than younger communities. (b) Larger communities have a 126.6% stronger desire to improve Safety than the smallest communities.

ties are more focused on common bonds than younger communities. Interestingly, this stronger desire to build Trust holds despite a lack of large difference in the perceived current state of Trust across older and younger communities (Fig. 3c). However, when examining community size, we find large communities with more than 50,000 contributors have a 126.6% (0.22 vs. 0.10) stronger desire to improve Safety than the smallest communities with less than 1,500 contributors (Fig. 3b), in contradiction of our hypothesis that smaller communities would value Safety more due to stronger interpersonal relations in smaller communities. Another potential explanation is that larger communities have poorer current Safety, as we find a 7.73% (6.27 vs. 5.78) decrease in perceived Safety amongst larger communities (Fig. 3d).

Implications. Different values are dramatically more or less important to different communities, which has profound implications, underlining that there is no 'one size fits all' approach to improving online communities (Weld, Zhang, and Althoff 2021). The relatively low importance placed upon Democracy may present challenges for the widespread adoption of systems that seek to implement participatory governance practices in online communities (Zhang, Hugh, and Bernstein 2020; Kelty 2017). We examine Democracy and governance further in §6. Our finding that some communities consider themselves fairly safe while others consider themselves to be very unsafe (Fig. 1) is consistent with previous findings that toxic behavior on reddit is extremely concentrated in a small number of subreddits (Weld, Glenski, and Althoff 2021), this could further support the practice of community level moderation interventions (Jhaver, Bruckman, and Gilbert 2019; Chandrasekharan et al. 2017; Habib et al. 2019; Shen and Rose 2019; Chandrasekharan et al. 2021). However, it is important to not only exclusively consider Safety by averaging over the values of all members of a community. Vulnerable minorities have an important perspective (Guinier 1994), yet inherently members of mi-

nority groups are too few to significantly influence the community average. We examine this further in §5. Finally, our results emphasize the importance of Common Identity and Common Bond Theory (§2), which can guide researchers in their future work on this topic.

5 RQ2: Within Communities, Where Is There Disagreement over Values?

Understanding where there is consensus on values, and where there is disagreement, is critical to building fair and equitable communities for everyone, including adequately protecting the needs and interests of vulnerable minority groups. Here, we begin by examining where there is the greatest disagreement on values (Fig. 4) before analyzing how different groups of reddit users disagree with others.

Informed by previous work on vulnerable members of online communities (Lenhart et al. 2016; Mahar, Zhang, and Karger 2018), we hypothesize that Safety will be especially disagreed over, as members who have personally felt unsafe online will perceive the current state of Safety as worse than others, and will rate Safety as more important and more urgent to change. We further hypothesize that newer and less popular community members will generally perceive their communities more negatively than older members, as previous work has found that incorporating newcomers into communities is a significant challenge (Kiene, Monroy-Hernández, and Hill 2016; Robert E Kraut et al. 2012).

Method. We measure disagreement by computing each response's difference from mean response for the corresponding community. We characterize overall disagreement by averaging across the absolute value of this deviation (MAD). We then further break down which types of community members tend to disagree in which direction by grouping users into approximate quartiles based on their seniority and popularity (with bin edges selected for interpretability), and computing the average deviation from the community mean amongst those groups. We operationalize members' seniority in the community by calculating the number of years since their account was created, and operationalize popularity as the sum of all upvotes received on their posts (called karma on reddit) (Glenski, Pennycuff, and Weninger 2017).

Results. We find that, in general, there is strongest consensus on the current state of the community (average MAD=0.17), with greater disagreement on the desired change (average MAD=0.20) and importance of different values (average MAD=0.22; all values adjusted for scale width to enable comparison). There is 13.3% (1.97 vs. 1.74) more disagreement over the importance of Safety than the importance of all other values, and 47.4% (2.67 vs. 1.81) more disagreement over the current state of Safety than all other values (Fig. 4b,c). Interestingly, there is relative consensus on the desire to improve Safety (Fig. 4c). There is strong consensus on the current state of Size (Fig. 4b), while there is relative disagreement over the importance and desired change of Size (Fig. 4a,c).
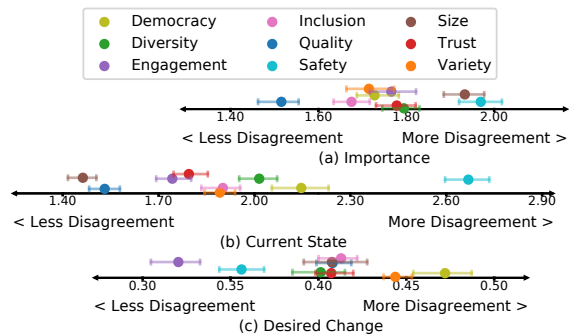
When examining differences between senior and junior

Perception of Inclusion Relative to Subreddit Mean

Figure 6: Differences in perceptions of Inclusion across less- and more-popular community members, as measured by account karma, divided into terciles. Relative to more popular users, less popular reddit users perceive Inclusion to be 0.36 points less important (a) and have 0.10 less desire to change Inclusion (b), yet perceive their communities to currently be 0.29 points more inclusive.

