Improving Online Community Governance at Web Scale

Galen Cassebeer Weld

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington
2025

Reading Committee:

Tim Althoff, Chair
Amy X. Zhang, Chair
Yulia Tsvetkov

Program Authorized to Offer Degree:
Computer Science and Engineering

University of Washington

Abstract

Improving Online Community Governance at Web Scale

Galen Cassebeer Weld

Co-chairs of the Supervisory Committee:

Tim Althoff

Paul G. Allen School of Computer Science and Engineering

Amy X. Zhang

Paul G. Allen School of Computer Science and Engineering

Nearly two out of every three people on the planet are members of an an online community, and this number is forecast to keep growing. These communities have an incredible diversity of topic, size, and structure, and they offer unique ways to connect their users and bring people together. Unfortunately, online communities have also been associated with significant offline harms, including the mental health crisis, abuse and harassment, interference with free and democratic elections, and radicalization and political polarization.

Almost all online communities rely on some form of governance to set and enforce rules, role model good behavior, and generally lead the community. The forms that this governance takes varies widely from community to community. On some platforms, moderators' work is conducted in the background, while in many others, community leaders are volunteers who take a more visible role. Many communities' governance also relies on a range of complex technical tools. Some communities operate on a pseudodemocratic basis, with nominations and regular elections, while others operate on a consensus model, and still others are effectively autocracies. It is very difficult to know how best to govern an online community, given different community needs, the enormous range of available governance strategies, and the challenge of empirically

measuring governance and outcomes.

In this dissertation, I conduct research that makes online communities better through data-driven analyses of community values, moderation practices, and experiments with new tools. My work focuses on three important research activities: (1) I *characterize* communities' values in community members' own words to build a foundational understanding of communities' needs and what 'better' actually means. (2) I *assess* existing moderation practices and community affordances such as voting at a massive scale across hundreds of thousands of communities in order to identify which practices are most promising. (3) I *deploy* interventions and best practices in partnership with community leaders to maximize real world impact. Much of my research is conducted on Reddit, one of the largest platforms for online communities, and a platform where I am a longtime moderator of several subreddits, and a member of the Reddit Moderator Council.

My dissertation makes several key contributions: My theoretical contributions include the first ever taxonomy of community values, based on the largest-to-date surveys of community members. My methodological contributions include a new method for scalably measuring community outcomes by quantifying how community members talk about their moderators, and a new method for classifying the rules enforced by communities. Finally, I make artifact contributions by publishing classifiers for discussions of moderators and rules, and datasets of anonymized survey results, community rules, and news sharing behavior.

Contents

List of Figures

List of Tables

Acknowledgments

The past seven years spent working on the research presented in this dissertation have truly been some of the best years of my life. I know that this is not the case for many PhD students, and the quality of my in graduate school is a testament to the amazing group of colleagues, friends, and family that I am honored to be a part of.

To my advisors, **Tim Althoff** and **Amy Zhang**, getting to work with both of you has been an honor, and also the biggest reason that I wake up looking forward to my work. I am relieved to know that I will continue to be able to collaborate with both of you in the future. You two may have very different mentorship styles, but I have felt so supported by both of you, and have learned so much. I don't know many PhD students in the Allen School who have been here for longer than both of their advisors, but getting to be one of the first students in both of your labs has meant the world to me.

Tim, I am endlessly appreciative of the ownership that you have enabled me to have over my research, which goes all the way back to day that asked you to advise me on a project studying online communities, in the second year of my PhD. Little did I know that this project would grow to become my entire dissertation! The way you encourage me to disagree with you, to push my own lines of thinking deeper, and to never take shortcuts will be a lifelong lesson.

Amy, your enthusiasm and energy for research is contagious! You've always encouraged me to chase after the questions that interest me, and you've shown me the value of working in a team while helping me grow my own mentorship skills. Thanks to you and Johnny for opening your home to your many students, despite the horrors we may have wrought in your kitchen—and you invited us back!

To my committee members, **Sanjay Kairam**, **Mako Hill**, and **Yulia Tsvetkov**. I appreciate you taking time out of your busy schedules to provide me with thoughtful feedback and the right number of tricky

questions.

I am lucky to have been mentored by a number of other great researchers during my time in graduate school. **Richard Anderson**, thank you for always being a friendly face around the Allen School, and for setting me on this path by inviting me into your ICTD Lab, way back in 2016. That feels like quite a long time ago now. **Jon Froehlich**, thank you for broadening my research perspectives with our work on Project Sidewalk, and for always having a kind word to say about my photography. **Maria Glenski**, thank you for your solid mentorship at PNNL during the uncertainty of the start of the COVID Pandemic. **Sanjay**, thank you for being such a proactive advocate for me at Reddit, even during your own exciting career transition, and **Carl Pearson**, thank you for helping me navigate the wilds of Reddit BigQuery tables.

My labmates in the Behavioral Data Science Lab and Social Futures Lab have provided me with invaluable feedback on my research and writing, and are some my best friends. **Mike Merrill**, **Ken Gu**, **Inna Lin**, **Ashish Sharma**, **Chris Rytting**, **Margaret Li**, **Advait Bhat Deniz Nazarova**, **Xinyi Zhou**, **Yasaman Sefidgar**, **Joe Breda**, **Jim Chen**, **Kevin Feng**, **Ruotong Wang**, and **Alicia Guo**, thanks to all of you. And a special thanks to **Leon Leibmann** for being such a delightful collaborator. Finally, thanks to the many members of the **ICTD Lab** for being my first home in the Allen School (back when there was only one building!)

The Allen School, I believe, is a truly special place, and I feel lucky to have been able to contribute to its unique culture. Thanks to everyone who came to all the (too many) TGIFs that I organized, to everyone who made Ski Day happen every year, and for all the writers, actors, and audience members who made the Holiday Party skits so memorable. There are also so many staff members whose work often goes unseen, but without whom, the Allen School would be chaos: the facilities folks, event staff, admins, and grants managers (special thanks to **Elaine Shelley**).

Thanks to all the friends who tolerated me walking into their offices, unannounced, so that I could chat. **Joe Breda**, I appreciate your perspectives on urbanism. **Moe Kayali**, I won't forget our many conversations held while walking around campus. **Artidoro Pagnoni**, you may have used stronger language at the time, but your admonishment of 'no more fussing around' is a lesson I have taken to heart. **Tina Yeung**, thank you for always looking out for me, especially during the final push to finish this dissertation, and for so many snacks. **Lancelot Wathieu**, I always enjoy hearing about your film photography and your backcountry

exploits. And a huge thanks to all the members of the **Security and Privacy Lab** for always opening the door for me.

A number of inanimate objects provided noteworthy support during graduate school. **Darkwing**, you faithfully burned countless CPU cycles as you crunched Spark jobs for me. When I cursed at you, I didn't mean it. **The Coca-Cola Company** may have dubiously moral business practices, but there's no denying the role that Costco-sized cases of Diet Coke had in powering my research (and, apparently, a number of other members of the bdata lab). The **Blue Moon Tavern** and its numerous regular customers provided a reliable Monday-night pitstop for me on my way home, but I also did more work from the back room during Andy Coe Band set breaks than I care to admit.

The incredible mountains surrounding Seattle were a frequent distraction during graduate school—a lower bound estimate is that I spent 455 days skiing in the past seven years, which is something to be thankful in and of itself. However, I spent almost as much time on a myriad of other activities: rock climbing, biking, and paragliding. These activities were immeasurably beneficial to my physical and mental health, but at the end of the day, it's the people who I spent time with doing these things that are my closest friends and to whom I am most grateful for their companionship. **Matt Schilling**, I may have known you for for more than ten years now, but it was only during graduate school that I feel our friendship really developed. Every day I am both inspired and intimidated by your work ethic and the intensity with which you approach everything you do. **Pieter van den Kieboom**, I appreciate your endless enthusiasm, and I owe you a particular thanks for introducing me to the unparalleled joy of the telemark turn. **Mike Merrill**, I am so glad that our friendship escaped the confines of Gates 386. The trips we went on together are some of my favorites, particularly our backpacking trip in the Sawtooth Mountains when both of us were healing from broken bones. **Serena Lotreck**, you may live on the other side of the country, but our countless phone calls are the highlight of many of my days.

I am biased, but I believe I have the best family anyone could ask for. **Dad**, in addition to be an amazing father and incredible friend, you were the first computer scientist I ever knew, and certainly the biggest reason that I chose to pursue a PhD. **Mom**, you inspire me with your passion for helping people with your research, and thanks for helping clear up some statistical questions as well. It's fun to have some common ground in our work. **Adam**, my brother, I am so lucky to have you be such a huge part of my life. Getting

to spend as much time getting into trouble with you has been, and will continue to be, a lifelong privilege. Lastly my grandfather **Arthur Rosenfeld** is perhaps the most preeminent Rosenweld scientist, and I am sad that he isn't here to ask me about my work. I know he would have been proud.

Finally, my wife **Becca**: you inspire me with your fierce dedication to your work, your friends, and your climbing. You are a role model in so many ways: you commit to making the world a better place through your research, you are a meticulous planner, and, topically, you earned a PhD! Your love is a blessing, and I am so happy to officially be a part of your family. I am so lucky to get to spend my life with you.

This dissertation research was supported by the Office of Naval Research (Grant N00014-21-1-2154), the National Science Foundation NSF (Grants IIS-1901386, CNS-2025022, and CAREER IIS-2142794, the Bill & Melinda Gates Foundation (Grant INV-004841), a Microsoft AI for Accessibility grant, and a Garvey Institute Innovation grant. This work was completed using computing resources on Hyak, UW's high performance computing cluster, which was funded by the UW student technology fee. Portions of this research were done at Reddit, Inc., and the Pacific Northwest National Laboratory.

DEDICATION

To my cat Sahara,

who taught me the importance of vocal self-advocacy

and the restorative power a good meal and a nice nap.

Chapter 1

Introduction

Online communities are a universal part of our society. Nearly two-thirds of the planet's population participates in an online community [59], and those community members spend an average of over 2 hours a day in those communities [58], with many groups spending even more time online. Nearly half of American teenagers say that they use social media 'almost constantly' [67]. Online communities have been around for even longer than the World Wide Web, with USENET, predecessor to the now-ubiquitous Internet forum, first established in 1980, thirteen years before the Web was opened to the public in 1993. Although USENET still exists today, its popularity was eclipsed as more and more people gained access to the Internet, and platforms that are now familiar were launched: Facebook in 2004, Reddit in 2005, Instagram in 2010, and TikTok in 2016. Today, online communities' ubiquity is a testament to their power. Online communities do an extraordinary job of connecting people, provide opportunities for learning and entertainment, and enable connection in a way that would not be feasible offline, for instance by supporting groups with marginalized identities or niche interests.

However, online communities are not perfect. For as long as online communities have existed, people have used them to argue, harass, and post spam. In fact, the commonly used terms 'spam' (unsolicited commercial content), 'trolling' (posting provocative content), and 'flamewar' (protracted and heated arguing) were first popularized on USENET [74], and these behaviors remain problematic in most online communities today [86].

With the increased popularity of online communities has come increased attention to their harms. Cy-

berbullying and cyberharassment affect millions of people a year [164, 30, 158, 173]. Serious concerns have been raised about the impact of online communities on their members' mental health, particularly for young people [82, 98]. The proposed harms of online communities aren't just limited to their direct users. Research has linked online communities to increased partisanship and political polarization [218, 152, 10, 194], and has even found that they may be able to influence elections [191, 49, 138, 94, 105, 26, 218]. Online communities have also contributed to the spread of mis- and dis-information [294, 214, 268].

To combat these challenges, as well as to simply keep the community on-topic and running smoothly, almost every online community relies on some form of governance. USENET introduced the concept of moderators, privileged users who had to approve submitted content before it was posted. Content moderation has evolved somewhat since USENET, but is still the most widely used framework of community governance today, adopted in some variant by every major (and most minor) community platforms [86, 150]. Under the content moderation framework, regular community members submit content (typically text, images or video, or links to websites), while only a small set of moderators have permissions to remove or approve that content.

However, there are countless details and decisions in the practice of content moderation that can have massive impacts on the community. On some platforms, moderators' work is conducted in the background, by moderators who are paid by the platform (such as on Twitter and Facebook), while in many others, moderators are volunteer from the community who take a more visible role, such as on Reddit and within Facebook Groups [181]. By one estimate, volunteer moderators contribute an absolute minimum of 3.4 million US dollars in uncompensated labor annually on Reddit alone [166], which has been problematized by some [181]. This reliance on volunteer labor also makes platforms vulnerable to collective action, as seen in the 2015 and 2023 Reddit moderation 'blackouts' [184]. On the other hand, some claim that their volunteer status makes moderators more impartial, and Reddit even prohibits their moderators from accepting compensation in their Moderator Code of Conduct [220]. However, even compensated professional moderators have reported being overworked, underpaid, and exposed to endless volumes of traumatizing content [245].

Platforms and communities also differ in their use of automated tools to assist in content moderation. While most platforms employ some degree of automation in their spam filtration and content moderation, these systems can be very simple moderator-configurable tools using regular expressions to match content,

as on Reddit [35, 121], or they can be extremely sophisticated machine learning systems with massive training corpora that perform the vast majority of a platform's content moderation actions [260, 64].

There is also dramatic variance in the degree to which democratic practices are employed (or not) across different communities. Platforms without volunteer moderators (like Facebook, Twitter, and Instagram) have very little transparency in their governance, making it challenging to draw parallels to offline governance models, but even on platforms with volunteer moderators, like Reddit and Facebook Groups, many different strategies are employed. On most platforms, the person to create the community automatically becomes its first moderator, and many communities remain above governed by their initial founder, who is often semi-humorously referred to as the 'benevolent dictator for life,' a title originating in the Open Source Software world, where it was initially applied to Guido Van Rossum, the creator of the Python programming language [232]. While most communities with volunteer moderators have some hierarchy of moderator seniority, the addition of new moderators, the removal of existing moderators, and promotions/demotions in the hierarchy are often made on an ad-hoc basis by the most senior moderators, who themselves were originally selected by the original founders, typically without formalized selection criteria [181]. Some communities, however, do provide opportunities for non-moderator community members to participate in the selection of new moderators, either through a consensus process, a public nomination process, or public elections. Due to the technical affordances provided by major platforms, these processes are frequently nonbinding and therefore voluntary. Academic efforts have been made to formally implement moderator elections on Reddit and a few other platforms [291], but these remain far from widespread adoption, although they are more common in peer production communities such as Stack Overflow and Wikipedia [175].

Communities also vary in the rules that they set, how those rules are presented to community members, and how they are enforced. Centralized platforms with hidden moderators, such as Twitter, Facebook, Instagram, and TikTok, each have centralized rules that are used to govern what can and cannot be posted by community members, and guide enforcement decisions [288, 190, 265]. Although communities with volunteer moderators typically must abide by platform-wide rules, such as Reddit's site-wide rules [222], moderators of these communities often have wide leeway in how they set, interpret, and enforce rules [72, 224]. Moderators can choose not only what rules are set, but how they are phrased. For example, a rule about community member conduct could be phrased either as saying what *to* do ('Be nice!') or what *not* to

do ('Don't be mean!'). Moderators also decide when and how rule-breaking content should be removed or otherwise sanctioned. Some communities have removal reasons [115], issue warnings, implement multiple-strike policies, or do any combination of some, all, or none of these things. Furthermore, there are differences in the strictness of rule enforcement, both within and between communities. Different communities with the same rule may choose to enforce that rule differently, and even within the moderator team for a single community, some moderators may be more permissive or lenient than others [147].

Content removal is not the only action that moderators can take. In most communities, moderators can also ban users, and bans are subject to all the same variability in implementation that content removal does [150]. Platforms can and occasionally do ban entire communities for consistent problematic behavior [264, 228].

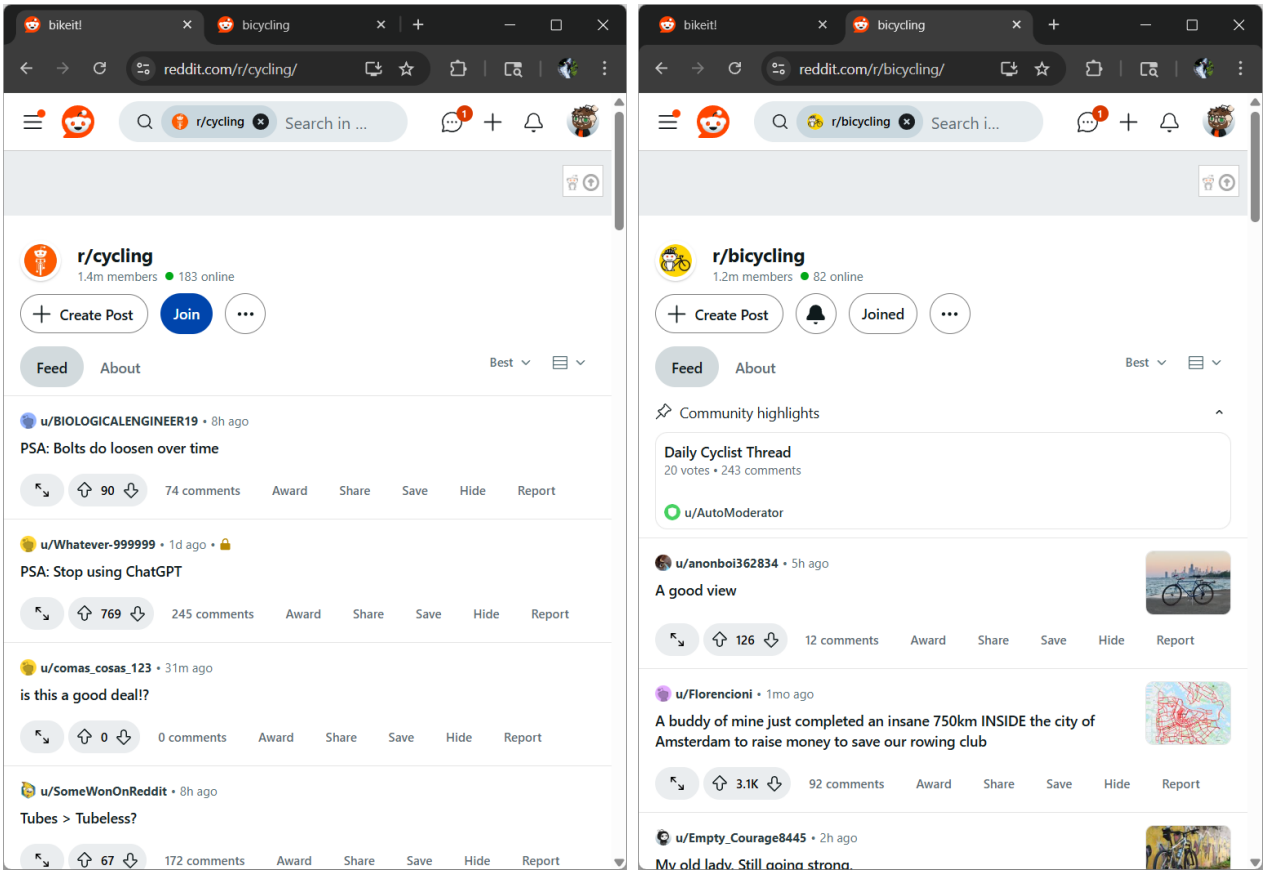
Content moderation is also only one part of community governance. Many other aspects of how the community is configured also constitute its governance, some of which can be out of the control of volunteer moderators, depending on the platform. What content is shown to community members, and how that content is ordered and prioritized, can play just as large of a role in determining community outcomes as content moderation [150]. All platforms use some sort of feed ranking algorithm, although that can be very simple (a simple chronological sort, for example), or a very complex and highly optimized algorithm with little transparency. Furthermore, some platforms, like Reddit and Hacker News, ensure that everyone who visits the community's homepage sees the same content, whereas others (Facebook, Instagram, TikTok) show everyone an individually personalized content feed that may increase engagement, but results in different members having different concepts of what is happening in the community—a breakdown in the concept of 'social translucence [84]. Many platforms use 'community curation' mechanisms where community feedback from upvotes, downvotes, or likes are used to increase the visibility of popular content by positioning it higher up in the feed, whereas disliked content is moved down or hidden entirely.

1.1 A Tale of Two Communities

By now, I've described the myriad ways in which online communities can differ from one another in their governance. Those potential differences, however, may still feel quite abstract. In this section, I will use an example to show how even communities with similar topics can differ dramatically from one another

in practice. We'll examine many of the features of online community governance and outcomes that we'll assess in greater detail in the rest of this dissertation.

As with much of the rest of this dissertation, we'll focus on Reddit, in particular, two subreddits dedicated to bicycling, which also happens to be a personal hobby of mine. Figure 1.1 shows the homepages of /r/cycling, on the left, and /r/bicycling, on the right. As you can see, these communities have similar topics, and similar types of content, with a mixture of pictures ('A good view'), questions ('is this a good deal!?'), and other bicycling related content. Both communities have similar sizes, with 1.4 million members in /r/cycling, and 1.2 million members in /r/bicycling. It's not visible in the screenshots, but both communities are almost exactly the same age, having both been established in May 2008.



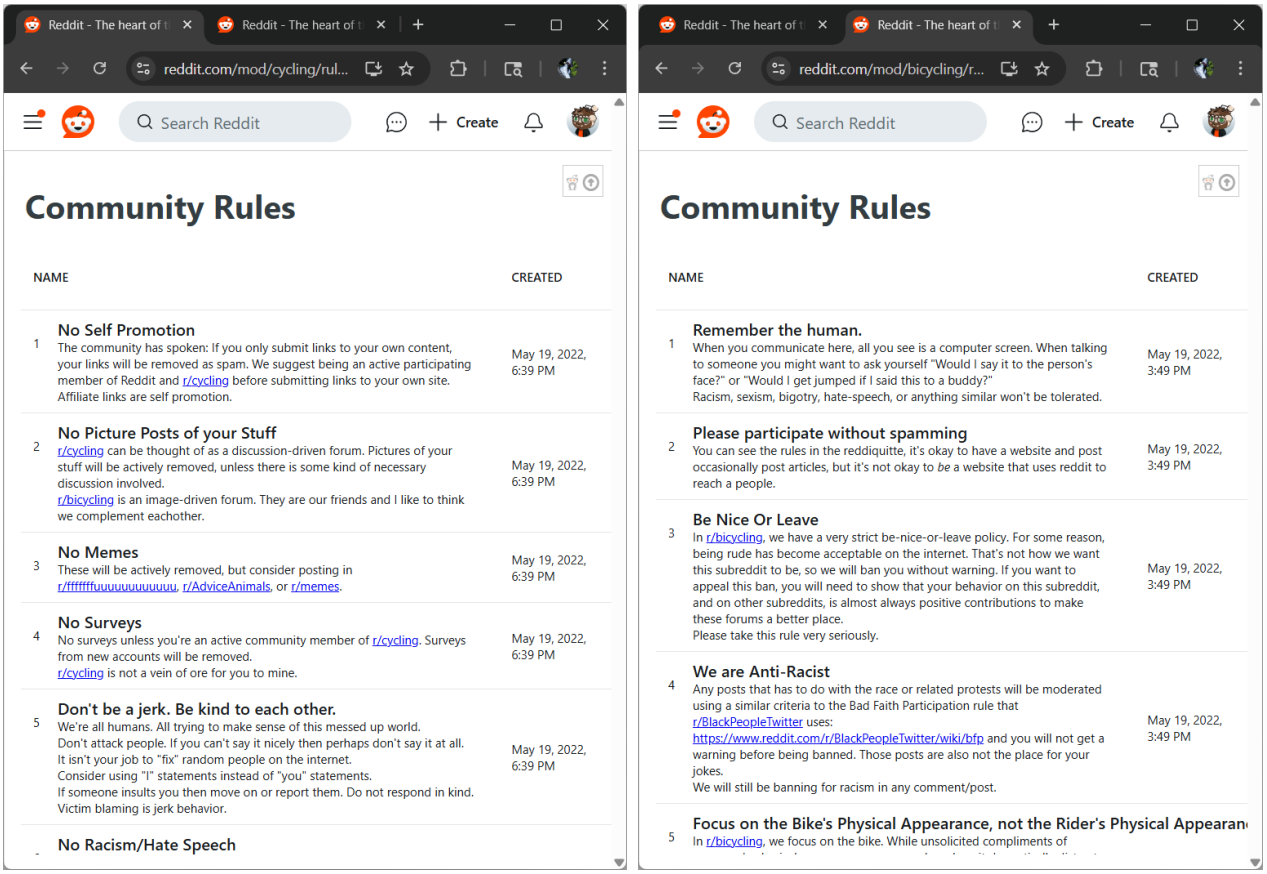
(a) The homepage of /r/cycling.

(b) The homepage of /r/bicycling.

Figure 1.1: Screenshots for the homepages of /r/cycling (left) and /r/bicycling (right). Both subreddits have similar topics, and similar sizes (1.4 *vs.* 1.2 million members).

However, even on their home pages, we can begin to see some differences. /r/bicycling has a 'Daily

Cyclist Thread,' an automatically posted daily discussion, with no such discussion thread in /r/cycling. If we click over to the rules pages for both of these communities, even more differences emerge (Figure 1.2. All of /r/cycling's rules explain what *not* to do, and they all follow the format 'No x' or 'Don't y.' /r/biycling, on the other hand, has more prescriptively phrased rules, telling community members how to behave. There are also many differences in the presence or abscence of specific rules. /r/cycling prohibits the posting of surveys, and of pictures without discussion involved, whereas /r/biycling has none of the prohibitions. /r/bicycling, through, addresses antiracism in their rules, which is not mentioned in /r/cycling.

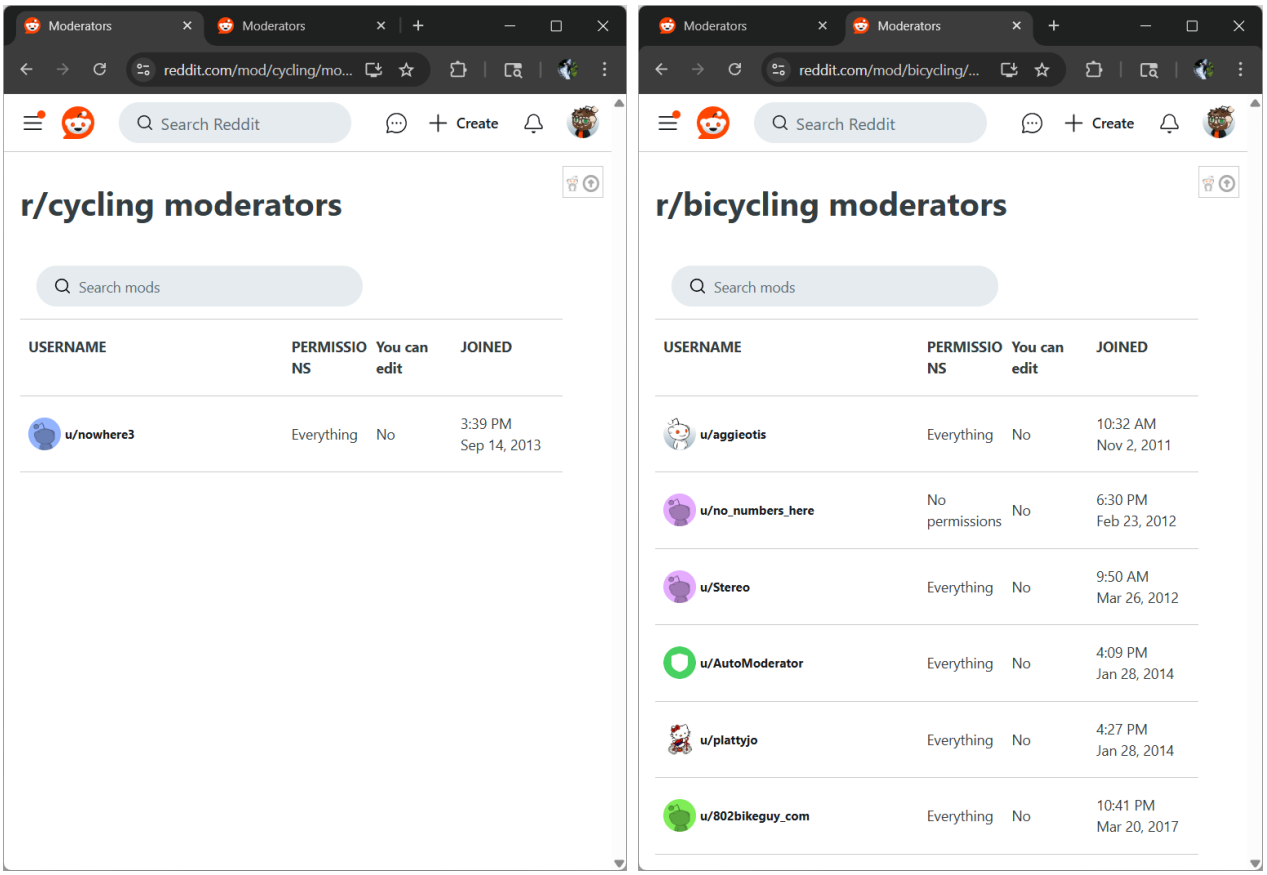


(a) /r/cycling's rules include prohibiting self promotion (1), **(b)** /r/bicycling asks members to 'Remember the human' pictures without discussion topics (2), and 'Don't be a jerk' (1) and 'Be Nice Or Leave.' (3) (5).

Figure 1.2: The rules of /r/cycling (left) and /r/bicycling (right). Both communities have some similar rules, as well is some key differences. /r/cycling prohibits surveys, with no equivalent rule in /r/bicycling. In general, /r/bicycling's rules focus more on what *to* do, while all of /r/cycling's rules say what *not* to do.

If we click over to each community's respective list of community moderators, which are publicly view-

able on Reddit, we see another dramatic difference. /r/cycling has just a single moderator who has been in that position since September 2013 (Figure 1.3a). Despite being slightly *smaller* than /r/cycling, /r/bicycling has far more moderators, with thirteen people listed on its moderator page. Most of /r/bicycling's moderators were appointed much more recently, with the two most junior having been added only two years ago (not visible in Figure 1.3).



(a) /r/cycling has just a singe moderator.

(b) /r/bicycling has too many mods to fit on a single page.

Figure 1.3: /r/cycling has just a single moderator, whereas /r/bicycling has thirteen moderators, too many to fit on a single page.

Do these differences in governance practices lead to differences in the experiences that community members have in these two communities? Establishing a definitive causal link is challenging (§1.5), but we can see that there are substantial differences in how members of these two communities feel about them. As part of the research presented in §2.2, we surveyed members of both /r/cycling and /r/bicycling to ask them how they felt about various aspects of their communities. We asked about the current state of the

community on an 11-point Likert scale, and found that members or /r/cycling felt their community had higher quality content (0.39 points more than /r/bicycling), had more varied content (2.61 points more than /r/bicycling), was more trustworthy (0.41 points more than /r/bicycling), was more engaged (2.41 points more than /r/bicycling), and was safer (1.41 points more than /r/bicycling).

These differences aren't limited to what community members say in private surveys. We also found differences in what sorts of things community members said publicly about their respective moderators (as we will do in §3.1). We found that, when saying something about their moderators, members of /r/cycling were $1.70\times$ as likely to express a negative sentiment about their moderators than members of /r/bicycling were, while members of /r/bicycling were $1.58\times$ more likely to express a positive sentiment.

1.2 Thesis Statement

Based on these data, it seems as though /r/cycling is the more successful of the two communities. Why is this? Are there lessons that could be learned from /r/cycling, and applied to /r/bicycling, in order to improve the experience there? Again, online communities are highly complex systems with many contributing factors at play, so just based on a simple comparison such as this, it's difficult to draw any definitive causal conclusions. However, as I hope to demonstrate in the rest of this dissertation, by *assessing* hundreds of thousands of communities, examining all their natural variance in both governance and outcomes, and using observational causal inference methods to control for confounding factors, it is possible to test far more hypotheses than would be possible through active experimentation, and to identify the most promising governance practices. However, just assessing current practices isn't enough. To ensure that we are assessing things that actually matter to communities, we must *characterize* their values, needs, and how these vary both between communities, and within individual communities. Finally, to ensure that our work has real impact, we must *deploy* interventions and data-driven best practices through collaborative partnerships with real world communities.

My dissertation research demonstrates the following thesis statement:

> *Informed by online communities' individual values and needs, large scale analyses of online communities can leverage existing variance in governance and outcomes to improve communi-*

ties through data-driven best practices, the development of novel tools, and real world collaborations.

By grounding our approach in community-centered values, leveraging existing datasets where possible, and collaborating with partners to put findings into practice, we can make online communities better.

1.3 Dissertation Overview

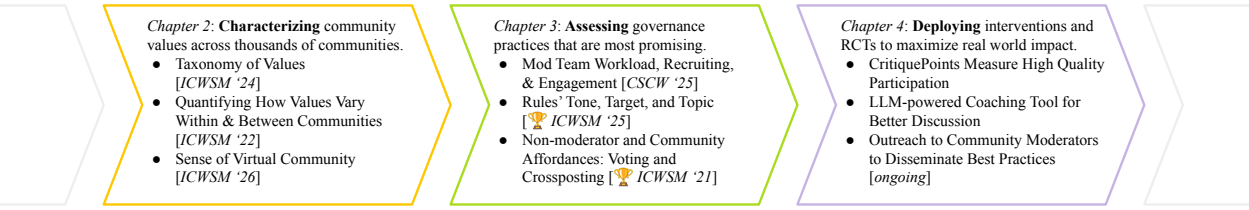


Figure 1.4: The rest of this dissertation is organized into three chapters, each focused on a different research activity, in order to empower communities and improve online community governance.

As is often the case with research, the work in this dissertation started when I attempted to look up the answer to what I thought was a relatively simple question, and discovered that no one had apparently answered it yet! I had done a little bit of work to quantifying aspects of moderation on Reddit, and I was trying to think of important metrics for online community health and success that I could measure automatically and at scale, to pair with the moderation data I had collected. I realized that, beyond some important but narrow topics like misinformation and online harassment, it wasn't actually clear what aspects of online communities were most important to community members, let alone how to measure them automatically. A lot of the existing literature focused on how to mitigate specific harms, not only misinformation and harassment, but also spam, hate speech, and other nefarious behaviors that are far too common in online communities. Of course, these harms are very important, and it is critical that we as a field continue to study how to diminish them. However, it's equally clear that the ideal online community isn't just one without misinformation, without harassment, and without harms. There are positive reasons, too, that provide strong motivating factors that cause billions of people across the planet to participate in online communities. I wanted to understand not just how to mitigate the harms, but how to actually make online communities 'better'—and what 'better' actually means! Furthermore, given how diverse online communities are in their

topics and memberships, it seemed quite likely that not all communities would have the same needs and values.

1.3.1 Chapter 2: Characterizing Community Values

An obvious place to start, then, is to *characterize* community members' values in their own words, by simply asking them. Ultimately, I wanted to survey many thousands of community members, but it's very difficult (especially in the pre-LLM era) to analyze thousands of open-ended, free-text responses. I needed some way to structure quantitative questions that I could analyze at a large scale, but no taxonomy of community values existed for base my survey questions on. In §2.1, I present the creation of such a taxonomy, designed to capture the full breadth of things that online community members value for their communities, using 1,481 open-ended responses from 212 members of 627 unique communities on Reddit. The resulting taxonomy contains nine high-level values, divided into 29 subcategories. I then discuss how tension between these values poses challenges for community moderators and platform designers, and highlight the values that are poorly understood and understudied. *This work was presented at ICWSM 2024 [283].*

Armed with this new understanding of *what* community members value, I wanted to empirically measure how these values vary among different members of the same communities, as well as between different communities. In §2.2, I present the largest-to-date survey of community members values on Reddit, including over 5,000 responses from 2,769 members of 2,151 unique communities on Reddit. I use these data to assess how values differ in importance across communities of different topics, and how community members currently feel about their communities, and want their communities to change. I examine how these perceptions and preferences vary across communities of different ages and sizes. By collecting multiple responses from different members of the same communities, I am able to measure where there is consensus on values, and where there is disagreement. I show that community members disagree about how safe their communities are, that longstanding 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 results have important implications for community design and moderation. *This work was presented at ICWSM 2022 [284].*

A strong Sense of Community (SoC) is perhaps the single most important ingredient for community

success, yet measuring Sense of Community in online communities is challenging, and how different community behaviors relate to SoC is poorly understood. In §2.3 I focus specifically on SoC in online communities, which I refer to specifically as Sense of *Virtual* Community (SOVC). I present the results of another large-scale survey of 2,826 Redditors, which introduced some new questions alongside my earlier survey, and was conducted in collaboration with researchers at Reddit. I identify three primary dimensions of SoC on Reddit: Membership & Belonging, Cooperation & Shared Values, and Connection & Influence. I develop a hierarchical model to predict how SOVC varies along with conversational structure and style across 281 different communities. This work highlights actionable strategies for fostering stronger community attachment, using an approach that can generalize readily across community topics, languages, and platforms. *This work was conducted as part of a Research Fellowship at Reddit, Inc. and will be presented at ICWSM 2026 [285].*

1.3.2 Chapter 3: Assessing Current Governance Practices

We now turn our attention to *assessing* which current governance practices are most effective, informed by our newfound knowledge of community values. By including hundreds of thousands of online communities in our analyses, we can leverage existing natural diversity in governance and outcomes to produce robust, data-driven insights and best practices, using causal inference methods to control for confounding factors without needing to resort to expensive and time consuming randomized controlled trials. We'd like to quantify both different community governance practices, as well as important community outcomes, in order to establish links between the two, but data is not readily available for everything we'd like to measure. With clever usage of tools such as the Wayback Machine [259], we can collect details about community rules and moderators, and most posts and comments in online communities are publicly available, at least on Reddit. The central challenge is this chapter is measuring community *outcomes* in a way that is feasible across thousands of communities, and billions of posts and comments. The availability of data here is a double edged sword: we can collect terabytes of posts, comments, and images, so much that it is quite computationally expensive to process all of it, yet at the same time, these data are unstructured and therefore it is difficult to tease out the relevant information from everything else.

In §3.1, I leverage a key insight to develop one of the first scalable measures of community governance.

That insight is that community members discuss the moderators of their communities publicly, saying things like 'the moderators of this subreddit are doing a great job!' or, perhaps more commonly, less positive remarks that are unfit for publication here. I develop a three step pipeline to automatically detect posts and comments discussing community moderators, and use this pipeline to identify 1.89 million statements discussing moderators on Reddit made over an 18 month period. I use these data, along with several causal inference methods, to assess how different moderator workloads, recruiting strategies, and degrees of community engagement are received by community members. I show that strict rule enforcement is linked to more favorable perceptions of moderators in communities dedicated to certain topics, such as news communities, than others. I find that moderators who are active community members before, during, and after their tenures as moderators are seen more favorably by their constituent community members. *This work was presented at CSCW 2025 [281].*

The setting and enforcing of rules is perhaps the most critical function performed by moderators, yet is it is difficult for moderators to make informed decisions about what rules to impose, how to phrase them, and how to enforce them. In §3.2, we take a deeper dive on community rules specifically. Using historical data collected from the Wayback Machine [259], I compute timelines of what rules were introduced to which communities when, examining 67,575 different rules from 5,225 different communities across a 5+ year period. I develop a method to classify rules according to their tone, target, and topic, conduct the largest-to-date assessment of rules on Reddit. After controlling for key confounding factors, I identify the rules most associated with positive perceptions of moderators, using the classification method developed in the preceding section. Finally using time-series data and a difference-in-difference analysis, I examine what happens when communities introduce new rules, finding that in the short term, new rules appear to improve community member perceptions of governance, but this effect seems to wear off after approximately six months. These results help inform best practices for community rule implementation. *This work was presented at ICWSM 2025, where it won the Best Paper award [163].*

In addition to content moderation, communities on Reddit have several affordances for non-moderators to influence the visibility of content, which is an important part of of democratizing those communities governance, broadly defined. Through the lens of news sharing, a common use for online communities, I examine how these non-moderator affordances influence the visibility of highly biased or low factual news

posts in §3.4. Using Media Bias/Fact Check, a non-partisan fact-checking service, I conduct the largest study of news sharing behavior on Reddit, labeling more than 550 million links to news sources across a four year period. I examine how upvoting, downvoting, and amplification via crossposting affect the visibility of highly biased content. I also assess how politically left- and right-leaning community members differ in their posting behavior, and explore the efficacy of deplatforming of entire communities by showing that exposures to highly biased and low factual content are extremely concentrated in a small number of potentially problematic communities. My results suggest that non-moderator affordances such as voting and crosspostsing are generally effective at reducing the relative visibility of highly biased and low factual news content. *This work was partially conducted during an internship at Pacific Northwest National Laboratory, and was presented at ICWSM 2021, where it won the Outstanding Analysis award [282].*

1.3.3 Chapter 4: Deploying Interventions to Improve Governance

Unless our strategies for better governance are *deployed* in real-world communities, this research will not have any impact. Data-driven best practices and tools for community governance, no matter how effective, do not have any impact unless they are actually widely adopted by communities. In this chapter, I partner with community leaders to deploy practices and novel systems to improve community outcomes. Deploying tools in realistic environments also enables me to validate their efficacy, as well as conduct randomized controlled trials, the gold standard method for establishing causality.

In §4.1, I introduce CritiquePoints, new system to gamify the delivery of high quality feedback on creative tasks in a large photography critique community with 1.7 million members. This system also enables the precise measurement and annotation of the quality of submissions in a community-specific fashion. Over a 3.5 year period, the system has been used 9,934 times by 3,356 different community members. I leverage these annotations to conduct an analysis of what makes for high quality feedback on creative tasks, using randomly sampled critiques of the same photographs which were *not* awarded points as a control. I find that high quality feedback is specific, actionable, addresses the questions asked by the recipient, and goes beyond what was originally asked.

Informed by these results, in §4.2, I deploy and evaluate an LLM powered tool to teach members of the same photography critique community how to give one another better feedback and improve the overall

community experience. Through an ongoing large scale deployment in a real online community, I compare two different pedagogical strategies to assess efficacy of coaching strategies. Results from semi-structured qualitative interviews with participants suggest that the system is effective at improving the quality of delivered feedback, compared to a non-LLM baseline, and that a Socratic coach that encourages participants to think more deeply about their comments without making specific suggestions requires more effort to use, but results in better learning of comment writing skills.

1.4 An Overview of Reddit

Before we dive into the bulk of this dissertation, I would like to take a moment to introduce in greater detail Reddit, a platform that I have already mentioned numerous times in this chapter, but we will focus on in even greater detail in the chapters to come. Indeed, almost all the research I present here uses Reddit as a natural laboratory to study community governance. This is deliberate: I focus on Reddit not because I don't believe my results generalize to other platforms, quite the opposite, but because I believe that Reddit is an ideal platform for studying community governance. Reddit is used by more than one out of every five Americans, making it the fifth most popular social media platform [90, 249], with over 100 million daily active users [34]. Reddit has also grown rapidly, doubling its userbase from 2019 to 2024 [90].

If you are unfamiliar with it, Reddit is a platform for online communities, which on Reddit are called subreddits. Although Reddit users (often called Redditors) use a single login and username for every subreddit they participate in, each subreddit is separate from one another, with its own unique topic, homepage, members, moderators, rules, culture, and identity. Reddit consists of several hundred thousand active subreddits, and anyone can create their own new subreddit. As a result, subreddits, which are prefixed with `/r/`, are famously available for nearly every conceivable topic, no matter how niche. For example, /r/sourdoh is an entire community dedicated to posting photographs of botched attempts at baking sourdough bread. However, not all subreddits are as narrowly focused as this, and many subreddits with broader topics are quite large. /r/science, for example, has almost 35 million community members as of the time of this writing. If /r/science were a sovereign nation, it would be only slightly less populous than Canada. The enormous number of subreddits, their relative independence, and the incredible diversity of subreddit size, topic, and governance practices all make Reddit a great platform to study online communities.

Compared to many other platforms, Reddit has an additional advantage that makes it an appealing research subject: it is relatively easy to access large quantities of Reddit data, and, in an era when more and more platforms are locking down access to their content, this is noteworthy. Reddit has also restricted access to their content in the past year [204], yet the vast majority of communities on Reddit are publicly accessible, and it is still possible to download Reddit posts and comments in bulk [20]. Reddit also has a relatively permissive API, and does not require a user to be signed in to access community metadata like lists of moderators or rules. With clever scraping techniques, it can even be possible to recover content that has been removed by moderators or deleted by users [37, 153], although such methods are not without their own ethical concerns. There is some data that is not possible to readily collect, however, particularly detailed moderation logs, and individual users' voting and browsing behavior. Data donation strategies have been used to study these phenomena, although this poses an additional barrier to researchers [165, 166, 88].