Figure 4: Average disagreement (measured with MAD) in perceptions of importance (a), current state (b) and desired change (c) across communities. Axes are adjusted for the widths of their respective scales, indicating greater disagreement over the importance of values than their current state and desired change. There is 13.3% and 47.4% more disagreement over the importance and current degree of Safety (light blue), respectively, relative to all other values, yet relative consensus on the desire to change Safety.
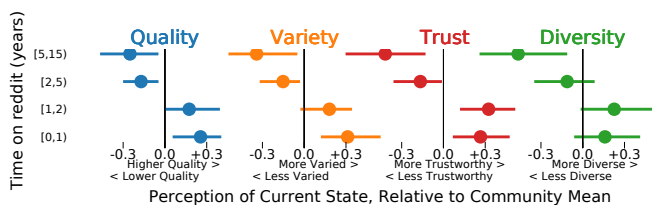

Figure 5: Differences in perception of the current state of communities between new reddit users and those who have been on reddit for longer. Generally, newer reddit users perceive their subs to be 0.55 points higher quality, 0.71 points more varied, 0.74 points more trustworthy, and 0.68 points more diverse compared to older reddit users.

community members, we demonstrate that junior redditors are generally more positive in their perception of the current state of their communities (Fig. 5), in contradiction of our hypothesis that new members' perceptions would be driven by the challenges of assimilation. Compared to the most senior community members (with at least 5 years of experience), those who joined reddit within the past year perceive their communities to be 0.55 points higher quality, 0.71 points more varied, 0.74 points more trustworthy, and 0.68 points more diverse. Note that the Current State scale is out of 11 points total, and thus the maximum possible MAD is half the scale width, *i.e.,* 5.5. However, the actual distribution of responses is more narrow (see Fig. 1).

We find significant differences in the perception of Inclusion between less- and more-popular community members (Fig. 6). These differences are especially stark between low-popularity users (in the bottom tercile of karma scores, with less than 67 karma), while differences between moderately and highly popular community members are statistically insignificant. Low-popularity community members place 0.36 points (out of 8) less importance on Inclusion, have 0.10 points (out of 2) less desire for Inclusion to change, and perceive the current state of Inclusion to be 0.29 points (out of 11) better than more popular users.

Implications. The disagreement over Safety (Fig. 4a,b) is a special concern that emphasizes the potential harm of community governance that only responds to the needs of the majority (Guinier 1994). While gathering data on past abuse is challenging as well as ethically fraught, it is distinctly probable that the community members most likely to feel that current community Safety is lacking and that Safety ought to be improved are those who have prior negative experiences that made them feel unsafe. While these community members may be a minority, is is critical to design communities that consider and protect their needs.

Our results also contradict our hypothesis that more senior and more popular users will have a more positive perception of their communities. Instead, we find evidence that it's actually the new reddit users who are most positive in their perception (Fig. 5), and correspondingly feel that their communities are the most inclusive (Fig. 6c). This is a noteworthy result that suggests that communities on reddit are generally effective in their practices to incorporate new members. However, as we only survey self-identified community
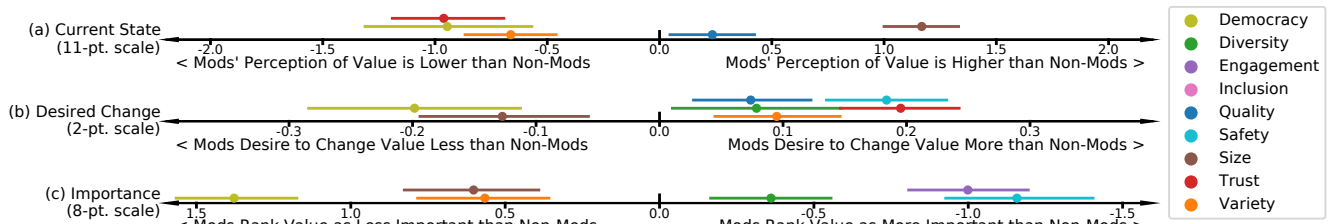
Figure 7: Differences in values between moderators and non-moderators. Moderators believe (a) their communities are 14.5% less democratic, (b) should be 56.7% less democratic, and (c) that Democracy is 23.6% less important, relative to the non-moderator mean in that community. Moderators rank Diversity, Engagement, and Safety as more important to their communities than non-moderator community members (c). Values with CIs overlapping 0 are removed.

members, additional work is needed to reach users who ultimately decided to *not* join a community.

6 RQ3: How do Moderators Differ in Their Values from Non-moderator Community Members?

Volunteer moderators are a key part of any community on reddit, as they bear the primary responsibility of setting rules and enforcing them, a task which frequently brings moderators into conflict with other reddit users (Matias 2019a; Seering et al. 2019; Seering 2020). Importantly, moderators also constitute a major part of the governance of communities on reddit (Jhaver, Frey, and Zhang 2021), making their perspective on Democracy especially important. As past work has shown both that much of moderators' interactions with community members are characterized by conflict (Matias 2019a) and that affordances for participatory governance are almost entirely absent from reddit (Zhang, Hugh, and Bernstein 2020), we hypothesize that moderators will have more negative perceptions of Democracy than non-moderators.

Method. We identify moderators within our survey responses by scraping users' reddit profile pages, which contain information on the communities each user moderates. We compute the differences between moderators and non-moderators by grouping responses by community, and, within each community, by taking the difference between all pairs of (mod, non-mod) responses. We compute test statistics and CIs from the resulting set of differences for analysis.

Results. We find substantial differences between moderators and non-moderators across all three dimensions of each value: current state, desired change, and importance (Fig. 7). Consistent with our hypothesis, we find that moderators believe their communities *are* 14.5% (5.57 vs. 6.51) less democratic, *should be* 56.7% (0.15 vs. 0.35) less democratic, and that Democracy is 23.6% (7.22 vs. 5.84[4]) *less important* than non-moderator members of the same community (relative to the non-moderator mean in that community). When examining all moderator responses (without adjusting for community mean), $2.15\times$ as many moderators report desiring their communities to be less democratic than non-moderators. These differences are not limited to Democ-

racy; moderators also more strongly desire to improve the Safety and trustworthiness of their communities than non-moderators (Fig. 7b), and rank Engagement and Safety as more important than non-moderators (Fig. 7c).

Implications. Moderators are directly able to control many aspects of Democracy in their subreddits (*e.g.,* by soliciting community feedback before implementing rule changes), and so their perspective on this value is of special interest. Native tools for enabling formalized community input into governance are lacking from almost all social media platforms, and while some research has attempted to develop such tools (Zhang, Hugh, and Bernstein 2020), under current governance paradigms, the adoption of such tools is completely limited by the desire of moderation teams to do so. Moderators also frequently feel overworked (Plackett 2018; Matias 2019a) and traumatized by exposure to offensive content (Solon 2017), which may contribute to our findings of a perception of larger community size and lower trustworthiness amongst moderators relative to non-moderators.

7 RQ4: To What Degree Can Community Values Be Predicted Based on Automatically Measurable Features?

Our survey responses (§3) contain more granular information about community members' values across a far greater set of communities than have been previously collected. However, survey responses are expensive and time consuming to collect and therefore require significant resources to scale. The ability to automatically and accurately predict the importance and desired change of values could be used to inform community design, rule changes, and the implementation of participatory governance practices, while the ability to automatically measure the current state of communities with regards to various values has numerous potential applications, including measuring the impact of interventions.

Throughout this paper, we have demonstrated that communities can vary significantly in their self-reported values, and highlighted general structure in this variation across several potentially generalizable factors. Here, we investigate how much variation the 14 factors discussed in §4-6, all of which can be automatically quantified from publicly available data, collectively capture.[5] A complete list of features

---
[4]For importance, lower rank values indicate higher importance.

[5]We further experimented with a much larger set of 74 features

Importance	Current State	Desired Change	Overall
0.660	0.673	0.666	**0.667**

Table 2: Quantile-preprocessed Logistic Regression results for the binary classification task on the test set, measured with ROC AUC. Best performance is achieved when predicting the importance of values. In all cases, the model exceeds the performance of a random baseline (0.5 ROC AUC).

used is given in AppendixTable 4.

Tasks. We formulate 27 (importance, current state, and desired change for each of the 9 values) binary classification tasks where the goal is to predict whether a given value is particularly important or unimportant, whether the current state is particularly high or low, and whether the desired change for that value is particularly high or low, for a given subreddit. Each task asks the model to distinguish between the top and bottom quartiles, as for most if not all values, a majority of communities differ only very slightly in their perception of the importance, current state, and desired change of the values. Particularly accurate prediction of small differences is less critical for understanding community values. As with previous analyses, we aggregate values for each subreddit by averaging across all responses received for that subreddit. To avoid extrapolating from a small number of data points, we filter out communities with fewer than three corresponding survey responses, resulting in 404 communities which are randomly divided with an 80/20 split to create a training and test set. Hyperparameters were chosen through cross-validation on the training data.

Models and Metrics. We report on an l_2-regularized Logistic Regression model with quantile preprocessing, a nonlinear quantile transformation which uses the distribution of the training set to spread the data evenly along each feature's axis in the feature space. Missing target values are dropped, and missing features are imputed with the training set mean. Categorical features are one-hot encoded. We also experimented with other models, including neural networks and support vector classifiers, as well as additional preprocessing schemes such as standardization and PCA. We report here on Logistic Regression as overall it performed the best.

Results. We find that a Logistic Regression with quantile preprocessing performs the best overall, with an ROC AUC of 0.667 averaged across all tasks. Performance is highest on current state, followed by desired change and then importance (Table 2). Furthermore, we find that performance is highly variable from value to value; the model is able to accurately predict users' perception of the current Size of the subreddit (ROC AUC 0.936) and the importance of Trust (ROC AUC 0.922), while performance at predicting the importance and current state of Safety is no better than baseline. This is partially due the presence of easily measured proxies for some values (*e.g.,* number of contributors is strongly correlated with perception of current Size), while others values, such as Safety, are more nuanced and chal-

lenging to automatically measure. A complete table of results for each value is given in Appendix Table 3.