All of these factors combine to make Reddit a wildly popular venue for online community research [189]. It is perhaps the most extensively studied social media platform in existence, with one 2021 review identifying 727 academic publications which make use of data from Reddit [213]. Reddit serves as a 'model organism' of sorts of computational social scientists, and Reddit has many commonalities with other platforms that make results from Reddit highly generalizable. Facebook Groups are the most similar widely used platform to Reddit, with many analogous affordances: explicitly defined communities with formalized membership, and volunteer moderators who can set and enforce rules.

In the design space of online community platforms, the primary cleavage is between Reddit and Facebook Groups *vs.* more user-focused platforms such as Twitter, Instagram, and TikTok. This latter class places a much greater emphasis on the 'following' of specific users, not specific communities. As such, on these platforms, while different communities clearly exist, they are not as explicitly formalized as on Reddit and Facebook Groups. Nonetheless, they still have many similar affordances (*e.g.,* voting mechanisms for content curation, the ability to reshare or crosspost content to other people/communities), and so findings from Reddit are still highly relevant to these other platforms.

There is one class of online community that deserves a mention here by way of differentiation, and that is online *peer production* communities. In this dissertation, I focus on social media communities, which exist to serve their members common interests or identities. Peer production communities, on the other hand, have

a different purpose: to support the group production of a large artifact or project, such as Wikipedia, or open source software (OSS) communities. There are many valuable insights to be learned from studying peer production communities, and indeed a large body literature exists focused on exactly that [142, 110, 134, 109, 145, 206, 293, 251], but, for the purposes of this dissertation, I consider peer production communities to be out of scope.

The last, but certainly not least important reason I choose to study Reddit, is my own personal experience on the platform. I began moderating my first subreddit, the photocritique subreddit, nearly 10 years ago, and it is this experience that pushed me to study community governance for my dissertation research. In the time that I've moderated /r/photocritique, it has grown from a community of fewer than 40 thousand members to the nearly 2 million strong community that it is today, and simultaneously, I have risen through the ranks of its moderators to my current position as the most senior active mod. During that time period, I also was invited to moderate a handful of other, smaller subreddits, all also photography related. As my work combines my personal and professional interests, I have always enjoyed sharing a summary of my research results with other moderators, and I find it very gratifying to hear their feedback and how their own moderation experience aligns with my results. In large part because of these interactions with other moderators, three years ago I was nominated to join the Reddit Moderator Council, a group of a few hundred experienced moderators who provide feedback to Reddit's internal teams and advise them on product decisions. Being on the Reddit Moderator Council has provided me with opportunities to connect with moderators from more than 87 communities with more than 115.4 million combined members, and the perspectives that I am exposed to on the Council daily continue to inform my research and enable me to disseminate my results to a key audience.

1.5 The Challenge of Studying Online Communities at Web Scale

Online communities are inherently complex systems with many technical and human factors involved in their functioning. As such, studying online communities at web scale poses a number of critical challenges. While simple theoretical models online community members' behavior are able to demonstrate interesting phenomena and derive insightful conclusions [143], yet the utility of theoretical analyses in this domain is fundamentally limited by the sheer complexity of online community behaviors, and therefore by the need

for empirical work to validate theoretical results.

Data availability is a core challenge when studying online communities empirically. There is simultaneously too much and not enough data describing online communities. It is easy to download terabytes worth of posts, comments, and images submitted to Reddit, in volumes that make processing such data on one's personal computer a fraught proposition. However, these data are unstructured, and thus actually finding the data relevant to the research question at hand can be very difficult. On the other hand, many specific types of data that would be very valuable to researchers are unavailable due to platform restrictions and/or ethical concerns. For example, access to removed or deleted content would enable much richer study of content moderation practices, yet accessing such content is both technically challenging and ethically fraught, particularly in cases where it was deleted by its author. Access to detailed moderation logs would similarly let researchers understand how moderation work is performed and distributed among members of moderator teams, yet access to such logs is heavily restricted by platforms and most moderators are unwilling to share such data with researchers. To overcome these challenges, I leverage my connections with moderators to collect additional data, such as moderation logs from subreddits where I deploy new tools (§4.1).

Directly surveying community moderators and community members can overcome some of these challenges, yet collecting survey responses at the scale necessary to comprehensively cover the volume and diversity of existing online communities is an extremely expensive, labor intensive, and time consuming affair. The repeated surveying necessary to enable longitudinal analyses only increases these costs, and is entirely impossible to apply retroactively, to study historical data. Despite these challenges, I conducted large scale surveys to collect valuable data and validate the performance of automatic methods (Chapter 2).

Even when data is available, identifying mere correlations between interventions and outcomes of interest is not enough to produce robust and actionable results, due to the presence of confounding factors. The gold standard for establishing causality is with a randomized controlled trial (RCT), yet the active experimentation necessary to conduct an RCT is once again often extremely impractical or even impossible due to ethical concerns, the inability to manipulate the treatment, or the scale necessary to study interventions at the community level (as opposed to user- or post level interventions).

Due the the challenges associated with RCTs, researchers often turn to methods which enable causal inferences with observational data. Two methods which I apply later in this work include difference in dif-

ference analyses (DID) and inverse propensity of treatment weighting (IPTW). DID leverages longitudinal data to examine how outcomes change over time, using the temporal changes in a control group to estimate a counterfactual. IPTW involves using observed covariates to model confounding factors, and then adjusts for these confounding factors by up- and down-weighting individual observations to simulate a randomized control where treatment assignment is independent of the covariances [13]. The challenge with these causal inference methods is that it is extremely difficult to evaluate their reliability in practice, as ground truth data is almost never available. Although it is out of scope for this dissertation, I have developed the first ever evaluation benchmark for causal inference methods using text data (as you would find in many online communities) [280]. In response to these challenges with causal inference methods, I collaborate with communities to run RCTs for interventions that I have developed, to ensure the validity of my results (§4.2).

Chapter 2

Characterizing the Needs and Values of Online Communities

We want to make online communities 'better,' but what does 'better' actually mean? Lots of research works on strategies to improve online communities, but much of this work focuses on mitigating specific harms, such as bullying, abuse, or harassment. These are important harms that they deserve the attention they receive, but there is more that draws members to online communities than just the absence of harms. In this chapter, we develop of foundational understanding of what community members want and need, by *characterizing* their values for their communities through surveys of thousands of community members.

To start, in §2.1, we develop a taxonomy to capture the full breadth of things that online community members value for their communities, in community members own words. Using 1,481 open-ended responses from 212 members of 627 unique communities on Reddit, we develop the taxonomy using iterative categorization. The resulting taxonomy contains nine high-level values, divided into 29 subcategories. We then discuss how tension between these values poses challenges for community moderators and platform designers, and highlight the values that are poorly understood and understudied.

With this new understanding of *what* community members value, next, we next empirically measure *how* these values vary among different members of the same communities, as well as between different communities. In §2.2, we present the largest-to-date survey of community member values on Reddit, encompassing over 5,000 responses from 2,769 members across 2,151 unique Reddit communities. We utilize these data to

evaluate how values differ in importance across communities of varying topics, and how community members currently perceive their communities and desire change within them. We analyze how these perceptions and preferences differ between communities of different ages and sizes. By gathering multiple responses from different members within the same communities, we are able to identify areas of consensus on values as well as points of disagreement. We demonstrate that community members have different perceptions of safe their communities are, that established communities place 30.1% greater importance on trustworthiness than newer communities, and that community moderators prefer their communities to be 56.7% less democratic than non-moderator community members. We discuss the implications that these results have for community design and moderation.

A strong Sense of Community (SoC) is perhaps the single most important ingredient for community success, yet measuring Sense of Community in online communities remains challenging, and how different community behaviors relate to SoC is poorly understood. In §2.3 we focus specifically on SoC in online communities, which we refer to specifically as Sense of *Virtual* Community (SOVC). We present the results of another large-scale survey of 2,826 Redditors, which incorporated some new questions alongside our earlier values survey, and was conducted in collaboration with researchers at Reddit. We identify three primary dimensions of SoC on Reddit: Membership & Belonging, Cooperation & Shared Values, and Connection & Influence. We develop a hierarchical linear model to predict how SOVC varies along with conversational structure and style across 281 different communities. This work highlights actionable strategies for fostering stronger community attachment, employing an approach that can be readily generalized across community topics, languages, and platforms.

2.1 A Taxonomy of Values for Online Communities

2.1.1 Introduction

There are many reasons why community members choose to participate in online communities [108]. Online communities can connect people across long distance, support shared interests or identities, and provide entertainment, education, and companionship. However, much research seeking to improve online communities focuses on specific harms such as spam, bullying [173, 164], or misinformation [294, 94]. However,

truly understanding how to make communities 'better' requires going beyond simply mitigating and minimizing harms and going towards an understanding of community *values*. Determining a community's values is a non-trivial challenge, as communities have many stakeholders with divergent preferences.

The term 'values' can have many meanings; here we use the definition from Value Sensitive Design: 'what a person or group of people consider important [to their online community]' [79, pg.2]. In this definition, values not only have a *topic*, *e.g.,* the diversity of the community, but also a *preference*, *e.g.,* a preference for *more* diversity. Value Sensitive Design is a design framework which underpins our work. At its core, Value Sensitive Design suggests that values must be considered when designing in contexts such as online communities [79].

Understanding community values is challenging because values can vary widely both between and within communities, and values can conflict with one another. A community focused on mental health support may want to foster inclusion more than a community focused on financial trading, a difference in values between communities. Within the same community, members may have different value preferences regarding the same value topic, *e.g.,* one community member may desire more diversity, while another may wish the community was more homogeneous. Someone who is a member of two (or more) communities may even have different preferences in different communities; that same person may prefer more diversity for one community and more homogeneity for the other. Lastly, values of different topics may conflict with one another as well. For example, while an online photography discussion community may desire to create a space that is welcoming to beginners, this value conflicts with the same community's desire to hear particularly from expert photographers who may be perceived as having the most to contribute to the conversation. These differences and conflicts are a critical consideration for researchers as well as community moderators and members.

Although much has been published on positive aspects of online communities [150], previous work which seeks to make online communities better has largely focused on specific aspects of online communities, especially those which are widely agreed to be harmful, such as harassment, rule-breaking, and misinformation. However, upon deeper inquiry, even these more commonly studied harms are quite complex, with substantial disagreement within and across communities regarding the extent to which these ostensibly harmful behaviors should be tolerated [123, 238].

Implicit in much of this prior work is an assumption of communities' and their members' values, yet exactly what values communities hold has not yet been comprehensively studied. While we do not argue that any one set of values is superior to others, we believe that a critical first step towards improving online communities is developing a comprehensive understanding of community members' own values, not just obvious ones that researchers have assumed all communities care about.

In this work, we survey Redditors to answer the research question "What values do community members hold for their communities?" Through a series of advertisements placed on Reddit, we recruit 212 people and collect 1,481 free response answers to questions about the values they hold for the 627 unique communities they consider themselves a part of (§2.2.3). To the best of our knowledge, this is the first such survey to gather community members' values in their own words.

We apply an iterative categorization methodology to produce our primary contribution: a taxonomy of community values with nine top-level categories and 29 subcategories (Table 2.1). We validate this taxonomy with a held-out set of 1,180 additional responses to demonstrate saturation, and achieve very high inter-rater agreement (Fleiss kappa $= 0.874$) using a codebook which we make public to support future research (Appendix A.5). Among the other key findings we contribute, we find that online social community members' values cover a broad range of categories from technical features to the diversity of community members, and that the quality of content submitted to communities is the most frequently reported value category (§2.1.3).

Our work enables new research to improve online communities (§2.1.4). In addition to working on what they consider valuable, researchers seeking to improve online communities should work on values we found to be frequently held yet understudied. Open research challenges with regard to understudied values include measuring the quality of content in the specific context of different communities, and the difficulty of growing communities' membership while maintaining meaningful community interaction (§2.1.4). Now that we have delineated a taxonomy of values, we can also begin to examine how values relate to each other. We argue that some widely held community values are inherently in tension, such as maintaining the quality of content in a community while simultaneously including new members, and call for research into community design with these tensions in mind (§2.1.4). Finally, we call for additional work on reconciling differences in values within communities, including community governance that protects the values of

vulnerable community members (§2.1.4).

2.1.2 Survey and Categorization Methodology

Reddit Background & Context

In this work, we focus on Reddit. Reddit is an ideal platform for conducting research on the values of online communities, as, unlike other social media platforms such as Twitter, Reddit is explicitly divided into thousands of unique communities, known as 'subreddits,' each with their own topics, rules, moderators, norms, and enforcement practices. Furthermore, almost all content on Reddit is publicly available [20], and Reddit has been widely examined by the research community [189]. On Reddit, subreddit's names are prefixed with `/r/`, and we adopt this notation in the rest of this paper.

Survey Instrument

All responses were gathered through an online survey hosted on the Qualtrics platform. The survey is summarized here and included in its entirety in Appendix A.4. The survey consists of five sections: (1) informed consent, (2) demographic questions, (3) general questions about the respondent's usage of Reddit, (4) subreddit specific questions, and (5) reflection questions. All questions except the informed consent question are optional.

Before any other questions are posed, the participant is shown a brief summary of the nature and aims of the survey, along with IRB information (for more details, see §2.1.2) and is asked for their consent. Then, in the demographics section, the participant is asked to describe their age, gender identity, and racial identity. Due to the challenge of obtaining parental consent over the internet, minors (people under the age of 18 in our jurisdiction) are excluded from participating.

Next, the participant is asked to optionally provide their Reddit username, which is used to query the Reddit API for a list of subreddits in which the participant recently posted or commented, and as a contact point for the raffle (§2.1.2).

Once the general Reddit questions are answered, the participant is shown a list of the five most recent subreddits they posted or commented in (Appendix A.1), and asked to remove any subreddits that they do not consider themselves a member of, and to add any subreddits they consider themselves a member of that

are missing from the list. Participants who decline to provide their username or whose username was not found on Reddit are presented with an empty list for them to populate themselves, and entered subreddits are checked in real time against the Reddit API to preclude spelling errors and to ensure the subreddit exists.

The next section is dedicated to subreddit specific questions. It consists of two free-response questions which are asked in turn for every subreddit listed by the participant in the previous step: *As it exists right now*, what are a few of the best aspects of the /r/<subreddit>community? *If you could change anything*, what are some aspects of the /r/<subreddit>community you would like to improve upon?

We chose to use these two questions to elicit participants' values from their specific, real experiences, rather than asking participants to speculate about abstract cases. To this goal, we chose to ask about positive aspects of the community as well as negative aspects, as these two questions elicit different values: the positive question offers insight into participants' values which are well-implemented by their communities, whereas responses to the negative question show which of the participants' values are less accommodated by current community practices. Asking both of these questions allows us to gather the broadest possible set of values for categorization, crucial to developing a comprehensive understanding of community values.

The survey was piloted with 13 participants from several departments in two large American universities. All 13 pilot participants denied having any difficulty understanding any of the survey questions or interacting with the online survey tool.

Participant Recruiting and Incentives

The 212 participants in this study were recruited primarily through Reddit advertisements. These advertisements display inline with other content in both individual subreddits and aggregated content views, and are shown on the Reddit website as well as the official Reddit mobile app. Appendix Figure A.2 shows the appearance of these advertisements. To increase the diversity of recruiting, the survey was also distributed to relevant university mailing lists and Slack channels at two large American universities, as well as posted to /r/SampleSize, a subreddit for the distribution of surveys, recruiting an additional 41 participants. Participation was incentivized with a raffle. Additional details on recruiting pipeline attrition and the raffle are located in Appendix B.3.

Participant Demographics

In general, we find that the demographics of our 212 survey respondents match overall platform demographics closely [19] (summary of respondent demographics in Appendix A.3). As with Reddit's overall userbase, our respondents skew young, white, and more commonly identify as men. Compared to Reddit's overall demographics, the age our of respondents is similar, with 71% of our respondents being under the age of thirty. People who identify as non-binary or do not identify with man or woman are slightly overrepresented in our survey results compared to Reddit as a whole. When it comes to racial and ethnic identities, white and Black users are slightly underrepresented in our responses, while Hispanic and Asian responses are slightly overrepresented.

Iterative Categorization

Once the survey responses were collected, all free-text responses were divided into idea units [252], where each idea unit represents a distinct thought. For example, the response 'The content is educational and I like how the community is engaged' would be divided into two idea units: 'The content is educational' and 'I like how the community is engaged.'

Our taxonomy was produced using an initial set of 301 idea units gathered from the first 39 respondents to the survey, with the remaining 1,180 idea units held out for validation. Using a grounded theory approach [87], a team of five researchers worked together to iteratively categorize the initial idea units using an inductive coding method [174]. While respondents often expressed value preferences in their responses, only value topics were used to categorize idea units. The researchers worked independently to initially cluster similar idea units, then came together to resolve differences in clustering until a consensus was reached. The initial tentative clusters were assigned names and definitions to produce a working taxonomy, then the researchers collaboratively recategorized all idea units, creating and removing categories under group consensus. This process was repeated three additional times until the iterative process converged and no further changes were needed to satisfactorily categorize all idea units. Once this was completed, the researchers worked together to write a codebook (Appendix A.5) describing the taxonomy, which is hierarchical, with top-level categories and subcategories. When possible, idea units are assigned to the more specific subcategories, with top-level categories reserved for broad or vague idea units.

Inter-Rater Reliability and Validation of the Taxonomy

To validate the codebook, three researchers were trained on the codebook and independently labeled 100 idea units randomly sampled from the held-out set. Inter-rater reliability was very high, with a Fleiss kappa of 0.874 when considering all 29 subcategories. When measuring agreement on only the nine top-level categories, the three raters were in even greater agreement (Fleiss kappa = 0.902). This level of agreement would typically be described as "almost perfect" [162] or "excellent" [96], and demonstrates that the taxonomy categorization using our codebook is robust and largely unambiguous.

To validate the taxonomy, the 1,180 held-out idea units were coded by a single researcher. Every single one of the idea units in the validation set was able to be categorized using the taxonomy codebook, demonstrating saturation.

Ethical Considerations & Broader Impacts

In order to ensure the anonymity of our participants, we keep all their responses confidential. Access to responses was limited to only the immediate research team, and all responses quoted in the following sections have been paraphrased to ensure participant anonymity. All data collected in this study was provided by participants who were informed of the nature of the study, the potential risks of participation, and who consented to participate. We do not make any use of participants' public Reddit histories for any purpose, as the use of such data has been shown to make many participants uncomfortable [71]. This research was reviewed and approved by the University of Washington IRB under ID STUDY00011457.

2.1.3 Taxonomy of Community Values

Our taxonomy has two hierarchical levels—the lower level consisting of 29 specific subcategories grouped into nine top-level categories. These categories reflect the diverse range of value topics reported by our participants, ranging from *Technical Features* to *Diversity* (Table 2.1). Categories are grouped according to value topic, regardless of value preference, and thus may contain conflicting value preferences, *e.g.,* the Size category contains idea units from respondents who prefer both larger and smaller sizes for their respective communities. *Quality of Content* is the most frequently mentioned value category, with 47% of all idea units, while *Technical Features* and *Trust* are the least frequently mentioned categories, each with less than

	Subcategory Name	Example Idea Unit	Frequency	
Quality of Content	Quality of Content	"I like the content in /r/HistoryMemes"	31	677 total (46.5%)
	a. Personal Preferences	"[r/StallmanWasRight] has too many articles and not enough personal stories"	293	
	b. Education, Entertainment	"it's interesting to learn about other countries [on /r/AskEurope]"	247	
	c. Curation, Recency, Discovery	"[I like that] I can keep up to date with new releases on /r/hiphopheads"	60	
	d. Spam, Reposts, Bots	"[There are] too many reposts of things people [on /r/Embroidery]"	46	
Community Engagement	Community Engagement	"the connection between users is great [on /r/vexillology]"	50	275 total (18.9%)
	a. Quality of Interaction or Community as a Whole	"[I like that] people [on /r/gaidhilig] are quick to respond with helpful comments"	170	
	b. Connection, Universalization	"being on /r/teenagers shows me that I am not alone in what I am going through"	55	
Size	Size	"[I like how] /r/stocks is big"	2	144 total (7.8%)
	a. Volume of Content	"because [/r/AskReddit] is so large, lots of interesting questions come up"	76	
	b. Size of Community	"[I like how /r/roguelikedev is a] tiny community with a clear purpose and scope"	36	
Participation & Inclusion	Participation & Inclusion	"[I like that /r/CrossStitch] is so inclusive"	16	93 total (6.4%)
	a. Offensive, Abusive, Harassing Content or Behaviors	"I don't like some of the sexist jokes [on /r/ProgrammerHumor]"	58	
	b. Outsiders, Demographics, Limits	"[I think] too many people from outside Seattle comment on posts [in /r/Seattle]"	12	
	c. Tools for Participation	"[/r/snails] should add an easily accessible 'beginners' questions' section"	7	
Diversity	Diversity	"[I like that /r/askscience is so] diverse"	1	76 total (5.2%)
	a. Variety of Content	"[I like how] I get to see different isopods [in /r/isopods], both wild and captive"	56	
	b. Diversity of People	"[/r/AskCulinary is] a nice mix of pros, experienced home cooks, and newcomers"	19	
Mods	Moderation & Moderators	"[I like how] mods of /r/goodanimememes don't change rules without a vote"	69	69 total (4.8%)
Norms	Norms	"[I dislike that] /r/alberta has weird rules"	3	59 total (4.1%)
	a. Adherence to Norms	"[I don't like that] people don't read the [/r/whatisthisplant] FAQ before posting"	46	
	b. Voting Behavior	"I wish people [on /r/ACMilan] wouldn't use downvote as a disagree button"	10	
Technical Features	Technical Features	"[It frustrates me that] posts [on /r/skiing] never load properly"	27	53 total (3.6%)
	a. Flair, Tags, NSFW labels	"flairs [on /r/ImmigrationCanada] are very useful here"	12	
	b. Search, Filters	"I wish I could filter by genre of game [on /r/ShouldIbuythisgame]"	11	
	c. Recommendation Systems	"I would like more recommendation measures [on /r/lotrmemes]"	3	
Trust	Trust		0	40 total (2.7%)
	a. Knowledgeable People	"[I like how /r/eyetriage has] real doctors in the community, so it's more reliable"	22	
	b. Trustworthy Content	"[I like how /r/nyu] is legit because students are honest in their opinions"	18	

Table 2.1: Summary of our taxonomy of values, along with example quotes from participants. Although the table is sorted by frequency, we note that the most frequent values are not necessarily the most important.

4% of the idea units.

While we report counts for the number of idea units falling into each category and subcategory, we do not make quantitative claims or comparisons between different categories or different communities. Communities, each with their own subject matter, rules, and membership, are different from one another, which is why our survey is designed to ask participants about each community individually. Rather than quantify the importance of any one category, our goal and contribution is to understand the broad diversity of value topics across communities in greater detail. The following subsections describe each taxonomy category in detail.

Quality of Content

The *Quality of Content* category is the largest of the categories that compose our taxonomy, containing 4 subcategories and 46.5% (677/1481) of all idea units derived from responses. It is unsurprising that most community members would have many values relating to the content of the communities they are a part of, as in most Reddit communities, sharing content is the primary mode of interaction with the community. We expect values in this category to be especially prevalent in communities that are less focused on engagement and connection with others and most focused on content itself (*e.g.,* meme sharing communities).

The four subcategories in this category (*Personal Preferences*, *Education/Entertainment*, *Curation/Recency/Discovery*, and *Spam/Reposts/Bots*) reflect the broad range of aspects of content and its presentation that are liked or disliked by community members. Many community members appreciate the curation and discovery of content that is provided by the community through mechanisms such as up- and down-voting, *e.g.,* "I keep up to date with new releases on /r/hiphopheads." The most frequently mentioned subcategory within the top-level *Quality of Content* category is *Personal Preferences*, which reflects how well (or poorly) the content aligns with personal preferences for specific types of content, such as memes ("the best part of /r/Terraria is the memes") or different subjects ("[in /r/BalticStates,] the posts about Estonia are the best because it's such a fabulous country"). These *Personal Preferences* are difficult to generalize because in most cases they are specific to the subject of the community as a whole.

Idea units regarding *Education/Entertainment* and *Spam/Reposts/Bots* tend to be more consistent from subreddit to subreddit. In general, content that is educational or entertaining is especially liked, *e.g.,* "/r/knitting teaches me new stitches and patterns", while *Spam/Reposts/Bots* are fairly universally disliked, and manifest more similarly in different communities. However, the exact nature of reposts and repeated content can vary somewhat from subreddit to subreddit, *e.g.,* "I wish users [of /r/AskReddit] would quit asking the same questions over and over just to get karma[1]."

Community Engagement & Interaction

The *Community Engagement* category contains 18.9% (275/1481) of the idea units derived from responses, making it the second most frequently reported value topic amongst respondents. We divide these idea

[1]Karma is Reddit's name for points gathered through the receipt of upvotes.

units into two subcategories: *Quality of Interaction or of the Community as a Whole*, and *Connection and Universalization*, *i.e.,* the realization that others exist with similar interests/feelings. The vast majority (256/275) of these responses were mentioned in a positive context, suggesting that Redditors mostly feel positively about community engagement.

Comments on *Quality of Interaction or of the Community as a Whole* mentioned both qualities of the individual interactions with community members (*e.g.,* "I often ask for help with language learning resources [on /r/gaidhlig], and everyone is always quick to respond") as well as qualities of the community as a whole (*e.g.,* "I appreciate the goodhearted nature of most of the people [on /r/Konosuba, a community dedicated to the eponymous Japanese novel series].") However, 19 participants were unhappy with the quality of the interactions in their communities, such as "[On /r/msu] there are often gatekeepers who comment on posts. However, these people tend to get downvoted quickly, so it's not that big of a deal." This community member then suggested that "a brief note about negativity in the sub's rules could help with this."

Connection and Universalization is especially oft-mentioned in communities for people with a common identity or interest who may be physically far apart or part of a minority group and therefore unable to connect as easily offline. One respondent wrote "[I love that /r/blackladies] is a community of black women coming together to discuss social issues that are prevalent and important to us." Another says "[/r/Glaucoma] lets me get in touch with people around the world [who are] dealing with a similar health issue." These quotes from participants reflect a body of literature that finds that online communities can be helpful venues for minority groups and those with special needs connect with similar people for support [215, 131].

Size

Many responses (144/1481) commented on the *Size* of communities, both regarding the *Volume of Content* submitted to the community and the number of people in that community (*Size of Community*). Participants were varied in their preferences, with some preferring a larger volume of content (*e.g.,* "I like that /r/assholedesign is regularly updated") and others preferring less content. One member of /r/SampleSize, a community for distributing surveys, wrote "The large volume of posts means that you need to time your submission very carefully in order to make sure that your survey doesn't get buried."

Similarly, respondents were fairly evenly split (21 preferring larger, 15 preferring smaller) on their

desire for larger or smaller communities. Those preferring larger communities perceived those larger communities as offering increased opportunities for interaction ("More active users [on /r/photocritique] would make it easier to engage with like minded individuals through their posts."), while those preferring smaller communities tended to like the specificity of subject matter and focus, *e.g.,* "[I like that /r/roguelikedev, a community dedicated to the development of a specific type of video game, is a] tiny community with a very clear purpose and scope." Community members' value preferences regarding community size seem likely to depend on many factors, especially including the subject of the community and the nature of interactions which occur within.

Participation & Inclusion

Participation and Inclusion in the community was a value topic reported by 93 respondents. We divide their responses into three subcategories: *Tools for Participation*, *Offensive/Abusive/Harassing Content or Behaviors*, and *Outsiders/Demographics/Limits*, which focuses on *who* participates.

Comments on *Offensive/Abusive/Harassing Content or Behaviors* were split (20 units to 38 units) between positive comments (praising the absence of offensive content), *e.g.,* "[I like that /r/WANDAVISION] has no homophobia, racism, or any discrimination," and comments that described the respondent's experience with such behavior that was detrimental to inclusion. One member of /r/sewing wrote "sometimes the commenters on a post will write personal things about a poster's appearance, which I don't think is appropriate."

Community members who commented on *who* participates in their communities (*Outsiders/Demographics/Limits*) were frequently concerned with the presence of outsiders, who are perceived as not belonging to the community by virtue of their lack of familiarity with the subject or even their physical location. One member of /r/Seattle wrote that "too many people from outside Seattle comment on posts in this subreddit." In any community with a specific subject, some degree of boundaries for membership are natural, yet too much insularity in online communities can be harmful [4].

Tools for Participation were suggested by seven participants who felt their communities were lacking such tools. One participant wrote "[/r/snails] should add an easily accessible 'beginners' questions' section." Some tools, such as automated posts and messages [289] and badges, have been found to reduce unwelcom-

ing reactions to new community members [235] as well as such members' compliance with rules [185].

Diversity

Many community members commented on the *Diversity* of their communities, which we divide into two subcategories: *Variety of Content* and *Diversity of People*. More respondents (56/75) commented on *Variety of Content*, *e.g.,* "[/r/CollegeBasketball] has an ideal blend of banter, rumors, statistical insights, and glorious shitposting." Those who commented on *Diversity of People* more frequently commented on aspects of the community they would like to change, *e.g.,* "I wish [/r/knitting] had more variety in skill. Right now it's mostly skilled knitters, whereas it would be appreciated to see some beginner knitters."

Moderation & Moderators

Moderation & Moderators are controversial subjects on Reddit, and many moderators feel strongly disliked by members of the community they moderate [181]. However, close to half (31/69) of idea units regarding moderation were positive, with respondents praising the moderator team in general ("mods [of /r/Phillipines] have done an excellent job of maintaining the community") as well as specific moderators ("<username redacted> is such a great mod [of /r/ApplyingToCollege] who is super helpful.").

Of the 38 negative comments on moderation, many of them were critical of perceived arbitrary rule enforcement, *e.g.,* "moderators arbitrarily remove posts because they're 'against the subreddit's rules'." Other idea units requested more active moderation (14 units), complained about perceived power abuses (8 units), and called for greater community involvement in rule making (2 units).

Norms, Voting Behavior, and Adherence

Almost every (54/59) response regarding *Norms* was on a negative aspect of the community, suggesting that norms are mostly noticeable in a community when they are violated. Many of the idea units relating to norms (46/59) were community specific and focused on *Adherence to Norms*, such as newcomers not reading the FAQ before posting, posters not providing adequate information and expecting a response (/r/whatisthisplant), or not including appropriate sources. Many respondents also requested additional rules or changes to norms (19 units), *e.g.,* "I would like to reduce the amount of people speaking English [on

/r/ich_iel, a German-speaking subreddit]."

The remainder of idea units (10/59) were complaints about *Voting Behavior*, particularly the use of the "downvote button as a disagree button."

Technical Features

53 idea units (3.6%) focused on *Technical Features*, with three subcategories: *Flairs/Tags/NSFW*[2] *Labels*, *Search and Filters*, and *Recommendation Systems*. Most (36/53) idea units in *Technical Features* focused on negative aspects of technical features, particularly such features' absence or failure to work properly. Most positive idea units praised their respective subreddits' use of flair and tags, *e.g.,* "NSFW tagging is perfectly used here [in /r/hemorrhoid]."

While flairs, tags, recommendations systems, filtering options, and quality search functionality can dramatically improve users' experiences in online communities, on Reddit, such technical features can be difficult for communities to implement and modify, as they are controlled by the Reddit administration, not community members. Frequently, incentives are misaligned between subreddit leadership and the Reddit administration, who may be unwilling to implement new technical features at the request of communities. As a workaround, many communities implement their own second-party technical features by modifying or repurposing existing features, such as custom community moderation bots [140] or reputation systems such those in /r/changemyview [119].

Trust

The *Trust* of a community was the least frequently commented on value category. This category consists of two subcategories, *Knowledgeable People* and *Trustworthy Content*, again differentiated by a focus on people *vs.* content. Of respondents who commented on the trustworthiness of people, one respondent appreciated the credentials of community members "There are many doctors [on /r/eyetriage] so it's a lot more reliable," whereas another respondent lamented how anonymity interfered with the trustworthiness of the /r/MachineLearning community: "Anonymity sometimes makes it so I don't know who is qualified to say what."

[2] A common Reddit acronym meaning 'not safe for work' and used to indicate content containing gore or nudity.

2.1.4 Implications & Discussion

Community members value a broad range of topics for their communities ranging from *Diversity* (§2.1.3) to *Technical Features* (§2.1.3). However, some of these topics, such as the *Quality of Content* (§2.1.3), are much more frequently reported by community members than others. We do not, however, suggest that these topics which are most frequently mentioned should be considered more important to *researchers* than less frequently mentioned topics, such as *Trust* (§2.1.3). Most existing research that seeks to make online communities 'better' focuses on implicit community values which are relatively narrow in scope, especially those with broader societal impact (*e.g.,* mis/disinformation and political polarization) and/or a negative impact on vulnerable populations (*e.g.,* abuse, harassment) (§2.1.5). While these research directions are critical to mitigating major harms associated with online communities, our findings suggest that many values held by community members are understudied, complex, and often in conflict with one another. In this section we discuss this complexity (§2.1.4) and conflict (§2.1.4), implications for moderation and governance practices (§2.1.4), and subsequently how our taxonomy relates to other taxonomies from different contexts (§2.1.5).

Understudied Community Values

Many values explicitly stated by community members have topics which are are not well studied by the research community. We find that *Quality of Content* is the most frequently mentioned value topic (§2.1.3), accounting for 46.5% of idea units in our responses, yet with the exception of spam and bot detection, this value is poorly understood. Quality of content is arguably the most challenging value topic to define and measure, as it is extraordinarily context-specific and dependent on the nature of the community in question. For example, content that is high quality in /r/catpictures would be woefully inappropriate for /r/science. As existing work in this space focuses mostly on understanding the quality of conversational (and other text-based) content [292, 18, 159], additional work which contributes to the understanding of the context-specific quality of image and video-based content is especially needed.

Community *Size* is another value category that is understudied and complex. Some work has studied the impact of rapid growth on communities [167, 141]; however, our findings suggest that members perceive a difference between changes in the size of the community (*e.g.,* the number of participants) and the volume of content (§2.1.3). While most respondents (47/76) prefer a larger volume of content, many respondents

(15/36) perceive smaller communities as better due to having stronger community engagement and interactions, *e.g.,* "as the community has grown [it] has lost its small and friendly community feel". Balancing this tension is a important area for future research, as is understanding how desired community size varies as a function of the community member's relationship with the community in question.

Conflicting Community Values

Our results show that community members hold a broad range of values, but these values are challenging to implement because values in different categories often conflict with one another.

Inclusion vs. Quality of Content and Norms. While the challenge of incorporating new members has been previously identified and studied [150, 48, 39], our findings permit the framing of this tension as a conflict between *Inclusion* (of new members) and *Quality of Content* and *Norms*. One participant expressed this sentiment in their response, saying "It's frustrating when [new members] tend to leave out information they're expected to include." In social media communities, some work has experimented with onboarding documents and mentorship [235, 185, 289] to improve new members' understanding of community norms and maintain the quality of content. Methods to mitigate the tension between inclusion, norms, and quality of content have been studied more deeply in the context of peer production communities such as Wikipedia [43, 100, 99], and there is great potential for future work to study how these findings generalize to more social communities and develop new tools [150, Ch.5]. To an extent, however, these values are inherently at odds.

Size vs. Community Engagement. Community engagement and size are also often in conflict with one another. While several studies have found that, by some metrics, communities' health is not harmed in the long term by increases in size [167, 141], growth is a frequent subject of complaint across many platforms. One of our participants wrote "[/r/formula1 is] such a large community that is hard to engage with other members." On the other hand, many participants also reported desiring more activity and more frequent content in their communities. An open and important research question is how to scale communities to larger size without sacrificing the sense of interpersonal engagement.

Community Governance & (Lack of) Consensus

While the previous section described conflicts between different value categories, we also find evidence that members of the same community can disagree with one another's value *preferences* for the same value category. For example, eight participants who are members of /r/AskReddit expressed differing preferences for what content is permitted in the community. Three idea units indicated a preference for an 'anything goes' strategy, while the remaining four wished for more restrictions on specific subjects, mainly sex and drug use. Other research has found similar evidence that community members often disagree with one another on matters such as the severity of harmful behavior [238, 123]. Strategies for reconciling these differences of opinion, however, are not well understood. In the context of peer production communities such as Wikipedia, some research has already explored some sources of internal disagreement, such as tensions between senior and junior members [99, 100, 251], however the extent to which these findings generalize to social media platforms such as Reddit may be limited. Our results suggest that even relatively small samples of members are adequate to surface some differences in values between community members. Future work should examine the degree of agreement or disagreement on values held by many members from the same communities, and what factors are predictive of such differences in opinion.

Affordances for Participatory Governance. Participatory governance practices can help build consensus and reconcile differences in opinion when considering and implementing rule changes. However, affordances for such practices are extremely limited on social media platforms, where the vast majority of communities' rules are determined exclusively by a small set of moderators with no formal input from the broader community [291]. We find (§2.1.3) that many community members perceive lack of transparency and arbitrary enforcement decisions as evidence of corruption ("mods will ban you without warning if you say something they disagree with"), and desire greater input into moderation ("moderators should consult the community about what we want"). Conflict between moderators and the broader community has been identified in some prior work [181], yet the exact differences between moderators' and nonmoderators' opinions have not yet been studied and quantified systematically. Increasing participatory governance practices in online communities may help alleviate some of this conflict, yet such practices are not a panacea, and can cause harm if not implemented carefully, *e.g.,* by increasing the burden of labor on certain groups, or by giving a veneer of legitimacy to unilateral decision making [135].

Methods for Managing Irreconcilable Differences. What happens when differences in values are too great to reconcile? On Reddit, it is not uncommon for some members to splinter and create an alternative subreddit in response to perceived grievances or frustration with the status quo, however, to the best of our knowledge, this 'exit' phenomenon has not been studied in depth [76]. Other communities, such as consensus-based peer production communities, have different practices to manage internal disputes such as formal arbitration committees [145], but it is unclear how such practices would generalize to social media communities. In some cases of divergent values, such as differing *Personal Preferences* for content, personalized filtering may be used so that each user does not see content they wish to avoid [122]. However, this undermines social translucence, a theory which suggests that making online behavior visible creates social spaces with shared accountability [63]. Some research has explored interventions to balance the trade-off between social translucence and the personalization afforded by filtering [84].

Power Structures and Protecting Vulnerable Community Members. Furthermore, it is likely that some of the categories of values we have identified in this work will be of special importance to vulnerable groups. For example, one member of /r/AskWomenOver30 said "everyone assumes I'm a white American, which really changes the dynamic when I ask about career, relationships, and more." In this case, only a small fraction of the community may recognize that these assumptions are harmful to the experience of some community members, but that minority perspective is still very important. If traditional participatory decision making practices such as voting were implemented naively, the majority (which may not have had personal experience with harms such as the aforementioned assumptions of members' background, harassment [183], *etc.*) may not support or even be opposed to measures designed to reduce these harms, which disproportionately affect women and other minority groups [164].

This phenomenon, known as 'Tyranny of the Majority,' is well known in the offline governance context [95], yet has not been studied in-depth in the online context. Some peer production communities, such as Wikipedia, emphasize consensus-based deliberation over vote-counting partly as a way to avoid this issue [286, 110], though issues with bias against women and other minority groups still persist [269]. Additional research into participatory governance practices for online communities which protect vulnerable groups is sorely needed.

At a higher level, much research on online communities studies these communities through the lens of

moderators [181] and other people in positions of power [150, 38, 258, 97]. This is especially the case for empirical work which relies upon rules and enforcement actions as concrete evidence of behavior and community norms [120, 115, 37, 72]. While these rules and enforcement actions are natural sources of empirical data, researchers must take care and recognize that many community members feel as though they do not have an adequate stake in rulemaking and enforcement (§2.1.3). Online platforms' power structures are complex [116], and while surveys such as this one are useful tools, additional work is needed on scalable methods for empirically studying community behavior and values that include the voices of the least empowered.

Limitations

While we made intentional efforts to recruit participants from a diverse set of backgrounds by using multiple recruiting methods, it is possible that selection effects impacted who responded to our survey, which may have resulted in a taxonomy that is not truly representative of all community members. Furthermore, we do not include participants younger than the age of 18. Lastly, while our taxonomy derived from responses relating to a large set of 627 diverse communities, we only have responses from a small fraction of each community's membership. These facts, along with our (in absolute terms) low click-through rate of 0.227%, are indicative of a fundamental reality for researchers of online communities: it is very difficult to effectively poll an online community. Additional research is needed to improve polling methods, *e.g.,* by reducing friction in online surveys. Improved methods could strengthen results' representativeness and susceptibility to bias from sources such as poorly defined community membership.

We believe the validity of our taxonomy is demonstrated both by the large (1,180 idea unit) validation set used to demonstrate saturation (§2.1.2), the high inter-rater reliability (Fleiss kappa = 0.874) of the codebook, as well as the correspondence between the values reported by our participants and those identified by sociologists and psychiatrists in the context of interpersonal interactions (§2.1.5). Making 'truly representative' samples of online communities' membership is especially difficult given that such membership is often not well-defined, yet this is an important area for future work.

Our participants were asked about their experiences on a single social platform, Reddit. While Reddit is large, popular, and contains many thousands of communities covering diverse subjects [189], Reddit has

important differences from other platforms. Unlike Reddit, platforms such as Twitter and Facebook lacks explicit communities with well-defined membership. reddit has a stronger focus on link-sharing than other platforms, in part due to technical differences and in part due to history and the culture on the platform. reddit is also almost entirely public, unlike private communities on Slack, Discord and some Facebook Groups.

Reviews of research conducted using Reddit data have found that the generalizability of results from Reddit to other contexts are promising, with certain caveats, and this is an important area for additional research [213]. However, we believe that the size and prominence of Reddit means that our work still has an important impact, regardless of generalizability.

reddit also differs dramatically from peer production communities such as Wikipedia and open source projects, where the community has a clear focus beyond social engagement and entertainment. Future work is needed to understand how community values may differ in these other contexts.

reddit also mostly consists of English-speaking users from Western cultures, and additional work is needed to understand how people from other cultures may hold different values. Some evidence for cultural differences in values has already been reported [123].

2.1.5 Related Work & Comparison to Existing Taxonomies

Community Rules, Norms, & Content Moderation. One way in which commonly-held community values manifest is in the formalized rules of communities and how those rules are enforced by content moderators. Rules on Reddit have been examined in prior work [72, 37], however, as subreddits' rules are written and enforced exclusively by community moderators, they may not be representative of the values of the broader community membership [181].

On Reddit, as on most social media platforms, more democratic self-moderation is relatively rare [240], with platforms' technical features built mostly around a strictly defined hierarchy of admins and moderators, encouraging an 'implicit feudalism' [239]. In these cases, conflict between non-moderator membership and the moderators over rules and their enforcement is relatively common [181, 120, 246]. While some third party tools to enable a broader range of governance practices have been proposed [291], these tools are not yet widely adopted. In this work, we directly ask all community members about their values, not just

This Work Iterative categorization of 212 Redditors' survey responses	**Bao et al. [18]** Survey of Prosocial Behavior literature	**Deri et al. [55] and Choi et al. [42]** Survey of Sociology & Psychology literature	**Fiesler et al. [72]** Manual coding of 300 rules from 18 subreddits
Quality of Content	Information Sharing	Knowledge, Fun	Content/Behavior, Format, Low-Quality Content, Off-topic, Reposting, Spam, Advertising/Commercialization, Images
Community Engagement	Social Cohesion, Social Support, Gratitude, Mentoring, Esteem Enhancement	Power, Status, Support	Personality
Diversity	Social Cohesion	Similarity, Identity	Off-topic
Size	Social Cohesion	N/A	N/A
Participation & Inclusion	Social Support, Mentoring, Absence of Antisocial Behavior	Trust, Identity, Conflict	Doxxing/Personal Info, Harassment, Hate Speech, Trolling, Links & Outside Content
Technical Features	N/A	N/A	NSFW
Moderation & Moderators	N/A	Power, Status	Consequences/Moderation/ Enforcement
Norms	Absence of Antisocial Behavior	Trust, Conflict	Format, Voting
Trust	Information Sharing	Trust	Content/Behavior
Not mentioned by community members	Fundraising & Donating	Romance	Copyright/Piracy, Personal Army, Politics, Sitewide

Table 2.2: A comparison of how categories from our empirically-derived taxonomy are mapped onto by taxonomies from prior work on small group interactions [18, 55, 42] and community rules [72]. All of our categories have analogues in these other taxonomies from different contexts.

moderators.