Implications. These results demonstrate that there is significant structure in how values vary from community to community, and that this structure is predictable using a small number of automatically quantifiable features. Prediction tasks based upon these features could be used to scale research which is informed by the values of the communities it impacts. However, the overall ROC AUC of $0.667 \ll 1$ indicates that there is significant remaining structure that is not explained by these features, and further research into what, if any, factors may capture this remaining structure is needed. Features which examine text-based content within subreddits, and graph-based features computed using subreddit membership are two promising avenues for future experimentation. We make our dataset public[6] to support further research.

8 Discussion & Conclusion

Diversity of Communities. Our study reveals that the set of communities surveyed have remarkably diverse values (Figs. 1,2,3). This underlines that there is no global set of values common to all online communities; what is important to one may be unimportant or even detrimental to another. Researchers, community leaders, and platforms alike must consider the specific context of the community and its needs before implementing changes.

Protecting Vulnerable Minorities. Community members are especially divided on the importance and current state of Safety (Fig. 4a,b), with 47.4% more disagreement over the current state of Safety than any other value in our survey. Because in many cases community members who feel their communities are unsafe are in the minority, care must be taken to protect the interests of these vulnerable groups. Although additional research is needed on this important topic, some work has shown that even simple interventions such as automated welcome messages can help support minority groups (Matias, Simko, and Reddan 2020).

Participatory Governance. Volunteer moderators play an important role in community governance on reddit (Matias 2019a). On the other hand, both formal and informal opportunities for non-moderators to influence decision making in their communities are quite rare (Zhang, Hugh, and Bernstein 2020). We find that in general, while non-moderators desire to have more Democracy in their communities, moderators are 56.7% less in favor of increased Democracy (Fig. 7b). This discrepancy could pose a challenge to increasing participatory governance; more research is needed on why moderators are less approving of Democracy, and what changes are needed to mitigate this difference.

Limitations. Our research is carried out only on reddit; additional work is needed to understand how our findings generalize to other platforms such as Twitter and Facebook. While we made every effort to recruit a diverse set of participants by using multiple recruiting methods, our work is

---

and found they did not lead to significant performance increases.

[6]https://behavioral-data.github.io/reddit_values_surveys_public/

still subject to some potential bias from groups of people who were not included in our study. One source of this bias is the limitations in who we can target ads to, as reddit restricts advertising to members of communities focused on porn and other controversial topics. We also recognize that in our analyses, when we filter out communities with fewer responses, we're disproportionately excluding smaller communities, however this filtering step is necessary to reliably assess consensus and disagreement (§3.3).

Conclusion. Online communities are extraordinarily varied, and the importance they place on different values reflects this variety. As such, what is good for one community may be harmful to another. In this work, we surveyed 2,796 reddit users to characterize how their values vary within and across 2,151 different communities. By combining these survey responses with publicly available reddit data, we examined differences between communities focused different topics, and measured where there is consensus and disagreement over different values. We compared moderators' values to non-moderators' values, identifying challenges for the implementation of participatory governance online. We make our dataset public to support future research.

References

Althoff, T.; Jindal, P.; and Leskovec, J. 2017. Online Actions with Offline Impact: How Online Social Networks Influence Online and Offline User Behavior. In *WSDM*, 537–546.

Anagnostopoulos, A.; Bessi, A.; Caldarelli, G.; Vicario, M. D.; Petroni, F.; Scala, A.; Zollo, F.; and Quattrociocchi, W. 2014. Viral Misinformation: The Role of Homophily and Polarization. *WWW '15 Companion* .

Bao, J.; Wu, J.; Zhang, Y.; Chandrasekharan, E.; and Jurgens, D. 2021. Conversations Gone Alright: Quantifying and Predicting Prosocial Outcomes in Online Conversations URL http://arxiv.org/abs/2102.08368.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The Pushshift Reddit Dataset URL http://arxiv.org/abs/2001.08435.

Bessi, A.; and Ferrara, E. 2016. Social Bots Distort the 2016 US Presidential Election Online Discussion. Technical Report ID 2982233, Social Science Research Network.

Bossetta, M. 2018. The Digital Architectures of Social Media: Comparing Political Campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. Election:. *Journalism & Mass Communication Quarterly* .

Bovet, A.; and Makse, H. A. 2019. Influence of Fake News in Twitter During the 2016 US Presidential Election. *Nature Communications* 10(1): 7. ISSN 2041-1723.

Bretschneider, U.; Wöhner, T.; and Peters, R. 2014. Detecting Online Harassment in Social Networks. In *International Conference on Information Systems*.

Burke Winkelman, S.; Oomen-Early, J.; Walker, A.; Chu, L.; and Yick-Flanagan, A. 2015. Exploring Cyber Harassment among Women Who Use Social Media. *Universal Journal of Public Health* 194–201.

Centivany, A.; and Glushko, B. 2016. "Popcorn Tastes Good": Participatory Policymaking and Reddit's. In *CHI '16*, CHI '16. New York, NY, USA: ACM.