Intracommunity Tension and Conflict.

Conflict between non-moderators and moderators is not the only form of tension within online communities. One major challenge in communities is that of integrating new members into an existing community [150, Ch.5]. This has been studied empirically on a wide range of platforms, finding that new members mostly learn community norms from experience [39], and that once established in a community, members are less likely to change their habits [48]. As a result, periods of massive growth can result in substantial change to the community [167, 100] and frustration amongst existing members [141]. In communities that are especially focused on topics requiring special knowledge or expertise, a related tension often occurs

between those with greater knowledge and those without; this has been studied on Reddit in case studies for science [126] and history communities [85]. In cases of extreme intracommunity conflict, such tensions can even lead to fragmentation as some members exit [76] and form new, alternative communities [70, 195]. Additional evidence for intracommunity tension can be found in work that examines how members define rulebreaking and how to fairly punish such behavior; this work has found substantial disagreement amongst community members [238, 123]. Our taxonomy enables us to frame these tensions as conflicts between different categories of values (§2.1.4), or as differences in members' value preferences in the same value category.

Implicit Values in Prior Research. Any research that seeks to improve the health of an online community implicitly (and often explicitly) values certain aspects of that community. For example, the abundance of research on reducing misinformation [26, 282, 94, 10] implicitly values the veracity of content. Similarly, research seeking to reduce harassment [30, 183] implicitly values safety. This 'implicit values' perspective can be used to identify what values of online communities are most studied by the research community, and conversely, what values are most understudied (§2.1.4).

In this work, we do not argue that the set of values derived from community members' responses is superior to any other set of values, but instead conduct a survey to outline the diversity of values held by community members, how they may be in conflict with one another, and how they relate to prior research and can inform future research.

Comparison to Existing Taxonomies. While we are not aware of any work prior to ours that directly asks members of online communities their values, aspects of online interactions have been studied in the context of 1:1 or small group interactions (Deri et al. [55, Table 1] provides an overview). We summarize how the major categories from our results (§2.1.3) relate to categories in taxonomies of aspects of small group interactions [18, 55, 42] and the taxonomy of subreddit rules from Fiesler et al. [72] in Table 2.2.

Every major category in our taxonomy has at least one analogue in the small group interaction and rule taxonomies. Naturally, given the different contexts, there are differences in how and where the categories overlap, and some components of the other taxonomies are not as relevant to the large online community context we study here. The small group taxonomies emphasize dynamics that affect 1:1 interactions, such as gratitude, mentorship, and asymmetrical power dynamics. In large communities, these translate to the

quality of community engagement and interaction, the inclusion of new members, and the power wielded by moderators, respectively. Romance and Fundraising/Donating, which appear in the small group taxonomies, do not appear in our taxonomy and were not mentioned by any of our 1,481 participants' idea units.

The taxonomy of different rules on Reddit [72] also overlaps substantially with our taxonomy, as rules are a formalized reflection of community values. As rules are primarily written with regards to content [72], our *Quality of Content* category is relevant to many categories from the rules taxonomy. Categories from the rules taxonomy which do not have analogues in our taxonomy are rules regarding copyright/piracy, politics, and 'personal armies' (brigading, when large numbers of members from one community temporarily participate in an often hostile manner in another community).

2.1.6 Conclusion

Online social communities are rich spaces that can bring people together in a healthy, productive, and enjoyable manner. Many researchers study how to make online communities 'better,' but understanding what 'better' means is a challenging problem, as there is no single set of values for online communities that can be used to inform research in this space. The values held by community members themselves are difficult to measure, and their perspectives have mostly not been included in existing research.

In this work, we surveyed 212 Redditors who are members of 627 unique communities. Using open ended questions (§2.1.2), we asked these Redditors what their values for their communities are, in their own words. Using an iterative categorization method based in grounded theory (§2.1.2), we contributed a taxonomy of 29 subcategories of community values across a broad range of topics from the diversity of the community to technical features (§2.1.3). Raters using our codebook demonstrate very high inter-rater agreement (Fleiss kappa = 0.874).

Our findings have important implications for future work on online social communities, and have already enabled followup work [284, 175]. We highlighted understudied and challenging-to-implement community values such as *Quality of Content*, *Size*, and *Community Engagement* (§2.1.4). We identified where community values conflict with one another (§2.1.4), and called for additional work on participatory governance for online communities that protects vulnerable groups of community members (§2.1.4).

2.2 Measuring Values, Consensus, and Conflict Across Thousands of Online Communities

2.2.1 Introduction

The values that community members have for their communities, and the value that those communities offer to their members, are extraordinarily varied [283]. Given the immense diversity of online communities, it follows that there is no 'one size fits all' approach to making communities 'better' [283]. 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 [72] or removed content [37]; however, these approaches only capture values as implemented by moderators [181], 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 (§2.2.3) previously developed from qualitative user studies [283], 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 (§2.2.3). 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 [283].

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 (§2.2.3). Using these data, we answer four research questions:

RQ1 What are communities' values, and how do they vary across communities? (§2.2.4)

RQ2 Within communities, where is there disagreement over values? (§2.2.5)

RQ3 How do moderators differ in their values from non-moderator community members? (§2.2.6)

RQ4 To what degree can community values be predicted based on automatically measurable features? (§2.2.7)

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 [211, 92, 226] 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 (§2.2.4). We examine differences between newcomers and senior community members in the context of literature on the challenges of managing community growth and new members [167, 48] and find that, on Reddit, new members are more positive in their perception of the current states of their communities than more senior members (§2.2.5). Given that governance on Reddit is often characterized by divisions between moderators and non-moderators [181], 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 [291] (§2.2.6).

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 (§2.2.7). 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[3].

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

2.2.2 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. [72], who produced a taxonomy of 24 different types of rules in use across 1,000 subreddits. These rules are enforced by volunteer moderators [181], and in some cases, content removed by moderators for violating the rules can be recovered and used to characterize community norms [37]. 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 [291]. 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 [246, 120]. In contrast, our method of explicitly surveying both moderators and non-moderators enables us to directly measure the differences between these two groups (§2.2.6).

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 [291, 181]. On Reddit, while moderators are beholden to platform administrators [116], 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 [100], or StackExchange, which holds formal elections. While some systems to incorporate democracy into Reddit have been developed [291], such systems have not been widely adopted, and moderators often face conflict and accusations of corruption [181]. In this work, we ask community members about their perceptions of democracy, and examine how moderators' and non-mods' responses differ (§2.2.6).

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 [150, 99]. 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 [48, 167], or a qualitative interview

with a small number of participants, focusing only on a single subreddit [141, 39]. 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 [211], 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 [226, 92]; in contrast, our work explicitly surveys community members on their values.

2.2.3 Methods

Measuring Community Values

Central to this work is the set of nine values around which we design our survey instrument (§2.2.3). The set of values we use is grounded in sociology literature on different dimensions of social relations [55, 18, 42] and drawn directly from the taxonomy developed by Weld, Zhang, and Althoff [283] via iterative categorization of unstructured survey responses from Redditors. The complete Weld, Zhang, and Althoff [283] 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 2.3. 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.

Reddit Background

Reddit is the fifth most popular social media site in the United States [249], and is an ideal platform for researching the values of online communities as Reddit is explicitly divided into many thousands of discrete

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 2.3: We leverage the taxonomy of widely-held values on Reddit by Weld, Zhang, and Althoff [283], which was developed through user studies and iterative categorization.

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 [20], and Reddit has been widely studied [189].

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[4]. 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 (§2.2.3) 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

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

listed themselves as a member of. For each subreddit, the participant is asked separately about nine different community values (§2.2.3). 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 May-July 2021, with a total of 2,769 people participating. Additional details on recruiting, participation, and compensation are included in Appendix B.3.

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.2) leveraging subreddits with a high number of responses, which indicated that subreddit averages have stabilized at this number of responses.

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 [20].

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.[5] For more details on this taxonomy, see Appendix B.1. The six topic categories we use are: **Hobby** communities *e.g.,* /r/nba, /r/bicycling, **Discussion** communities *e.g.,* /r/AskReddit, /r/relation-

[5]We additionally experimented with pre-computed subreddit embeddings [155, 179, 275], but these did not explain significant variation in values.

ship_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.

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 [71], which we do not publish. This study was approved by the University of Washington IRB under ID number STUDY00011457.

2.2.4 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 (§2.2.2), 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 Quality and Variety of Content and Size.

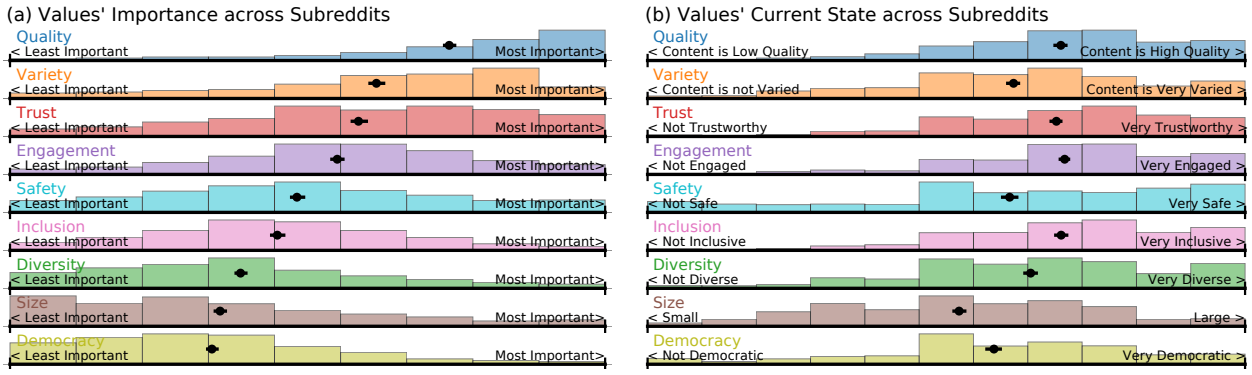


Figure 2.1: To understand what communities' values are, we average all responses for each community. (a) shows the distribution of the relative importance of each value across communities. Quality of Content is most frequently considered the most important value, while Size and Democracy are generally 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.

Method

To analyze how values vary across communities, we group communities based on their topical category (see Appendix B.1 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 (Figure 2.1). On average, Quality of Content is the most important value, with Size and Democracy generally considered the least important (Figure 2.1a). Safety is especially varied with regards to both its importance (Figure 2.1a) and current state (Figure 2.1b), 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.

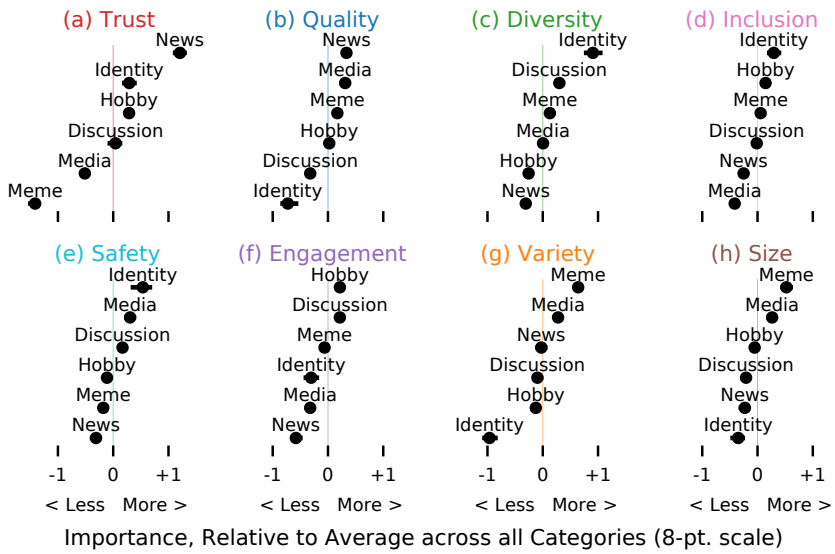


Figure 2.2: Differences in value importance across communities of different topics. News Communities rate Trust as 2.62 points more important than Meme Communities (out of an 8 point scale). Diversity and Inclusion are especially important to Identity-based Communities. Variety of Content and Size are especially important to Meme and Media-sharing Communities.

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 (Figure 2.2c,d), while Hobby and News Communities place greater importance on Quality (Figure 2.2b). 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 2.2g,h). Identity-based Communities rate Inclusion as 1.21 points (out of 8) more important than Media Communities (Figure 2.2d). It is important to note that Common Identity and Common Bond Theory [211] 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 (Figure 2.2a).

When examining the differences between new and older communities, and between small and large communities, differences are especially pronounced for Trust, Size, and Safety (Figure 2.3). 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 (Figure 2.3a), which is consistent with our hypothesis that older

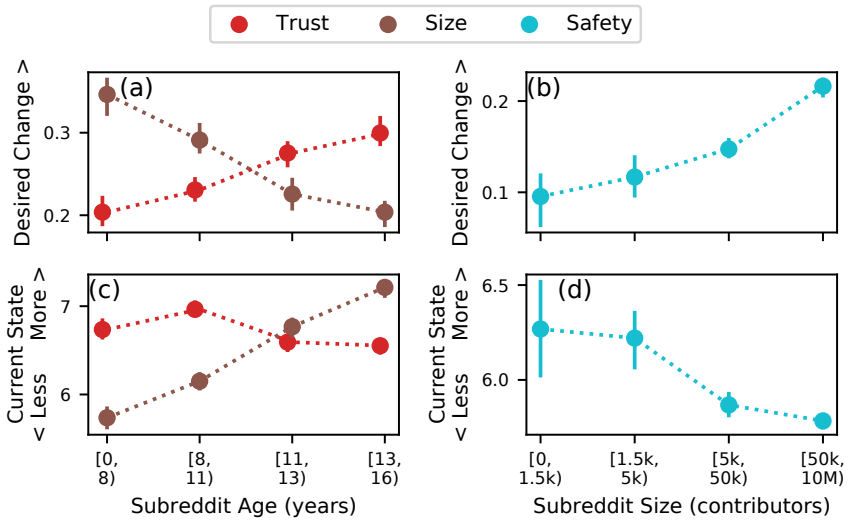


Figure 2.3: Average importance and desired change across community, binned into approximate quartiles by the age (since founding) and size. (a) Older communities 30.1% more strongly desire increased Trust than younger communities. (b) Larger communities have a 126.6% stronger desire to improve Safety than the smallest communities.

communities 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 (Figure 2.3c). 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 (Figure 2.3b), 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 (Figure 2.3d).

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 [283]. 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 [291, 135]. We examine Democracy and governance further in §2.2.6. Our finding that some communities consider themselves fairly safe while others consider themselves to be very unsafe (Figure 2.1) is consistent with previous

findings that toxic behavior on Reddit is extremely concentrated in a small number of subreddits [282], this could further support the practice of community level moderation interventions [115, 38, 97, 258, 36]. 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 [95], yet inherently members of minority groups are too few to significantly influence the community average. We examine this further in §2.2.5. Finally, our results emphasize the importance of Common Identity and Common Bond Theory (§2.2.2), which can guide researchers in their future work on this topic.

2.2.5 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 (Figure 2.4) before analyzing how different groups of Reddit users disagree with others.

Informed by previous work on vulnerable members of online communities [164, 176], 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 [141, 150].

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

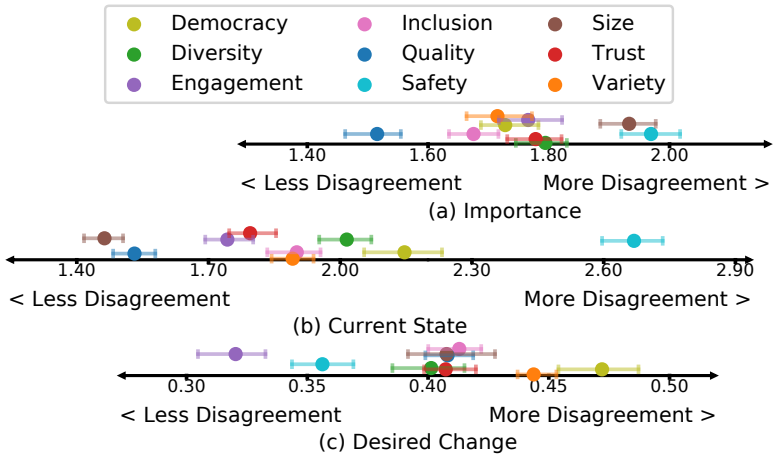


Figure 2.4: Average disagreement (measured with MAD) in perceptions of importance (a), current state (b) and desired change (c) across communities. Axes are adjusted for the widths of their respective scales, indicating greater disagreement over the importance of values than their current state and desired 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.

(called karma on Reddit) [88].

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 (Figure 2.4b,c). Interestingly, there is relative consensus on the desire to improve Safety (Figure 2.4c). There is strong consensus on the current state of Size (Figure 2.4b), while there is relative disagreement over the importance and desired change of Size (Figure 2.4a,c).

When examining differences between senior and junior community members, we demonstrate that junior Redditors are generally more positive in their perception of the current state of their communities (Figure 2.5), 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,

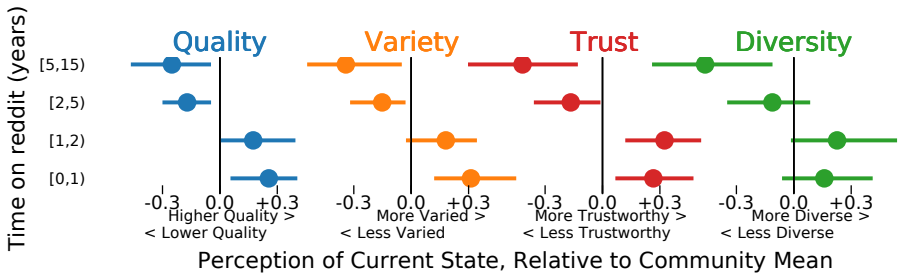


Figure 2.5: Differences in perception of the current state of communities between new Reddit users and those who have been on Reddit for longer. Generally, newer Reddit users perceive their subs to be 0.55 points higher quality, 0.71 points more varied, 0.74 points more trustworthy, and 0.68 points more diverse compared to older Reddit users.

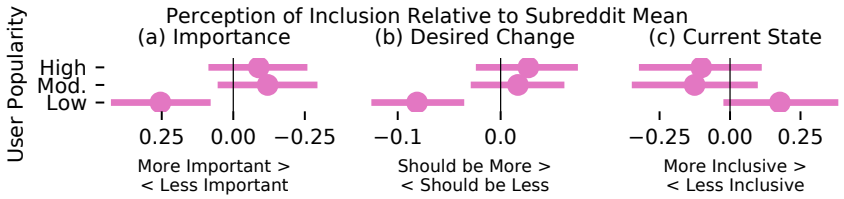


Figure 2.6: Differences in perceptions of Inclusion across less- and more-popular community members, as measured by account karma, divided into terciles. Relative to more popular users, less popular Reddit users perceive Inclusion to be 0.36 points less important (a) and have 0.10 less desire to change Inclusion (b), yet perceive their communities to currently be 0.29 points more inclusive.

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 Figure 2.1).

We find significant differences in the perception of Inclusion between less- and more-popular community members (Figure 2.6). 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 (Figure 2.4a,b) is a special concern that emphasizes the potential harm of community governance that only responds to the needs of the majority [95]. 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 (Figure 2.5), and correspondingly feel that their communities are the most inclusive (Figure 2.6c). 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 members, additional work is needed to reach users who ultimately decided to *not* join a community.

2.2.6 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 [181, 242, 240]. Importantly, moderators also constitute a major part of the governance of communities on Reddit [116], 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 [181] and that affordances for participatory governance are almost entirely absent from Reddit [291], 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.

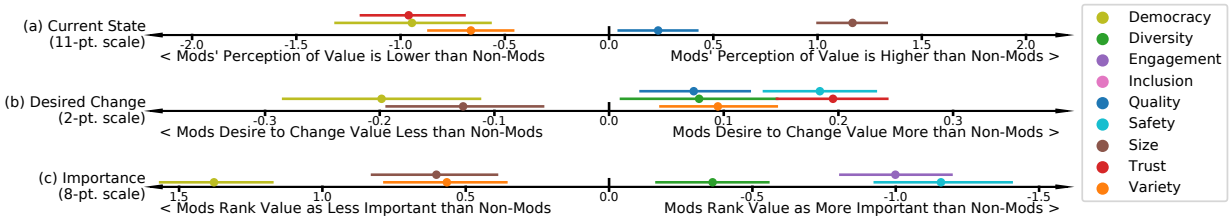


Figure 2.7: Differences in values between moderators and non-moderators. Moderators believe (a) their communities are 14.5% less democratic, (b) should be 56.7% less democratic, and (c) that Democracy is 23.6% less important, relative to the non-moderator mean in that community. Moderators rank Diversity, Engagement, and Safety as more important to their communities than non-moderator community members (c). Values with CIs overlapping 0 are removed.

Results

We find substantial differences between moderators and non-moderators across all three dimensions of each value: current state, desired change, and importance (Figure 2.7). 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[6]) *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× as many moderators report desiring their communities to be less democratic than non-moderators. These differences are not limited to Democracy; moderators also more strongly desire to improve the Safety and trustworthiness of their communities than non-moderators (Figure 2.7b), and rank Engagement and Safety as more important than non-moderators (Figure 2.7c).

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 [291], 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 [210, 181] and traumatized by exposure to offensive

[6]For importance, lower rank values indicate higher importance.

content [245], which may contribute to our findings of a perception of larger community size and lower trustworthiness amongst moderators relative to non-moderators.

2.2.7 To What Degree Can Community Values Be Predicted Based on Automatically Measurable Features?

Our survey responses (§2.2.3) 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 §2.2.4-2.2.6, all of which can be automatically quantified from publicly available data, collectively capture.[7] A complete list of features used is given in Appendix Table B.2.

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 extrapo-

[7] We further experimented with a much larger set of 74 features and found they did not lead to significant performance increases.

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

Table 2.4: 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).

lating 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 non-linear 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 2.4). 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 challenging to automatically measure. A complete table of results for each value is given in Appendix Table B.1.

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[8] to support further research.

2.2.8 Discussion & Conclusion

Diversity of Communities

Our study reveals that the set of communities surveyed have remarkably diverse values (Figs. 2.1,2.2,2.3). 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 (Figure 2.4a,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 [183].

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

Participatory Governance

Volunteer moderators play an important role in community governance on Reddit [181]. On the other hand, both formal and informal opportunities for non-moderators to influence decision making in their communities are quite rare [291]. 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 (Figure 2.7b). 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 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 (§2.2.3).

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.

2.3 How Conversational Structure and Style Shape Sense of Virtual Community

2.3.1 Introduction

Sense of Community, the bond between individuals and their social groups, plays a critical role in individual and collective well-being [188]. Stronger community attachment, both offline and online, is predictive of increased individual well-being [66, 136, 253], resilience against stress and high-risk behaviors [46, 69, 279, 51, 62, 227, 66, 106], and community resilience and problem-solving [2, 31, 14].

As billions of people have moved their social interactions into online spaces, understanding the dynamics of Sense of Virtual Community (SOVC) is essential to fostering these positive effects within online environments [144]. Prior work has relied primarily on surveys to understand SOVC across listservs and newsgroups [25, 24], blogs [22, 75, 102], discussion forums [278, 279, 266, 1, 83, 244, 144], and livestreaming communities [128]. These studies offer valuable insight into how SOVC manifests but have largely overlooked how specific aspects of community behavior shape these perceptions. This gap limits our understanding of how community behaviors translate into the feelings of belonging, cooperation, and connection that characterize SOVC.

Our study bridges this gap by examining how SOVC manifests across diverse communities on Reddit, as well as how conversational structure and linguistic style predict SOVC. Conversational structure captures the patterns and dynamics of user interactions, such as thread depth, reciprocal exchanges, and voting behaviors [11, 15, 40, 154]. These structural elements can predict conversation- or thread-level outcomes, such as engagement and virality [29, 237, 45].

In contrast, linguistic style captures the expressive and emotional characteristics of user communication, such as who are the chosen targets of speech acts, whether speech is past- or future-oriented, or whether speakers use emotional or socially-oriented language. Prior work using LIWC (Linguistic Inquiry & Word Count), a well-validated dictionary-based tool for grouping words into 'psychologically meaningful categories' [256], has shown how linguistic, content-agnostic signals like affect or cognitive processes can predict various outcomes in online communities [7, 60, 186, 187, 12, 33]. Providing insight into the tone and intent of user interactions, these stylistic markers offer a complementary lens to the structural perspec-

tive. No prior work has sought to understand how conversational style and structure relate to the broader community-level construct of SOVC.

In this study, we evaluate how community behavior, as measured by both conversational structure and linguistic style categories of features, can predict differences in the SOVC experienced by members of Reddit communities. Reddit's diverse communities, which span a broad range of topics, provide an ideal context for exploring these relationships. Specifically, we address two key research questions:

RQ1 What are the dimensions along which SOVC varies across communities on Reddit (§2.4.1)?

RQ2 How do patterns of conversational structure and linguistic style explain differences in SOVC between and within communities (§2.4.1)?

To answer these questions, we combined survey responses from 2,826 Reddit users with detailed behavioral data capturing interactions within 281 communities. We answer the first research question using exploratory factor analysis to identify meaningful dimensions which describe how the community experience varies across channels (§2.4). We answer the second question by constructing a hierarchical model to predict individual-level and community-level scores for each of these dimensions, based on behavioral trace data (§2.4) which characterize these communities in terms of their conversational structure and linguistic style. We intentionally choose automatically quantifiable features which are minimally related to communities' topic and content in order to maximize the degree to which our findings generalize across communities of different topics, languages, and potentially different platforms.

Contributions. This paper provides the following contributions to the area of computer-mediated communication:

- We provide the first large-scale study of SOVC in Reddit, finding that SOVC varies along three dimensions (*Membership/Belonging*, *Cooperation/Shared Values*, and *Connection/Influence*) which we relate to existing research (§2.4.1).
- We provide the first quantitative evidence linking both conversational structure and linguistic style to differences in SOVC (§2.4.1). These generalizable features capture a substantial fraction of the variance of SOVC in our hierarchical model, highlighting the extent to which these patterns of interaction shape the subjective experience of online communities.

- We identify actionable features, such as deeper reply chains and prosocial language, that platforms may be able to leverage to foster engagement, improve moderation, and strengthen community bonds (§2.4.2).

By modeling the relationship between automatically measurable conversational features and SOVC, this work lays the foundation for future studies aimed at enhancing online community health and resilience. Our approach enables the study of SOVC across diverse communities and platforms by deliberately excluding features that capture the topic of discussions, and our methods could extend to non-English and smaller online communities. Our findings open new avenues for research into how behavior patterns shape – and are shaped by – social dynamics and offer practical opportunities for improving community experiences, and their associated benefits, at scale.

2.3.2 Related Work

Predicting SOVC in Online Communities.

Extensive prior research on Sense of Virtual Community (SOVC) has identified a set of core dimensions that overlap with and extend those identified in studies of offline communities. In study of offline communities, McMillan and Chavis [188] initially proposed a definition of 'Sense of Community' (SOC) encompassing four dimensions: *membership* (feeling of belonging or personal relatedness), *influence* (a sense of mattering to the group or other members), *integration and fulfillment of needs* (a feeling that members will provide for each other), and *shared emotional connection* (a sense of shared history and experiences).

Online communities afford mechanisms for interaction that can produce different styles of community attachment. Prior studies of SOVC have found that, while the concept of membership translates well to online communities [1, 23, 24, 22, 144, 266, 128], SOVC can also manifest along different dimensions, such as immersion [144], recognition [22], identification [23, 22, 266], emotional feelings [23, 266], or cohesion [128].

Previous research has examined how activity patterns in online communities relate to Sense of Virtual Community (SOVC). Studies across a broad range of contexts, such as online health forums, online games, and livestreaming communities, has demonstrated strong links between the volume and types of participation and the feelings of SOVC that members experience [47, 287, 128, 270, 113]. In the specific context of

Reddit, survey and interview studies have identified factors like interaction volume and quality, member tenure, norm enforcement, and bot usage as factors that influence members' experiences [244, 212, 283] and examined how these factors vary between different communities [284]. However, these studies have not quantified SOVC nor directly connected such factors to measured SOVC, which is the primary contribution of this work.

Markers of Linguistic Style.

The conversations which make up online communities exhibit distinctive stylistic features that prior work has linked to differences in individual and collective outcomes. Many studies have relied on LIWC (Linguistic Inquiry and Word Count), a dictionary-based method for grouping words into 'psychologically meaningful' categories [256]. LIWC can be used to identify patterns of linguistic style that can be used to reliably differentiate content from different communities, even if they share topics [137], and to profile the distribution of individual personalities or values across a community [156, 247]. While other methods have been used to measure how relative differences in language usage between individuals and community pairs [48, 112, 276], we focus on absolute measurements of language, and we use LIWC for its robustness and generalizability.

Variation in specific categories within LIWC, such as words expressing group affiliation and those expressing positive/negative emotion, correlate with differences in individual and collective outcomes, including individual retention [172], subjective well-being [60], community satisfaction [186], and community stability [187]. A small number of studies have applied LIWC specifically to understanding group identity and cohesion [12, 33]. In this work, we directly use LIWC features to model SOVC across a diverse set of communities and members.

Conversational Structure.

Modeling the structure of online conversations – as threads, trees, or interaction networks – reveals important patterns of user engagement, influence and interaction. Key metrics include conversation volume (e.g. total comments) [40, 154], depth (reply chain length) [40, 154, 290], breadth (reply chain branching) [290], and degree distribution [40, 154, 290]. Network-based approaches have also modeled conversational dynamics

to understand influence and community formation [29, 148, 237].

A particularly useful approach involves analyzing local interaction patterns, such as recurring motifs within conversation graphs. Coletto et al. [45] used network motifs – small, recurring subgraphs of user interactions – to detect controversial discussions, finding that specific dyadic and triadic patterns reliably identify divisive exchanges. Paranjape, Benson, and Leskovec [208] introduced temporal motifs to capture dynamic elements, such as response delays, that static models often miss. Kušen and Strembeck [157] used temporal motifs to differentiate controversial and non-controversial exchanges. Zhang et al. [292] expanded on structural motifs to include engagement signals (*e.g.,* likes), an approach that we leverage in the present study.

These structural features have been successfully used across platforms to predict various conversational outcomes without content signals. Examples include forecasting discussion thread lengths on Facebook and Wikipedia [15] and on Reddit [103], and modeling the controversiality [45], toxicity [237], and emotional tenor [157] of Twitter conversations. Zhang et al. [292] presents the prior work most similar to our own, using conversational motifs to characterize communities along various dimensions (e.g. focused vs. expansionary, civil vs. uncivil). However, no prior work has used structural patterns to measure or predict members' community-level attitudes, such as SOVC.

Drawing on this related work, this paper fills a key research gap by modeling the relationship between community-level perceptions of SOVC and conversation-level markers of structure and style.

2.4 Methods

We conducted our study on Reddit, a large, global platform hosting over 100,000 active communities around various topics and interests [223], independently created and managed by users called moderators. The platform's threaded, comment-driven discussion structure and voting features enable rich conversational interactions. Combined with the platform's diversity of communities, this makes Reddit an ideal context for studying SOVC.

Weekly Subscribed Visitors	Everyone	Mature
1-800	22.3%	3.3%
800-2000	22.7%	3.7%
2000-6400	21.0%	3.3%
6400+	20.3%	3.3%

Table 2.5: The 300 target subreddits were selected to be broadly representative of the broader target subreddit population, based on segments by size (weekly subscribed visitors) and rating (Everyone / Mature).

Measuring Sense of Virtual Community

Survey Design and Participants. We measured SOVC using a survey distributed via Qualtrics in 2023 to 2,826 Reddit users across 281 subreddits. Participants were selected randomly from subscribers to a sample of 300 public, English-language, safe-for-work (not containing sexually-explicit or gore content) subreddits, stratified based on activity and subscriber count. Eligible participants were over 18, located in the US, Canada, Australia, or the UK, with accounts in good standing and eligible to be contacted for a survey. Respondents were compensated with $10 gift cards.

The survey asked about (1) informed consent, (2) demographics, and (3) community-specific questions. The full survey instrument is included in Appendix C.1 and summarized here. Users self-reporting that they were under 18 years old were skipped to the end and excluded from participation. The community-specific measures included 14 SOVC items, introduced by Perkins et al. [209] and extensively evaluated in prior work [199, 200, 201, 23, 1, 128]). Participants responded to each using a 5-point Likert scale (-2 = 'Strongly disagree' to +2 = 'Strongly agree').

Respondents self-reported their age using predefined ranges; the median age was '25-34', with 46% (1,305) reporting that they were 35 or older. 91% opted to self-disclosure their gender identity; of these, 65% identified as men, 21% as women, and 4% as non-binary (the remainder choosing to self-describe). These proportions of gender identity fairly closely resemble that of Reddit more broadly [19]. 89% self-reported that they had been active on the platform for at least a year; just 3% reported a tenure of 3 months or less. 81% indicated that they visited Reddit at least once per day. Just 4% reported that they visited Reddit less often than weekly.

Exploratory Factor Analysis. We removed 112 respondents (4.0%) who 'straight-lined' through the 14

SOVC measures,[9] yielding 2,714 responses. We analyzed responses using polychoric correlations to account for ordinal data (*cf.* [16]). A Kaiser-Meyer-Olkin test ($MSA = 0.92$) and Bartlett's test ($p < 0.0001$) confirmed suitability for Exploratory Factor Analysis (EFA). Given the close expected relationship among factors, we used `promax` rotation, yielding a three-factor solution that explained 44% of the variance in responses. McDonald's Omega indicates that a substantial portion of this variance (73%) is explained by a common underlying dimension, supporting our expectation that these three distinct factors align with a shared construct of SOVC [130].

Capturing Online Behavioral Activity

To explore relationships between community interactions and SOVC, we collected a range of individual- and community-level behavioral trace data over a 30-day period preceding the survey. These data were used to construct our features capturing conversational structure and linguistic style, along with control variables to account for large-scale differences in community participation and size.

Conversational Patterns. To capture the dynamics of user interactions within communities, we computed conversational patterns derived from chains of interactions initiated by posts or comments in the target subreddit. Following prior work on interaction motifs (*e.g.,* [45]), we identified recurring patterns of engagement that reflect different types of participation and response behaviors. For each post or comment by a user, we tracked subsequent interactions to form conversational patterns.

Using this approach, we identified eight distinct conversational patterns, representing variations in interaction dynamics (examples in Figure 2.8, for a full list of patterns, see Appendix C.2), including differentiation between patterns that start with a post (a top-level submission on Reddit) and those that start with a comment (a comment that is a reply to a post or another comment). For each subreddit, we calculated the total occurrences of these patterns over a 30-day observation period occurring immediately prior to the SOVC survey. Conversational pattern counts were normalized by the total number of interactions within each subreddit, yielding an 8-element proportion vector that ensures comparability across communities with varying activity levels.

We selected this method for computing conversational patterns as it is both established in prior work

[9] We removed all responses where the respondent provided the same answer to every single survey question.

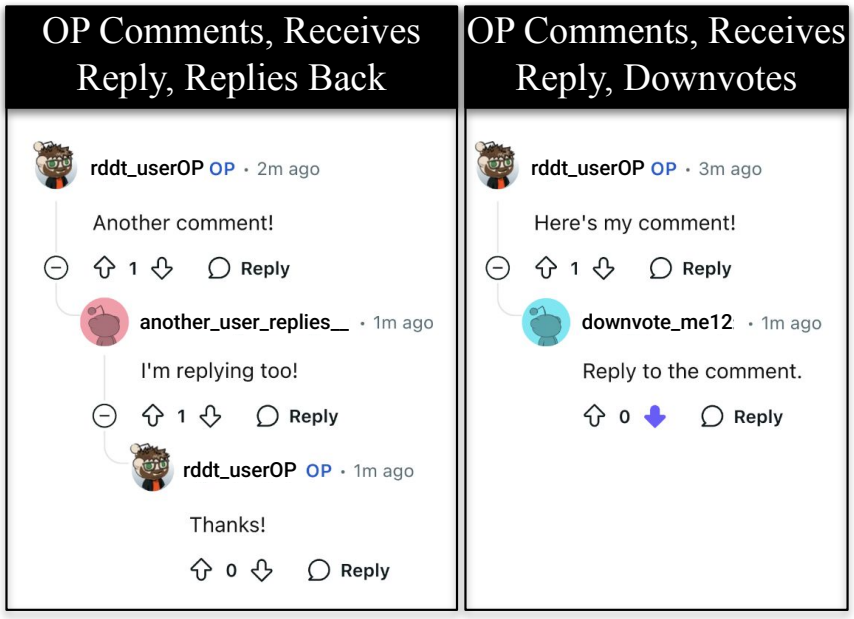


Figure 2.8: Conversational patterns include different interactions between community members. The left example here shows OP commenting, receiving a reply, and replying back, while the right example shows OP commenting, receiving a reply, and downvoting (blue arrow). A complete list of interactions is given in Appendix C.2.

[45] and feasible to implement at a very large scale. Our analyses included 2,342,651 posts and comments across our 281 participating subreddit. While such methods have been used to predict conversational level outcomes [45], ours is the first work that we are aware of to use conversational patterns to predict outcomes at the community level.

Linguistic Style. Linguistic style was analyzed using LIWC-22, a validated tool for computational text analysis that derives stylistic and psychological features from language [27]. We apply LIWC-22 to all posts and comments in each community from the target 30-day period, mapping words to one or more of 118 different categories, such as pronouns, emotion words, and perception-related words. Following precedent and guidance from prior work (*e.g.,* Khalid and Srinivasan [137] and Matthews et al. [186]), we include all LIWC features except 19 topic-specific LIWC categories (the culture, lifestyle, and physical dictionaries from the LIWC extended dictionary) to avoid topical leakage and focus on stylistic features, which are more generalizable across domains and strongly linked to social phenomena.

LIWC was selected for quantifying linguistic style as it has been extensively used and validated for a

broad range applications [27] as well as being suitable to apply to text at a massive scale. Our analysis of linguistic style included more than 70 million tokens, a scale which renders many alternative methods computationally intractable. While methods such as word embeddings can be used to compare language between different users or communities [276, 48], in this work we use an absolute measurement of language at the community level to predict SOVC, an important community outcome.

Control Variables. At the individual level, we captured measures of activity, both across all of Reddit (*e.g.,* such as the number of unique communities visited, the frequency of contribution) – and within the target community (*e.g.,* duration of membership tenure, whether or not they had recently voted). At the community level, we collected aggregate measures including the total number of visitors, contributors, and daily active users. These control features were included in our models to adjust for baseline differences across users and subreddits.

Modeling Approach

To predict our measures of SOVC, we adopted a hierarchical linear model approach with a random slope, integrating individual-level and community-level features while accommodating the nested data structure. We applied LASSO (Least Absolute Shrinkage and Selection Operator) for feature selection, which identifies the most predictive behavioral features by setting irrelevant coefficients to zero. We used 5-fold cross-validation to select the best alpha penalty for the L1 norm of the coefficients. Features were standardized (zero mean, unit variance) to facilitate interpretation of coefficients and ensure comparability across predictors.

The model itself was constructed with two levels:

- **Individual-Level.** A mixed-effects model used individual-level features (*e.g.,*, voting behavior, communities visited) to predict SOVC scores, with random intercepts accounting for subreddit-specific variability.

- **Community-Level.** A LASSO-regularized linear model used community-level features (*e.g.,*, community age, number of visitors) to predict the subreddit random intercepts obtained from the first model.

We selected a hierarchal linear model due to the natural bilevel nature of our features, which describe

both individuals (respondents to our SOVC survey) and communities, each of which had several respondents.

More formally, the individual level SOVC score $\hat{y}_{ij}$ is given by:

$$\hat{y}_{ij} = \beta_1 x_{ij1} + \beta_2 x_{ij2} + \cdots + \beta_n x_{ijn} + \alpha_i$$

The community-level random intercept $\hat{\alpha}_i$ is given by:

$$\hat{\alpha}_i = \gamma_1 x_{i1} + \gamma_2 x_{i2} + \cdots + \gamma_m x_{im}$$

Where x_{ijn} represents that n-th value of community i's member j's feature vector, x_{im} represents the m-th value of community i's feature vector, and β_n and γ_m represent the learned linear coefficients selected using LASSO-regularization.

Model performance was assessed using adjusted R^2 and variance partitioning to evaluate the relative importance of the individual- and community-level predictors. An ablation study quantified the relative importance of conversational structure and linguistic style features, confirming that both contributed significantly to SOVC predictions.

Ethical and Privacy Considerations

Our study followed strict ethical guidelines to protect participant privacy and ensure responsible data use, as confirmed by an internal review of our study procedures at Reddit, Inc. Surveys were targeted to users whose selected settings made them eligible to be contacted, and survey participants provided informed consent. All data were collected and processed under Reddit's privacy policy. Behavioral trace data were anonymized and aggregated to prevent reidentification and analysis focused exclusively on either high-level individual-level features (*e.g.,* vote counts *vs.* individual votes) or aggregated community-level patterns (*e.g.,* LIWC category counts vs. individual text strings). We believe the potential for negative social impacts and misuse from our work is very limited.

Scale / Item	Factor 1	Factor 2	Factor 3
MB: Membership & Belonging ($\alpha = 0.698$)			
I expect to be a part of this community for a long time.	**0.79**	-0.04	-0.04
I think this community is a good thing for me to be a part of.	**0.64**	0.19	-0.09
It is important to me to be a part of this community.	**0.49**	0.01	0.34
I feel at home in this community.	**0.44**	0.18	0.14
CSV: Cooperation & Shared Values ($\alpha = 0.795$)			
If there is a problem in this community, members can get it solved.	-0.07	**0.70**	-0.01
Members of this community can be counted on to help others.	0.01	**0.70**	-0.04
I want the same things from this community as other members.	0.21	**0.43**	-0.11
Members of this community share the same values.	0.07	**0.50**	0.00
CI: Connection & Influence ($\alpha = 0.792$)			
Most members of this community know me.	-0.07	-0.20	**0.88**
I have friends in this community that I can depend on.	-0.03	0.04	**0.73**
I recognize the screen names of most participants in this community.	0.04	-0.07	**0.67**
I feel like I have influence over what this community is like.	0.00	0.12	**0.53**
I care about what other community members think of me.	0.13	0.03	**0.45**
If I have a personal problem, I can turn to members of this community.	-0.12	0.38	**0.41**

Table 2.6: Dimensions of SOVC across subreddits as identified through EFA, with factor loadings for items and Cronbach's α for factors. All items were tested using a 5-point Likert scale ranging from "-2: Strongly Disagree" to "+2: Strongly Agree".

2.4.1 Results

Characterizing SOVC on Reddit

To characterize SOVC on Reddit, we performed an Exploratory Factor Analysis (EFA, described in §2.4) on our survey data to identify three dimensions of Sense of Virtual Community (SOVC) on Reddit, which we term **Membership & Belonging** (MB), **Cooperation & Shared Values** (CSV), and **Connection & Influence** (CI). These dimensions, summarized in Table 2.6, align with established constructs in the literature and capture distinct, complementary aspects of community experience.

Membership & Belonging reflects the extent to which individuals feel embedded within a community. Items loading on this factor have previously been categorized as as 'belonging' [199, 200, 128], 'membership' [201], 'conscious identification' [200], and 'shared emotional connection' [1].

Cooperation & Shared Values captures perceptions of collective alignment and mutual support. Items loading on this factor have been categorized in prior work under various constructs, including 'cooperation/shared values', emotional connection, 'influence', 'integration/needs fulfillment', and 'cohesion'.

Connection & Influence relates to interpersonal relationships and perceived social influence within the community. Items in this factor have been categorized in prior work as a variety of constructs related to SOVC, including 'leadership/influence', 'friendship & social support', and 'integration/needs fulfillment'.

As shown in Table 2.6, the inter-item reliability for these scales is moderate ($\alpha_{MB} = 0.698$; $\alpha_{CSV} = 0.795$; $\alpha_{CI} = 0.792$). While α_{MB} can be improved substantially by dropping the least-correlated item *'I feel at home in this community'*, we have retained this item given its strong alignment with this construct in prior work.

Quantifying SOVC on Reddit

We quantify SOVC for inclusion into our models by mapping responses from -2 (Strongly Disagree) to +2 (Strongly Agree) and averaging scores across items in a scale. This lets us compute for each respondent a score for each factor. These scores represent the dependent variables in our viewer-level and community-level SOVC models. As shown in Figure 2.9, all three factors follow a roughly normal distribution.