Chandrasekharan, E.; Jhaver, S.; Bruckman, A.; and Gilbert, E. 2021. Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit. *arXiv:2009.11483* .

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech 1(CSCW).

Chandrasekharan, E.; Samory, M.; Jhaver, S.; Charvat, H.; Bruckman, A.; Lampe, C.; Eisenstein, J.; and Gilbert, E. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *CSCW '18* 2: 1–25.

Chang, J. P.; Cheng, J.; and Danescu-Niculescu-Mizil, C. 2020. Don't Let Me be Misunderstood: Comparing Intentions and Perceptions in Online Discussions. In *WWW '20*, 2066–2077.

Cho, J.; and Wash, R. 2021. How Potential New Members Approach an Online Community. *Computer Supported Cooperative Work (CSCW)* 1–43.

Choi, M.; Aiello, L. M.; Varga, K. Z.; and Quercia, D. 2020. Ten Social Dimensions of Conversations and Relationships. In *WWW '20*, 1514–1525. New York, NY, USA.

Danescu-Niculescu-Mizil, C.; West, R.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities. In *WWW '13*.

Deri, S.; Rappaz, J.; Aiello, L. M.; and Quercia, D. 2018. Coloring in the Links: Capturing Social Ties as They are Perceived. *CSCW '18* 2: 43:1–43:18.

Fiesler, C.; Jiang, J. A.; McCann, J.; Frye, K.; and Brubaker, J. 2018. Reddit Rules! Characterizing an Ecosystem of Governance. In *ICWSM*.

Fiesler, C.; and Proferes, N. 2018. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society* 4.

Geiger, A. 2019. Key Findings About the Online News Landscape in America. https://www.pewresearch.org/fact-tank/2019/09/11/key-findings-about-the-online-news-landscape-in-america/. Accessed: 2022-04-19.

Glenski, M.; Pennycuff, C.; and Weninger, T. 2017. Consumers and Curators: Browsing and Voting Patterns on Reddit. *IEEE Transactions on Computational Social Systems* 4(4): 196–206.

Grabowicz, P. A.; Aiello, L. M.; Eguiluz, V. M.; and Jaimes, A. 2013. Distinguishing Topical and Social Groups Based on Common Identity and Bond Theory. In *WSDM '13*, 627–636.

Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; and Lazer, D. 2019. Fake News on Twitter During the 2016 U.S. Presidential Election. *Science* 363: 374–378.

Guinier, L. 1994. *The Tyranny of the Majority : Fundamental Fairness in Representative Democracy*. Free Press.

Habib, H.; Musa, M. B.; Zaffar, F.; and Nithyanand, R. 2019. To Act or React: Investigating Proactive Strategies For Online Community Moderation. *arXiv:1906.11932* .

Halfaker, A.; Geiger, R. S.; Morgan, J. T.; and Riedl, J. 2013. The Rise and Decline of an Open Collaboration System: How Wikipedia's Reaction to Popularity Is Causing Its Decline. *American Behavioral Scientist* 57(5): 664–688.

Halfaker, A.; Kittur, A.; and Riedl, J. 2011. Don't Bite the Newbies: How Reverts Affect the Quantity and Quality of Wikipedia Work. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, 163–172.

Howard, P. N.; Kollanyi, B.; Bradshaw, S.; and Neudert, L.-M. 2018. Social Media, News and Political Information during the US Election: Was Polarizing Content Concentrated in Swing States? *arXiv:1802.03573 [cs]* URL http://arxiv.org/abs/1802.03573.

Jahanbakhsh, F.; Zhang, A. X.; Berinsky, A. J.; Pennycook, G.; Rand, D. G.; and Karger, D. R. 2021. Exploring Lightweight Interventions at Posting Time to Reduce the Sharing of Misinformation on Social Media. *arXiv:2101.11824* .

Jhaver, S.; Appling, D. S.; Gilbert, E.; and Bruckman, A. 2019. "Did You Suspect the Post Would Be Removed?": Understanding User Reactions to Content Removals on Reddit. *Proc. ACM Hum.-Comput. Interact.* 3(CSCW).

Jhaver, S.; Bruckman, A.; and Gilbert, E. 2019. Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW): 150:1–150:27.

Jhaver, S.; Frey, S.; and Zhang, A. 2021. Designing for Multiple Centers of Power: A Taxonomy of Multi-level Governance in Online Social Platforms. *arXiv:2108.12529* .

Jhaver, S.; Ghoshal, S.; Bruckman, A.; and Gilbert, E. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput.-Hum. Interact.* 25(2). ISSN 1073-0516.

Kelty, C. 2017. Too Much Democracy in All the Wrong Places: Toward a Grammar of Participation. *Current Anthropology* 58.

Kiene, C.; Monroy-Hernández, A.; and Hill, B. M. 2016. Surviving an "Eternal September": How an Online Community Managed a Surge of Newcomers. *CHI '16* .

Kumar, S.; Hamilton, W. L.; Leskovec, J.; and Jurafsky, D. 2018. Community Interaction and Conflict on the Web. WWW '18, 933–943. Lyon, France.