SOVC scores varied significantly across subreddits and respondents (Figure 2.9). Communities scored highest on **Cooperation & Shared Values** ($\mu = 0.50, \sigma = 0.61$) and **Membership & Belonging** ($\mu = 0.35, \sigma = 0.72$), with relatively lower scores on **Connection & Influence** ($\mu = -0.81, \sigma = 0.70$). These results suggest that personal connections and influence among members may be less prominent than generalized feelings of belonging or cooperation in the communities studied.

While studying SOVC independently from the topic of a community is the primary goal of this work, it's important to understand how SOVC varies across communities of different topics. To do so, we use Reddit's publicly available topics. More details on topic classification and extended results are included in Appendix C.3.

Subreddits in our sample which scored highest on Connection & Influence include Business & Finance Communities and Nature & Outdoors communities. Health communities and those focused on Vehicles had the highest sense of Cooperation & Shared Values, while Health and Nature & Outdoors communities had some of the strongest senses of Membership & Belonging (Appendix C.3). However, the primary focus of this work is on how SOVC varies across communities based on their conversational structure and linguistic style. Future work could specifically examine SOVC between communities with different topics (§2.4.2).

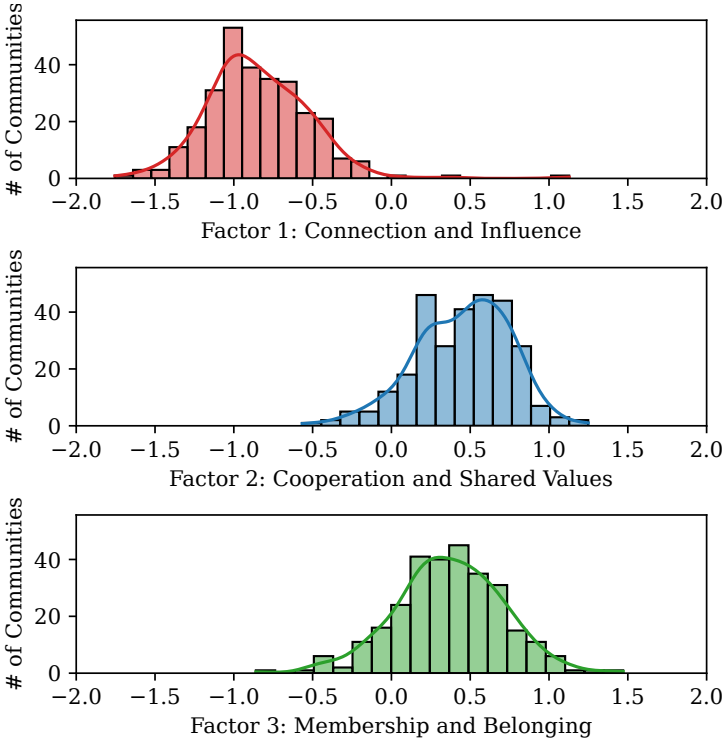


Figure 2.9: Community members generally rate their connection and influence (Factor 1) as lower than their cooperation and shared values (Factor 2) or membership and belonging (Factor 3). This figure shows the distribution of communities' SOVC scores for each factor, computed by averaging the responses for each community.

Model	CI	CSV	MB	Overall
Control Only	0.152	0.095	0.054	0.100
+ Style	0.311	0.326	0.272	0.303
+ Structure	0.171	0.197	0.113	0.160
+ Style & Structure	0.341	0.348	0.305	0.331

Table 2.7: Ablation Study Results. R^2 values capturing explained variance for each SOVC dimension (CI: Connection & Influence, CSV: Cooperation & Shared Values, MB: Membership & Belonging) and overall, for models with different feature combinations. The "Control Only" model includes no structural or stylistic features. Stylistic features contribute more to R^2 than structural features alone, with the full model (Control + Style & Structure) achieving the highest overall performance.

Modeling SOVC with On-Platform Behavior

Using our hierarchical linear modeling approach, we predicted each of the three SOVC dimensions, as measured by the EFA factor loadings computed above, using individual- and community-level measures of conversation structure and linguistic style (§2.4). We find that both conversational structure and linguistic style separately capture more SOVC than control features only, and that linguistic style features contribute more to model performance than conversational structure features (Table 2.7). Linguistic style features on their own capture more than 30% of the variance in overall SOVC.

Our full SOVC model, including both conversational structure and linguistic style features, captures 33.1% of the variance in SOVC across communities and individuals. As communities vary widely in their topic, membership, and values [284], we believe that it is noteworthy that our 19 readily quantifiable features capture nearly one-third of variance in SOVC.

The coefficients for our linear model, shown in Table 2.8, show the strength of the relationship between each measure of community activity and self-reported SOVC. We illustrate the relationship between features discussed below and model predictions for SOVC scores in Figure 2.10.

Among individual-level features, **days active** in the community positively predicted both Connection & Influence and Membership & Belonging, while longer **membership tenure** was uniquely associated with higher Membership & Belonging. **Active participation**, such as posting/commenting, was positively associated with all three SOVC dimensions, highlighting its centrality to community experience. Interestingly, greater **cross-subreddit participation** tended to be associated with lower Cooperation & Shared Values, suggesting that cross-community engagement may dilute perceptions of alignment within a single subred-

Model Level	Feature	CI (β, p)	CSV (β, p)	MB (β, p)
Individual Level	*Within-Community Features*			
	Days Active (past 30)	$0.005**$	0.001	$0.006**$
	Length of Membership ($\log_2$)	-0.001	0.006	$0.017**$
	Has Voted (T/F)	0.049	$0.072*$	$0.077*$
	Has Posted or Commented (T/F)	$0.108**$	0.023	$0.092*$
	Sitewide Features			
	Communities Visited ($\log_2$)	$-0.030**$	-0.017	$-0.029**$
	Number of Votes ($\log_2$)	$-0.023***$	-0.005	-0.004
	Number of Posts and Comments ($\log_2$)	$0.021**$	$-0.013*$	-0.006
Community Level	*Control Features*			
	Community Age (days)	$-0.121*$	-0.050	–
	Visitors (past 30 days, $\log_2$)	$-0.185*$	–	–
	Conversational Structure Features			
	OP Posts, Receives Reply, Replies Back	–	$0.182*$	0.090
	OP Comments, Receives Reply, Replies Back	$0.234**$	–	–
	OP Comments, Receives Reply, Upvotes	–	–	$0.141*$
	Linguistic Style Features (LIWC)			
	Third Person Plural Pronouns (they)	$0.127*$	$0.122*$	0.118
	Prosocial Behavior (care, thank, help)	$0.166*$	0.096	0.124
	Female References (she, girl, woman)	$-0.171*$	$-0.204*$	$-0.284***$
	Lacks (don't have, didn't have, less)	$-0.127*$	–	-0.093
	Feeling (feel, hard, cool, felt)	$0.211**$	$0.138*$	0.099
	Future Focused (will, going to, have to)	–	0.016	$0.187**$
	Netspeak (u, lol, haha, emoji)	–	$0.156*$	$0.167*$

Table 2.8: β coefficients for models predicting Connection & Influence (CI), Cooperation & Shared Values (CSV), and Membership & Belonging (MB) scores, based on structure & style features, along with individual- and community-level controls. Features are included in the model only if significantly associated, on their own, with the outcome variable. Many coefficients are shrunk to zero via Lasso regularization. Indicators for p-values are as follows: * : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$

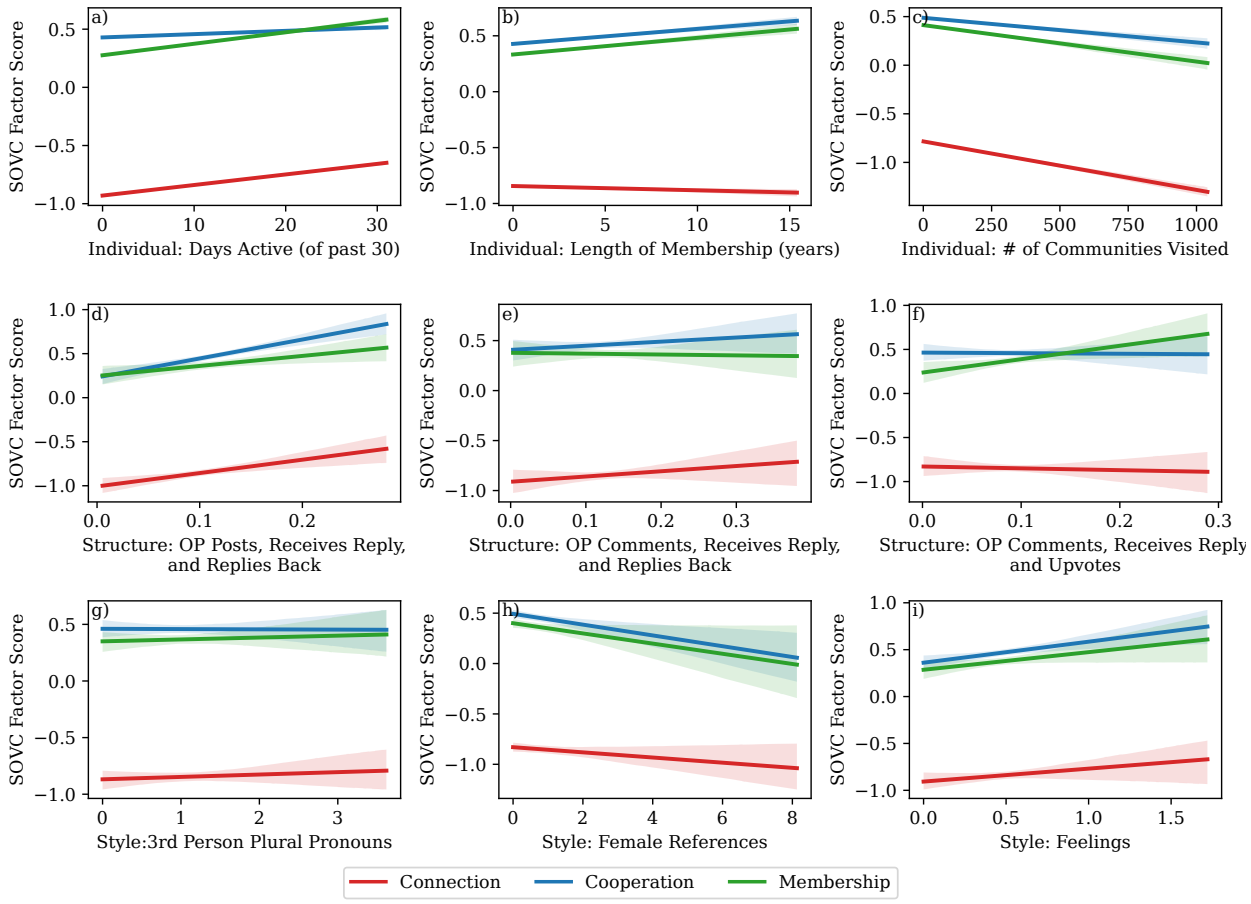


Figure 2.10: Relationships between selected individual and community level features and model-predicted SOVC. The top row shows individual-level features, while the bottom two rows show community-level conversational structure and linguistic style, respectively. Conversational structure axes show the proportion of all patterns that match the pattern on the X axis label. Linguistic Style axes units are taken directly from LIWC output.

dit.

Conversational Structure. Our analysis reveals nuanced relationships between conversational structures and the dimensions of Sense of Virtual Community (SOVC). Notably, communities with more reciprocal reply chains receive higher scores for Cooperation & Shared Values and Membership & Belonging (when they start on posts) and Connection & Influence (when they start on comments). This result emphasizes the role of sustained dialogue in fostering interpersonal relationships and collective alignment. Increased reciprocal conversation has been linked to various positive community outcomes, including discussion quality [74, 114] and provision of support [7].

On the other hand, we find relatively limited relationships between voting behavior and SOVC. We do not find that downvoting behavior is significantly associated with any dimension of SOVC, while more frequent upvotes to replies are only associated with one of the three dimensions of SOVC; increased Membership & Belonging. Upvotes on Reddit serve two functions which may impact membership/belonging: (1) providing positive feedback to the upvoted commenter that enhances their pro-community attitudes [52, 161], and (2) aligning community expectations by making desirable content more visible [91].

Linguistic Style. Features capturing community-wide differences in linguistic style were highly informative, explaining more than 30% of the variance in self-reported SOVC scores. Stronger feelings of Connection & Influence are associated with the use of more third-person plural pronouns (*e.g., they*). Prosocial language (*e.g.,*, 'thank', 'help') is positively associated with Connection & Influence and Cooperation & Shared Values, aligning with prior work suggesting that gratitude language in online communities can create positive feedback loops, prompting future prosocial behavior [178, 7]. However, more references to wants or lack of things (*e.g., don't have*, *less*) are associated with *lower* Connection & Influence and Membership & Belonging.

Expressions of sensation or emotion (*e.g.,*, 'feel', 'hard') align positively with Connection & Influence and Cooperation & Shared Values. Positive expressions of emotion can certainly indicate effective functioning as a collective [73, 151]; this may also reflect how strong, anonymous communities empower members to share negative emotional experiences to obtain support [53, 9].

We note a positive relationship between Netspeak (*e.g., u*, *lol*, emoji) and 'future-focused' language and feelings of Membership & Belonging. Prior distinctions between 'common identity' and 'common bond'

groups define the former as focused more on the goal and purpose of the group than connections to other members [236]. Fewer references to 'girls' or 'women' corresponded to stronger SOVC across all three dimensions; this may reflect that communities which adopt more inclusive language may tend to function better overall, although additional research is needed to understand the role that gendered language plays in community members' SOVC.

In Figure 2.10, we visualize the relationships between selected individual and community level features and the three factors of SOVC. Figure 2.10b shows how community members with longer lengths of membership have higher senses of cooperation and membership, but lower senses of connection. Figure 2.10d,e,f shows how greater conversational depth and more upvotes by OP are associated with higher SOVC, especially connection and cooperation. Figure 2.10h shows how more female references are associated with lower SOVC, while more discussion of feelings are associated with higher SOVC across all three factors (Figure 2.10i).

2.4.2 Discussion

In this work, we sought to understand how *Sense of Virtual Community* (SOVC) varies across online communities, and identify structural and stylistic markers of interactions associated with stronger or weaker community attachment. Using self-reported community perceptions from 2,826 active Reddit users, regarding 281 unique communities, in which they participated, our hierarchical model allowed us to identify how differences in conversational structure and linguistic style can predict community-level variation in SOVC. Below, we present a brief summary of our results and discuss some of the theoretical and practical implications for researchers, designers, and practitioners who share an interest in fostering more positive social experiences online.

Summary of Results

Dimensions of SOVC on Reddit. Through exploratory factor analysis, we identified three dimensions of SOVC – **Membership & Belonging** (MB), **Cooperation & Shared Values** (CSV), and **Connection & Influence** (CI) – that align well with established constructs from prior work. Membership & Belonging captures the extent to which individuals feel integrated within their community, aligning with notions of

belonging and shared emotional connection. Cooperation & Shared Values reflects collective alignment and mutual support, Connection & Influence represents interpersonal relationships and the perceived ability to impact the community. Researchers should consider how these different dimensions of SOVC may be in tension with one another, requiring trade-offs to be made when working to improve communities [283].

Distribution of SOVC Scores. SOVC scores varied substantially across communities and respondents. The higher scores observed for Cooperation & Shared Values and Membership & Belonging than for Connection & Influence suggest that community experience on Reddit has more to do with fostering shared goals and a broader sense of belonging than prompting strong interpersonal connections or influence as an individual on the community. We note that our strategy for sampling communities for this study necessarily means that these findings are not representative of the broader population of Reddit communities. In particular, we do not include small communities in which interpersonal connection may be more likely to occur. Future work could look explicitly at small communities.

Modeling SOVC with Linguistic Style and Conversational Structure. Our hierarchical model is able to predict 33.1% of variance in SOVC using automatically quantifiable features. While both linguistic style and conversational structure features are important to our model's predictive performance, linguistic style plays a larger role in modeling SOVC. Future work using LLM-based measurement of features may capture even more variance in SOVC. Regarding conversational structure, deeper conversational depth (more back-and-forth dialogue) is more strongly associated with stronger SOVC than voting behavior. We found downvoting behavior is not significantly associated with any dimension of SOVC, while upvoting behavior is only associated with Membership & Belonging. Regarding linguistic style, more third person pronouns (they) and more discussion of prosocial behavior are especially strongly associated with increased SOVC, being positively associated with all three dimensions.

Practical Implications

Our findings offer actionable insights for community moderators and platform designers to foster stronger communities.

Implications for Community Moderators. Given the outsized role that moderators play in shaping community experiences on Reddit [127], these findings suggest certain actions that moderators could take within

their communities to foster behaviors associated with higher SOVC. Moderators might foster more reciprocal reply chains by initiating and promoting dialogues through Q&A threads (*e.g.,* AMAs, 'ask me anything') or collaborative discussion formats. Moderators can emphasize norms of appreciation and recognition, such as rewarding thoughtful responses with upvotes or featuring top contributors [161]. Engaging community members in shared projects, such as those supported by Reddit Community Funds [219], could engage community members in goal-oriented, future-tense discussion that build a sense of Membership & Belonging. Moderators could easily support the use of more inclusive language through rules set up in Post Guidance [229] or similar proactive moderation tools on other platforms. Moderators are also uniquely familiar with their community's specific attributes and needs [284]; future work could explore how moderators can support SOVC in community-specific ways.

Implications for Platform Product Teams. Our findings also invite design thinking around products and features that could foster stronger communities. Platform designers could foster more reciprocal engagement and positive reinforcement via prompts/notifications which encourage community members to reply to or upvote messages, especially those from first-time posters. Automated summaries or highlights from active threads could be used to help direct users to opportunities to reply and foster reciprocal engagement. Designers could encourage positive feedback by explicitly rewarding users who promote belonging by upvoting others' contributions through badges or flair.

Design of Community Features. These findings also inform the design of product features that can increase reciprocal interaction and positive engagement. On-platform prompts or off-platform notifications encouraging community members to reply to messages, especially those from first-time posters, might help to increase SOVC for both the posters and the repliers. Automated summaries or highlights from active threads could be used to help direct users to opportunities to reply and foster reciprocal engagement. Designers could encourage positive feedback by explicitly rewarding users who promote belonging by upvoting others' contributions through badges or flair. Hiding content that has received a certain number of downvotes could prevent the accumulation of additional downvotes, potentially limiting further impacts on feelings of cooperation and membership.

Leveraging SOVC scores. Finally, this work offers a content-agnostic way to compute SOVC scores per community on an ongoing basis, which could be leveraged for purposes outside of the communities them-

selves. These signals can be included in community recommendation algorithms to help drive new users towards communities that offer more potential to drive feelings of connection, cooperation, and belonging depending on user needs & preferences. Tracking SOVC over time could offer insight into how community health is changing over time, for individual communities and in aggregate across the platform, enabling social platforms to intervene, when needed. SOVC scores could be used as metrics in experiments to evaluate feature rollouts to determine if new features are successfully driving behaviors associated with stronger community attitudes.

Limitations and Future Work

We acknowledge several limitations of our study, which may also offer opportunities for future research. Our analysis focused on a subset of Reddit communities selected based on activity and size, which may not represent the broader spectrum of smaller or less active subreddits; prior work has found that communities with features characteristic of 'common-bond' groups may naturally limit their growth [129], which would leave them underrepresented in our sample. Future work should explore whether the identified patterns generalize to these contexts, as well as other platforms with distinct community norms and structures, such as different voting affordances (*e.g.,* Likes in Facebook Groups *vs.* upvotes on Reddit). While our content-agnostic approach supports generalization across community topics, it remains limited to English-language data. Cross-cultural studies are needed to assess how linguistic and conversational patterns predict SOVC in non-English or multilingual communities, potentially revealing cultural nuances in community dynamics. Our modeling approach identifies meaningful associations between community behavior and SOVC, but cannot establish causality; future experimental studies could clarify causal relationships. We deliberately use relatively simple methods, such as LIWC, to quantify conversational structure and linguistic style, as these methods are wide used, well validated, and computationally tractable given the large scale of our study. However, future work could incorporate additional and more sophisticated features, such as more complex conversational patterns, LLM-based measurements of linguistic style, and additional control features to measure community topic directly. Such methods may be able to predict a larger fraction of the variance in SOVC.

2.4.3 Conclusion

This study advances the understanding of Sense of Virtual Community (SOVC) by providing the first large-scale analysis of how conversational structure and linguistic style shape community experiences across diverse Reddit communities. Using a content-agnostic approach, we identified three dimensions of SOVC—Membership & Belonging, Cooperation & Shared Values, and Connection & Influence—and demonstrated that interaction patterns and linguistic features explain a substantial portion of their variance.

Our findings emphasize the importance of reciprocal interactions, prosocial language, and future-oriented communication in fostering stronger online communities. These insights deepen theoretical understanding of how online communities function and offer actionable strategies for moderators and platform designers to enhance community engagement, cohesion, and inclusivity.

In contrast to prior work focused on specific content or platforms, this study establishes generalizable principles that are applicable across a diverse set of communities and community members. By demonstrating how online interactions influence community attachment, we highlight the transformative potential of digital communities to promote individual and collective well-being. Future research should build on these findings to explore causal mechanisms and extend their applicability to broader contexts.

2.5 Summary of Contributions to Thesis

In this chapter, surveyed the members of thousands of online communities in order to develop a deep understanding of community members' values for their own communities, and how these values vary both between different members of the same community, as well as between communities of different ages, topics, and sizes. We found that community members value an enormously broad set of attributes of the communities they participate in, and some of these values, such as high quality content, are poorly understood by researchers. Most importantly, we found that there is no 'one size fits all' set of community values: what's good for one community may not be good for another community, and researchers need to consider the individual needs of communities, and to focus not only on minimizing harms, but also on enabling communities to maximize the good parts. We showed that community members broadly agree that safety should be improved in their communities, but different community members have vastly different experiences with

how safe their communities currently are, even within the same community. This suggests that community leaders need to take steps to ensure that minority viewpoints are considered when making decisions that affect their entire community. We also found that moderators have different perspectives on democracy than non-moderators, with moderators feeling that their community *are* less democratic, *should be* less democratic, and *want less* democracy than non-moderators. This suggests that non-moderators may want to be more involved in community governance than they currently are. Finally, we examined community members Sense of Virtual Community (SOVC) and how this varies with different community behaviors. We found that reciprocal commenting and upvoting, and more frequent expressions of gratitude were all associated with stronger SOVC, while increased downvoting, surprisingly, were not associated with decreased SOVC.

These contributions characterize the different needs and values of online communities, and highlight the importance of high quality community governance for maintaining healthy communities. In the next chapter, we will examine a broad range of governance practices, moderation strategies, and non-moderator affordances, in order to identify which are associated with positive community outcomes and seem most effective at making communities better.

Chapter 3

Assessing the Impact of Current Governance Practices and Community Affordances

We can leverage existing data to identify the most practicing current governance practices. This includes moderation stuff and community affordances like voting, crossposting, etc.

Informed by our newfound knowledge of community values, we now turn our attention to *assessing* which current governance practices are most effective. By including hundreds of thousands of online communities in our analyses, we will leverage existing natural diversity in governance and outcomes to produce robust, data-driven insights and best practices, using causal inference methods to control for confounding factors without needing to resort to expensive and time consuming randomized controlled trials. We will quantify both different community governance practices, as well as important community outcomes, in order to establish links between the two, but data is not readily available for everything we'd like to measure. With clever usage of tools such as the Wayback Machine [259], we will collect details about community rules and moderators. The central challenge in this chapter is measuring community *outcomes* in a way that is feasible across thousands of communities, and billions of posts and comments.

In §3.1, we develop one of the first scalable outcome measures of community governance. The key insight behind this method is that community members discuss the moderators of their communities publicly,

saying things like 'the moderators of this subreddit are doing a great job!' or, more commonly, 'the moderators in this subreddit suck!' We develop a three-step pipeline to automatically detect posts and comments discussing community moderators, and employ this pipeline to identify 1.89 million statements discussing moderators on Reddit made over an 18-month period. We utilize these data, along with IPTW causal inference methods, to examine how different moderator workloads, recruiting strategies, and degrees of community engagement are received by community members. We find that different types of communities respond differently to strict rule enforcement, which seems especially effective in News sharing communities. We also find that moderators who are active community members before, during, and after their tenures as moderators are viewed more favorably by their constituent community members.

One of moderators most critical functions is to set and enforce rules, yet is it is difficult for moderators to make informed decisions about what rules to impose, how to phrase them, and how to enforce them. In §3.2, we take examine community rules in greater detail. Using historical data collected from the Wayback Machine [259], we compute timelines of what rules were introduced to which communities when, examining 67,575 different rules from 5,225 different communities across a 5+ year period. We develop a method to classify rules according to their tone (is the rule telling members what to do, or what *not* to do?), target (is the rule directed at who participates, what they say, or how they format their message?), and topic, and conduct the largest-to-date assessment of rules on Reddit. After controlling for key confounding factors using IPTW and DID analyses, we identify the rules most strongly associated with positive perceptions of moderators, utilizing the classification method developed for our assessment of broader moderation strategies in the preceding section. Finally, using time-series data and a difference-in-differences analysis, we examine what happens when communities introduce new rules, finding that in the short term, new rules appear to improve community member perceptions of governance, but this effect seems to diminish after approximately six months. We then discuss how these findings help inform best practices for community rule implementation.

Community governance also includes non-moderation affordances that influence how content is viewed and shared in and between online communities, such as voting and reposting mechanisms. In §3.4, we use the lens of news sharing to study how these non-moderator affordances influence the visibility of highly biased or low factual news posts (misinformation and fake news). Using Media Bias/Fact Check, a non-partisan fact-checking service, we conduct the largest study of news sharing behavior on Reddit, labeling

more than 550 million links to news sources across a four year period. We examine how upvoting, downvoting, and amplification via crossposting affect the visibility of potentially problematic content. We assess how politically left- and right-leaning community members differ in the content they post, and we explore the efficacy of deplatforming of entire communities by showing that exposures to highly biased and low factual content are extremely concentrated in a small number of potentially problematic communities. These results suggest that non-moderator affordances such as voting and crosspostsing are generally effective at reducing the relative visibility of highly biased and low factual news content.

3.1 Perceptions of Moderators as a Large-Scale Measure of Online Community Governance

3.1.1 Introduction

Mod teams must determine how many moderators to have, how to recruit new mods, what rules to set, and answer many more questions. With so many decisions to make, it's challenging to determine what governance practices are most effective at ensuring high quality outcomes for online communities. There are so many different communities, with different sizes and topics, that there is no 'one-size-fits-all' solution [283, 284]. However, the variety of existing communities presents an opportunity: if we can develop a method to assess the success of communities' governance practices, we can leverage the natural diversity of these practices to identify the most promising strategies for moderators. Doing so is challenging. Surveys can be used to ask community members about governance, but surveys are expensive to deploy and typically only capture one point in time rather than a longer history. Notions of 'success' are multifaceted, and while automated measurements using classifiers have been used to detect specific harms such as misinformation [282] or specific aspects of governance such as rules enforcement [37], current automated methods are unable to quantify broader notions of successful governance.

In this work, we measure online community governance by examining how community members themselves perceive their moderators. We develop a method to classify how community members discuss their moderators, publicly, within their communities (§3.1.3). We use this method to gather and characterize community members' perceptions of moderators at a massive scale, enabling the largest study-to-date of

governance practices that we are aware of. We label 1.89 million posts and comments discussing moderators from 8,477 unique subreddits which account for more than $2/3$ of all content on Reddit, covering an 18-month period from January 2020–June 2021. We relate these perceptions of moderators data to different kinds of online communities, and to different actions that community mod teams can take.

Our analyses address two key research questions:

RQ1 How are moderators of different communities perceived differently by their communities? (§3.1.4)

RQ2 What moderation strategies might improve perceptions of moderators? (§3.1.6)

We find that community members' perceptions of their mods vary substantially from community to community (§3.1.4). Hobby communities have the most positive perceptions of their mods, with meme and news communities having the poorest. Communities that perceive themselves as having high quality content, and being trustworthy, engaged, inclusive, and safe all perceive their mod teams more positively than communities at consider themselves low quality, untrustworthy, disengaged, uninclusive, or unsafe. Community size is also a major differentiator—tiny communities (1-10 posts+comments/day) use $6.1\times$ as much positive language to describe their mods as huge communities (>10k posts+comments/day).

We identify actionable strategies for community moderators, using IPTW and DID causal inference methods to control for confounding factors including the topic and size of communities (§3.1.5). We show that communities with fewer than 5 daily posts+comments per moderator use $2.5\times$ more positive language to describe their mods as communities with $20\times$ more posts+comments per mod. Yet our findings do not suggest that more strict moderation improves perception of moderators; in fact, for most types of communities, a 3 percentage point increase in removed content is associated with a 9pp increase in negative language used to describe mods, although news-sharing communities are a notable exception to this trend (§3.1.6). We measure changes in perceptions when different types of moderators are added to a community, and find that moderators who are also active community members before and during their tenure as moderators are most strongly associated with improved perceptions of mods (§3.1.6).

We discuss the limitations of our methods, and identify key areas for future work (§3.1.7). We make our models, anonymized datasets, and code public[1] to enable further research on this important topic.

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

3.1.2 Related Work

Measuring Community Governance

Comprehensively measuring community governance is challenging, as data access to moderator actions is often restricted, except for some types of communities, such as Wikipedia [206] and some gaming servers [77]. On Reddit and most popular social media platforms, however access to moderation actions is difficult to obtain, and requires separate permission from the mods of every community to be studied [165, 166]. In contrast, by utilizing public discourse around moderators, our method works for every community whose content is public.

Because of these challenges, many researchers turn to surveys to study governance, sometimes qualitatively [181], and sometimes at a larger scale, for instance to quantify harassment of mods [5] or moderator recruiting practices [241]. However, regardless of the measure or method used, because governance is a fundamentally complex topic, there is no single ground truth for the 'quality' of a community's governance. Our method complements survey based methods by directly measuring community members' publicly stated perceptions of their moderators in a automated, scalable manner.

Governance Through the Lens of Rules

Rules are often used as a lens to study governance. Large-scale analyses of rules have been used to characterize governance at the platform level on Reddit [224, 72] and on Wikipedia [109], however it is challenging to infer much about governance in specific communities, as rules constitute only a small portion of governance activity. Surveys of community members' attitudes towards rules have been used to understand members' attitudes towards governance more broadly [146], however this method is challenging to scale beyond a single community. Our method can be applied to thousands of communities.

Moderation Strategies on Reddit

Rules and their enforcement can also be evaluated in terms of their impact on the community as a whole [115, 246, 118] and their embodiment of communitywide norms [37]. Beyond rules and their enforcement, researchers have also studied other moderator strategies on Reddit, including platform level decisions such as community bans and quarantines [36, 228], and community level actions such as user bans [264] and stickied

posts [185]. Our work quantifies many of these strategies, including content removal, mod interactions with the community, and associates them with positive and negative moderation discourse.

Measuring Outcomes in Online Communities

Researchers have studied the range of values that community members hold for their communities [283], yet it is challenging to accurately predict these values automatically [284]. Large-scale embeddings can be used to understand community culture [276], yet not governance directly. Longitudinal work has examined how communities fare with massive growth [167] and the lifecycles of their members[48], while some research focuses on smaller scale outcomes at the conversation level [18]. Our work complements this literature with a large scale method for studying governance-specific outcomes.

3.1.3 Methods

Reddit Overview and Data Sources

In this work, we focus primarily on Reddit, one of the largest social media platforms, and one is frequently studied in the computational social sciences [213, 282, 283, 284, 181, 37, 246]. Reddit is composed thousands of communities, known as subreddits, each with its own community norms, rules, membership, and moderators. These attributes, along with the the fact that almost all content on Reddit is publicly available, make it an ideal platform for studying community governance at scale.

In this work we make extensive use of publicly available Reddit data from the Pushshift [20]. From these data, we detect and classify posts and comments which discuss moderators to produce our dataset of mod discourse (§3.1.3). In addition, we collect supplementary data about moderators (§3.1.3) and communities (§3.1.3).

Computing Moderator Timelines

An understanding of *who* moderates *which* subreddits *when* is critical to our analyses. We reconstruct timelines of the 10,000 largest subreddit's moderators using Reddit's API and snapshots from the Wayback Machine, a web archiving service provided by the Internet Archive [259]. We start by scraping current moderator info pages for the 10,000 largest subreddits on the platform, directly from Reddit. Each such

page contains a list of the current moderators, in order of seniority, along with the exact timestamp that each moderator was added as a moderator to the subreddit. We then use the publicly accessible API from the Internet Archive's Wayback Machine to scrape every archived copy of every subreddits' moderator info page going back to 2010. We scrape 30,302 historical archived copies of moderator info pages, which, combined with the 10,000 present-day copies (one per subreddit), give us detailed timelines of who moderated each subreddit when, and when they started as a moderator. Due to the functionality of the Wayback Machine, we have higher temporal resolution (more archived snapshots) for larger and more popular subreddits, yet, because each moderator info page encodes the exact timestamp for each moderator's start date, having fewer snapshots results only decreased accuracy of when moderators resign their posts, not when they are appointed — which is the primary focus of many of our analyses. Since the exact end time of a moderator's tenure can only be inferred from examining when they were removed from the list of moderators for each subreddit, we adopt a conservative strategy which deliberately underestimates the length of moderators' tenures: we consider the end date of their tenure as that of the last snapshot for which they were still listed as a moderator. Our method only misses moderators whose appointment, entire tenure, and resignation all occur within the 'gap' between snapshots[2]. Note that to preserve moderator privacy, we do not make our moderator timelines public.

Community Topic and Health Measures

Understanding a community's topic and health is critical for understanding that community's perceptions of their moderators. We used a few-shot GPT-4-based classifier to classify the topic of every subreddit included in our analyses, based on their names, into six topical categories taken from existing work: Discussion communities, hobby communities, meme communities, news communities, and video/picture-sharing communities [284]. The prompt used for this class is given in Appendix D.2.1. We used the manually-labeled dataset of 123 subreddits from Weld, Zhang, and Althoff [284] for our few-shot examples, as well as to evaluate the performance of the classifier, which has 86.1% accuracy on the test set, with a macro-average F_1 score of 0.858. To understand communities' health, we leverage a recent survey of members of 2,151 different subreddits. Survey participants were asked to rate the current state of their community on an

[2]We believe this is rare, as the average mod tenure we can detect is longer than four years, and the mean gap between snapshots is 56 days.

11-point Likert scale with regards to nine aspects of community health such as the quality of content, the trustworthiness of the community, and the safety of the community [284]. We average across all survey responses for each subreddit to compute an overall subreddit score for each value. Our analyses of differences between communities with different topics and health aspects are in §3.1.4.

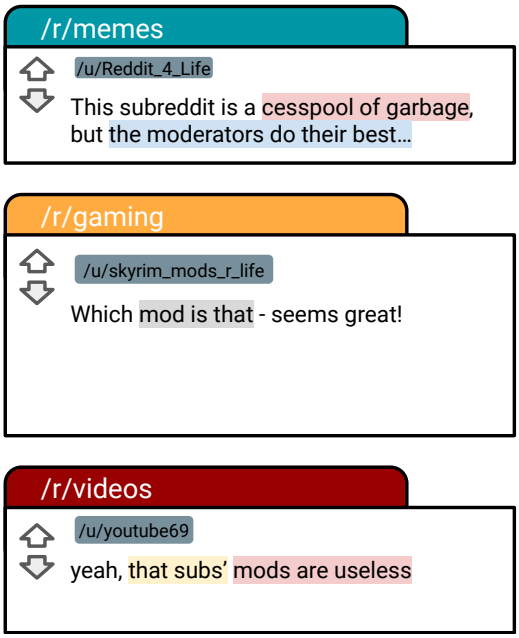
Detecting and Classifying Community Members' Perceptions of Moderators

In this work, we quantify community members' publicly stated *perceptions* of the quality of moderation by automatically detecting and classifying public posts and comments which discuss moderators. We focus on public discussion of moderation for several reasons. First and foremost, public discussion of governance is a key component of civic discourse in an offline context [6], and our work is the first to show that discussion of moderation in online communities can be measured to make meaningful insights. Furthermore, while there may be discussion of moderation in private, attempting to extract these private discussions is ethically nonviable, and surveying community members for their private thoughts on moderation, is much more expensive and less scalable than our approach.

Our detection and classification pipeline to detect and classify community members' perceptions of moderators consists of three stages: (1) a prefilter step which uses regular expressions to identify content written by non-moderators which include the words 'mod(s),' 'moderator(s);' (2) a detection step, which detects content discussing moderators (differentiating them from those which make other use of 'mod', *e.g.,* 'video game mods'); and (3) a classification step, which classifies the sentiment of this content into positive, neutral, negative, and exclude classes. We only include content written by non-moderators as the goal of our work us to examine how non-moderators perceive their mod teams, not how mods discuss themselves in public. We apply our detection and classification pipeline to all Reddit posts and comments made from January 2020 to June 2021, and we make the resulting dataset public[3].

Prefilter Step Details (Step 1). The prefilter step efficiently identifies posts and comments which discuss moderators. We use a regular expression-based filter to find all posts and comments which include the words 'mod(s),' 'moderator(s), and use our moderator tenure timelines to exclude posts and comments written by moderators during their appointment, as our goal is to measure how non-moderator community members

[3] `https://behavioral-data.github.io/moderator_perceptions_public/`



1. Distracting expressions have strong sentiment, but not aimed at the moderators. The correct classification here is positive.

2. Ambiguous terms require context, in this case the name of the community, to interpret that (in this example) 'mods' means video game modifications, not moderators. The correct classification here is exclude.

3. Definite articles must be dereferenced to determining that the author is referring to the mods of a different community. The correct classification here is exclude.

Figure 3.1: Determining the sentiment with regards to the moderators of comments can be very challenging.

discuss moderators, not how moderators discuss themselves.

Detection Step Details (Step 2). On Reddit, the term 'mods' often refers to 'moderators,' but often is used as shorthand for 'modifications,' as in 'video game mods' or 'car mods.' To differentiate posts and comments which discuss moderators from those which use 'mod(s)' in other senses, we fine-tuned a RoBERTa-based binary classifier [169] using a manually-labeled dataset of 1,155 posts and comments, randomly sampled from the prefilter step results and further divided into a training set of 655 and a test set of 500 posts and comments. RoBERTa was chosen for its high-performance, and its relative ease of training and minimal compute requirements. To improve performance and provide additional context to the model, we input not only the body of comments (or title and selftext of posts), but also the name of the subreddit the comment/post was made in, as well as the parent comment that the comment being classified was in reply to, when applicable. Our fine-tuned detector model has a precision of 0.82 and a recall of 0.94 on the test set, with an F_1 score of 0.88.

Classification Step Details (Step 3). The classification step classifies posts and comments based on their sentiment *with regards to the community moderators.* A comment with an overall-negative sentiment may have a positive sentiment with regards to the moderators, and vice versa (Figure 3.1). To label posts and com-

	Human	Random	VADER	GPT-4	Our Model
Positive F_1	*0.85*	0.14	0.30	0.70	**0.71**
Neutral F_1	*0.89*	0.50	0.34	0.71	**0.73**
Negative F_1	*0.84*	0.13	0.34	0.61	**0.71**
Exclude F_1	*0.98*	0.21	0.00	0.66	**0.73**
Test Set Acc.	*89.0%*	27.0%	30.6%	66.2%	**72.4%**

Table 3.1: Our Classification Step model, a LLaMA 2 model fine tuned with QLoRA [267, 56], exceeds the performance of a retrieval-based few-shot classifier using GPT-4. Our model is also more scalable (it can be deployed locally), more affordable, and more reliable (it is not subject to prompt filtering), than GPT-4. This table compares F_1 scores for expert human labelers, retrieval-based few-shot GPT-4, and our model, alongside an empirical class distribution random baseline and a VADER-based classifier [107].

ments for the classification task, two annotators worked together to iteratively refine a codebook (Appendix A.5), then independently labeled 200 randomly sampled posts/comments. The annotators had 'almost perfect' inter-annotator reliability (0.85 Fleiss' kappa) [162]. To produce a 'gold standard' test set, the same two annotators independently labeled a random sample of 500 posts and comments, then discussed their disagreements until consensus was reached. A single annotator then labeled an additional 734 posts/comments for use as a training set, of which 484 were randomly sampled, and 250 were sampled based on their proximity to the decision boundary of a simpler RoBERTa model trained for the classification task.

As the purpose of this method is to identify perceived moderation quality, we must use care to identify the moderators that posts/comments are discussing. On Reddit, community members occasionally discuss moderators of *other* communities, *e.g.,* a member of /r/gaming praising the moderators of a specific Discord server, or a member of /r/nfl complaining about the moderators of /r/seahawks. To ensure correct attribution, we decided to limit our analyses to community members talking about their *own communities'* moderators (*e.g.,* discussion of the /r/cats moderators taking place on /r/cats). For this, we specifically trained our model to identify posts and comments which discuss the moderators of other communities, along with content that erroneously passed the detection step. We trained on this 'exclude' class in addition to those used for downstream analyses: positive, neutral, and negative.[4] Note that distinguishing between positive, neutral, and negative comments requires capturing a broad range of nuanced comments discussing specific aspects of moderation, community specific concerns, and non-straightforward language such as sarcasm. For example, the comment '*I've reported your comment to the mods but I don't expect them to do anything because transmisogyny is totally allowed here.*' posted in a trans rights community is correctly labeled as

[4]Additionally, we attempted to classify specific complaints about moderation: excessive moderation, insufficient moderation, and biased moderation, but found classifier performance not to be amenable for high-confidence downstream analyses.

having negative sentiment towards the moderators by our final model, despite referring to a specific aspect of moderation (addressing reports) and a community specific topic (in the context of this community, misogyny is unwanted by community members, whereas other communities may have different attitudes).

To further enhance the performance of our model, we performed data augmentation using a retrieval-based few-shot classifier built with GPT-4 [202]. We used this classifier to label an additional sample of 10,000 posts and comments, which were combined with our manually-labeled training set to fine-tune our final classification model, a 13-billion parameter LLaMA 2 model [267] fine-tuned with QLoRA [56], using the prompt shown in Appendix D.2.2. LLaMA 2 and QLoRA were selected for their very high performance while still being feasible to fine-tune and deploy on a massive dataset. Our model was fine-tuned on an internal university HPC cluster with $2\times$ NVIDIA a40 GPUs, which took $\approx$ 13 hours. Our final model exceeds the performance of GPT-4 on all classes (Table 3.1). Applying the finalized Classification Step to the results from the Detection Step left us with a labeled dataset from 8,477 communities of 196,231 posts and 1,694,551 comments which discuss moderators: 175,296 with positive sentiment, 968,235 with neutral sentiment, and 747,251 with negative sentiment, and we make this dataset public[5].

Accurately classifying the sentiment with regards to the moderators of a post or comment is an extremely challenging task that is often heavily reliant on context and background knowledge of the community, and the sentiment towards the moderators specifically often differs from the overall sentiment of the entire comment (Figure 3.1). As such, off-the-shelf sentiment classifiers, such as VADER, perform poorly, with accuracy close to that of a random classifier (Table 3.1). Even general purpose LLMs such as GPT-4, which obtain state-of-the-art results on standard sentiment analysis benchmarks [139], perform worse than our fine-tuned LLaMA2 QLoRA model, even when prompted using a retrieval-based few-shot in-context learning method.

Interpreting Mod Discourse. Finally, to enable downstream analyses, we define several aggregate values for each community, used in §3.1.4 & §3.1.6: The *Amount of Mod Discourse* for a given community is the fraction of all posts and comments discussing mods in that community, regardless of their sentiment. The *Composition* of Mod Discourse with positive/negative sentiment is the percent of all posts and comments discussing mods in a given community that have positive/negative sentiment. To ensure a meaningful

[5] `https://behavioral-data.github.io/moderator_perceptions_public/`

amount of data, we exclude all communities that did not have at least one positive, neutral, and negative post or comment discussing moderators during our 18-month analysis period. Finally, we exclude nine communities explicitly devoted to discussing moderators, as it is infeasible to differentiate discussion of those communities' moderators from other moderators. This filtering leaves us with 5,282 subreddits used in downstream analyses.

Validating our Measures of Perceptions of Moderators

Communities attitudes towards their moderators are complex and nuanced, and therefore impossible to ever measure perfectly. As such, there is no 'ground truth' metric for governance outcomes that we could compare our method to. Every method for measuring governance outcomes has its own advantages, disadvantages, and biases. For example, even survey-based methods which directly ask community members about their perceptions of their moderators suffer from selection biases as not everyone will respond to a survey [284].