Lenhart, A.; Ybarra, M.; Zickuhr, K.; and Price-Feeney, M. 2016. Online Harassment, Digital Abuse, and Cyberstalking in America. *Data & Society Research Institute* .

Lin, Z.; Salehi, N.; Yao, B.; Chen, Y.; and Bernstein, M. S. 2017. Better When It Was Smaller? Community Content and Behavior After Massive Growth. In *ICWSM*.

Ling, C.; AbuHilal, I.; Blackburn, J.; De Cristofaro, E.; Zannettou, S.; and Stringhini, G. 2021. Dissecting the Meme Magic: Understanding Indicators of Virality in Image Memes. *Proc. ACM Hum.-Comput. Interact.* 5(CSCW1).

Mahar, K.; Zhang, A. X.; and Karger, D. 2018. *Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation*, 1–13.

Martin, T. 2017. community2vec: Vector representations of online communities encode semantic relationships. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, 27–31. Vancouver, Canada: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W17-2904.

Matias, J. N. 2019a. The Civic Labor of Volunteer Moderators Online. *Social Media + Society* .

Matias, J. N. 2019b. Preventing Harassment and Increasing Group Participation through Social Norms in 2,190 Online Science Discussions. *PNAS* 116(20): 9785–9789.

Matias, J. N.; Simko, T.; and Reddan, M. 2020. Reducing the Silencing Role of Harassment in Online Feminism Discussions. https://citizensandtech.org/2020/06/reducing-harassment-impacts-in-feminism-online/. Accessed: 2020-07-22.

Medvedev, A. N.; Lambiotte, R.; and Delvenne, J.-C. 2019. The anatomy of Reddit: An overview of academic research. *arXiv:1810.10881* 183–204.

Pater, J. A.; Kim, M. K.; Mynatt, E. D.; and Fiesler, C. 2016. Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. GROUP '16, 369–374. New York, NY, USA.

Plackett, B. J. 2018. Unpaid and abused: Moderators speak out against Reddit. https://www.engadget.com/2018-08-31-reddit-moderators-speak-out.html. Accessed: 2019-07-08.

Prentice, D. A.; Miller, D. T.; and Lightdale, J. R. 1994. Asymmetries in Attachments to Groups and to their Members: Distinguishing between Common-Identity and Common-Bond Groups. *Personality and Social Psychology Bulletin* 20(5): 484–493.

Ren, Y.; Kraut, R.; and Kiesler, S. 2007. Applying Common Identity and Bond Theory to Design of Online Communities. *Organization Studies* 28(3): 377–408. ISSN 0170-8406.

Robert E Kraut; Paul Resnick; Sara Kiesler; Moira Burke; Yan Chen; Niki Kittur; Joseph Konstan; Yuqing Ren; and John Riedl. 2012. *Building Successful Online Communities: Evidence-Based Social Design*. The MIT Press.

Seering, J. 2020. Reconsidering Self-Moderation: The Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 4(CSCW2).

Seering, J.; Wang, T.; Yoon, J.; and Kaufman, G. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* 21(7): 1417–1443. ISSN 1461-4448.

Shameli, A.; Althoff, T.; Saberi, A.; and Leskovec, J. 2017. How gamification affects physical activity: Large-scale analysis of walking challenges in a mobile application. In *WWW '17*, 455–463.

Shao, C.; Hui, P.-M.; Wang, L.; Jiang, X.; Flammini, A.; Menczer, F.; and Ciampaglia, G. L. 2018. Anatomy of an online misinformation network. *PloS One* 13(4): e0196087.

Sharma, A.; Choudhury, M.; Althoff, T.; and Sharma, A. 2020a. Engagement Patterns of Peer-to-Peer Interactions on Mental Health Platforms. *ICWSM* 14: 614–625.

Sharma, A.; Miner, A. S.; Atkins, D. C.; and Althoff, T. 2020b. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *EMNLP*.

Shen, Q.; and Rose, C. 2019. The Discourse of Online Content Moderation: Investigating Polarized User Responses to Changes in Reddit's Quarantine Policy. In *Proceedings of the Third Workshop on Abusive Language Online*, 58–69. Florence, Italy: Association for Computational Linguistics.

Solon, O. 2017. Underpaid and Overburdened: the Life of a Facebook Moderator. *The Guardian* .

Srinivasan, K. B.; Danescu-Niculescu-Mizil, C.; Lee, L.; and Tan, C. 2019. Content Removal as a Moderation Strategy: Compliance and Other Outcomes in the ChangeMyView Community. *Proc. ACM Hum.-Comput. Interact.* 3(CSCW).

Statista. 2021. Most popular mobile social networking apps in the United States as of September 2019, by monthly users. https://www.statista.com/statistics/248074/most-popular-us-social-networking-apps-ranked-by-audience/. Accessed: 2022-04-07.

Tan, C.; Niculae, V.; Danescu-Niculescu-Mizil, C.; and Lee, L. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *WWW '16*, 613–624.