Our methods for detecting and classifying community members' perceptions of moderators (§3.1.3) measure when community members publicly post or comment in their own communities to discuss moderators. For ethical as well as practical reasons, we cannot measure when community members discuss their moderators *privately*. However, there are two other instances where community members may attempt to discuss their moderators publicly but would not be included in our analyses. In this section, we show that these two cases are very rare relative to the public moderator discourse that we *do* measure.

For purposes of computational efficiency, our initial prefiltering step only includes posts and comments which use the phrase 'mod(s)' or 'moderator(s)' for subsequent classification. Occasionally, however, community members may discuss their moderators by mentioning their usernames specifically. For example, instead of writing 'the mods of this subreddit are unfair,' which would be captured by our pipeline, a community member might write '/u/exampleModerator is unfair.' To assess the frequency of this form of moderator discourse, for a sample period of one month (January 2018, the first month of our analysis period), we compute how frequently the usernames of *any active moderator* is mentioned directly in their own subreddit, in addition to the frequency of uses of 'mod(s)' or 'moderator(s).' We find that such specific moderator mentions are very rare, with the average subreddit having 64.11 generic uses of 'mod(s)' or 'moderator(s)' for

every time a moderator is mentioned specifically by their username. We assessed the impact that including these specific mentions would have on downstream analyses, and concluded qualitatively that they did not make a substantial difference in our results. Thus, for purposes of computational efficiency, we chose to exclude specific moderator mentions from our moderator discourse classification pipeline.

An additional potential source of bias in our analyses is that of removed posts and comments. On Reddit, moderators are occasionally accused of removing content in their communities that is critical of the moderators [181]. As this content could be removed before entering our moderator discourse classification pipeline, it is conceivable that our results undercount moderator discourse in communities where the moderators remove such content. Assessing this source of bias requires knowledge of the content that is removed, which is technically infeasible to collect at a large scale. However, prior researchers have collected content before it is removed by moderators by scraping content from specific subreddits as it is posted, then checking later to see if the content has been removed [37, 153]. We used the Pushshift API (before it was taken offline) in the same manner, to collect the text of content that was removed from 100 randomly sampled subreddits over a two year period (2017-2018, inclusive). Using this method, we collected 263,657 pieces of removed content from 100 different communities. Of the content that was removed by moderators in this sample, we find that it is extremely rare for moderators to remove content that mentions the moderators: The average community in our sample removes 102 posts/comments which do not mention the moderators for every post or comment removed that does mention the moderators. Once again, we assessed the impact that including removed comments would have on our downstream analyses, and concluded that their inclusion would not make a substantial difference in our results (in addition to be infeasible at the scale of the rest of this work).

3.1.4 How are moderators of different communities perceived differently by their communities?

Online communities exist for nearly every conceivable topic, and range in size from just a few members to many millions. In this section, we examine how community members' perceptions of their moderators vary across communities with different topics, different community health metrics, and different sizes.

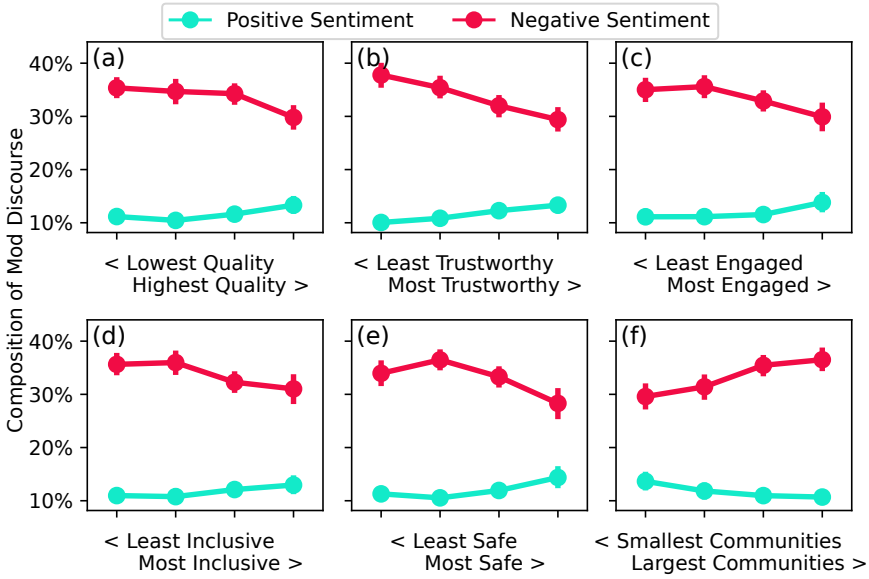


Figure 3.2: Communities that consider themselves higher quality (a), more trustworthy (b), more engaged (c), more inclusive (d), and more safe (e) all use more positive and less negative sentiment to describe their moderators.

Here, communities are grouped into quartiles based on their community members' self-reported perceptions of the current state of the community. This effect is most pronounced for communities' self-reported trustworthiness (b), with the top-25% most trustworthy communities using 34% more positive and 22% less negative language to describe their mods. Communities that rate themselves as feeling smaller (f) have a more positive perception of their mods. In this and all other figures, points represent mean estimates alongside bootstrapped 95% confidence intervals.

Method

To quantify a community's health, we divide communities into quartiles based on their community members' responses to a recent survey (§3.1.3). Survey responses were collected between May-July 2021, a period overlapping the the end of our mod discourse data time range. For this analysis only, we exclude communities which were not surveyed. Community topic is classified into one of six topical categories using our topic classifier (§3.1.3). We quantify community size as the volume of submitted content: the average number of posts and comments per day over our study period, which we use to group communities with similar size. For each group, we compute the amount of mod discourse as well as the composition of that discourse.

Results

Different aspects of community health are associated with better and worse perceptions of moderators. The smallest-feeling 25% of communities, based on member self-reports, use 27% more positive and 16% less negative sentiment to discuss their moderators than the largest-feeling 25% of communities (Figure 3.2a). Figure 3.2a-e shows that communities which rate themselves as having higher quality content and being more trustworthy, more inclusive, and more safe all use more positive and less negative to discuss their moderators, as well.

Direct measurement of communities' size allows us to further investigate the relationship between the volume of content submitted to a community, and its members' perceptions of its moderators. In general, smaller communities have more mod discourse (Figure 3.3a), with tiny communities with fewer than ten posts+comments per day having $7.6\times$ as many posts and comments discussing moderators as huge communities (those with more than 10k posts and comments per day), relative to the total amount of content. Tiny communities also have $6.0\times$ more positive mod discourse, and $0.46\times$ as much negative discourse, than huge communities.

Examining communities of different topics, discussion communities have the most mod discourse, on average (Figure 3.3b). Hobby communities have generally the most positive mod discourse, with meme and news-sharing communities having 46% more of their mod discourse have negative sentiment than communities of other topics (Figure 3.3d).

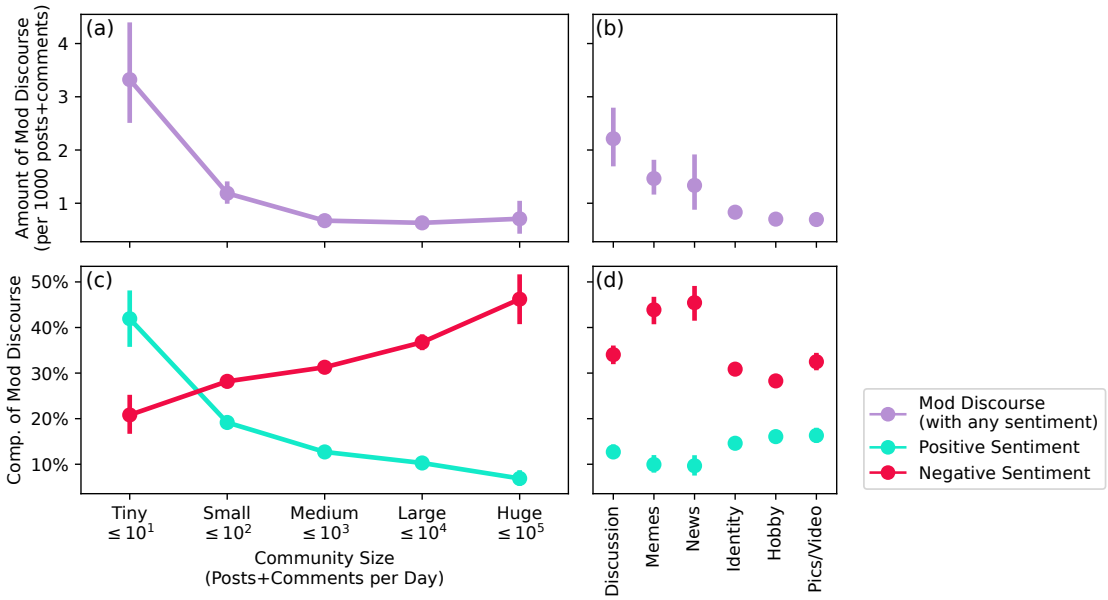


Figure 3.3: Perceptions of moderators vary significantly across communities with different sizes (a,c) and topics (b,d). In general, smaller communities devote a relatively larger proportion of their content to discussing their moderators (a), and smaller communities express more positive and less negative sentiment towards their mods (c). Discussion, meme-sharing, and news communities have proportionally more mod discourse (b), while meme and news-sharing communities exhibit the most negative sentiment towards their moderators (d).

Implications

Surveys of Redditors have shown that they consider a wide range of factors to be important to the overall 'health' of their communities [283]. Here we show that communities that are doing well with regards to factors widely considered to be important, such as safety and quality of content, tend to perceive their moderators more positively. Many of these factors are only indirectly controlled by the moderators, for example quality of content—while moderators can set and enforce rules aimed at improve the quality of content in their communities, quality of content is ultimately a function of the content submitted by community members, not moderators. This can lead to moderators being 'blamed' for problems largely outside of their control, which has been previously shown in small-n surveys [181, 117] and is further supported empirically here, as shown in Figure 3.2.

Small communities appear to both discuss their moderators more, as well as use more positive sentiment in their discussions (Figure 3.3a,c). Several factors may contribute to this, including that the increased anonymity that comes with participating in a larger community may make people feel more comfortable

speaking negatively about the moderators, and that smaller communities are more likely to be newly formed and thus still establishing moderation norms, leading to more mod discourse [108, 241].

3.1.5 Adjusting for Confounders using IPTW & DID

While we are ideally interested in causal, actionable insights, our study is a retrospective observational study, and like any other, subject to potential confounding. In §3.1.4 we identified two important confounders, community size and and community topic, which are correlated with perceptions of moderators (Figure 3.3). These confounding factors are particularly important to adjust for because they are relevant to every community, are correlated with perceptions of moderators, and are out of the direct control of moderators. The community topic in particular is inherent to each community and is with few exceptions unable to change after a community is formed. It is also associated with many deeper aspects of community interactions that are more challenging to directly measure, such as the form that community interactions take (*e.g.,* text-based discussion *vs.* image sharing). Next we describe two methods from the causal inference literature which we use to adjust for community size and topic in subsequent analyses.

Inverse Probability of Treatment Weighting (IPTW). IPTW is a statistical method to adjust for imbalance in potential confounding factors when making comparisons between different observed groups (*e.g.,* a treatment and a control group). For example, consider the moderator workload analysis from §3.1.6. Here, we wish to assess the association between a moderator team's workload and that community's perceptions over their moderators. The straightforward, naive approach would be to simply take communities with high moderator workloads and communities with low moderator workloads, and look at the difference between these two groups' average perceptions of their moderators. However, this naive approach does not adjust for confounding factors, such as community size. Larger communities generally have higher moderator workloads than smaller communities, and larger communities also generally have more negative perceptions of their moderators (Figure 3.3b). This potential confounder is visible as an *imbalance* between the two groups: the average size of a community with a low moderator workload is smaller than the average size of a community with a high moderator workload. Intuitively, IPTW works by *reweighting* individual observations within each group in order to counter this imbalance. In the group of low moderator workload communities, larger communities are weighted more heavily, as they are relatively uncommon within this group. In the

other group, the opposite effect occurs: in high workload communities, *smaller* communities are weighted more heavily, while larger communities are weighted less heavily. The end result is that, after reweighting, both groups are balanced, with similar distributions of all covariates (including potential confounding factors). For each analysis, we manually verified balance to check the efficacy of reweighting (Appendix D.4). We select IPTW over other methods, such as stratification, due to its excellent efficiency [13].

More precisely, in §3.1.6, we compute the probability of treatment (propensity score, $P(Z|\mathbf{X})$) by applying a logistic regression model to community covariates $\mathbf{X}$ including the topic and size of the community. The precise set of covariates used varies from analysis to analysis. A complete list of covariates for each analysis is given in Appendix D.4. We generalize IPTW to non-binary treatments to estimate a dose-response curve using the method from Althoff et al. [8] in which the treatment is discretized into bins and a different propensity score is computed for each treatment bin in a 'one-versus-rest' scheme. For final analyses, each observation is weighted by the inverse of the probability of the treatment it received, such that the weight for observation i is $w_i = \frac{Z_i}{P(Z_i=1|\mathbf{X}_i)} + \frac{1-Z_i}{P(Z_i=0|\mathbf{X}_i)}$, where Z_i is the treatment received by observation i (0 for control, 1 for treated), and $\mathbf{X}_i$ is a vector of the covariates. In analyses where we use IPTW, we also include the non-IPTW adjusted estimate in figures with a gray color. In these analyses, reweighting does not dramatically impact our findings, suggesting that the impact of these confounders is moderate at best.

An important validity check for IPTW is to assess the balance of covariates for each treatment group after weighting is applied [13]. Two groups are often considered 'balanced' or 'indistinguishable' if all covariates are within a standardized mean difference (SMD) of 0.25 standard deviations [8]. For each treatment group, we compute the difference between each covariate's weighted mean value and the *reference distribution*, consisting of the entire population. We compute the standardized mean difference (SMD) by normalizing the difference in means by the standard deviation of the values of the reference distribution. Our reweighting method achieves balance for every covariate substantially related to the treatment. SMDs after weighting are given for each covariate and each analysis in Appendix D.4.

Difference in Difference Analyses (DID). Difference in difference analyses enable the estimation of the impact of an intervention by comparing the outcome *before vs. after* the intervention was applied to the treated group, and comparing this before *vs.* after difference to an untreated control group. For our analyses of the impact of moderators' community engagement (§3.1.6), we use modified DID analyses with multiple

time periods [32] to compare the difference in communities' receptions of new moderators who are engaged with the community (the 'treatment') *vs.* those who are not (the 'control'). Two four week long periods immediately preceding and following each moderators' appointment were used to compute the first-order difference. Since mod discourse naturally varies over time, averaging over many DIDs from many new moderator appointments serves to reduce confounding background temporal trends by comparing the difference to an untreated control group.

3.1.6 What moderation strategies might improve perceptions of moderators?

Moderators have a great deal of autonomy to run their communities as they see fit, including enforcing rules by removing content and growing the mod team by recruiting or appointing new moderators. Moderators also may (or may not) participate in the community as 'regular community members' in addition to their mod duties. In this section, we identify promising suggestions for moderators by comparing communities whose moderators have different workloads, rule enforcement strategies, and degrees of community engagement.

Moderator Workload and Content Removal

Method. We can identify content removed by moderators in each community by counting the occurrences of '[removed]' posts and comments within each community. By dividing by the total amount of content submitted to that community, we can compute the total percentage of content removed by mods. Using our mod timelines (§3.1.3), we can compute the total number of moderator-tenures in any given subreddit over the course of our analysis time period. We also approximate 'workload' of each mod by dividing the total amount of content submitted to each community by the number of mods available to review that content; while in reality it is unlikely that all moderators share the work of reviewing content evenly [165]; this metric nonetheless helps us understand the ratio of moderators to content within each community. If workload was shared unevenly, this would only make the workload even higher for a subset of moderators, leading to at worst to overly conservative estimates (higher workload to be addressed by fewer moderators).

Results. We find that communities with higher mod workloads have more negative mod discourse (Figure 3.4). Communities with fewer than five pieces of daily content for each mod (8.8% of all communities) have approximately equal amounts of mod discourse with positive and negative sentiment, a rarity on a platform

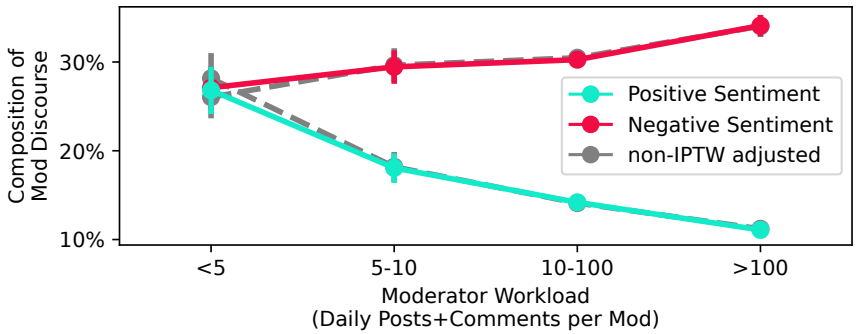


Figure 3.4: Moderators in communities with lower workloads are perceived more positively and less negatively than moderators in communities with high workloads. Communities with lower moderator workloads (more moderators relative to the amount of content submitted) tend to have more more positive sentiment in their discussion of the moderators, and less negative sentiment. Communities with fewer than five posts and comments per mod per day use $2.5\times$ as much positive sentiment in their mod discourse compared to communities with more than 100 posts and comments per mod per day.

where mods are far more commonly discussed negatively. In communities with higher mod workload, the composition of mod discourse is much less positive, with communities with more than 100 posts and comments per mod per day using positive sentiment to describe their mods only $0.39\times$ as often as the 8.8% of communities with the lowest mod workload.

Generally, we find that the composition of mod discourse is more negative in communities with more removed content (Supplementary Figure D.1). However, these trends vary depending on the topic of the community, with news communities in particular showing the opposite trend (Figure 3.5).[6] Amongst news communities that remove $\leq 1\%$ of their content, 48% of mod discourse has negative sentiment, which *drops* 11pp to 39% for news communities whose mods remove between 2% and 3% of their content. For the same amounts of removed content for non-news communities, the fraction of mod discourse with negative sentiment *increases* 6pp, from 26% to 33%.

Implications. Our results suggest that adding additional moderators to a community (and therefore reducing the effective moderator workload) may improve community members' perceptions of their mods (Figure 3.4). In §3.1.6 we report on additional longitudinal evidence that recruiting additional moderators can have a positive effect. However, simply increasing the amount of content which is removed does not, in general, appear to be associated with more positive moderator discourse sentiment—in fact, the opposite appears to

[6]Similar figures for each of the six community topics are included in Supplementary Figure D.2.

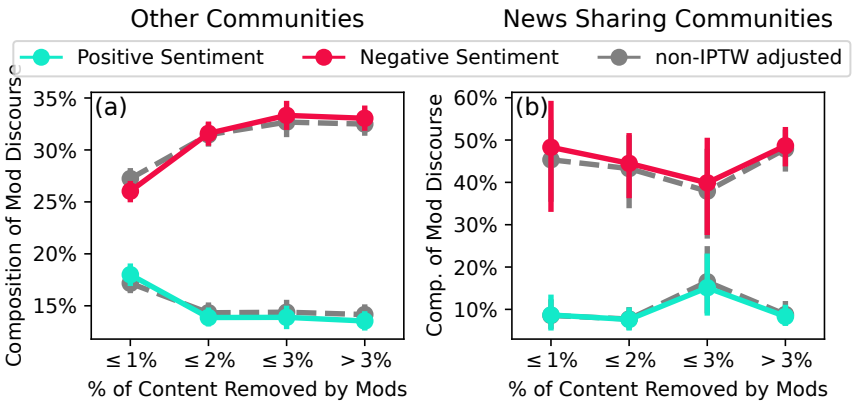


Figure 3.5: For most topics, communities where moderators remove more content exhibit *more* negative sentiment (a). News communities, however, buck this trend, with the fraction of mod discourse with negative sentiment 11 percentage points *lower* in news communities whose mods remove less than 1% of content compared to communities whose mods remove 2% - 3% of content (b).

be the case for communities that are not focused on sharing news (Figure 3.5). Taken together, these results imply that there are topic-specific nuances to content removal, and that mods should use care when deciding how strictly to enforce rules, and how much content to remove. Our results suggest that certain topics are more amenable to stricter rule enforcement than others.

Community Engagement

Method. Some moderators are actively engaged with community, soliciting feedback from non-moderators, updating community members on moderation-related news, and contributing to regular community content in addition to their official moderator duties. Other moderators are far less visible, opting to remove content and change rules without participating in the community more broadly. Furthermore, many moderators also participate in other communities beyond just the one(s) they moderate. Using moderators' public posts and comments, for each moderator-appointment to a community, we compute the number of posts and comments they made in that community and in other communities before and during their tenure as a moderator. We use DID analyses (§3.1.5) to estimate the impact of appointing new mods with different degrees of community engagement (or lack thereof).

Results. We find that, regardless of the moderators' previous engagement with the community, adding a new moderator to a community is associated with a decrease in mod discourse with negative sentiment

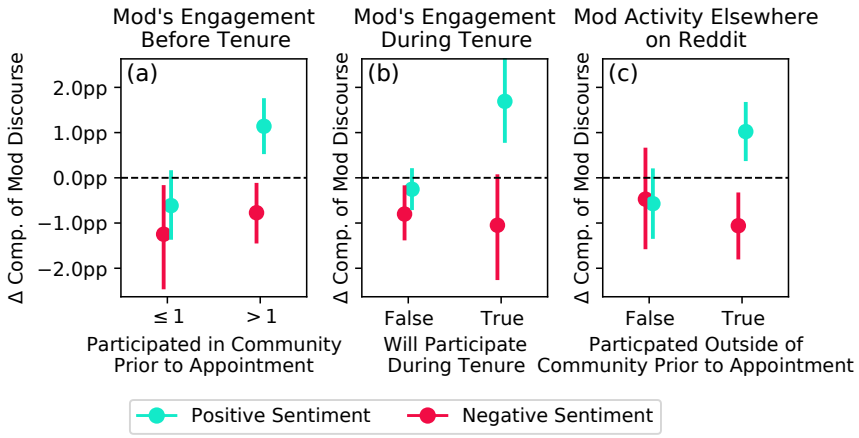


Figure 3.6: Newly appointed mods are associated with a greater improvement in mod perceptions if they are engaged in the community and elsewhere on Reddit before their tenure, and if they are engaged during their tenure (a-c). Adding a moderator who already has or will engage with the community is associated with in an increase in the fraction of mod discourse in the community with positive sentiment (aqua), and that increase is 32.5% larger when adding a mod who will engage with the community going forward (b) than for one who already has (a). Adding a moderator who is an active member of communities other than the one they are becoming a mod of is associated with an increase in positive, and a decrease in negative, sentiment in mod discourse (c).

(Figure 3.6). However, some moderators are associated with bigger changes than others. Moderators who are engaged with the community prior to their appointment (Figure 3.6a), will continue to engage with the community *during* their tenures (Figure 3.6b), and were also active in *other* communities prior to their appointment (Figure 3.6c) are all associated with the largest improvements in perceptions of moderators. New mods who do all three of these things are associated with a 2.5pp increase in positive mod discourse, whereas mods who do none of these things are associated with a 0.5pp *decrease*.

Implications. Our results suggest that moderators who actively engage with the community both before and during their tenures are may have a more positive impact on the community's perception of their mod team, which may be because these moderators are more familiar with the community, and are better mods as a result. Another plausible mechanism for this effect is that non-moderator community members value the transparency and accountability that may stem from increased moderator engagement. Our results also suggest that moderators who are active in other subreddits beyond the one(s) they moderate may have a more positive impact on the communities they moderate, perhaps because participating in a broader range of communities makes them more effective moderators [293]. Lastly, our results suggest that of these factors, engagement with the community *during* the moderator's tenure is associated with the largest improvement

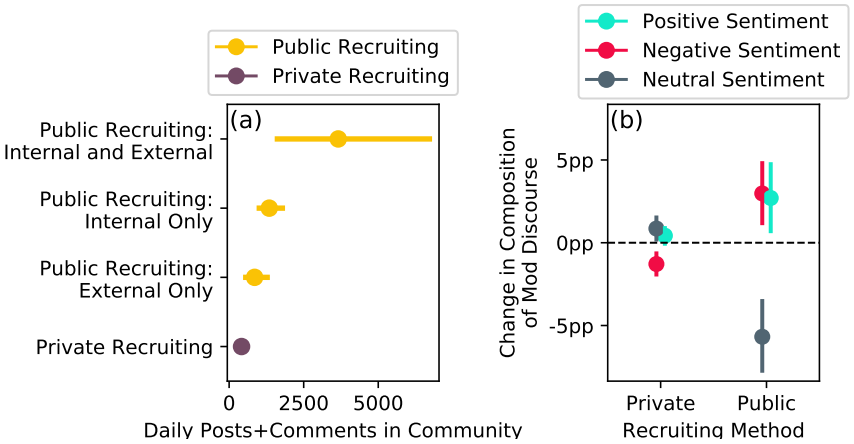


Figure 3.7: Communities that recruit moderators publicly are $8.78\times$ larger than the average community which recruits only privately, in terms of the community's daily volume of content (a). Small communities lean towards private recruiting. (b) Compared to private recruiting, recruiting moderators publicly is polarizing: it is associated with an increase in *both* positive and negative fractions of mod discourse, and a corresponding decrease in neutral sentiment.

in perceptions of moderators.

Moderator Recruiting

When it's time to grow the mod team, existing moderators have a wide array of options for who to recruit, and how to recruit them, and little guidance, official or otherwise, for how to select new mods. In practice, moderator recruiting falls into two different strategies: *Public Recruiting*, where moderators post publicly *internally* in their own subreddit that they are looking for moderators, and solicit applications, nominations, or hold elections. Sometimes, moderators make use of special *external* moderator-recruiting subreddits, such as /r/needamod, where they can post 'job listings' to prospective applicants. By contrast, *Private Recruiting* is the recruiting of new moderators in the absence of any public recruiting activity. Moderators recruited privately are either invited to join directly by an existing mod, make an offer to moderate themselves, or are recruited via other backchannel means.

Method. We identify instances of public recruiting using a regular expression-based search for posts made by sitting moderators which use 'recruiting new mods,' 'applications open for new mods,' 'holding mod elections,' or similar phrases. We also apply a regular expression-based filter to the complete set of posts from /r/needamod to identify which subreddits are recruiting, restricting the results to only posts made by current moderators of the subreddit that is recruiting. We then take our dataset of moderator timelines

(§3.1.3) and match moderators who were appointed to a community within 8 weeks of a public recruiting post as having been publicly recruited. Moderators appointed without a recent public recruiting post are considered privately recruited.

Results. We find that, while communities of all sizes use public recruiting methods, larger communities are the most likely to make use of public recruiting (Figure 3.7a). In terms of its daily volume of content, the average community that uses both Internal and External Public Recruiting is $8.78\times$ larger than the average community that only recruits privately. Examining the impact that adding a single moderator recruited privately vs. publicly has to a community in Figure 3.7b, we find that public recruiting appears to be polarizing, with the fractions of both positive *and* negative mod discourse increasing by 2.7pp and 3.0pp, respectively, after a publicly recruited mod is added, on average.

Implications. Our findings suggest that public moderator recruiting has the potential be a powerful tool to improve community perceptions of moderators when used carefully, and can be harmful when used without regard to the community's preferences. A plausible explanation for the increased polarization resulting from public recruiting is that in some circumstances, the public nature of the recruiting exacerbates existing frustrations with mod teams, for example if due process was not followed during moderator elections, or if a perceived-outside was brought in via external recruiting when community members themselves preferred someone with more experience in the community. More work is needed to assess the differences between different moderator recruiting strategies.

Novice and Experienced Mods

Method. When selecting new moderators, existing moderators may be inclined to favor candidates who already have some moderation experience in other communities. Using Moderator Timelines (§3.1.3), we can accurately assess how much experience a new moderator has at the time the are appointed, if any.

Results. We find that large communities are much more likely to recruit mods who already have moderation experience (Figure 3.8). 74% of mods appointed to communities with fewer than 10 mods are first time mods, while 94% of mods appointed to communities with > 100 mods have more than 2 years of experience.

Implications. Large communities with large moderator teams rarely appoint novice moderators, perhaps

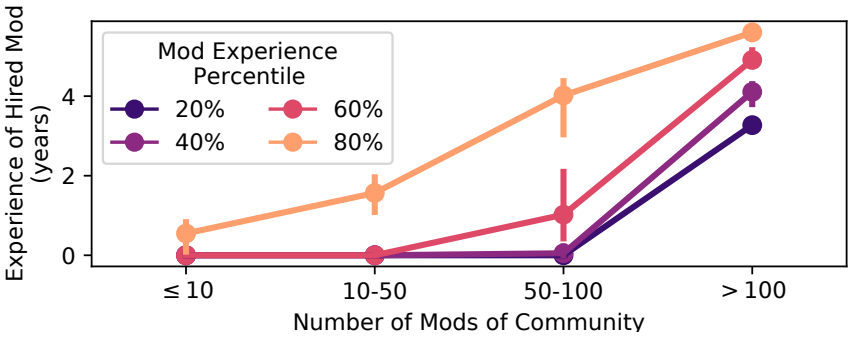


Figure 3.8: In contrast to small moderator teams, large teams are much more likely to appoint new moderators who already have moderation experience. 94% of mods recruited to join mod teams with more than 100 mods have at least 2 years of experience, while 74% of mods who join small teams with fewer than 10 mods have no previous experience at all.

due to a perception that a large community is not an appropriate place for new moderators to gain experience. As such, it seems that the most common pattern on Reddit is for new moderators to start off moderating small communities, and then work their way up to larger ones. This may contribute to a perception, especially amongst non-moderators, that moderators are motivated to increase their number of appointments and to moderate larger and larger communities, potentially biasing the performance of their duties.

3.1.7 Discussion

The massive diversity of online communities offers enormous potential to empirically study how to make online communities better. Any such studies, however, require robust and scalable methods to quantify both the outcomes (which communities are doing 'better' than others), as well as the independent variables (the aspects of communities that might make them better). Measuring the success of a community's moderators is more challenging than many other aspects of making a community 'better,' as unlike, say, misinformation, governance is far less visible in a community, and its success is far less well defined. The primary contribution of our work is the use of community members' own *perceptions* of their moderators, leveraging millions of people's perceptions of good governance, rather than attempting to define good governance ourselves.

We are tremendously excited about the potential synergies this line of work enables. While the findings presented in this paper are impactful by themselves, combining our measure of governance with many other important measures of community outcomes (safety, inclusion, and discussion quality, for example) will enable studies of how to make communities better that are both more comprehensive and more robust than

previously possible.

Diversity of Communities' Mod Perceptions

Our results show that communities' perceptions of their moderators are highly varied (Figure 3.2 & 3.3) but are associated with other aspects of community health and size. This suggests that not only do different kinds of communities have different norms for their mod discourse, but also that community members' perceptions of overall health of their community, which includes many factors outside of the moderators' direct control, also influences their mod discourse. Researchers and moderators must carefully consider the specific needs of their communities.

Topic Contention Affects Governance

Our results suggest communities with different topics have different perceptions of the strictness of moderators' removal of content and rule enforcement (Figure 3.5, Appendix Figure D.2). For example, we find that in news communities where moderators remove more content, community members' perception of moderators is more positive (up until $\approx 3\%$ of content is removed), while for communities with other topics, perceptions of moderators are more negative in communities where mods remove more content. This suggests that communities with certain topics, perhaps more contentious ones, are more appropriate to moderate fairly strictly than others.Future work should specifically examine the impact of community topic on community members' perceptions of content removal.

Community Overestimates of Moderators' Power

There are many aspects of every community that moderators have limited-to-no control over. For example, our results show that perceptions of moderators are most positive in smaller communities (Figure 3.2f). However, although they can set and enforce policies with the goal of growing or shrinking a community, moderators cannot directly control a community's size. Thus, our results support previous research that moderators are frequently 'blamed' for problems that they cannot easily resolve [181]. Future work could address this tension directly, perhaps by attempting to educate community members on the roles and powers of moderators, as well as the powers that moderators *do not* have.

Mod Recruiting and Engagement

We examine instances of moderators being added to moderator teams to compute the changes in perceptions of moderators that occur when different moderators are added to a mod team. We find that more engaged moderators are associated with an increase in the amount of positive mod discourse, perhaps as a result of an increased sense of transparency and/or accountability for community governance (Figure 3.6a-c). We also look at differences between moderators whose recruiting was discussed publicly, versus those who were recruited in private. We find that publicly recruited moderators appear to be more controversial, with both negative *and* positive discourse increasing after their appointment (Figure 3.7b). This suggests that moderators must use caution when recruiting moderators publicly, and be sure to consider community preferences.

Implications for Platforms and Moderator Tools

Our results show that perceptions of moderators vary widely across different communities (Figure 3.3) as well as within the same community over time (§3.1.6). Our method for quantifying perceptions of moderators offers the potential for platforms to automatically identify when specific communities could use additional moderation help, which could help them target additional resources, such as via an updated version of Reddit's Moderator Reserves program [221], which provides additional experienced moderators to communities experiencing a surge in traffic.

Communities themselves could also use our method to better understand the impact of changes to the community. For example, tools could be built to provide dashboards that let communities monitor perceptions of moderation over time, or even to conduct A/B tests for changes such as rule additions and measure perceptions of moderators as an outcome, in a manner similar to CivilServant [182].

Generalizability Beyond Reddit

While our work focuses on Reddit communities, our method could be applied to any community with where moderators are discussed, such as Facebook Groups and Discord servers. For these applications, straightforward changes to the regular expressions used in the Prefilter Step (§3.1.3) may be necessary, for example as Facebook Groups call their mods 'admins.' More broadly, future work could apply similar

methods to study how governance is discussed in other contexts, such as peer production communities like open source software projects and wikis.

3.1.8 Limitations

Our work measures community governance through community members' public discussion of moderators. While these signals enable insights about different governance strategies, they do not capture every aspect of the success of a community's governance. What people say publicly does not always reflect their actual beliefs, and even if it did, minimizing community unhappiness is not necessarily the best objective function for community moderators. Although we conclude that excluding content that mentions specific moderators by name and removed content does not substantially impact our results (§3.1.3), differences in community members' behavior and values also may still bias our results, as different community may have different norms around mod discourse, and different community members may feel more or less comfortable expressing their opinions publicly. Additional study is needed to ensure that scalable methods for measuring governance reflect the needs of all community members, not just the noisiest. Surveys could be used to directly ask community members about their perceptions of moderators, instead of relying on publicly made comments. However, even with this strategy, it is essentially impossible to avoid selection effects related to who responds to the surveys.

Even though our pipeline for detecting and classifying perceptions of moderators exceeds the performance of state-of-the-art methods (§3.1.3), it is not perfect. The large volume of data of data used in our analyses minimizes the impact of any single misclassification by the model, and all of our figures include bootstrapped confidence intervals to better understand the robustness of our findings. Despite these robustness checks, a more fine-grained and accurate classification pipeline may be capable of providing additional insights. As NLP techniques are rapidly evolving, future work should continue to integrate new NLP techniques to improve the performance of pipelines such as ours.

Systems as complex as online communities have countless confounding factors that can bias analyses such as ours. While we attempt to control for several key confounders by using IPTW and DID causal inference methods (§3.1.5), ours are fundamentally observational results, and we cannot rule out than other confounders may have been unobserved. Future work using observational methods could include additional

covariates that are more difficult to quantify, such as how community members interact with one another, and the inherent contentiousness of the material discussed within the community. We identify many aspects of governance that are associated with more positive perceptions of moderation, and are therefore promising targets for future study, but we cannot make any truly causal claims. Future work should examine these promising interventions and make use of active experimentation, such as randomized controlled trials (RCTs) for gold-standard causal estimates.

Validating measures of online governance is challenging, as there is minimal 'ground truth' to use for assessment. Every method for measuring governance outcomes has its own biases. While it exceeds the scope of this work, a large-scale survey of many community members' perceptions of their moderators could be used to refine and validate future models. Future models could go beyond just positive and negative sentiment to identity specific critiques of moderation, such as biased, overly-strict, or too-permissive moderation, for example, and theoretical analyses of measurement approaches could be used to bound errors.

While our methods for quantifying communities' practices are sophisticated, they also miss many key aspects of governance. Our analyses of moderators' engagement with communities (§3.1.6) only consider how much each moderator posts in their communities, not *what* they post. Future analyses could examine the types of contributions that moderators make. Our analyses of moderator team dynamics are also limited by the lack of publicly available data about which moderators are active; many communities' mod teams contain mods who do not contribute to the day-to-day governance of the community. Future work should incorporate detailed information about specific moderators' actions, although data collection is a substantial challenge [165].

Our analyses also focus on Reddit communities specifically. While we study 8,477 highly varied communities which account for more than $2/3$ of content on Reddit, and thus believe our results generalize to communities of a broad range of topics, formats, and sizes, additional work is needed to understand the impact of analogous governance practices in communities hosted on other platforms, such as Facebook Groups or Discord servers.

3.1.9 Conclusion

Good governance is critical to the functioning of online communities, yet it is difficult to know what governance practices are most effective, as it is challenging to measure the 'success' of community governance. In this work, we developed a method to quantify community members' perceptions of their moderators across thousands of communities. We relate these perceptions to different aspects of governance including community size and topic (§3.1.4) as well as to different actions that moderators can take, including rule enforcement (§3.1.6), community engagement (§3.1.6), and moderator recruiting practices (§3.1.6-3.1.6). We empirically identify promising strategies for community moderators, including tailoring the strictness of rule enforcement to the community topic, and recruiting engaged mods. We make our models and anonymized datasets public to support future research.

3.1.10 Ethical Considerations

We believe this work will have a positive broader impact by informing better moderation practices in online communities, as well as providing researchers with better tools to study community members' perceptions of moderators. As we only make use of public data, we believe our work has minimal risk to participants' privacy. As research has shown that some online community users are uncomfortable with their data being used for research, even when posted publicly [71], we take further steps to reduce potential harms and misuse potential of our mod discourse dataset: we do not publish usernames or identifiable information, only predicted sentiment with regards to the moderators. We publish moderator timelines, including moderators' usernames, however these usernames are already publicly listed on communities' 'about' pages. Upon publication of this dataset, we will provide affordances for users to have their data removed at their request. We comply with relevant licenses for NLP models we use or modify. This study was approved by the University of Washington IRB under ID number STUDY00011457.

3.2 Quantifying the Impact of Community Rules on Perceptions of Moderation

3.2.1 Introduction

Rules are critical to the safe and healthy functioning of online communities, and setting and enforcing rules is central part of their governance [181, 72]. However, on many platforms, including Facebook Groups, Discord, and Reddit, community leaders have enormous leeway and minimal guidance in the rules they choose to set [150]. Furthermore, measuring the relationship between rules and community outcomes (such as bullying, misinformation, or community members' attitudes about governance) is challenging due to the difficulty of quantifying both [72, 281], as well as the abundance of confounding factors. As such, it's challenging for community moderators to know what rules would result in the best outcomes for their communities.

Studying rules in online communities poses several key difficulties. Making sense of the dozens of thousands of unique rules that are posted by thousands of varied communities requires a robust system for classifying rules. Furthermore, communities change their rules as they grow and in response to events, meaning that point-in-time assessments of rules fail to capture temporal dynamics [224]. Previous work has examined rules across many communities [72, 109, 65], and how rules are added as communities grow [224, 134], yet how these rules relate to community outcomes is poorly understood. Community outcomes themselves are difficult to measure, and are often assessed with surveys [284, 146, 5, 241, 128]. Recently, methods have been proposed to automatically quantify different community outcomes [281, 18], yet this work does not directly address rules.

In this paper, we present the largest-to-date study of rules in online communities. We collect 67,545 unique rules from 5,225 communities constituting 67.58% of all activity on Reddit, and identify how these rules were changed over a 5+ year study period, using data from the Wayback Machine (§3.2.3). We develop a pipeline to classify these rules using a taxonomy of 17 different attributes, extended from previous work [72]. Further, we use our existing method to quantify communities' public discussion of their governance [281] (§3.2.3) along with causal inference methods (§3.2.3) and assess how rules are associated with community members' perceptions of their community's governance, an important outcome.

Our analyses address three key research questions:

RQ1 What rules do communities have on Reddit? (§3.2.4)

RQ2 What rules are associated with positive community perceptions of governance? (§3.2.5)

RQ3 What is the impact of adding new rules? (§3.2.6)

We find that the rules set by communities vary widely, but important patterns emerge. Rules targeting post content are by far the most common, appearing in 89.74% of communities, while rules about who can participate are relatively rare, appearing in only 20.89% of communities (§3.2.4). Communities are also more likely to phrase their rules in terms of what is *prohibited* as opposed to permitted (§3.2.4). However, not all communities are the same: Discussion communities and communities for specific identity groups are $1.36\times$ more likely to have rules restricting who can participate (§3.2.4).

Using IPTW causal inference methods [13], we adjust for confounding factors while assessing differences between communities *with* vs. *without* different rules. We find that rules about how content is formatted and who participates are associated with significantly more positive perceptions of governance (§3.2.5). Furthermore, we find that rules emphasizing what behavior is *permitted* as opposed to *prohibited* are also associated with more positive perceptions of governance. Finally, we conduct a longitudinal study of what happens in communities after new rules are added (§3.2.6), and find that rule additions are associated with an immediate improvement in communities' perceptions of their governance, although this effect typically diminishes after approximately six months.

We discuss important implications of our results for researchers, platforms, and community moderators, and highlight key opportunities for future work (§3.3). We make our rule classification codebook, model, and rules data public[7] to support additional research on this important topic.

3.2.2 Related Work

Rules On and Off Reddit

Rules have been previously studied on Reddit, including analyses categorizing rules that are present [72, 65], and analyses of how rules have changed [224, 171, 65]. However, our work is the first to directly connect

[7] `bdata.cs.washington.edu/mod-perceptions`

rules to community outcomes by examining how community members perceive their governance (§3.2.3). Studies that assess rules' changes over time have used the Wayback Machine [65, 224] or manual scraping [171] to reconstruct timelines of communities' rules. In contrast to these studies, ours covers both a much longer time period (5+ years vs. 3 and 1.5) and a much larger set of communities (5,225 vs. 967 and 467). Lloyd et al. [171] includes rules from a larger set of communities than we do, yet they only measure if rules address AI generated content. In contrast, we assess 17 different attributes of rules. Two taxonomies have been proposed for classifying Reddit rules [72, 65]. We extend these taxonomies in our work, contributing a simpler taxonomy that captures a broader set of rule attributes (§3.2.3).

Off of Reddit, rules have been studied on peer production communities such as Wikipedia & Wikia [109, 134, 142], the fediverse [197], and gaming communities [77, 78]. While these are important other contexts, our work empirically evaluates rules specific to online social communities.

Content Moderation and Governance

Rules are also an important component of content moderation, which has been studied extensively on Reddit [213, 165, 181]. Some of this work has focused on specific aspects of content moderation, including banning users [44, 264], content removal [120, 246, 228] and removal reasons [115]. Other work has developed systems for participatory rule making [291] or has studied governance and norms through the lens of removed content [37]. Empirical analyses of removed content almost always do not enable the assessment of specific rules and their impact. In contrast, in this work we examine rules directly, along with their connection with perceptions of community governance.

Community Values and Outcomes

Understanding what it means to make communities 'better' and what values are held by community members is a challenging problem [283]. Previous work has shown that values vary dramatically between different communities, and that what's good for one community is not necessarily good for another [284, 212]. Other research has taken an ecological approach to understand how related communities overlap in their membership [293] and rely upon one another [108, 257]. Surveys are often used to assess community governance and community health [146, 5, 241, 128]. As surveys do not scale well, some methods have been

proposed to automatically quantify aspects of community outcomes [18]. In this work we use an existing dataset of community members' stated perceptions of governance in their communities [281] to empirically assess rules at a scale not feasible with surveys or qualitative methods.

3.2.3 Methods & Data Collection

To understand rules on Reddit, we first must know what communities had which rules when. We collect timelines of how communities' rules change using data from the Wayback Machine (§3.2.3), and classify those rules according to their tone, target, and topic using a GPT-4o-based retrieval-augmented few-shot multilabel classifier that achieves near-human performance (§3.2.3). Our work is the first to assess associations between what rules a community has, and how a community perceives its governance. To measure communities' perceptions of their governance, we use a classification pipeline developed by Weld et al. [281] to identify posts and comments discussing community governance (§3.2.3). Finally, after identifying community size and topic as key factors associated with a community's rules in §3.2.4, we use causal inference methods (§3.2.3) to adjust for these confounding factors in subsequent analyses to further identify how rules impact a community's perceptions of its governance. Throughout our analyses, in addition to the data collected using methods described below, we also use metadata (*e.g.,* for community size) computed from Pushshift data [20] and community topics classified by Weld et al. [281].

Computing Timelines of How Rules Change

To collect timelines of how communities' rules change over time, we utilize the Wayback Machine [259]. Communities on Reddit post their rules on the sidebar of their homepages, where they are archived by the Wayback Machine[8] We sought to collect rules timelines for as large and varied of a set of communities as possible. We downloaded snapshots for approximately every English-language community with at least one snapshot every six months, for a set of 67,545 unique rules from 5,225 communities that collectively account for 67.58% of all activity on Reddit during our 5+ year study period from April 2018 to December 2023. As the Wayback Machine archives more popular communities more often, most communities in our

[8]In 2019, Reddit introduced additional Rules pages which supplement the Rules posted in the sidebar. As these pages largely duplicate the information presented in the sidebar, are not archived by the Wayback Machine, and not available for the entire study period, we focus on sidebar content in this work.

study are snapshotted more frequently than than every six months: 2,316 communities have at least monthly snapshots, and these communities account for 83.36% of the community content included our study.

Next, we computed differences between snapshots to identify **Rules Periods**, blocks of time where a given community has an constant, unchanged set of rules. Each rule period consists of a set of one or more rules present during that period, along with time that the period began and ended. Due to temporal gaps between snapshots, we have limited temporal resolution. To understand how this may impact our analyses, using the time between subsequent snapshots, we compute the temporal uncertainty around when exactly each rules period began and ended. The average start/end uncertainty is ± 17 days, relatively small compared to the average period length of 378 days. Therefore, we believe the limited temporal resolution for some communities has a negligible impact on our analyses.

To understand how common different rules are, we compute **Rule Prevalence**, which is the fraction of rules periods containing at least one rule of a given type, weighted by the duration of the rule period. Thus, rule prevalence measures the average likelihood of encountering a rule across all communities[9].

Automatically Classifying Rules' Tone, Target, and Topic

On Reddit, rules are arbitrary text strings that are highly varied across communities. To perform our analyses of 67,545 unique rules, we developed a system to automatically classify rules' tone, target, and topic.

Taxonomy Development. Starting with a random sample of 100 rules as a development set, a team of two researchers used a grounded theory approach to iteratively categorize the rules, using an inductive coding method [174]. The researchers independently clustered similar rules, then came together to resolve differences and reach consensus. Tentative clusters were assigned names and definitions to produce a working taxonomy, then the development set was recategorized, creating and removing categories by consensus. After two rounds of iteration, the process converged. At this point, the researchers independently labeled a separate test set of 200 randomly sampled rules, which was used to evaluate human IRR (Fleiss' kappa between 0.61 'substantial' and 0.91 'almost perfect' was achieved for all categories, with a Macro average of 0.83 'almost perfect') [162]. The resulting taxonomy consists of 17 different rule attributes across three

[9]For simplicity, we refer to this throughout the manuscript as 'the fraction of communities with a rule of type x. However, as communities occasionally change their rules, a more precise interpretation would be 'the fraction of communities with a rule of type x when sampled uniformly across all communities and time across our study period.

aspects of rules: tone, target, and topic. Table 3.3 gives a description of each class. After resolving disagreements, our set of 200 rules labeled by two annotators was used as test set for the evaluation of our automated classification pipeline.

Rule Tone captures how the rule is phrased: prescriptive rules tell community members what *to* do (*e.g.,* 'Be nice!') while restrictive rules tell community members what *not* to do (*e.g.,* 'Don't be mean!'). **Rule Target** distinguishes between what type of interaction the rule is targeting: Rules can address the content of what is posted in a community (Post Content, *e.g.,* 'no off-topic content allowed'), how that content is formatted (Post Format, *e.g.,* 'Posts must begin with 'ELI5"), or the users who post the content (User-Related, *e.g.,* 'You must be approved by the mods to post'). Finally, the **Rule topic** classes focus on specific topics addressed by rules, such as Tagging & Flaring, Spam, and Respect for Others.

Comparison with Previous Taxonomies. Our taxonomy extends two previous taxonomies of rules [72, 65]. Our taxonomy covers all codes from both previous taxonomies while adding two new attributes not directly covered (User-related rules and rules addressing Karma & Score). Our taxonomy simplifies the Fiesler et al. [72] taxonomy, with with a slightly smaller set of attributes (17 vs. 24). To achieve this simplification, we merge several of the previous taxonomy's codes. The Fang, Yang, and Zhu [65] taxonomy consists of 15 specific rule types. In contrast, our taxonomy covers a broader set of attributes (Table 3.2). Compared to the rules taxonomy from Fiesler et al. [72] while adopting a hierarchical structure. Our 17 rule attributes are divided into three higher-level aspects of rules: tone, target, and topic. Compared to the other taxonomy, our taxonomy directly addresses rules targeting *who* is permitted to participate, as well as rules about Karma and Score. We consolidate several codes from Fiesler et al. [72] into a smaller number of rule topics: Spam, Low-Quality Content, Reposting, and Off-topic become a single rule topic. We use a single topic for Respect for Others, which includes Trolling and Harassment from Fiesler et al. [72]. Likewise, we include a single topic for Illegal Content and for Brigading.

With regards to the taxonomy outlined by Fang, Yang, and Zhu [65], our taxonomy adopts a broader perspective while aligning with two of their three overarching categories: Post Content and Post Format. We identify four categories (Images & Video, Illegal Content, Tagging & Flairing, Brigading) not present in the other taxonomy, and collapse the 3 format codes (Minimum Text, Templates, Title vs Description) proposed into a single Post Format label. Our taxonomy additionally identifies individual rule tone, *how* a

	Our Taxonomy	Fiesler et al. [72]	Fang, Yang, and Zhu [65]
Tone	Restrictive	Restrictive	
	Prescriptive	Prescriptive	
Rule Target	Post Content	Content/Behavior	
	Post Format	Format	Requires a minimum text Requires template for post Limits use of title vs. description text
	User-Related		
Rule Topic	Spam & Low Quality	Spam Low-Quality Content Reposting Off-topic	No jokes Posts must be high-quality Content must be original Posts must be on-topic
	Respect for Others	Trolling Hate Speech Harassment	Enforces respect for others No promotion of bad behavior
	Commercialization	Advertising & Commercialization	No promotional content
	Links & External	Links & Outside Content	Posts must be verifiable
	Peer Engagement	Personality	Enable peer feedback
	Images & Video	Images	
	Illegal Content	Copyright/Piracy Doxxing/Personal Info Reddiquette/Sitewide	
	Tagging & Flaring	NSFW Spoilers	
	Divisive Content	Politics	Discourages divisiveness Avoid distressing material
	Brigading	Personal Army Voting	
	Ban Mentioned	Consequences/Moderation/Enforcement	
	Karma/Score		

Table 3.2: A comparison of our taxonomy with those from Fiesler et al. [72] and Fang, Yang, and Zhu [65]. Our taxonomy captures elements of rules not captured by previous taxonomies (*e.g.,* User-related rules and Karma/Score), while combining together some categories from previous taxonomies (*e.g.,* Fiesler et al. [72] provides four categories for Spam, Low-Quality Content, Reposting, and Off-topic, while we combine these into a single Spam & Low Quality attribute).

rule is phrased (prescriptive vs. restrictive), as well as explicit references to Bans and to Karma/Score.

Classification Pipeline. Given the many thousands of rules we need to label, human labeling is infeasible. As such, we developed a retrieval-augmented few-shot classification pipeline using on GPT-4o [203]. After the codebook was finalized, one annotator labeled an additional 200 rule training set used to provide few shot examples, retrieved using embedding-similarity provided by Nomic [198] as examples for the LLM (6-shot classification). This pipeline approaches human performance, achieving near-human performance with a Cohen's Kappa of 0.71 and Macro-F1 of 0.74 on our test set (Table 3.4). Our GPT-4o model also exceeds the performance of all other models we evaluated, including other top general purpose large-language models: GPT-4 [202], Mistral [192], and Llama-3 [170], as well as a fine-tuned version of RoBERTa [169]. Previous

	Name	Description	Model F1	Unique Rules	Prevalence
Tone	Restrictive	A rule which explicitly limits or forbids certain actions	0.87	40,448	87.98%
	Prescriptive	A rule expressing general guidelines or desires	0.74	17,621	71.52%
Rule Target	Post Content	Any rule explicitly stating desired or undesired content within the subreddit	0.88	42,460	89.74%
	Post Format	Any rule prescribing post structure, formatting, titling, tagging or referencing a location to post	0.76	7,685	43.60%
	User-Related	Any rule that applies to users differently for identical content, including verification and prior approval rules	0.61	3,710	20.89%
Rule Topic	Spam & Low Quality	Any rule covering low-quality, off-topic, spam, reposted, or otherwise generally-undesired content	0.78	17,360	74.52%
	Respect for Others	Rules explicitly discouraging hate speech, antagonization, harassment or encouraging respect for others	0.82	10,782	72.15%
	Commercialization	Any rule regulating advertisement, self-promotion, referrals and referral links, or buying/selling	0.93	7,031	42.65%
	Links & External Content	Any rule regulating links to or the content of external resources	0.73	5,659	35.55%
	Peer Engagement	Any rule encouraging or discouraging the quantity of on-site peer engagement, including reporting, commenting, voting, and general activity	0.60	4,001	28.79%
	Images & Video	Any rule pertaining to the content or quality of images or videos	0.78	4,567	25.78%
	Illegal Content	Rules about illegal content, including copyright infringement and piracy	0.75	2,637	23.10%
	Tagging & Flaring	Any rule pertaining to labeling/flairing, marking nsfw, using tags, etc	0.91	2,705	22.56%
	Divisive Content	Rules regulating content which is inherently divisive, such as politics, current events, or community-specific hot topics	0.75	3,169	20.14%
	Brigading	Any rule regulating large group actions, including raiding or mass-voting	0.89	874	9.20%
	Ban Mentioned	Any rule that explicitly mentions banning	1.00	1,103	3.84%
	Karma/Score Mentioned	Any rule that explicitly mentions karma, user scores, or voting	1.00	559	3.60%
	Overall	Macro Average:	**0.81**	Total: **67,545**	

Table 3.3: Our taxonomy of rules consists of 17 rule attributes across three different aspects of rules: tone, target, and topic. This breakdown of rule labels shows the performance of our retrieval-augmented GPT-4o classification model (Macro F1 = 0.81) and how likely to be encountered each rule type is (prevalence) across our 5+ year study period. Post Content is the most prevalent rule target, while Spam & Low Quality and Respect for Others are the most prevalent rule topics.

studies classifying rules on Reddit have used logistic regression classification methods [72, 65]. However, these studies both predate the introduction of LLMs. We instead use a few-shot LLM-based pipeline, which results in higher performance with much less training data.

To further reduce the cost of labeling, we assigned the same label to rules that were nearly identical (differing only in capitalization or having an edit distance of two or fewer characters). This edit distance threshold was selected such that the addition of 'no ' would cause the rules to be labeled separately. This reduced the number of rules to classify to 51,404. Labeling these rules using our pipeline took six hours at a cost of $491.39 worth of OpenAI API credits.

	Kappa	Macro F1
Human Expert Performance	0.83	0.80
RoBERTa [169] on 300 human labeled	0.18	0.23
RoBERTa on 3k GPT-4o RAG labeled	0.41	0.52
Mistral [192] retrieval-augmented	-0.07	0.39
Llama-3 [170] retrieval-augmented	-0.07	0.40
GPT-4 [202] zero-shot	0.66	0.63
GPT-4 [202] retrieval-augmented	0.64	0.70
GPT-4o [203] retrieval-augmented	**0.71**	**0.74**

Table 3.4: Our rule classification pipeline using GPT-4o exceeds the performance of all other models evaluated and approaches human performance. An identical six-shot prompting pipeline was used for Mistral [192], Llama-3 [170], GPT-4 [202], and GPT-4o [203]. Complete prompts are given in Appendix E.2.

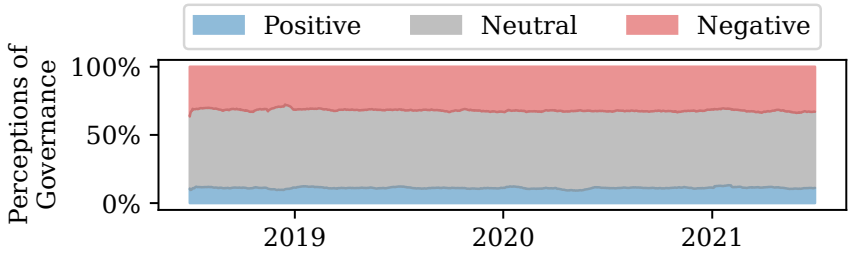


Figure 3.9: Across all communities on Reddit, community members' publicly expressed perceptions of their governance are approximately constant over time. Of posts and comments discussing governance, on average 11% have positive sentiment, 57% have neutral sentiment, and 32% have negative sentiment.

Measuring Community Perceptions of Governance

Our work is the first to examine how online communities' rules are associated with how communities perceive their governance. To measure community members' perceptions of governance, we use the method developed by Weld et al. [281] to classify public posts and comments made on Reddit which discuss governance and moderators. This method consists of a three step pipeline to identify posts and comments discussing governance, filter out posts and comments addressing other communities, and classify the sentiment of each post and comment with regards to the governance using a using a QLoRA-tuned Llama2 model [56]. To address the possibility that moderators' removal of comments critical of them might bias the validity of this pipeline, Weld et al. [281, §3.4] examine a sample of removed comments, and conclude that this is highly unlikely: content discussing governance accounts for $< 1\%$ of removed content.

We apply this method to all posts and comments made during our study period in communities for which we have rules timelines, collecting 3.8 million posts and comments discussing governance: 347 thousand

posts and comments with positive sentiment, 1.9 million with neutral sentiment, and 1.5 million with negative sentiment. For each community's rules periods, we compute the **Community's Perceptions of Governance** as the fraction of posts and comments discussing governance made in a given community during a given rules period having positive, neutral, and negative sentiment towards the community's governance.

Adjusting for Confounding Factors

In §3.2.4, we compute the prevalence of rules and identify two key factors that are associated with the rules set by each community: community topic and size. In subsequent sections, we examine which different types of rules are associated with more positive community perceptions of governance (§3.2.5), and what happens when new rules are added (§3.2.6). In each of these sections, we use a different method to adjust for these factors in order to better identify the potential impact of rules.

In §3.2.5, we use **Inverse Probability of Treatment Weighting (IPTW)**, a statistical method to adjust for confounding when comparing between different observed groups (*e.g.,* communities *with* and *without* rules of a given type). We use IPTW over other similar methods, such as stratification, due to its superior efficiency [13].

We begin with an intuitive example of how IPTW works. Consider an analysis of rules about links and external content. Some communities have such rules, whereas others do not have such rules. We can compute the average perception of governance in both communities with and without links and external content rules, then take the difference between these averages to determine possible associations. However, news-focused communities are more likely to have rules about links and external content (53.40% vs. 33.79% prevalence, see Figure 3.12) and would thus be overrepresented in the set of communities with rules on links and underrepresented in the set of communities without these rules. To control for this selection effect, IPTW upweights underepresented communities and downweights overrepresented communities when computing average perceptions of governance across these groups. After weighting is performed, both groups have similar distributions of topics and size.

More formally, we compute IPTW weights using the probability of treatment (probability of having a rule of a given type), known as the propensity score, $P(Z|\mathbf{X})$ by applying a logistic regression model to community covariates $\mathbf{X}$ including the topic (one-hot encoded) and size of the community. For final

analyses, each observation is weighted by the inverse of the probability of the treatment it received, such that the weight for observation i is $w_i = \frac{Z_i}{P(Z_i=1|\mathbf{X}_i)} + \frac{1-Z_i}{P(Z_i=0|\mathbf{X}_i)}$, where Z_i is the treatment received by observation i (0 for control, 1 for treated), and $\mathbf{X}_i$ is a vector of the covariates. This weighting makes the distributions of covariates among the treatment and control groups more similar to the overall population.

An important validity check for IPTW is to assess the balance of covariates for each treatment group after weighting [13]. Two groups are often considered 'balanced' or 'indistinguishable' if all covariates are within a standardized mean difference (SMD) of 0.25 standard deviations [8]. Using the same procedure as in Weld et al. [281], for each treatment/control group, we compute the difference between each covariate's weighted mean value and the *reference distribution*, consisting of the entire population of communities. We compute the SMD by normalizing the difference in means by the standard deviation of the values of the reference distribution. Across 238 condition-covariate-rule type pairs (17 experiments $\times$ 2 treatment/control conditions $\times$ 7 covariates), our method achieves balance in 235 cases (98.74%).

In the three cases where our method fails to achieve balance, no experiment has more than a single unbalanced covariate (out of seven), and no SMD exceeds 1.00. A complete list of covariates and SMDs for each experiment is given in Appendix E.3, along with a comparison of weighted vs. unweighted results. Appendix E.3 also shows the difference between IPTW results and those without any adjustment for confounding. We find that after performing IPTW, most confidence intervals do not overlap 0, suggesting that rules play an important role in communities' perceptions of governance, although community topic and size partially account for some of the differences between communities.

In §3.2.6, to measure what happens after rules are added, we use time series data and longitudinal analyses to quantify perceptions of governance before vs. after a rule is added. As such, because we are not comparing between different groups, IPTW is not feasible. Instead, the time-based format of our analyses inherently control for the potential community topic and size confounds, as these factors do not meaningfully change over a relatively short time period.

Ethical Considerations

We believe this work will have a positive broader impact by informing rule setting in online communities. We believe our work has minimal risk to participants' privacy as we only use public data, however, we take

further steps to reduce potential harms and misuse potential of work: we do not publish moderator information, participant usernames or any identifiable information. This study was approved by the University of Washington IRB under ID number STUDY00011457.

3.2.4 What Rules Do Communities Have?

Number, Diversity, and Prevalence of Rules

Results. We find that communities on Reddit vary dramatically in the number and types of rules they declare. 95% of communities declare between two and 14 rules (median 7 rules), although in general, the larger the community, the more rules they declare (Figure 3.11). On average, communities with < 10 daily posts and comments declare 4.32 rules, while huge communities (which account for 29.87% of all posts and comments) have an average of 9.26 rules.

Some rules are much more frequently encountered than others. 89.74% of communities have rules targeting post content (Table 3.3), while rules targeting *who* participates (User-Related rules) are present in only 20.89% of communities. Rules about brigading, bans, and karma/score are the least common rule topics, occurring in only 9.20%, 3.84%, and 3.60% of communities, respectively.

Implications. As communities grow and age, they tend to add rules [224]. Moderators on Reddit and other platforms often take a reactive approach [181], meaning that rules may be added only after an incident in a community makes it clear that such a rule is needed. The ubiquity of rules about spam, low quality content, and respect for others suggest that these are topics that most communities struggle with, and efforts by platforms and researchers to improve these aspects of communities are likely broadly useful.

Prescriptive vs. Restrictive Rule Tone

Results. Rules can be phrased with prescriptive tone ('Be nice.') or restrictive tone ('Don't be mean.'), and occasionally phrased both prescriptively and restrictively ('Be nice. Being mean is not allowed.'). We classify the tone of rules (§3.2.3), and find that on the whole, restrictive rules are more common than prescriptive rules, but both tones are widely used (Figure 3.10), and 66.58% of communities have both prescriptive and restrictive rules. The use of both tones simultaneously is relatively rare, with 94.22% of

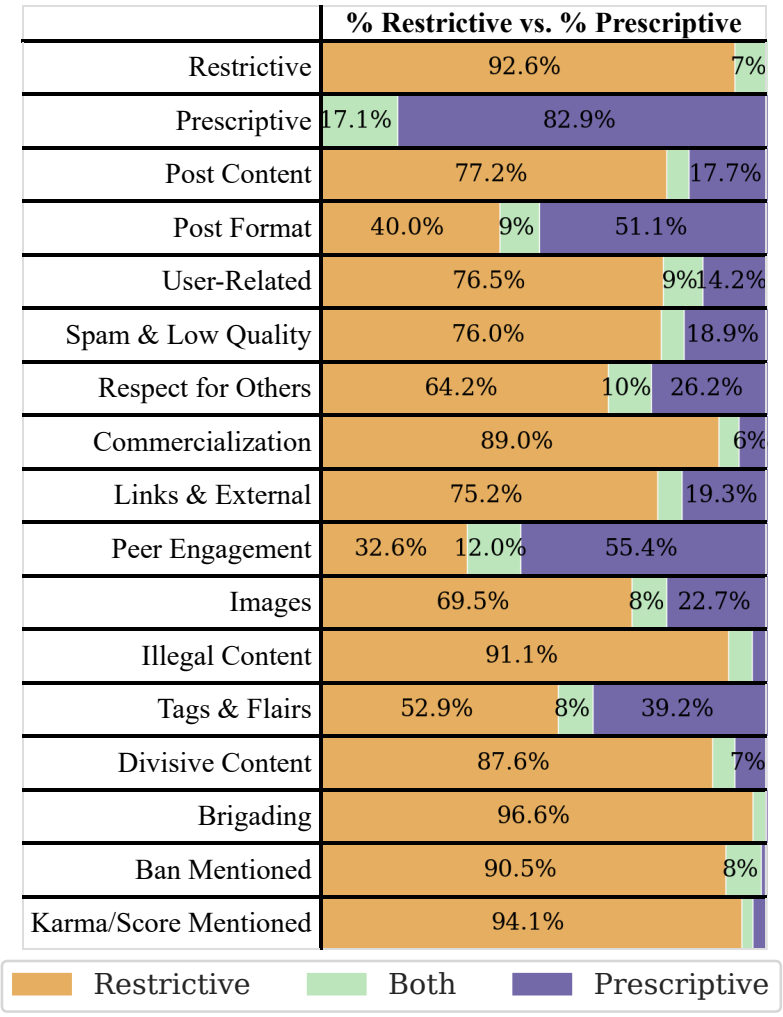


Figure 3.10: Rules vary with regards to their tone. On the whole, restrictive rules are more commonly encountered than prescriptive rules on Reddit, although both are ubiquitous: 87% of communities have at least one restrictive rule, while 70% of communities have at least one prescriptive rule (prevalence). Rules addressing post format and peer engagement are both more likely to be prescriptive ('Be Nice') than restrictive ('Don't be mean'), while all other types of rules are more commonly phrased with restrictive tone.

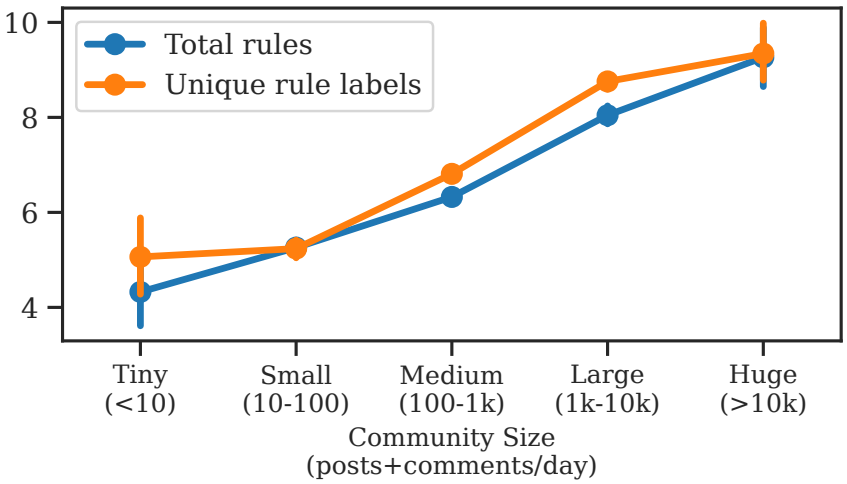


Figure 3.11: Larger communities have both more rules and more diverse rules. Tiny communities have on average 4.32 rules, while huge communities (the 0.77% largest) have on average 9.26 rules. In this and subsequent figures, the bars shown represent bootstrapped 95% confidence intervals.

rules being phrased only prescriptively or only restrictively.

We find that certain types of rules are more frequently phrased restrictively, and others are more frequently phrased prescriptively. Rules about post format and peer engagement are both more frequently phrased using prescriptive tone (*e.g.,* 'posts titles must be descriptive'). On the other hand, rules mentioning bans, karma/score, and brigading are almost often written using restrictive tone.

Implications. Offline criminology literature has explored differences between telling people what they *should* vs. *shouldn't* do, which we examine in the online community context. In §3.2.5 we examine associations between rule tone and communities' perceptions of their governance, and find that communities with prescriptive rules tend to use more positive language to discuss their governance than those without prescriptive rules, whereas communities with restrictive rules tend to use more *negative* language than those without restrictive rules (Figure 3.13).

Furthermore, our findings align with previous analyses of rules that are more likely to be phrased prescriptively [72].

Differences Between Communities with Different Topics

Results. Using an existing classification of communities based on their topic [281], we examine how likely communities with different topics are to have different rules. While the prevalence of some rules varies

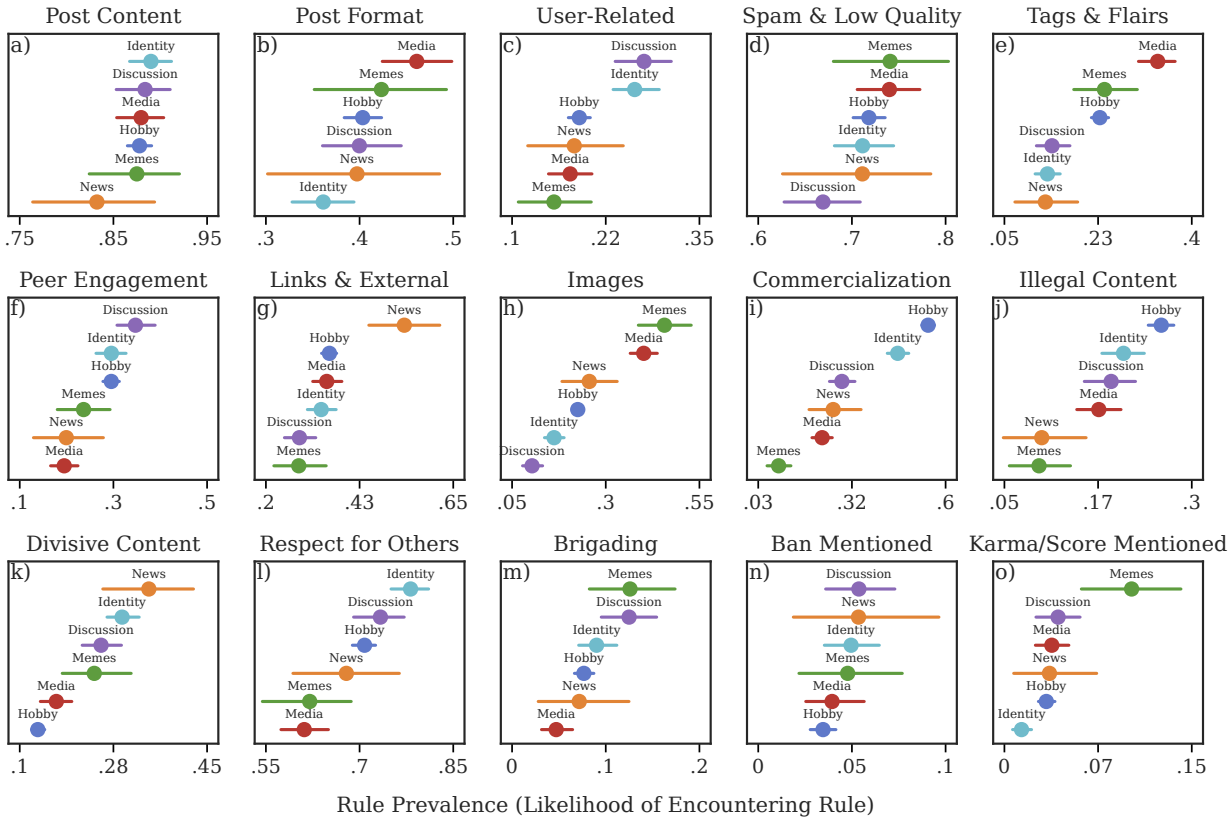


Figure 3.12: The ubiquity of different types of rules differs greatly based on community topic. Discussion and Identity communities are especially likely to have rules about who participates (c). News communities are almost twice as likely to have rules about links and external content (g), on average, while Hobby & Identity communities are more likely to have rules on Commercialization (i). Rules about Images are much more common in Meme and Media communities and very rare in Discussion (often text-based) communities (h).

substantially based on topic, other rules are applied relatively consistently. Discussion communities and communities focused on Identity (such as those for LGBTQ groups) are $1.36\times$ more likely to have User-related rules than other communities (Figure 3.12c). News-sharing communities are $1.58\times$ more likely to have rules about Links & External Content (Figure 3.12g), while Meme communities are $2.75\times$ more likely to have rules about Karma (Figure 3.12o). On the other hand, we do not find a significant difference between communities of different topics in their likelihood of having rules targeting post content (Figure 3.12a), which are present in 89.74% of communities. Similarly, only 3.84% of communities have rules mentioning bans, and we do not find that this varies significantly based on the topic of the community (Fig. 3.12n).

Implications. Our results suggest that certain types of rules, like those targeting the content of posts as well as specifically addressing Spam & Low Quality content, are considered necessary by a majority of communities of every topic. Other types of rules are more necessary to communities with specific topics, such as rules about Commercialization, which are especially prevalent in Hobby communities. Communities should consider the specific needs of their community when determining what rules to post and enforce.

3.2.5 What Rules Are Associated With Positive Community Perceptions of Governance?

Method

To understand what rules are associated with positive community perceptions of governance, we make comparisons between communities *with* and *without* rules of various types. We measure perceptions of governance using an existing method to classify community members' discussion of their community's governance [281]. This dataset consists of posts and comments classified into positive, neutral, or negative sentiment with regards to community governance. For each community, we compute the fraction of posts and comments with each of these three sentiment classes (for more detail, see §3.2.3) made during the entirety of each rules period.

When making comparisons between different communities, we seek to identify the differences in perceptions of governance that are associated with the presence or absence of different rules. Towards this goal, we adjust for two key confounding factors, community size and topic, using Inverse Probability of Treatment Weighting (IPTW) [13]. For more details on IPTW, see §3.2.3 and Appendix E.3.

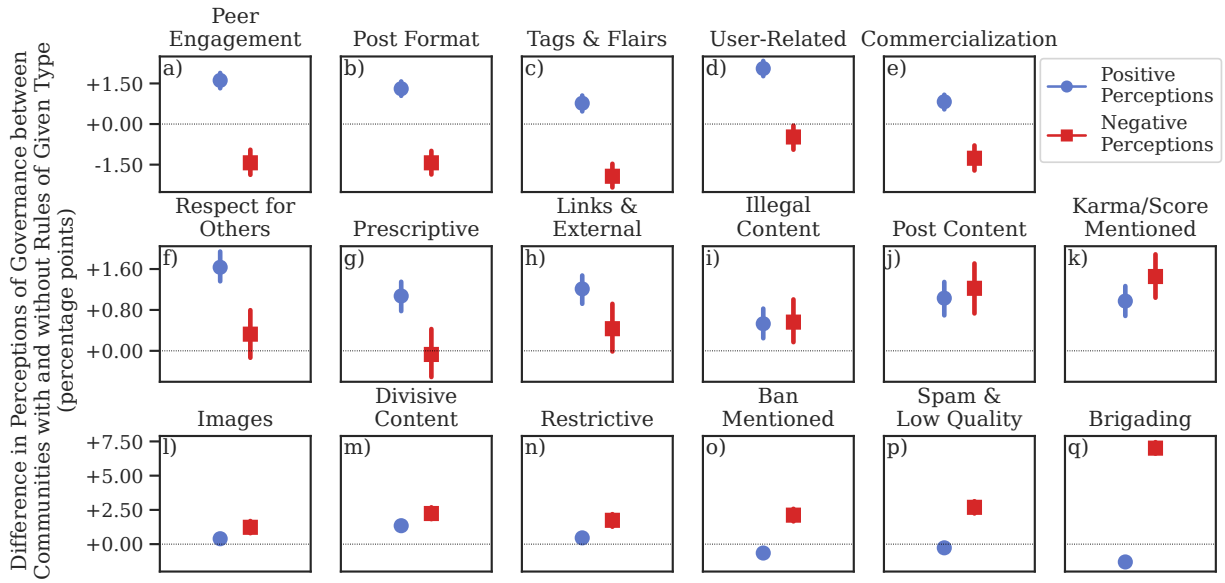


Figure 3.13: Perceptions of moderators vary between communities with and without different types of rules, even after adjusting for confounding factors (§3.2.3). Rules about Peer Engagement (a), Post Format (b), Tags & Flairs (c) , and Commercialization (e) are all associated with higher positive perceptions and lower negative perceptions of moderators than communities without those rules. On the other hand, communities with rules about Illegal Content (i), Post Content (j), and Karma/Score (k) have more polarized perceptions of moderation, with positive *and* negative perceptions *both* higher than in other communities (which have higher neutral perceptions of moderation, not shown here).

Results

We find substantial differences in communities' perceptions of their governance between communities with and without various rules (Figure 3.13). Several rule types are associated with more positive and less negative perceptions of governance, particularly rules about Peer Engagement, Post Format, Tags & Flairs, and Commercialization (Figure 3.13a-c,e). We find that even though rules addressing *who* participates are less common than rules addressing Post Content or Format, communities with User-Related rules have 1.83 percentage points more posts and comments expressing positive sentiment towards their governance after adjusting for confounding factors (Figure 3.13d). As only 11.12% of posts and comments discussing governance in the average community have positive sentiment, (Figure 3.9), a 1.83 percentage point increase would result in a 16.45% increase in positive sentiment in a typical community. Put another way, after 1.83 pp increase, a huge[10] community could expect almost 24 more comments per week expressing positive attitudes, while a large community could expect almost 6 more such comments per month.

However, some types of rules have more complex associations with perceptions of governance. Rules about Illegal Content, Post Content, and Karma/Score are all associated with more *polarized* perceptions of governance, having *both* higher positive and higher negative perceptions of governance than communities without such rules (Figure 3.13i,j,k). The correspondingly smaller fractions of posts and comments and comments with neutral sentiment towards governance are not shown.

Implications

Our results show substantial differences in community perceptions of governance between communities with different rules. In particular, rules that promote positive interactions (Peer Engagement, User-Related, and Commercialization) and rules that structure contributions (Post Format, Tagging & Flairing) are associated with the most positive perceptions of governance. Community moderators should consider how their rules address positive interactions and contribution structure.

We also find differences between rules with different tone. Rules with prescriptive tone ('do this') are associated with more positive perceptions of governance, whereas rules with restrictive tone ('don't do this'), are associated with more negative perceptions of governance (Figure 3.13g,n).

[10]We use the same community sizes bins as in Figure 3.2.4: Large communities have between 10^3 and 10^4 posts & comments per day, and Huge communities have $> 10^4$.

We find that rules mentioning bans and brigading are associated with more negative and less positive perceptions of governance (Figure 3.13o,q). These are both emotionally charged actions, which may contribute to the association with more negative perceptions of governance [264, 44, 50].

Overall, we find that many (13/15) rule types are associated with more positive perceptions of governance. Does adding new rules improve community perceptions of governance? We examine this question directly in §3.2.6.

3.2.6 What Is the Impact of Adding New Rules?

In §3.2.5 we assessed what types of rules are most associated with favorable community perceptions of governance. But what happens when communities add new rules? We use our timelines of rules (§3.2.3) to conduct longitudinal analyses of how perceptions of governance (§3.2.3) vary immediately after new rules are added to a community.[11]

Method

Starting with our set of rule periods, we compute **rule change** events for each community, where each rule change consists of two adjacent rule periods. We examine only rule changes where a rule that was not previously present is added, and, as this analysis examines a narrower time range and therefore requires higher temporal precision, we only include rule changes whose time of change is certain within ± 1 week precision. Across our 5+ year study period and 5,225 communities, this leaves a set of 6,645 rule change events matching these criteria. For each change event, we identify a 12 month long pre-change baseline window, and divide the 12 months post-change into $6\times$ two month long comparison windows, enabling analysis of how perceptions change over time[12].

For the pre-change baseline and each of the post-change comparison windows, we compute the perceptions of governance (fractions of posts+comments discussing governance having positive, neutral, and negative sentiment, for more details see §3.2.3) for the relevant community. We then take the difference

[11] Rule additions are relatively rare events, with only 6,645 occurring during our 5 year study period. Although our data permit evaluation of different rule types, for these longitudinal analyses, confidence intervals are too large to be conclusive when analyzing the response to additions of rules of different types. As such, here we consider the addition of rules of *any type*.

[12] We experimented with a range of window widths and found that the exact window width does not make a qualitative difference in the results; thus we selected a two month width for the post-change comparison window to balance temporal resolution with statistical power.

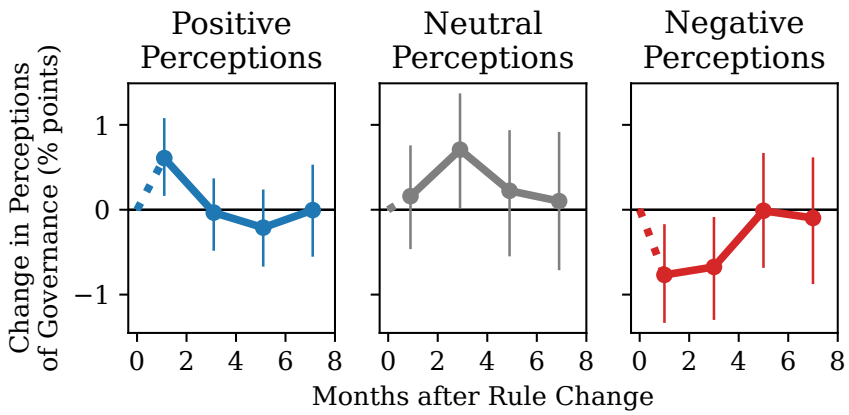


Figure 3.14: Immediately after a new rule is added, on average, positive perceptions of governance increases while negative perceptions of governance decrease. After approximately 6 months, this effect diminishes and community perceptions of governance are not significantly different than before the rule change.

between the baseline and the comparison windows to see how perceptions of governance changed after the new rule was added. We average across all rule change events and bootstrap confidence intervals (as in all other analyses) to quantify our statistical power.

Results

We find that, on average, immediately after a new rule is added, perceptions of governance become more favorable: The fraction of posts and comments discussion governance positively *increases* by 0.61 percentage points, while the fraction with negative sentiment *decreases* by 0.77 percentage points (Figure 3.14). For a typical huge community, this effect would be equivalent to almost 17 comments per week expressing negative perceptions of governance switching to expressing positive perceptions. We also find that rule additions have a depolarizing effect on perceptions of governance, with the fraction of posts and comments discussing governance with neutral sentiment increasing by 0.71 percentage points in the 2-4 month period after a new rule is added. Importantly, however, we find that the effects of new rule additions appear to 'wear off' after approximately six months post-rule change, as the perceptions of governance returns to a state not significantly different from the pre-change baseline (Figure 3.14).

Implications

Our results show the addition of new rules is associated with a small but statistically significant improvement in community members' publicly stated perceptions of the governance in their communities. This is consistent with the hypothesis that rules are often added as a response to an incident within a community that makes it clear such a rule is necessary. Our finding that the impact of a new rule on perceptions of governance at least partially diminishes after approximately six months is noteworthy. Part of the impact of a new rule comes not from the associated enforcement of that rule, but from the signal that the new rule provides that the moderator team is adjusting their strategy to the needs of the community. We speculate that this signaling aspect of rule changes contributes to the observed diminishing effect.

3.3 Discussion & Conclusion

Rules, their communication, and their enforcement, are critical parts of the governance of nearly every online community. Our results have important implications for platforms, community moderators, and researchers.

Our results inform what rules communities should have. We find that rules about Post Content, Spam & Low Quality content, and Respect for Others are especially common (§3.2.4). Platforms could consider providing rule 'starter packs' for new communities based on common rules across the platform. We find that certain rules are more ubiquitous in communities of certain types, for example, Discussion and Identity communities are more likely to have User-related rules than other communities (§3.2.4). An interesting future research idea is a tool that recommends rules for moderators to consider implementing based on community size, topic, and other factors. Our results also show that differences in rules explain some of the differences between different communities' perceptions of their governance (§3.2.5). Rules about Illegal Content, Post Content, and Karma/Score are associated with more polarized perceptions of governance. These rules maybe more polarizing because they relate to the boundaries of what behaviors and content are acceptable in a community, rather than how that content is presented. Rules that are polarizing are not necessarily bad for communities, but these are important impacts to consider.

Our results also have implications for how rules should be phrased. In many cases, moderators can choose to phrase rules to describe what community members *should* do (prescriptive) vs. what they *should*

not do (restrictive). Our results suggest that certain rules are more easily phrased restrictively (§3.2.4), but that prescriptive rules are associated with more positive perceptions of governance, whereas restrictive rules are associated with more negative perceptions of governance (§3.2.5). Future research could evaluate a system that helps moderators consider alternate phrasings or presentations for their rules [185].

Our results show that new rules are associated with an immediate improvement in communities' perceptions of governance (§3.2.6). While we cannot know *why* moderators added rules, any rule additions are a sign that moderators are engaging with their community, and our results are consistent with results from other studies showing that moderator engagement with community members is received positively [281]. As it is neither feasible nor necessarily a good idea for communities to continuously add new rules, our finding that the positive impact of adding a new rule diminishes after 6 months (§3.2.6) highlights the importance of other means for moderators to engage with their communities, such as educating members about the rules [115, 185] and soliciting feedback [93, 150].

For researchers, we make our dataset of rules, our taxonomy of rule attributes, and our classification methods public to support future research.

3.3.1 Limitations

Our work establishes connections between the rules a community has and that community's perceptions of its governance, as measured using public discussion of governance and moderators (§3.2.3). While we take great care to ensure the robustness of our results, including quantifying temporal uncertainty and reporting confidence intervals throughout, our work has several important limitations. The Wayback Machine, which we use to measure how rules change over time, has infrequent or absent snapshots for less popular and very small communities, thus potentially biasing our results. Additional research is needed to understand how very small communities and communities in languages other than English differ. We also only consider the rules that are posted by communities, not how these rules are enforced or how visible they are [185]. Collecting information about rule enforcement is difficult, but not impossible [37].

Rule modifications can occur for a variety of reasons. While our work examines many thousands of rule changes, our methods are unable to track exactly why a given rule was changed. Moderator conflicts, platform-wide policy shifts, or real-world events can all cause rule changes for reasons outside a single

community's control. Future work could attempt to differentiate between rules changed as a result of internal *vs.* external factors.

We measure perceptions of governance using public discussion threads, however public discussion threads do not necessarily align with privately held attitudes about governance. Furthermore, community attitudes are not and should not be the only objective when considering rule changes. We also measure discussions about governance and moderators generally, not rules specifically. Future work could develop a more granular classification system.

Finally, while we use causal inference methods to adjust for two key confounding factors (§3.2.3), it is highly likely that there are additional confounding factors that we do not adjust for. Future research could control for some of these confounding factors, such as content removal rates and moderator workload. Furthermore, future research could use even more robust methods such as randomized controlled trials to avoid bias due to unobserved confounding. Additional data collection could enable time-series based difference-in-difference analyses of rule additions.

3.3.2 Conclusion

Rules and their enforcement are a critical component of the governance of online communities, yet it is difficult for community moderators to know which rules to select for their community. In this work, we conducted the largest-to-date analysis of rules on Reddit, examining 67,545 unique rules across 5,225 communities over a 5+ year period. We reconstructed timelines of how rules change over time (§3.2.3) and classify rules according to their tone, target, and topic (§3.2.3). We assessed what types of rules are most prevalent, and how these rules vary across communities of different types (§3.2.4). Ours is the first study to connect rules to community outcomes, using data about how community members perceive the governance of their communities. Using these data, we identified the rules most strongly associated with positive community perceptions of governance: Prescriptive Rules ('do this'), User-related rules, and rules about Tags & Flairs and Peer Engagement (§3.2.5). We also conducted a longitudinal study of the impact of adding new rules to communities, finding that after a rule is added, community perceptions of governance immediately improve, yet this effect 'wears off' after approximately six months (§3.2.6). Our results have important implications for moderators and community leaders (§3.3), and we make our datasets public to support future

research on this topic.