Tran, T.; Valecha, R.; Rad, P.; and Rao, H. 2020. An Investigation of Misinformation Harms Related to Social Media during Two Humanitarian Crises. *Information Systems Frontiers* 1 – 9.

van Laer, T. 2014. The Means to Justify the End: Combating Cyber Harassment in Social Media. *Journal of Business Ethics* 123: 85–98.

Volkova, S.; Shaffer, K.; Jang, J. Y.; and Hodas, N. 2017. Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter. In *ACL Short Papers*, 647–653. Vancouver, Canada.

Wadden, D.; August, T.; Li, Q.; and Althoff, T. 2021. The Effect of Moderation on Online Mental Health Conversations. In *ICWSM '21*, volume 15, 751–763.

Waller, I.; and Anderson, A. 2019. Generalists and Specialists: Using Community Embeddings to Quantify Activity Diversity in Online Platforms. In *The World Wide Web Conference*, WWW '19, 1954–1964. San Francisco, CA, USA: Association for Computing Machinery. ISBN 978-1-4503-6674-8. doi:10.1145/3308558.3313729.

Weld, G.; Glenski, M.; and Althoff, T. 2021. Political Bias and Factualness in News Sharing across more than 100,000 Online Communities. *ICWSM '21* 15(1): 796–807. URL https://ojs.aaai.org/index.php/ICWSM/article/view/18104.

Weld, G.; Zhang, A. X.; and Althoff, T. 2021. Making Online Communities 'Better': A Taxonomy of Community Values on Reddit. *arXiv:2109.05152* .

Ybarra, M. L.; and Mitchell, K. J. 2008. How Risky Are Social Networking Sites? A Comparison of Places Online Where Youth Sexual Solicitation and Harassment Occurs. *Pediatrics* 121(2).

Zhang, A. X.; Hugh, G.; and Bernstein, M. S. 2020. PolicyKit: Building Governance in Online Communities. UIST '20, 365–378. New York, NY, USA.

Zhang, J.; Danescu-Niculescu-Mizil, C.; Sauper, C.; and Taylor, S. J. 2018. Characterizing Online Public Discussions through Patterns of Participant Interactions. *CSCW '18* 2: 1–27.

Zollo, F.; and Quattrociocchi, W. 2018. Misinformation spreading on Facebook. *ArXiv* doi:10.1007/978-3-319-77332-2_10.

A Categorizing Subreddit Topics

To operationalize higher-level notions of community topic and focus,[7] we manually investigated each of the 122 subreddits for which we received responses from at least 10 different community members. Among the author team, we iteratively clustered these communities until there was agreement on 6 different and mutually exclusive categories. We were unable to come up with other categories that were relevant to a significant fraction of these communities.

- **Hobby** Communities for people interested in specific games and hobbies (53 communities, *e.g.,* /r/nba, /r/bicycling)
- **Discussion** Communities which focus on question answering and discussion (18 communities, *e.g.,* /r/AskReddit, /r/relationship_advice)
- **Media-sharing** Communities for posting pictures and video of different things (17 communities, *e.g.,* /r/pics, /r/CrappyDesign)
- **News** Communities which share news, research, and data (15 communities, *e.g.,* /r/worldnews, /r/science)
- **Meme** Communities which are primarily for memes and shitposting (11 communities, *e.g.,* /r/dankmemes, /r/me_irl)
- **Identity-based** Communities which are primarily for specific groups of people (8 communities, *e.g.,* /r/india, /r/teenagers)

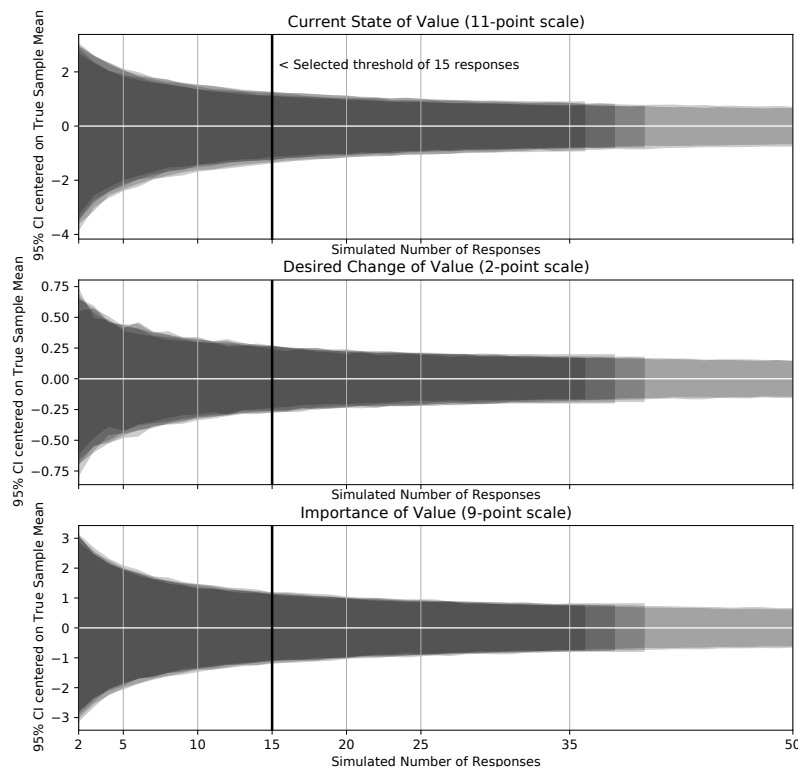# B Power Analysis to Determine Validity of Responses

Figure 8: Using responses from the 5 subreddits with more than 35 responses each, we randomly downsampled (1,000-fold bootstrapping) responses to estimate the sample variance when collecting fewer responses. We found that beyond 15 responses per subreddit, sample variance does not decrease significantly, and so we select this threshold for our analyses.