3.4 Informing Community Governance Through an Assessment of News Sharing Behavior at Web Scale

3.4.1 Introduction

The trustworthiness of content shared online is an important community value [284]. Mis- and dis-information, in particular, have been shown to have substantial offline harms. Understanding how news content is shared online enables us to better understand the impact of current governance practices, and inform improved strategies going forward. However, measuring how news is shared online is extremely challenging because of the immense scale of communities where news is shared and the difficulty of fact checking. In this work, we conduct the largest study of news sharing on reddit to date, analyzing more than 550 million links spanning 4 years. We use non-partisan news source ratings from Media Bias/Fact Check to annotate links to news sources with their political bias and factualness. Through the lens of the trustworthiness of news shared, We examine several key aspects of current governance practices that impact the visibility of problematic news content: user lifespan (§3.4.5), voting practices and community acceptance (§3.4.5), and amplification via crossposting (§3.4.5).

Biased and inaccurate news shared online are major concerns that have risen to the forefront of public discourse regarding social media in recent years. Two thirds of Americans get at least some of their news content from social media, but less than half expect this content to be accurate [243]. Globally, only 22% of survey respondents trust the news in social media "most of the time" [196]. Internet platforms such as Twitter, Facebook, and Reddit account for an ever-increasing share of the dissemination and discussion of news [81].

Harms caused by biased and false news have substantial impact across our society. Polarized content on Twitter and Facebook has been shown to play a role in the outcome of elections [218, 138]; and misinformation related to COVID-19 has been found to have a negative impact on public health responses to the pandemic [255, 149]. Developing methods for reducing these harms requires a broad understanding of the political bias and factualness of news content shared online, but studying news sharing is challenging

for three reasons: (1) the scale is immense, with billions of news links shared annually, (2) it is difficult to automatically quantify bias and factualness at scales where human labeling is often infeasible [216], and (3) the distribution of links is complex, with these links shared by many millions of users and thousands of communities.

While previous research has led to important insights on specific aspects of news sharing, such as user engagement [231], fact checking [273, 41], specific communities [216], and specific rumors [272, 214], large scale studies of news sharing are critical to understanding polarization and misinformation more broadly, and can inform community design, governance, and moderation interventions.

In this work, we present the largest study to date of news sharing behavior on Reddit, one of the most popular social media websites. We analyze all 559 million links submitted to Reddit from 2015-2019[13], including 35 million news links submitted by 1.3 million users to 135 thousand communities. We rate the bias and factualness of linked-to news sources using Media Bias/Fact Check (MBFC),[14] which considers how news sources favor different sides of the left-right political spectrum (bias), and the veracity of claims made in specific news stories (factualness) (§3.4.3).

In our analyses, we examine: the *diversity of news within communities* (§3.4.4), and how this diversity is composed of both the differences between community members and individual members' diversity of submissions; the *impact of current curation and amplification behaviors* on news' visibility and spread (§3.4.5); and the *concentration of extremely biased and low factual content* (§3.4.6), examining the distribution of links from the perspectives of who submitted them and what community they were submitted to.

We show that communities on Reddit exist across the left-right political spectrum, as measured by MBFC, but 74% are ideologically center left. We find that the diversity of left-leaning communities' membership is similar to that of equivalently right-leaning communities, but right-leaning communities have 105% more politically varied news sources, as their members individually post more varied links. This variance comes from the presence of links that are different from the community average, and in right-leaning communities, 74% of such links are to relatively-more biased news sources, 35% more than in left-leaning communities (§3.4.4).

[13] August 2019 was the most recent month of data available at the time of this study.

[14] While bias and factualness may vary from story to story, news source-level ratings maximize the number of links that can be rated, and are commonly used in research [28].

We demonstrate that, regardless of the political leaning of the community, community members' voting and crossposting (re-sharing) behavior reduces the impact of extremely biased and low factual news sources. Links to these news sources receive 20% fewer upvotes (§3.4.5) and 30% fewer exposures from crossposts compared to more neutral and higher factual content (§3.4.5). Furthermore, we find that users who submit such content leave Reddit 68% more quickly than others (§3.4.5). These findings suggest that low factual content spreads more slowly and is amplified less on Reddit than has been reported for Twitter [273, 26], although we do not directly compare behavior across the two platforms. Differences between Reddit and Twitter may stem from Reddit's explicit division into communities, or users' ability to downvote content, both of which help control content exposure.

Extremely biased and low factual content can be challenging to manage, as it is spread through many users, news sources, and communities. We find that extremely biased and low factual content is spread by an even broader set of users and communities relative to news content as a whole, exacerbating this challenge (§3.4.6). However, we find that 99% of extremely biased or low factual content is still concentrated in 0.5% of communities, lending credence to recent interventions at the community level [38, 36, 233, 228].

Our work demonstrates that additional research on news sharing online is especially needed on the topics of why users depart platforms and where they go, why false news appears to spread more quickly on Twitter than on Reddit, and how curation and amplification practices can manage influxes of extremely biased and low factual content.

Finally, we make all of our data and analyses publicly available to encourage future work on this important topic.

3.4.2 Related Work

Misinformation and Deceptive News

Social news platforms have seen a continued increase in use and a simultaneous increase in concern regarding biased news and misinformation [193, 180]. Recent studies have used network spread [273, 68, 26], content consumer [3], and content producer [168] approaches to assess the spread of misinformation. In this work, we examine news sharing behavior from news sources who publish content with varied degrees of bias or factualness, building on related work that has analyzed social news based on the characteristics of a

new source's audience [234] or the type of content posted [89].

Polarization and Political Bias

Many papers have recently been published on detecting political bias of online content either automatically [17, 54] or manually [80, 28]. Others have examined bias in moderation of content, as opposed to biased content or news sources themselves [124, 125]. Echo chambers are a major consideration in understanding polarization, with papers focusing on their development [4] and the role of news sources in echo chambers [104]. Others have examined who shares what content with what political bias, but did so using implicit community structure [234]. In this work, we examine thousands of explicit communities on Reddit, characterizing their polarization by examining the political diversity of news sources shared within, and the diversity of the community members who contribute.

Moderation and Governance

A large body of work has examined the role of moderation interventions such explanations [115], content removal [37], community bans [38, 36, 233] on outcomes such as migration [228], harassment [185] and harmful language use [274]. Others have focused on moderators themselves [181, 61], and technological tools to assist them [121, 291, 35], as well as self-moderation through voting [88, 231] and community norms [72]. In contrast, our work informs the viability of different moderation strategies, specifically by examining the sharing and visibility of news content across thousands of communities.

3.4.3 Dataset & Validation

We analyze all Reddit submissions to extract links, and annotate links to news sources with their political bias and factualness using ratings from Media Bias/Fact Check.

Reddit Content

Reddit is the sixth most visited website in the world, and is widely studied due to its size, diversity of communities, and the public availability of its content [189]. Users can submit links or text (known as "selfposts") to specific communities, known as "subreddits." Users may view submissions for a single com-

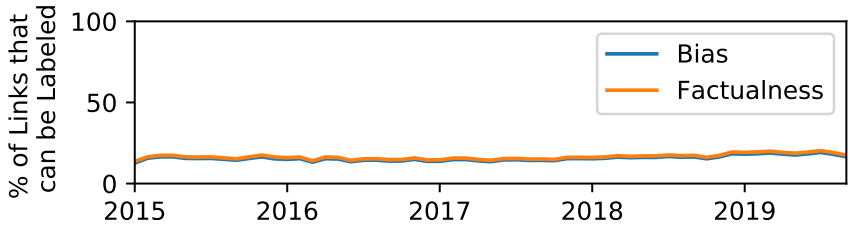


Figure 3.15: The percentage of links that can be annotated using the MBFC labels is very consistent (± 3.3%) over time, suggesting that comparisons over time are not significantly impacted by changes in annotation coverage.

munity, or create a "front page" which aggregates submissions from all communities the user "subscribes" to. Here, we focus on submissions over comments, as submissions are the primary mechanism for sharing content on Reddit, and users spend most of their time engaging with submissions [88].

To create our dataset, we downloaded all public Reddit submissions from Pushshift [20] posted between January 2015 and August 2019[15], inclusive, for a total of 56 months of content (580 million submissions, 35 million unique authors, 3.4 million unique subreddits). For each submission, we extract the URLs of each linked-to website, which resulted in 559 million links[16]. Additional summary statistics are included in the Appendix.

Ethical Considerations. We value and respect the privacy and agency of all people potentially impacted by this work. All Reddit content analyzed in this study is publicly accessible, and Pushshift, from which we source our Reddit content, permits any user to request removal of their submissions at any time. We take specific steps to protect the privacy of people included in our study [71]: we do not identify specific users, and we exclusively analyze data and report our results in aggregate. All analysis of data in this study was conducted in accordance with the Institutional Review Board at the University of Washington under identification number STUDY00011457.

Annotation of Links' News Sources

To identify and annotate links to news sources, we make use of Media Bias/Fact Check (hereafter MBFC), an independently run news source rating service. Bozarth, Saraf, and Budak [28] find that "the choice of traditional news lists [for fact checking] seems to not matter," when comparing 5 different news lists

[15] August 2019 was the most recent month available at the time of this study.
[16] While link submissions by definition contain exactly one link, text submissions (selfposts) can include 0 or more links.

including MBFC. Therefore, we selected MBFC as it offers the largest set of labels of any news source rating service [28]. MBFC provides ratings of the political bias (left to right) and factualness (low to high) of news outlets around the world, along with additional details and justifications for ratings, using a rigorous public methodology[17]. MBFC is widely used for labelling bias and factualness of news sources for downstream analysis [101, 177, 248, 49, 194] and as ground truth for prediction tasks [57, 250].

From MBFC's public reports on each news source, we extract the name of the news source, its website, and the political bias and factualness ratings. Bias is measured on a 7-point scale of 'extreme left,' 'left,' 'center left,' 'center,' 'center right,' 'right,' and 'extreme right,' and is reported for 2,440 news sources. Factualness is measured on a 6-point scale of 'very low factual,' 'low factual,' 'mixed factual,' 'mostly factual,' 'high factual,' and 'very high factual,' and is reported for 2,676 news sources (as of April 2020). For brevity, in the following analyses, we occasionally use the term 'left leaning' to indicate a news source with a bias rating of 'extreme left,' 'left,' or 'center left,' and the term 'right leaning' to indicate a news source with a bias rating of 'center right,' 'right,' or 'extreme right.'

We then annotate the links extracted from Reddit submissions with the MBFC ratings using regular expressions to match the URL of the link with the domain of the corresponding news source. For example, a link to www.rt.com/news/covid/ would be matched with the rt.com domain of RT, the Russian-funded television network, and annotated with a bias of 'center right' and a factualness of 'very low.' Links to URL shorteners such as bit.ly were excluded from labeling. We find that links to center left and high factual news sources are most common, accounting for 53% and 64% of all news links, respectively. Extreme left news source links are much less common, with 22.2 extreme right links for every 1 extreme left link (Figure 3.16).

Validation of MBFC Annotations. The use of fact checking sources such as MBFC is common practice for large scale studies, and MBFC in particular is widely used [57, 250, 101, 177, 248, 49, 194]. Additional confidence in MFBC annotations comes from the results of Bozarth, Saraf, and Budak [28], who find that (1) MBFC offers the largest set of biased and low factual news sources when compared among 5 fact checking datasets, and (2) the selection of a specific fact checking source has little impact on the evaluation of online content. Furthermore, we find that the coverage (the percentage of links that can be annotated using MBFC, excluding links to obvious non-news sources such as links to elsewhere on Reddit, to shopping sites, *etc.*) is

[17] https://mediabiasfactcheck.com/methodology/

very consistent ($\pm$ 3.3%) over the 4 year span of our dataset (Figure 3.15). Additionally, Bozarth, Saraf, and Budak [28] find that it is very rare for a news source's bias or factualness to change over time, suggesting that the potential 'drift' of ratings over time should not affect our results.

Robustness Checks with Different Set of Annotations. Lastly, we use an additional fact checking dataset from Volkova et al. [271], consisting of 251 'verified' news sources and 210 'suspicious' news sources, as an additional point of comparison for validation. While the exact classes in the Volkova *et al.* dataset are not directly comparable to MBFC, we can create a comparable class by comparing links with a MBFC factualness rating of 'mostly factual' or higher with Volkova *et al.*'s 'verified' news sources. In this case, when considering links that can be annotated using both datasets, MBFC and Volkova *et al.* have a Cohen's kappa coefficient of 0.82, indicating "almost perfect" inter-rater reliability [162]. We examined if these differences could have an impact of downstream analysis and found this to be unlikely. For example, results computed separately using MBFC and Volkova *et al.* agree with one another with a Pearson's correlation of 0.98 on the task of identifying the number of 'mostly factual' or higher links posted to a community.

Computing Mean Bias/Factualness. As described above, MBFC labels for bias and factualness are ordinal, yet for many analyses, it is useful to have numeric labels (*e.g.,* computing the variance of links in a community). To convert from MBFC's categorical labels to a numeric scale, we use a mapping of (-3, -2, -1, 0, 1, 2, 3) to assign 'extreme left' links a numeric bias value of -3, 'left' links a value of -2, 'center left' links a value of -1, 'center' links a value of 0, and positive values to map to the equivalent categories on the right. While this choice is somewhat arbitrary, it is consistent with the linear spacing between bias levels given by MBFC. Furthermore, we explored different mappings, including nonlinear ones, and found that our results are robust to different mappings. As such, we use the mapping given above as it is easiest to interpret. We use a similar mapping of (0, 1, 2, 3, 4, 5) to assign 'very low factual' links a numeric value of 0, 'low factual' links a value of 1, *etc.*, with 'very high factualness' links assigned a value of 5.

These numeric values are used to compute users' and communities' *mean bias* and *mean factualness*, central constructs in our analyses. To do so, we simply take the average of the numeric bias and factualness values of the links by each user or in each community. For many of our analyses, we group users by rounding their mean bias/factualness to the nearest integer. Thus, when we describe a user as having a 'left center bias,' we are indicating that the mean bias of the links they submitted is between -1.5 and -0.5.

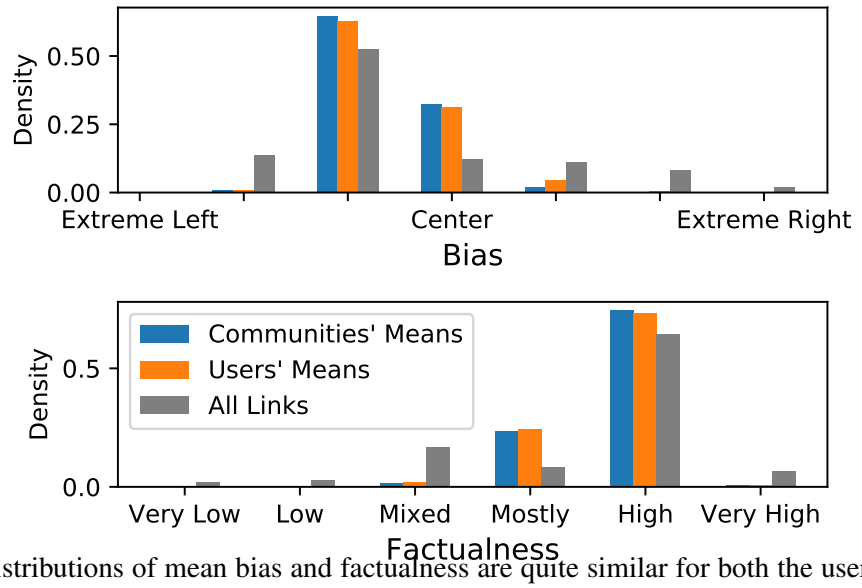


Figure 3.16: Distributions of mean bias and factualness are quite similar for both the user and community units of analysis. Grey bars show the normalized total counts of links of each type across all of Reddit.

The distributions of means are very similar for users and communities, with both closely following the overall distribution of news links on Reddit, shown with grey bars (Figure 3.16). 74% of communities and 73% of users have a mean bias of approximately center left, and 65% of communities and 62% of users have a mean factualness of 'high factual' (among users/communities with more than 10 links).

Similarly, we define *user variance of bias* as the variance of the bias values of the links submitted by a user, and similarly *community variance of bias* is defined as the variance of of the bias values of links submitted to a community. As with mean bias, we find that the distributions of user and community variance of bias are very similar to one another. The median user has a variance of 0.85, approximately the variance of a user with center bias who submits 62% center links, 22% center-left or center-right links, and 16% left or right links. The median community has a variance of 0.91, approximately that of a community where 62% of the content submitted has center bias, 20% of the content has center-left or center-right bias, and 18% of the content has left or right bias. Of course, a substantial amount of a community's variance comes from the variance of its userbase. We explore sources of this variance in §3.4.4.

Estimating Potential Exposures to Content

Links on Reddit do not have equal impact; some links are viewed by far more people than others. To understand the impact of certain types of content, we would like to understand how broadly that content is

viewed. As view counts are not publicly available, we use the number of subscribers to the community that a link was posted to as an estimate for the number of *potential exposures* to community members that this content may have had. While some users, especially those without accounts, view content from communities they are not subscribed to, subscription counts capture both active contributors and passive consumers within the community, which motivated our use of this proxy over other alternatives, such as the number of votes.

As communities are constantly growing, we define the number of potential exposures to a link as the number of subscribers to the community the link was posted to *at the time it was posted*. To estimate historic subscriber counts, we make use of archived Wayback machine snapshots of subreddit about pages, which provide the number of subscribers at the time of the snapshot. These snapshots are available for the $\sim$3,500 largest subreddits. In addition, we collected the current (as of Dec. 29, 2020) subscriber count for the 25,000 largest subreddits, as well as the date the subreddit was created (at which point it had 0 subscribers). We use the present subscriber count, archived subscriber counts (if available), and the creation date, and linearly interpolate between these data points to create a historical estimate of the subscriber counts over time for each of the 25,000 largest (by number of posts) subreddits in our dataset. The resulting set of subscriber count data, when joined with our set of Reddit content, provides potential exposure estimates for 93.8% of submissions. For the remaining 6.2% of submissions, we intentionally, conservatively *overestimate* the potential exposures by using the first percentile value (4 subscribers) from our subscriber count data. The effect of this imputation on our results is very minor as these only occur in communities with extremely little activity.

3.4.4 Diversity of News within Communities

In this section, we examine the factors that contribute to a community's variance of bias. This variance can come from a combination of two sources: (1) community members who are individually ideologically diverse (*user diversity*), and (2) a diverse group of users with different mean biases (*group diversity*). High user diversity corresponds to a community whose members have high user variance (*e.g.,* users who are ideologically diverse individually), and high group diversity corresponds to a community with high variance of its members' mean bias (*e.g.,* a diverse group of users who may be ideologically consistent individually). Of course, these sources of variance are not mutually exclusive; *overall community variance* is maximized

when both user diversity *and* group diversity are large.

Method

This intuition can be formalized using the Law of Total Variance, which states that total community variance is exactly the sum of User Diversity (within-user variance) and Group Diversity (between-user variance):

$$\mathrm{Var}(\mathcal{B}_c) = \mathrm{E}[\mathrm{Var}(\mathcal{B}_c|\mathcal{U})] + \mathrm{Var}(\mathrm{E}[\mathcal{B}_c|\mathcal{U}])$$

where $\mathcal{B}_c$ is a random variable representing the bias of a link submitted to community c, and $\mathcal{U}$ is a random variable representing the user who submitted the link.

We compute user diversity and group diversity for each community. User diversity is given by taking the mean of each user's variance of bias, weighted by the number of labeled bias links that user submitted. Group diversity is given by taking the variance of each community members' mean user bias, again weighted by their number of labeled links. We then sum the user and group diversity values to compute the overall community variance of political bias.

To understand *how* communities vary relative to their mean, we compute the balance of links in the adjacent relatively more- and less- biased categories. For example, a community with 'left' mean bias has two adjacent categories: 'extreme left' and 'center left,' with 'extreme left' being the relatively-more biased category, and 'center left' being the relatively-less biased category.

Results

Across all of Reddit, we find most (82%) communities' group diversity constitutes a majority of their overall variance of bias. When binned by their mean bias, we find that communities with extreme bias have, on average, lower total variance than communities closer to the middle of the spectrum (Figure 3.17). A community with mean bias of 'extreme left' would be expected to have a lower total variance as there are no links with bias further left than 'extreme left.' To control for this dynamic, we only compare symmetric labels: 'extreme left' to 'extreme right,' 'left' to 'right,' and 'center left' to 'center right.'

We find that right- and left-leaning communities have similar group diversity (Figure 3.17, right), but right-leaning communities (red) have 341% more user diversity than equivalently left-leaning communities,

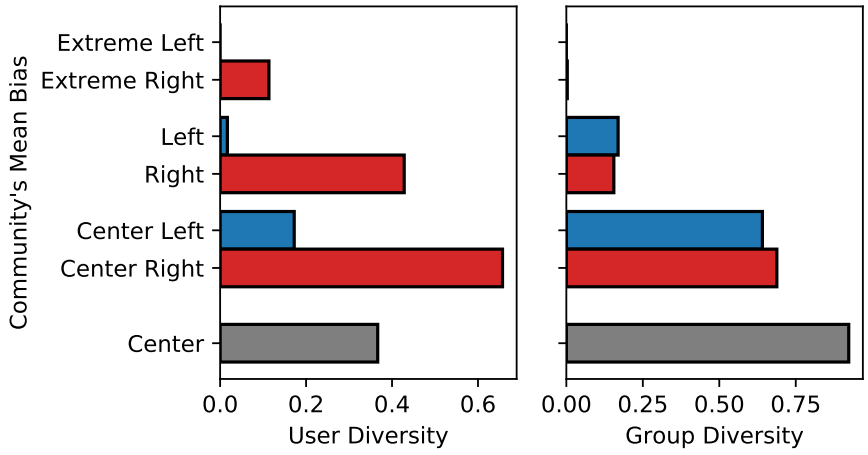


Figure 3.17: While group diversity is similar between left- and right-leaning communities with a similar degree of bias (right panel), right-leaning communities have higher user diversity than equivalently biased communities on the left (left panel). As a result, right-leaning communities have higher overall variance around their community mean. Right-leaning communities also favor relatively-more biased links, when compared to left-leaning communities.

on average (Figure 3.17, left). As a result, the average overall variance is 105% greater for right-leaning communities than left-leaning communities. Interestingly, we find that a larger share of right-leaning communities' variance is in more biased categories, relative to the community mean. 74% of right-leaning communities' adjacent links are relatively-more biased, compared to 55% for left-leaning communities, in other words, an increase of 35% $\left(\frac{74\%}{55\%}\right)$.

Implications

These results suggest that members of communities on the left and right have comparable group diversity, indicating the range of users are equally similar to one another. However, right-leaning communities have higher user diversity, indicating that the individual users themselves tend to submit links to news sources with a larger variety of political leaning. This creates higher overall variance of political bias in right-leaning communities, however these right-leaning communities also contain more links with higher bias, relative to the community mean, as opposed to more relatively-neutral news sources.

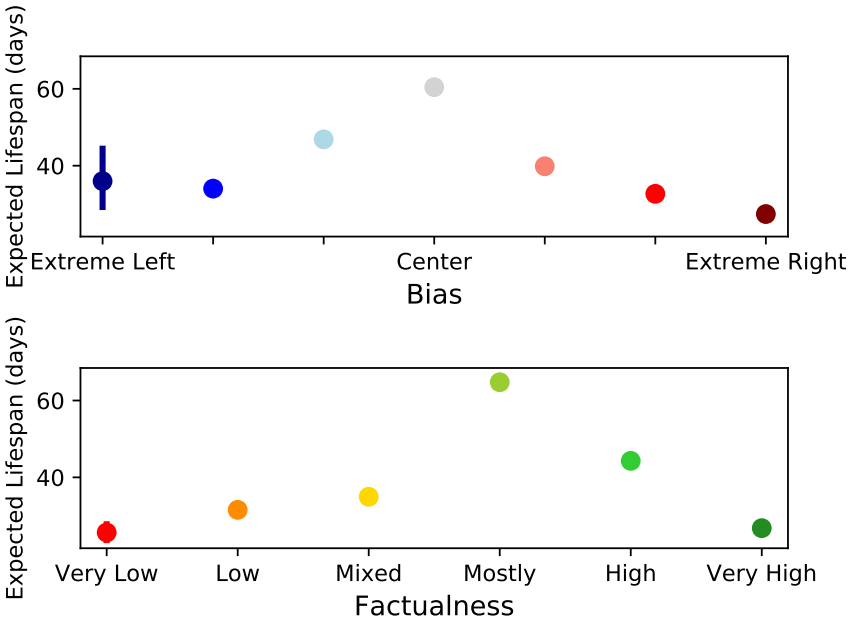


Figure 3.18: Users with extreme mean bias stay on Reddit less than half as long as users with center mean bias. Users with low and very low mean factualness also leave more quickly, but expected lifespan decreases as users' mean factualness increases past 'mixed factual'. Across all figures, error bars correspond to bootstrapped 95% confidence intervals (and may be too small to be visible).

3.4.5 Impact of Current Curation and Amplification Behaviors

The impact of content on Reddit is affected by users' behavior: how long they stay on the platform, how they vote, and how they amplify. In this section, we examine user longevity and turnover, community acceptance of biased and low factual content, and amplification through crossposting.

User Lifespan

Do users who post extremely biased or low factual content stay on Reddit as long as other users?

Method. We compute each user's lifespan on the platform by measuring how long they stay active on the platform after their first submission. We define "active" as posting at least once every 30 days, as in Waller and Anderson [275]. We group users by their mean bias and factualness, and for each group, compute the expected lifespan of the group members.

Results. We find that expected lifespan is longer for users who typically submit less politically biased content, with users whose mean bias is near center remaining on Reddit for approximately twice as long as users

with extreme or moderate mean bias, on average (Figure 3.18, top). This result holds regardless of whether users are left- or right-leaning. Users with a mean factualness close to 'mixed factual' or lower leave Reddit 68% faster than users whose mean factualness is near 'mostly factual' (Figure 3.18, bottom). However, we also find that users' expected lifespan decreases dramatically as their mean factualness increases to 'high' or 'very high' levels of factualness.

Implications. These results suggest that users who mostly post links to extremely biased or low factual news sources leave Reddit more quickly than other users. We can only speculate as to the causes of this faster turnover, but we note that users who stay on Reddit the longest tend to post links to the types of news sources that are most prevalent (grey bars in Figure 3.16 show overall prevalence of each type of link).

The faster turnover suggests that users sharing this type of content leave relatively early, limiting their impact on their communities. However, faster turnover also may make user-level interventions such as bans less effective, as these sanctions have shorter-lived impact when the users they are made against leave the site more quickly. Future research could examine why users leave, whether they rejoin with new accounts in violation of Reddit policy, and the efficacy of restrictions of new accounts.

Acceptance of Biased or Low Factual Content

How do communities respond to politically biased or low factual content?

Method. On Reddit, community members curate content in their communities by voting submissions up or down, which affects its position on the community feed [88]. A submission's 'score' is defined by Reddit as approximately the number of upvotes minus the number of downvotes that post receives. The score has been used in previous work as a proxy for a link's reception by a community [275, 50]. Links submitted to larger communities are seen by more users and therefore receive more votes. Therefore, we normalize each link's score by dividing by the mean score of all submissions in that community; links with a normalized score over 1 are more accepted than average, and links with a score under 1 are less accepted than average. In accordance with Reddit's ranking algorithm, submissions with higher normalized score appear higher in the feed viewed by community members, and stay in this position for longer [189].

To compute the *community acceptance* of links of a given bias or factualness, we average the normalized score of all links of that type in that community. We then take the median community acceptance across

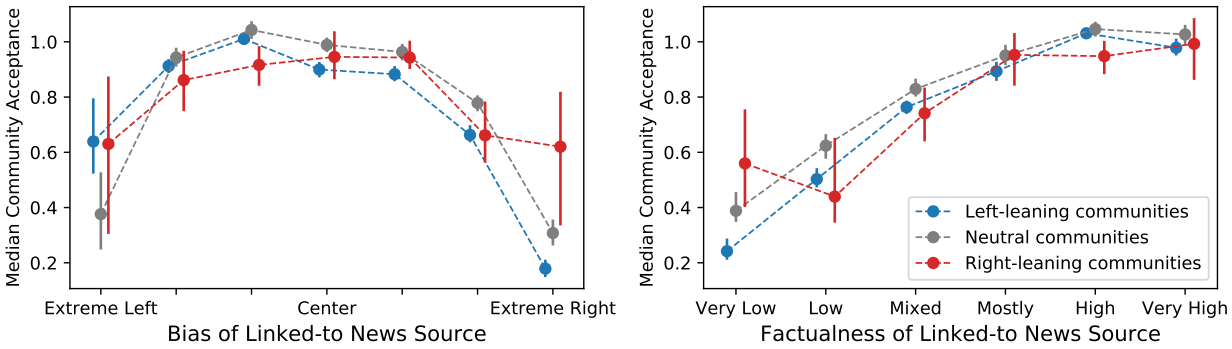


Figure 3.19: Regardless of the political leaning of the community, extremely biased content is less accepted by communities than content closer to center. Similarly, low and very low factual content is less accepted than higher factual content. Points perturbed on the x-axis to aid readability.

all left-leaning, right-leaning, and neutral communities. Here we use the median as it is more resilient to outliers than the mean.

Results. We find that, regardless of the community's political leaning, median expected community acceptance is 18% lower for extremely biased content than other content (Figure 3.19). For left-leaning and neutral communities, community acceptance decreases monotonically as factualness drops below 'high.' However, we observe that right leaning communities are 167% ($p = 0.0002$) more accepting of extreme right biased and 85% ($p = 0.004$) more accepting of very low factual content than left-leaning and neutral communities (Mann–Whitney U significance tests).

Implications. This suggests that across Reddit, communities are sensitive to extremely biased and low factual content, and users' voting behavior is fairly effective at reducing the acceptance of this content. However, curation does not seem to result in better-than-average acceptance for any content—no median acceptance values are significantly ($p < 0.05$) above 1, as non-news content tends to receive higher community acceptance than news content.

Previous research has found that on Twitter, news that failed fact-checking spread more quickly and was seen more widely than news that passed a fact-check [273]. Interestingly, we find evidence that behavior on Reddit is somewhat different, with median left-leaning, right-leaning, and neutral communities all being less accepting of low and very low factual content. Importantly, our methodology differs from Vosoughi, Roy, and Aral [273] in that we use bias and factualness evaluations that were applied to entire news sources, as opposed to the fact checking of specific news articles, limiting direct comparisons. Furthermore, we do not

analyze the time between an initial post and its subsequent amplification, and so cannot directly comment on the 'speed' of amplification. We do find evidence, however, that highly biased content on Reddit is less upvoted than more neutral content.

These difference may in part be explained by differences between reddit's and Twitter's mechanisms for impacting the visibility of content. Whereas Twitter users are only able to increase visibility by retweeting, liking, replying to, or quoting content, on Reddit, users may downvote to decrease visibility of content they object to. We speculate that this may partially explain the differences in acceptance that we find between Reddit and Twitter.

Selective Amplification of News Content

How does amplification of content affect exposure to biased and low factual content? On Reddit, users are not only able to submit links to *external* content (such as news sites), but users are also able to submit links to *internal* content elsewhere on Reddit, effectively re-sharing and therefore *amplifying* content by increasing its visibility on the site. This is commonly known as 'crossposting,' and often occurs when a user submits a post from one subreddit to another subreddit, although such re-sharing of internal content can happen within a single community as well. Here, we seek to understand the role that amplification through crossposts has on Reddit user's exposure to various kinds of content.

Method. To identify the political bias and factualness of crossposted content, we identify all crossposted links to news sources, and propagate the label of the crossposted link. Then, we compute the fraction of total potential exposures from crossposts for each bias/factualness category.

Results. We find that amplification via crossposting has an overall small effect on the potential exposures of news content. While 10% of all news links are crossposts, only 1% of potential exposures to news links are due to crossposts. This suggests that the majority of crossposts are content posted in relatively larger communities re-shared to relatively smaller communities with relatively fewer subscribers, diminishing the impact of amplification via crossposting. As such, *direct* links to news sites have a far greater bearing on Reddit users' exposure to news content than crossposts.

However, the role of crossposts in exposing users to new content is still important, as crossposts account for more than 750 billion potential exposures. We find that extremely biased and low factual content is

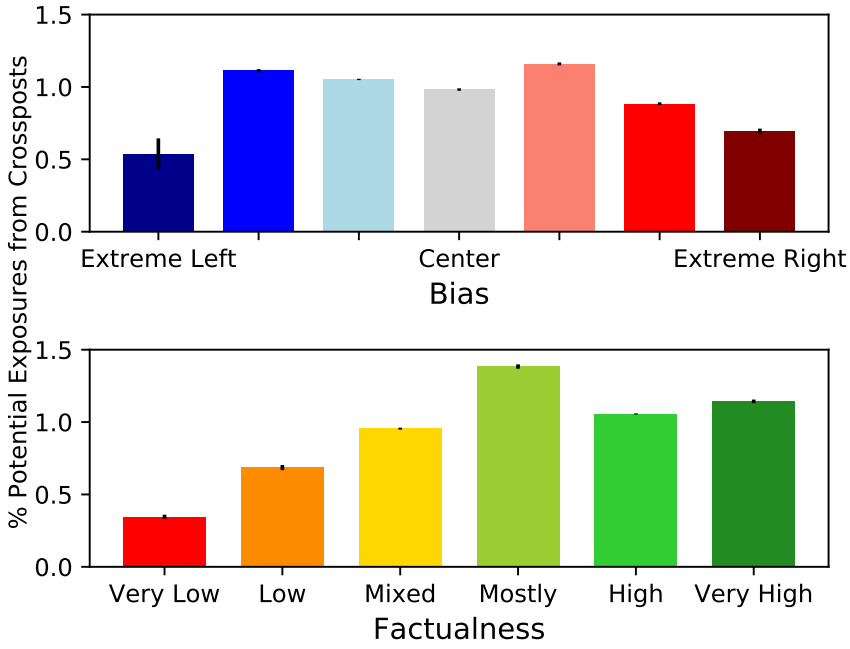


Figure 3.20: Extremely biased and low factual content is amplified by crossposts relatively less than other content. Regardless of the bias or factualness of the content, while crossposts are responsible for more than 750 billion potential exposures, they make up only 1% of total potential exposures, suggesting that direct links to news sources play an especially important role in content distribution.

amplified less than other content, as shown in Figure 3.20, which illustrates the percentage of total potential exposures that come from crossposts for each bias/factualness category. Reddit users exposed to center left biased, center biased, or center right biased content are 53% more likely to be exposed to this content via amplification than Reddit users exposed to extremely biased content. Similarly, Reddit users exposed to 'mostly factual' or higher factualness content are 217% more likely to be exposed to such content via amplification than Reddit users exposed to very low factual content.

Implications. Given that only 1% of potential exposures are from amplifications, understanding the way that *direct* links to external content are shared is critical to understanding the sharing of news content on Reddit more broadly.

The relative lower amplification of extremely biased and very low factual content suggests users' sensitivity to the bias and factualness of the content they are re-sharing. As in §3.4.5, this suggests differences between Reddit and Twitter, where content that failed a fact-check has been found to spread more quickly than fact-checked content [273]. We speculate that this may be due to structural differences between the

two platforms. On Reddit, users primarily consume content through subscriptions to communities, not other users. This may explain the diminished impact of re-sharing on Reddit compared to Twitter.

3.4.6 Concentrations of Extremely Biased or Low Factual News Content

It is critical to understand where different news content is concentrated in order to best inform strategies for monitoring and managing its spread online. In this section, we examine how extremely biased and low factual content is distributed across users, communities, and news sources. We also compare the concentration of extremely biased and low factual content to all content.

Method

We consider three types of content: (1) news content with extreme bias or low factualness, (2) all news content, and (3) all content (including non-news). We group each of these types of content by three perspectives: the user who posted the content, the community it was posted to, and the news source (or domain, in the case of all content) linked to. We then take the cumulative sum of potential exposures across the users, communities, and news sources, to compute the fraction of potential exposures contributed by the top $n\%$ of users, communities, and news sources. We repeat this process, replacing the number of potential exposures with the total number of links, to consider the concentration of links being submitted, regardless of visibility.

Results

We find that overall, extremely biased and low factual content is highly concentrated across all three perspectives, but is especially concentrated in a small number of communities, where 99% of potential exposures stem from a mere 109 (0.5%) communities (Gini coefficient=0.997) (Figure 3.21a). No matter the perspective, exposures to extremely biased or low factual content (solid line) are less concentrated than all content (dotted line) (Figure 3.21abc).

Under the community and news source perspectives, exposures (Figure 3.21ac) are more concentrated than links (Figure 3.21df). While links are already concentrated in a small share of communities, some communities are especially large, and therefore content from these communities receives a disproportionate share of potential exposures. This is not the case for users, as the distributions of exposures (Figure 3.21b)

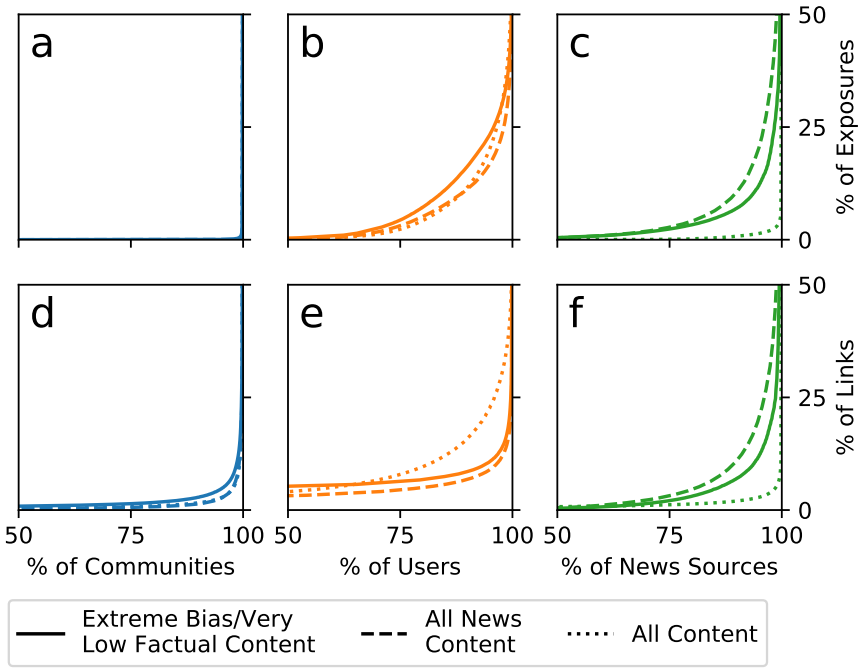


Figure 3.21: When compared to all content on Reddit (dotted line), extremely biased or low factual content (solid line) is more broadly distributed, making it harder to detect, regardless of the community, user, or news source perspective. However, 99% of potential exposures to extremely biased or low factual content are restricted to only 0.5% of communities. Here, a curve closer to the lower-right corner indicates a more extreme concentration. Note that axis limits do not extend from 0 to 100%.

are less concentrated than the distributions of links (Figure 3.21e). This indicates that while some users submit a disproportionate share of links, these are not the users whose links receive the largest potential exposure, as potential exposure is primarily a function of submitting links to large communities.

Implications

The extreme concentration of extremely biased or low factual content amongst a tiny fraction of communities supports Reddit's recent and high profile decision to take sanctions against entire communities, not just specific users [111]. These decisions have been extensively studied [38, 36, 264, 233, 228]. While this content is relatively less concentrated amongst users, in absolute terms, this content is still fairly concentrated, with 10% of users contributing 84% of potential exposures. As such, moderation sanctions against users can still be effective [181]. We note that the concentration of extremely biased or low factual content amongst a small fraction of users is similar to what has been found on Twitter [94], although methodological differences preclude a direct comparison.

3.4.7 Discussion

Summary & Implications

In this work, we analyze all 580 million submissions to Reddit from 2015-2019, and annotate 35 million links to news sources with their political bias and factualness using Media Bias/Fact Check. We find:

- Right-leaning communities' links to news sources have 105% greater variance in their political bias than left-leaning communities. When right-leaning communities link to news sources that are different than the community average, they link to relatively-more biased sources 35% more often than left-leaning communities (§3.4.4).

- Existing curation and amplification behaviors moderately reduce the impact of highly biased and low factual content. This suggests that Reddit differs somewhat from Twitter, perhaps due to its explicit community structure, or the ability for users to downvote content (§3.4.5).

- Highly biased and low factual content tends to be shared by a broader set of users and in a broader set of communities than news content as a whole. Furthermore, the distribution of this content is more

concentrated in a small number of communities than a small number of users, as 99% of exposures to extremely biased or low factual content stem from only 0.5% or 109 communities (§3.4.6). This lends credence to recent Reddit interventions at the community level, including bans and quarantines.

Limitations

One limitation of our analyses is the use of a single news source rating service, MBFC. However, the selection of news source rating annotation sets has been found to have a minimal impact on research results [28]. MBFC is the largest (that we know of) dataset of news sources' bias and factualness, and is widely used [57, 250, 101, 248, 49]. More robust approaches could combine annotations from multiple sources, and we find that MBFC annotations agree with the Volkova et al. [271] dataset with a Pearson Correlation of 0.96 on an example downstream task (§3.4.3).

Our focus is on the bias and factualness of news sources shared online. We do not consider factors such as the content of links (*e.g.,* shared images, specific details of news stories), or the context in which links are shared (*e.g.,* sentiment of a submission's comments). These factors are important areas for future work, and are outside the scope of this paper.

While MBFC (and by extension, our annotations) includes news sources from around the world, our analyses, especially the left-right political spectrum and associated colors, takes a US-centric approach. Polarization and misinformation are challenges across the globe [196], and more work is needed on other cultural contexts.

Our paper explores the impact of curation and amplification practices, but not the impact of community moderators who are a critical component of Reddit's moderation pipeline [181]. Future work could examine news content removed by moderators.

Finally, we are limited by the unavailability of data on which users view what content. While we use subreddits' subscriber counts to estimate exposures to content, more granular data would enable us to better understand the impact of shared news articles, for example, the percentage of users who are exposed to extremely biased or low factual content [94].

Conclusion

Biased and inaccurate news shared online are significant problems, with real harms across our society. Large-scale studies of news sharing online are critical for understanding the scale and dynamics of these problems. We presented the largest study to date of news sharing behavior on Reddit, and found that right-leaning communities have more politically varied and relatively-more biased links than left-leaning communities, current voting and re-sharing behaviors are moderately effective at reducing the impact of extremely biased and low factual content, and that such content is extremely concentrated in a small number of communities. We make our dataset of news sharing on Reddit public, in order to support further research.

3.5 Summary of Contributions to Thesis

In this chapter, we examined a broad range of moderation practices, rules, and non-moderation affordances, like voting and reposting, across hundreds of thousands of communities, using causal inference methods to control for confounding factors and identify which are most likely to improve two important community outcomes: how community members perceive the governance of their communities, and how likely they are to be exposed to untrustworthy content. First, we leveraged the insight that community members discuss their moderators to develop an automated pipeline to detect and classify such statements. We analyzed 1.89 million statements about moderators across an 18-month period, demonstrating that strict rule enforcement was more favorably received in certain community types than others, and that moderators who maintained active community engagement before, during, and after their tenure were viewed more positively by their constituent community members. Next, we examined community rules specifically. We conducted the most comprehensive analysis of community rules on Reddit to date, examining 67,575 rules from 5,225 communities over five years using historical data from the Wayback Machine. We developed a classification system to classify rules' tone, target, and topic, and revealing that prescriptively phrased rules ('Be nice') are associated with more positive community perceptions of governance than restrictively phrased rules ('Don't be mean'), and that while new rules initially improved member perceptions of governance, this positive effect diminished after approximately six months. Last, we examined voting and reposting behavior, two important non-moderator affordances that enable community members to enact some control over the

visibility of content in their communities. Through the lens of news sharing, we examined the associations between these affordances and the visibility of highly biased or low factual news in the largest study of news sharing on Reddit. Our analysis demonstrated that community-driven affordances such as upvoting, downvoting, and crossposting are generally effective at reducing the relative visibility of misinformation and fake news.

These large-scale assessments of existing governance practices, which leverage the natural diversity in governance and outcomes across hundreds of thousands of communities, contribute a robust and data-driven set of best practices for governance that generalizes to a broad set of communities both on and off Reddit. In the next chapter, we will work to maximize the positive impact of this work by actually deploying promising interventions in partnership with community moderators and the Reddit Moderator Council.