---

[7]We additionally experimented with pre-computed subreddit embeddings (Kumar et al. 2018; Martin 2017; Waller and Anderson 2019). We found these representations were unable to differentiate between communities based on their values.
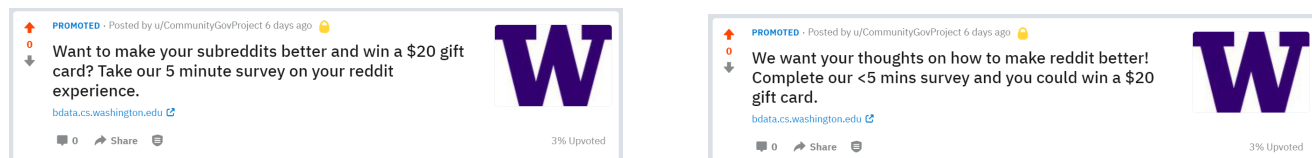
C Participant Recruiting and Incentives

Figure 9: Reddit advertisements used to recruit participants.

Survey respondents were recruited primarily through reddit advertisements and private messages (PMs), which are displayed to reddit users on both the website as well as the reddit mobile app. We used several different titles for the ads, Appendix Fig. 9 shows examples. We ran three different recruitment campaigns: (1) a general campaign targeted at all reddit users and designed to capture responses from members of a wide range of subreddits, (2) a specific campaign intended to increase the number of responses received for the most popular subreddits, conducted by creating separate ads for each of the 300 largest subreddits, and (3) a moderator recruitment campaign to encourage participation specifically from community moderators, who were recruited via PMs sent to each of the 100 largest subreddits).

Survey responses were gathered from May-July 2021. In total, 2,769 people participated, with the participants answering questions for 2.15 subreddits on average, for a total of 5,962 subreddit-responses across 2,151 unique subreddits. 562 responses (20.30%) were recruited via the general campaign, 2,022 (73.02%) were recruited the specific campaign, 81 (2.93%) were recruited via the moderator campaign, and 104 (3.80%) were recruited via friend referrals and word of mouth. The median completion time was approximately 8 minutes. 97.33% of respondents provided their username and consented to the inclusion of their post and comment history in our research.

To incentivize participation, we raffled off a $100 Amazon gift card and $5\times$ $20 Amazon gift cards to participants who completed the survey. Participants were offered additional raffle tickets for recruiting their friends to participate as well. Winners were contacted via reddit PM, and those outside of the US were offered a gift card of equivalent value in their local currency.

D Details on Prediction Tasks

Value	Dimension	ROC AUC
Democracy	Current State	0.622
	Desired Change	0.684
	Importance	0.541
Diversity	Current State	0.634
	Desired Change	0.800
	Importance	0.716
Engagement	Current State	0.635
	Desired Change	0.532
	Importance	0.642
Inclusion	Current State	0.730
	Desired Change	0.708
	Importance	0.555
Quality	Current State	0.725
	Desired Change	0.624
	Importance	0.677
Safety	Current State	0.441
	Desired Change	0.714
	Importance	0.391
Size	Current State	0.936
	Desired Change	0.655
	Importance	0.661
Trust	Current State	0.709
	Desired Change	0.688
	Importance	0.922
Variety	Current State	0.625
	Desired Change	0.589
	Importance	0.838

Table 3: Task-level results (ROC AUC) for the Logistic Regression model on our 27 prediction tasks.

sub_num_posts	The number of posts in the subreddit.
sub_num_removed_posts	The number of posts removed by a moderator in the subreddit.
sub_num_deleted_posts	The number of posts deleted by their author in the subreddit.
sub_num_selfposts	The number of selfposts (text-posts) in the subreddit.
sub_num_linkposts	The number of posts which link to external websites.
sub_num_comments	The number of comments in the subreddit.
sub_num_removed_comments	The number of comments removed by a moderator.
sub_num_deleted_comments	The number of comments deleted by their author.
sub_distinct_users	The number of distinct contributors to the subreddit.
sub_num_subscribers	The number of users who 'subscribe' to the subreddit.
sub_age	The number of days since the subreddit was founded.
sub_topic_specificity	The manually-categorized specificity of the topic of the subreddit, on an 3-point scale.
sub_topic_category	The manually-categorized (see §3.4) topic of the subreddit.

Table 4: Descriptions of features used in the prediction tasks (§7).