Chapter 4

Deploying Systems to Increase Engagement and Maximize Impact

Data-driven best practices and tools for community governance, no matter how effective, do not have any impact unless they are actually widely adopted by communities. In this chapter, we partner with community leaders to deploy practices and novel systems to improve community outcomes.

As a member of the Reddit Moderator Council (§1.4), I am fortunate to have opportunities to connect with hundreds of other moderators on Reddit. Throughout my dissertation research, I have shared my results with members of the Council and solicited their feedback, along with other 'meta communities' popular on Reddit with moderators, such as /r/TheoryOfReddit. The feedback I have received has been overwhelmingly positive, with many moderators remarking that my results align with their own personal experiences in their communities. By sharing best practices from my research results with the Reddit Moderator Council, I have informed moderator recruiting strategies in over 87 communities with more than 115.4 million combined member.

Online community governance can be further improved by the development and deployment of new tools. In §4.1, we introduce CritiquePoints, new system to gamify the delivery of high quality feedback on creative tasks in a large photography critique community with 1.7 million members. This system also enables the precise measurement and annotation of the quality of submissions in a community-specific fashion. Over a 3.5 year period, the system has been used 9,934 times by 3,356 different community members. We

leverage these annotations to conduct an analysis of what makes for high quality feedback on creative tasks, using randomly sampled critiques of the same photographs which were *not* awarded points as a control. We find that high quality feedback is specific, actionable, addresses the questions asked by the recipient, and goes beyond what was originally asked. We use these results to inform the design of our LLM coaching intervention, presented in the subsequent section.

In §4.2, we describe the deployment and evaluation of an LLM-powered tool to teach members of the same photography critique community how to give one another better feedback and improve the overall community experience. We evaluate the system with community members, comparing comparing two different pedagogical strategies: an Assistant that makes easy-to-accept writing suggestions, and a Socratic Coach that encourages participants to think more deeply about their writing without making specific suggestions. Semi-structured qualitative interviews with a participants show that both LLM powered strategies are effective at improving the quality of delivered feedback when compared with a simple control. Comparing between the Assistant and the Coach, participants report that the Assistant is easier to use, with lower cognitive burden, but the Coach results in deeper learning of writing skills.

4.1 CritiquePoints: System to Measure and Incentivize High Quality Participation in a Photography Feedback Community

4.1.1 Introduction

Online communities vary dramatically in their topics, culture, and values [284, 283]. Our previous surveys of community members across thousands of communities has shown that the Quality of Content in an online community is one of the most important aspects of a community [284], yet what makes for high quality content is different in one community than another [283]: clearly, a picture of a cat may be a great fit in /r/cats, but less so in /r/dogs. Because the nature of high quality content is so community dependent, it is very difficult to measure the quality of content for most online communities, and, of course, it is challenging to improve something that cannot be measured.

In this work, we develop and deploy CritiquePoints, a system that allows community members to give one another virtual points to recognize high quality contributions. Through badges [207, 235] and leaderboards, CritiquePoints encourages participation while simultaneously enabling the measurement of characteristics of content that is positively received by community members. We partner with the moderators of /r/photocritique, a large photography critique community on Reddit, which has over 1.7 million members, to deploy CritiquePoints. In the first 3.5 years of an ongoing deployment, 3,356 members of /r/photocritique have awarded 9,934 CritiquePoints.

We use CritiquePoints to conduct one of the largest studies of creative task feedback delivery in online communities. We sample critiques that did not receive CritiquePoints to create a control set for comparison with those that did receive points, and find that critiques most likely to be awarded points have four key characteristics: (1) they are specific and actionable, (2) they use images to exemplify concpets and suggestions, (3) they directly address the questions asked by the critiquee, and (4) they go beyond what was asked-for, providing additional details and suggestions.

We use these results to inform the development of a system to teach community members to give one another better feedback, improving the experience in the community. We will make an anonymized dataset of the critiques that received CritiquePoints public upon the completion of this project. Furthermore, we open source the code underlying CritiquePoints, enabling its adoption in other communities and other context.

4.1.2 Related Work

Measuring the Quality of Content

What makes for high quality content in an online community is highly dependent on the community context, and studying high quality content requires deep familiarity with the community in question [283, 284]. Voting affordances (such as upvoting on Reddit and liking on Facebook) can be used to quantify a community's reception of a piece of content, but these mechanisms can be unreliable, as they can be muddled with people's personal preferences and other factors [88, 91]. Manually selected high quality content has been used for moderation interventions, but these processes are difficult to scale [277]. In contrast, in this work we introduce a scalable system to measure high quality content using crowdsourced annotations, a practice which is extremely rare, having only been deployed in one other community on Reddit [246, 119, 146].

Gamification in Online Communities

Gamification involves applying game-like elements, like points, awards, badges, leaderboards, and competitions, to non-game contexts [207, 225]. Gamification can increase engagement [133, 205, 217] and improve the quality of participation [261, 160, 277, 21], although concerns have been raised about the sustainability this increased engagement over time [217, 263, 262] and the impacts this may have upon users [230, 132]. In this work, we deploy a gamification system in a large online community not only to incentivize high quality participation, but primarily to enable the precise measurement of the quality of said content.

Point-based Systems on Reddit

Beyond the built in upvote system, point based systems to specifically measure content quality and gamify participation are rare, with /r/ChangeMyView being the only community that we are aware of having deployed one. /r/ChangeMyView, a highly-structured discussion platform, encourages users to debate their opinions, and users can give the most effective debators 'Deltas,' the name used in /r/ChangeMyView to describe their community specific points. /r/ChangeMyView has been studied by researchers to understand what makes for convincing argument [254], and to evaluate content moderation strategies [119, 146, 246]. In this work, we introduce CritiquePoints, a system inspired by /r/ChangeMyView's Deltas, but applied to a unique and unstudied context: online feedback on creative tasks.

4.1.3 An Overview of the CritiquePoint System

Background on /r/photocritique

This work was conducted in partnership with the moderators of /r/photocritique, one of whom is the author of this dissertation. All the moderators of the subreddit agreed to deploy the CritiquePoint system.

/r/photocritique is the largest photography feedback community on Reddit, with over 1.7 million members as of Summer 2025. /r/photocritique was founded in 2009, and has grown steadily since then. Compared to other photography communities on Reddit, /r/photocritique is relatively strictly moderated and has a specific, narrow purpose: to provide feedback and critique on members' photography. Only the photographer is permitted to submit their own work for critique, submitting others' images is a violation of community rules.

To submit an image for critique, the photographer creates a new top-level post in the community, containing their image. Alongside their image, the photographer is required to submit a succinct title describing the image, and a comment that is at least a few sentences in length, describing their intentions for the image, their process, and asking specific questions (*e.g.,* 'is this too dark?') or asking for the specific feedback they are looking for (*e.g.,* 'how's my editing?').

Once a submission satisfies all these requirements, it is automatically approved by the community's automated moderation software [121] and becomes visible to other members of the community, who can see a list of recently submitted images on the community's 'front page.' Other community members then post their critiques as comments on the image, and the original photographer can reply, along with other community members, to enable further, deeper discussion.

Awarding CritiquePoints

The CritiquePoint system uses a bot to track points and interact with community members. Anyone can give a CritiquePoint to the author of a critique by simply replying to that critique with a comment containing the CritiquePoint macro, `!CritiquePoint`, and there is no limit to the number of points someone can give. Behind the scenes, the CritiquePoints bot, which has the username /u/CritiquePointBot, is streaming comments made in /r/photocritique, using regular expressions to detect uses of this macro. Once an awarded CritiquePoint is detected, the bot replies to the awarding comment with an acknowledgment (Figure 4.1).

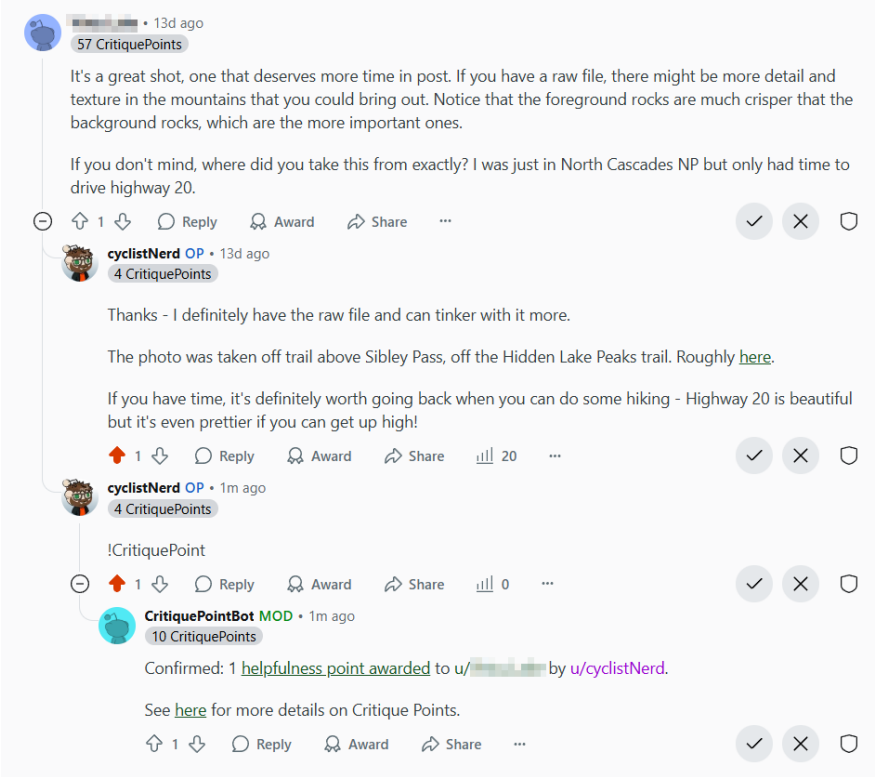


Figure 4.1: An example of the awarding of a CritiquePoint, with some usernames redacted to protect privacy. The top comment is the awarded comment. The awarding comment is made by user /u/cyclistNerd, second from the bottom, consisting in this case of just the macro `!CritiquePoint`. /u/CritiquePointBot has immediately replied with an acknowledgment comment at the bottom of the screenshot. Note the badges (known on Reddit as flairs) indicating how many CritiquePoints each user involved in the exchange has previously received.

After the awarding comment is submitted to Reddit's servers, both the author of the awarded comment and the author of the awarding comment are sent a Reddit notification of the CritiquePoint exchange.

After each CritiquePoint is awarded, the system makes a few more actions. It logs the details of the point, along with the text of the awarding and awarded comments, and the usernames of the community members involved, to an internal database. Then, the system gives the author of the awarded comment a publicly visible badge denoting that they received a point. If the author has received CritiquePoints in the past, then the system increments the number of points shown on their badge. The appearance of these badges, known as 'flair' on Reddit, is visible in Figure 4.1.

The system logs the full text of every critique that is awarded a CritiquePoint, but many critiques are long, with several paragraphs of text. To increase the precision of measurement, we would like to know

which parts of the critique are most helpful. Towards this goal, shortly after someone awards a CritiquePoint, /u/CritiquePointBot sends the awarder a private Reddit Chat message asking them to select the sentences from the awarded comment that they thought were most helpful. We call this process Segmentation. Depicted in Figure 4.2, Segmentation is entirely optional, and users can permanently opt-out of receiving reminder messages. To segment the critique, the system uses regular expressions to divide the text of the awarded critique up into sentences, and inserts index numbers into the text before sending a copy in the chat, along with instructions. The user then responds with the indices of the most-helpful sentences, which the system parses out and logs internally.

To encourage participation, the CritiquePoint system reminds photographers to award CritiquePoints to helpful comments. If, 48 hours after submitting an image for critique, the photographer has received at least four critiques [1] but has not yet awarded any critiques a CritiquePoint, /u/CritiquePointBot sense them a private Reddit Chat reminder, asking them to consider awarding helpful comments. Recipients of these reminders can permanently opt-out of receiving them if they prefer. The text of the reminders is:

> You recently posted a photo for critique <link to post> in /r/photocritique.
>
> <Title of Submission>
>
> I hope you received some useful feedback! If you did, you can reply to any comment with `!CritiquePoint` to give a Critique Point to the person who provided you with a high quality critique. Giving Critique Points is easy, and helps improve the whole community. See here <Link to /r/photocritique Wiki page> for more information on Critique Points.
>
> This message was sent by a bot. To opt out of future reminders such as this one, reply to this PM with `optout reminders`.

To raise awareness of the CritiquePoints system, and to further incentivize participation, at the start of each calendar month, /u/CritiquePointBot posts a monthly leaderboard to the front page of the community, which is pinned to the top of the page for all to see. This leaderboard mentions the usernames of the three community members who received the most CritiquePoints in the previous month, and encourages community members to use the comments to discuss their ideas or concerns for the community (Figure 4.3).

[1] We chose this threshold to avoid encouraging the awarding of CritiquePoints to unhelpful critiques if a photographer only received a few critiques.

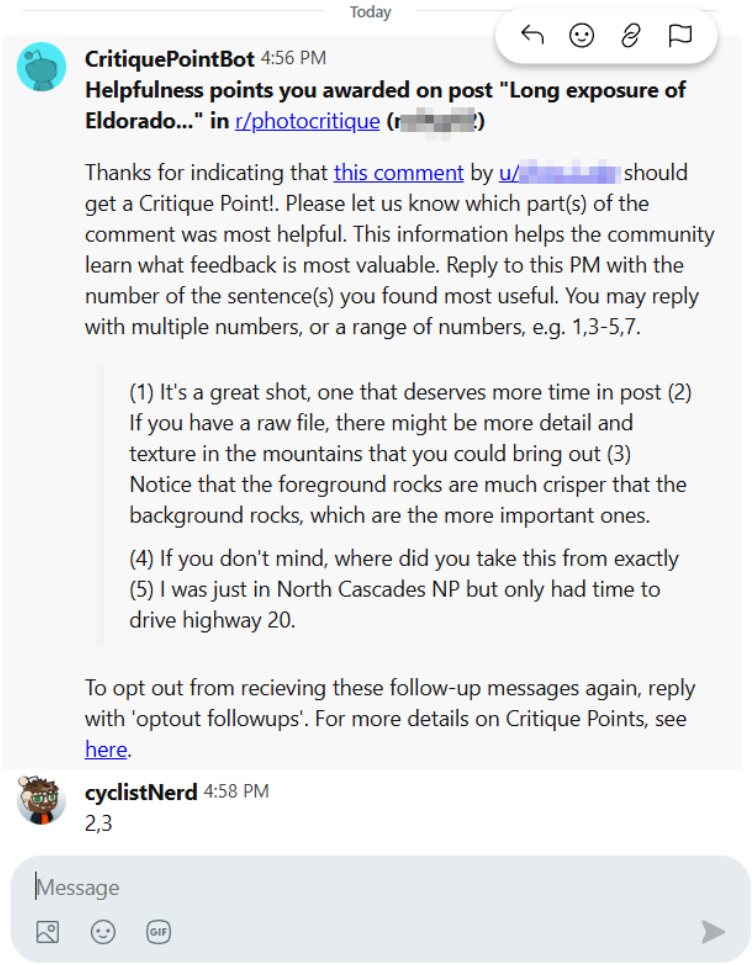
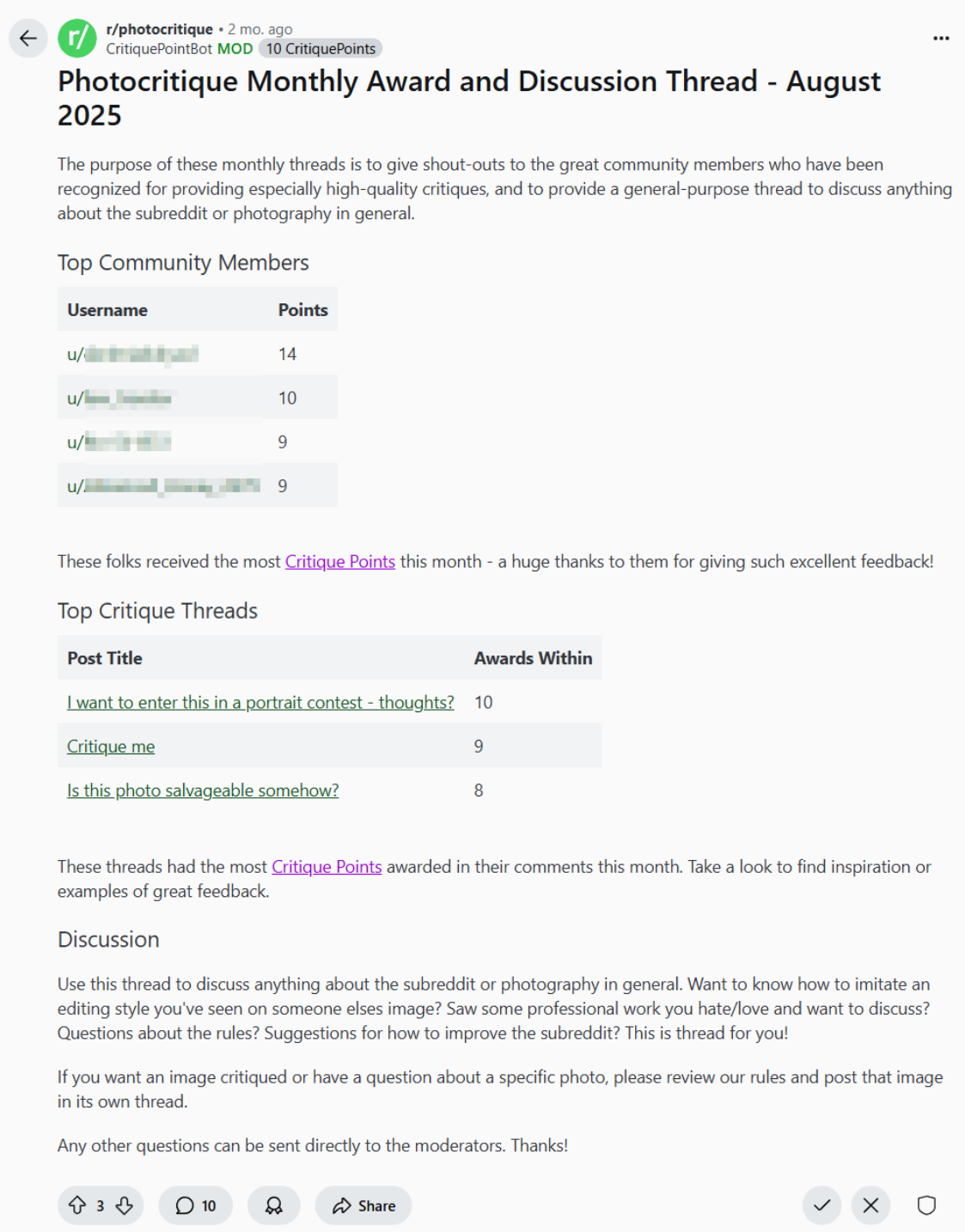


Figure 4.2: A demonstration of the Segmentation process, with usernames redacted to protect privacy. /u/CritiquePointBot sends a private Reddit Chat message asking the awarder of a CritiquePoint to identify the sentences within the awarded critique that are most helpful. Here, the awarder has responded that they thought sentences number 2 and 3 were most helpful. These numbers correspond to the index numbers (1) in parenthesis inserted into the text of the critique.



r/photocritique • 2 mo. ago
CritiquePointBot MOD 10 CritiquePoints

Photocritique Monthly Award and Discussion Thread - August 2025

The purpose of these monthly threads is to give shout-outs to the great community members who have been recognized for providing especially high-quality critiques, and to provide a general-purpose thread to discuss anything about the subreddit or photography in general.

Top Community Members

Username	Points
u/[redacted]	14
u/[redacted]	10
u/[redacted]	9
u/[redacted]	9

These folks received the most Critique Points this month - a huge thanks to them for giving such excellent feedback!

Top Critique Threads

Post Title	Awards Within
I want to enter this in a portrait contest - thoughts?	10
Critique me	9
Is this photo salvageable somehow?	8

These threads had the most Critique Points awarded in their comments this month. Take a look to find inspiration or examples of great feedback.

Discussion

Use this thread to discuss anything about the subreddit or photography in general. Want to know how to imitate an editing style you've seen on someone elses image? Saw some professional work you hate/love and want to discuss? Questions about the rules? Suggestions for how to improve the subreddit? This is thread for you!

If you want an image critiqued or have a question about a specific photo, please review our rules and post that image in its own thread.

Any other questions can be sent directly to the moderators. Thanks!

3 10 Share

Figure 4.3: An example of a monthly leaderboard and discussion thread, with usernames redacted to protect privacy.

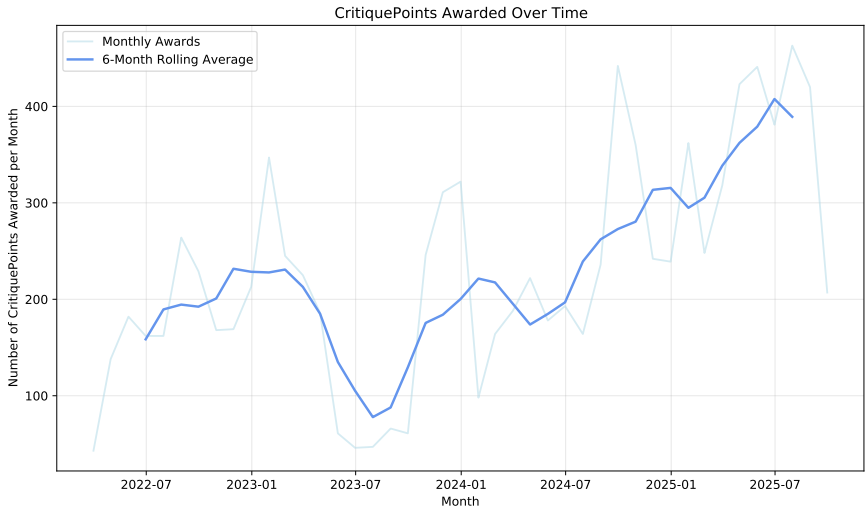


Figure 4.4: Although there is some noise, usage of the CritiquePoint system has increased over time. The substantial drop in usage around July 2023 corresponds to the period where the /r/photocritique community was 'blacked out' as part of the 2023 Reddit API Change Protests.

The PhotoCritique system is implemented using the Python and the Python Reddit API Wrapper (PRAW). It is hosted on a University of Washington webserver, and, on the client side, is implemented entirely using Reddit, so users can use any Reddit interface they choose without needing to visit any external webpage or install any additional apps. We make the code for the system open source[2] for transparency, and so that other communities can adopt it for their own purposes.

Launch and Adoption. We first launched the CritiquePoint system in March 2022, and it has been online continuously since then. In the first 3.5 years of its deployment, 3,356 different members of /r/photocritique have awarded a cumulative 9,934 CritiquePoints to critiques on 5,130 different images. While anyone can award a CritiquePoint, CritiquePoints are by far most frequently awarded by the photographer who requested the critique. Only 5.67% of CritiquePoints have been awarded by someone other than the photographer soliciting feedback.

[2] `https://github.com/galenweld/photocritique_feedback_bot/`

4.1.4 What makes for high quality feedback?

We use the awarding of CritiquePoints to measure what makes for high quality feedback. To do so, we must make comparisons between critiques that *do* and *do not* receive CritiquePoints. We do this by randomly sampling un-awarded critiques of the same images that also received awarded critiques, in order to exclude critiques of images that did not receive any awarded critiques. This lets us correct for differences in critiques between images that do and not receive any high quality critique. For each image that received an awarded critique, we sample an equal number of unawarded critiques, to maintain class balance.

Length of Critique

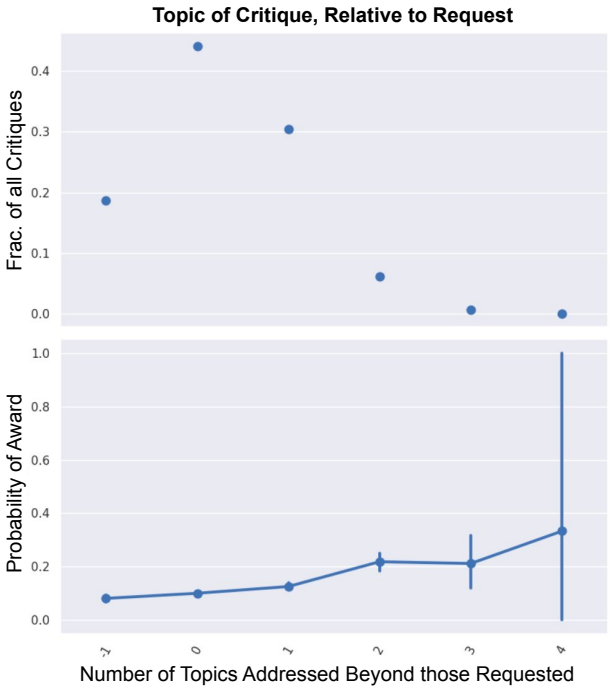
We examine how the length of critiques is associated with their likelihood of receiving a CritiquePoint. Using the number of characters in a critique as our metric for its length, we find that longer critiques are much more likely to receive a CritiquePoint (Figure 4.5b). Long critiques, with more than 700 characters, are approximately $10\times$ as likely to receive a CritiquePoint than critiques under 100 characters, and furthermore, the majority of critiques over 600 characters receive awards.

Use of Examples

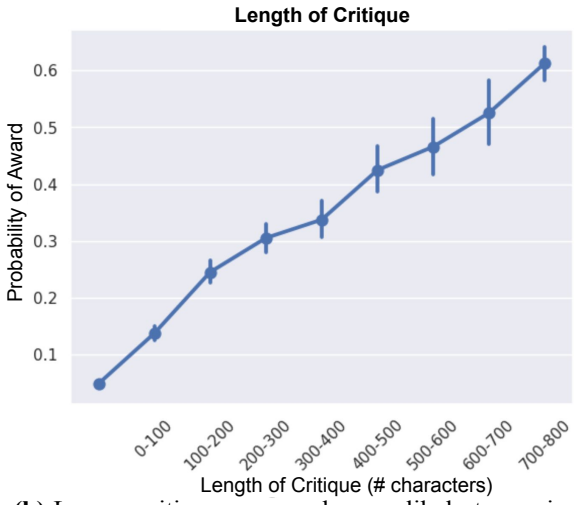
Reddit enables users to include images in their comments, and on /r/photocritique, community members occasionally include images in their critiques as examples. Community members may make a suggestion for a crop by demonstrating the crop themselves, or might edit a copy of the photographer's image and upload the copy as a part of their critique to show their suggestion. By examining if a critique includes an image, we can measure the difference in quality between critiques with and without example images. Here, we find that critiques that include examples are more than twice as likely as those without examples to be awarded a CritiquePoint (Figure 4.5c).

Alignment with the Photographer's Request

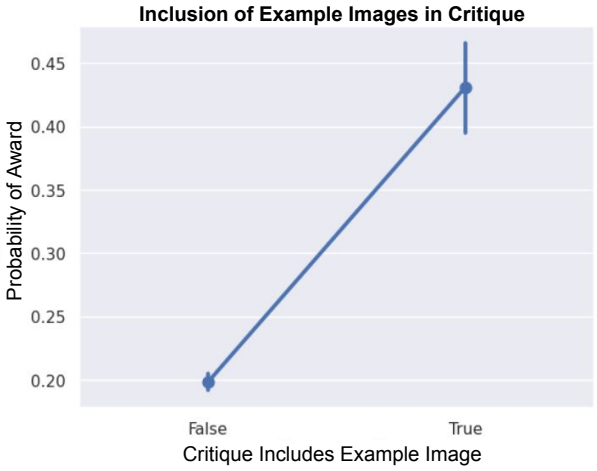
Every photographer who submits an image for critique on /r/photocritique submits an accompanying request for critique, a few sentences describing their intent for the image and asking feedback on specific aspects of the work. A natural question to ask if directly addressing the topics requested by the photographer improves



(a) 70% of critiques do not address any topics not requested by the photographer (top), yet critiques that *do* go beyond the requested topic are much more likely to be awarded a CritiquePoint (bottom).



(b) Longer critiques are much more likely to receive CritiquePoints. The majority of critiques over 600 characters receive CritiquePoints.



(c) Critiques that include images as examples or suggestions are twice as likely to receive awards.

Figure 4.5: Results from our analysis of how CritiquePoints are awarded. We find that critiques that address more than what the photographer requested (a), are longer (b), and include images as examples (c) are all more likely to receive CritiquePoints.

the quality of the critique. Intuitively, it seems as though good critiques should provide feedback on all the topics requested by the photographer, but what about addressing topics that the photographer did not ask about. Here, there are compelling arguments on both sides: on the one hand, providing feedback on something that was not asked about could feel patronizing or overbearing, yet on the other hand, a critique that goes beyond what was requested could help the photographer consider new perspectives or ideas that hadn't occurred to them.

To measure alignment between the request for critique and the critique itself, we use a vocabulary-based approach. Two authors, both experts in photography, iteratively developed a vocabulary of 241 photography terms that are each associated with exactly one of four photography topics: Editing, Composition, Camera Settings, and Emotion. We identify the request for critique associated with each submitted image, and compute the set of topics for which feedback is being requested, where topic is considered to be requested if the request for critique text uses at least one vocabulary word associated with that topic. We use the same procedure to identify which topics are being addressed by each received critique, and then compute the number of topics addressed beyond those requested. A negative value indicates that the critique left some requested topics unaddressed, a value of zero indicates that the critique addressed exactly the requested topics, and a positive value indicates that the critique went beyond what was requested.

We find that the most common critique (approximately 45% of all critiques) addresses exactly the topics requested by the photographer: no more, no less (Figure 4.5a, top). Just under 20% of critiques do not address the topic(s) requested by the photographer. Importantly, though, we find that the more a critique goes beyond the requested topics, the more likely it is to receive a CritiquePoint. Compared to critiques that leave topics unaddressed, critiques that address every one of the four topics are between $2\times$ and $3\times$ as likely to receive a CritiquePoint (Figure 4.5a, bottom). However, as these critiques are vanishingly rare, some uncertainty surrounds these results.

4.1.5 Discussion & Implications

In this work, we developed and deployed a point-based system to gamify participation in a large online photography critique community. This system, CritiquePoints, enables community members to award high quality critiques, and enables the precise measurement of what constitutes high quality content in this

community-specific context. During the first 3.5 years of its ongoing deployment, CritiquePoints were awarded nearly 10,000 times. We used the data collected with CritiquePoints to assess the attributes of successful feedback on creative tasks. We found that critiques which are specific, contain examples, and not only answer the questions posed by the photographer, but go beyond, are considered to be the most helpful. These insights drive our development of an LLM-powered coaching system, which we deploy in the same community. This work is described in the next section.

4.2 LLM-Powered Coaching to Improve Discussion Quality

4.2.1 Introduction & Related Work

Online communities are reliant upon their membership to contribute high quality content, yet learning to participate effectively in online communities can be quite challenging, especially for new members and in communities focused on technical topics or with specific participation requirements [150]. Scalable teaching of participation skills to members is essential to the long-term health of every community, enabling the community to grow sustainably and improve the experiences of its members.

However, teaching members how to participate in communities is very challenging, and, as each community is different, such instruction must be tailored to the community in question. Existing strategies like mentorship can be effective, but are very labor intensive and therefore difficult to scale [150, 141, 167, 129]. Other communities simply let community members fail in their attempts to participate until they eventually learn on their own, but this strategy is discouraging to newcomers [167, 99], and necessarily involve injecting low quality content into the community [141]. Some platforms have introduced automated affordances to increase the visibility of community norms, such as Reddit's recently-launched Post Guidance feature [229], yet these systems require a great deal of manual configuration on the part of the moderators to be effective, and are quite limited in their ability to instruct community members.

In this work, we develop a new, LLM-powered tool to coach community members how to participate effectively. Our coaching tool, a web browser extension that modifies the comment authoring process on Reddit, is informed by an empirically derived, community-specific assessment of effective participation strategies. Through a partnership with the moderators of /r/photocritique, a large community with 1.7 million members, we deploy our tool in a realistic setting, teaching an important and generalizable skill: how to deliver effective feedback on creative tasks.

We evaluate two different LLM-powered pedagogical strategies, a Socratic Coach and an Assistant, against a third non-LLM Control interface (§4.2.2). The Coach and Assistant both use language models to provide feedback on draft comments. The Coach encourages users to think more deeply about how to offer the most useful feedback possible, suggesting strategies and topics to consider, but does not suggest specific changes. The Assistant, in contrast, provides specific suggestions for changes that users can click to accept

or reject, along with a rationale for the suggested changes. The Control shows static, generic tips for writing feedback that are not customized to the user's writing.

These interfaces enable us to evaluate the differences between different uses of LLMs. While the ability of LLMs to assist with writing tasks by making suggestions or even writing substantial amounts of text themselves has been relatively well-demonstrated, much less is known about the utility of LLMs to instead challenge their users to put *more* effort into their writing, rather than less. As part of an ongoing deployment of this system, we recruited five participants from the /r/photocritique community to try out all three interfaces, in a random order.

In semi-structured qualitative pilot with five participants, we find that both participants strongly prefer the Coach and Assistant over the Control, and believe that while the Assistant interface is easier and faster to use than the Coach, the Coach is more effective than the Assistant at teaching participants writing skills and encouraging them to think deeply about their work. These results suggest that LLMs are able to teach skills by challenging their users, and raises concerns about the of writing assistants on skill retention.

4.2.2 Description of Coaching Interventions

Our study design includes three different interfaces: (1) a Socratic 'Coach' that suggests strategies for participants to improve their comments and nudges participants to do that work themselves, (2) an 'Assistant' that makes specific suggestions that can be adopted the click of a button, and (3) a control that simply shows a selection of generic tips without any LLM involvement. All three of these interfaces are implemented in a single Google Chrome browser extension available on the Chrome App Store for easy installation by participants. Once a participant installs the extension, which only functions on /r/photocritique, they are shown an enrollment popup which asks for informed consent. Upon clicking clicking 'accept,' the participant is automatically randomized into one of the three interfaces, and the page is automatically reloaded, triggering the display of the assigned intervention. Logging and LLM requests are handled via POST requests by a simple webserver, hosted by the University of Washington, which uses a local database to log information about participant's usage of the system, and which wraps client requests in a prompt before sending messages to the OpenAI API, and relaying the response back to the client. In addition to the draft critique and messages from the participant, every LLM prompt also includes the critique request from the photographer,

and the image being critiqued, using GPT-5's multimodal capabilities.

Both the Coach and Assistant interventions use a common starting interface before diverging to their respective interfaces. This starting interface (Figure 4.6a) serves two purposes: first, it shows the critique author a summary of the critique request made by the photographer, in order to inform the critique, and second, it requires the author to write at least 50 characters before permitting them to interact with the Coach or Assistant. A key design goal for all interventions was to maintain the participant's voice and control over the process, and this 50 character requirement insures that the participants do some writing before receiving LLM assistance, as well as providing the LLM with some of the participant's writing to provide suggestions. After the participant has written 50 characters, a previously greyed-out 'Get Suggestions' button is enabled. When the participant clicks this button, they receive suggestions from their selected interface.

Socratic Coach Interface

The Coach is designed to encourage participants to think more deeply about their critique, and requires them to edit their critiques themselves, rather than making specific suggestions. The Coach uses a chat-style interface to tell participants what they might consider including or changing about their critique (Figure 4.6b). Participants can chat back-and-forth with the coach to ask questions for for elaboration, and the Coach can see the changes they make to their draft critique.

Assistant Interface

The Assistant, in contrast to the Coach, is happy to edit participants' critiques for them. It makes specific suggestions that can easily be accepted or rejected with a single click. These suggestions are displayed using a track-changes style diff interface (Figure 4.6c), and each suggestion includes a rationale for why it may be a good idea, viewable as a tooltip. The Assistant generates at most three suggestions at a time, and after editing their draft, participants can click a button to see new suggestions.

Static Guidance Control Interface

The Control interface simply displays generic writing tips, informed by our assessment of high quality critiques (§4.1). These tips are specific to the /r/photocritique community, but are generic in the sense that

(a) This starting interface is used by both Coach and Assistant. It shows a single-sentence summary of the photographer's (OP) request, with a clickable button to see the full text. After the participant has written at least 50 characters, the 'Get Suggestions' button changes is enabled.

(b) The Coach uses a chat style interface to initialize the conversation after the 'Get Suggestions' button is clicked. The initial message 'What do you think of my comment draft' is included in all conversations to hint to the participant that they can chat with the Coach.

(c) The Assistant makes specific suggestions that can be accepted or rejected with a single click. The blue ⓘ symbol, when hovered over, displays a tooltip with a brief explanation of the rationale for the suggestion. The ◯ button at upper right generates new suggestions. Up to three suggestions are generated at a time.

(d) The static control interface simply shows generic writing tips. The 'New Tip' button cycles through a series of tips, and the next tip is automatically displayed with each new page load.

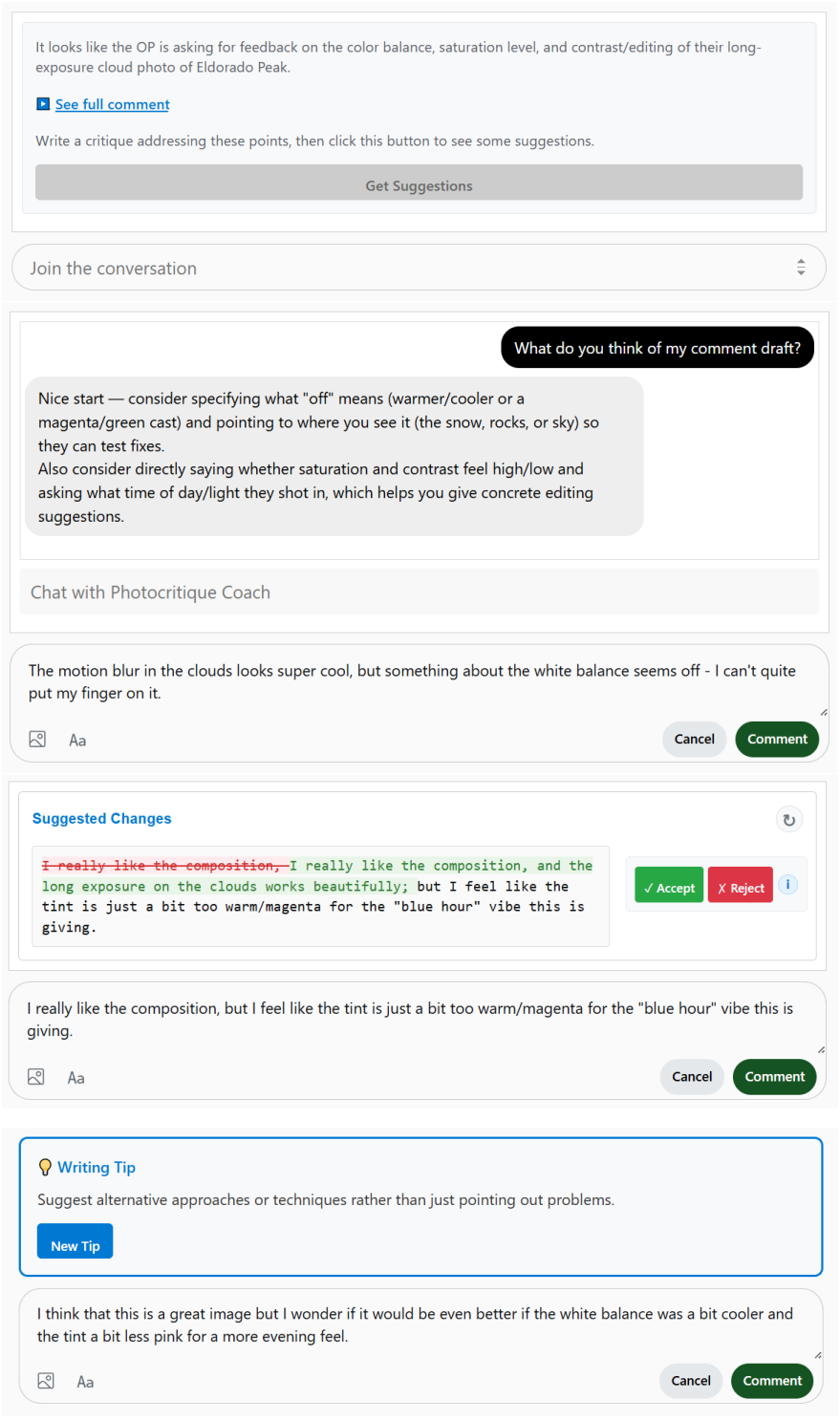


Figure 4.6: Screenshotted examples of (a) the common starting interface for Coach and Assistant, (b) the Coach, (c) the Assistant, and (d) the control interface.

they are not customized to the image being critiqued, or the critique draft. The interface provides a 'New Tip' button which cycles through a series of prewritten tips (Figure 4.6d). In addition, the tip is automatically advanced every time a new page is loaded, so that the participant will see new tips even without clicking the button.

4.2.3 Evaluation Method

To evaluate the efficacy of the different interventions and learn which participants prefer, we conducted a qualitative study, consisting of 30 minute semi-structured interviews with five participants. Participants were able to select between all three interfaces, and were randomly assigned an order in which to use all three and make explicit comparisons between them. One researcher conducted the interviews, starting first by asking about the participants' background experience in the community, next asking them to make explicit comparisons between the interfaces and discuss which they preferred, then finishing by discussing each interface separately.

Participant Recruiting, Incentives, and Background

Participants were recruited from a stickied post in the /r/photocritique subreddit and were each compensated with $20 Amazon gift cards. Each participant completed a screening form to ensure that they were able to use Reddit from Chrome and therefore able to install the extension. The screening form also collected demographic information and asked participants to self-assess their critiquing skill level on a 5-point Likert scale. Written instructions describing how to install and use the extension were distributed to participants, who then used the extension for several days, trying out all three interventions before participating in the semi-structured interview.

There was a broad range in critique skill and community experience across the participants, although none of them rated themselves as complete novices. While two participants had previously posted their own photographs for critique, most participants had not, and all participants primarily participated in the community to give feedback on others' photos. The length of time that participants had been members of the community ranged from several years to just over a month. Every participant wrote and submitted at least four comments with the extension before their interview; one participant wrote over 40. Participants

have continued to use the extension since interviews were completed, they have collectively authored 517 comments using the extension as of this writing.

4.2.4 Results

All five participants agreed that the Coach and Assistant interfaces were far more helpful than the Control interface, and spoke generally highly of the two LLM-infused treatments. P1 said '[the system as a whole] is a triumph! I really enjoyed using it, particularly the Coach.' P6 also stated that he found the treatments helpful, although he preferred the Assistant due to its ease of use.

Increased Friction Contributes to Better Learning Outcomes

Participants said that they found using the Coach taught them more about critiquing, but also was slower and required them to exert more effort. P10 said that the coach encouraged her to make substantial rewrites to her comments, and having to do the work herself was slow but she learned from the rationales provided by the Coach. P1 described being impressed by how the Coach encouraged him to begin his critiques with a complement, which P1 said is not his 'usual style.' Regarding these suggestions, P1 said 'I was surprised how much better [my critique] was. It was less offensive, nicely put, patted them on the behind and made them feel good.' However, P1 added that he did not always take the advice provided by the Coach: 'I ignored some of the suggestions because they weren't me. I am not a big believer in the Oreo [starting and ending with a complement], I think it's possible to be kind and gentle and not blow smoke up someone's shirt.' P5 concurred, saying that he disliked the positivity. P6 described how the Coach encouraged him to try translating his comment into French, a language he does not speak, when writing a critique for a French-speaking photographer. P6 used Google Translate to translate his critique, which he wrote in English, into French before submitting. P1 stated that he preferred the Coach, as it let him maintain more control over what he wrote, compared to the Assistant. P1 said '[using the Coach] was very close to getting personal feedback [on my critique]. I was surprised by how human like that was. It was almost frightening.' P10 said that the 'Coach was the most educational, because it gave me ideas about what strategies I should be including in my comments.'

Easy-to-accept Suggestions Reduce Cognitive Load

Participants agreed that the Assistant is faster, easier to use, and more straightforward than the Coach, which P6 preferred, as it didn't require doing as much writing himself. P6 mentioned being frustrated with the 50 character minimum draft length before seeing suggestions, as P6 found it burdensome to write this initial comment. P6 said they made extensive use of the ability to regenerate new suggestions, and enjoyed being able to choose between the suggestions he was presented with, saying '[The suggestions generated by the Assistant] were helpful because it would explain why. It added detail which is very authentic and looks genuine, and is more specific than "this picture looks amazing."' P6 felt that the Assistant was particularly helpful at tailoring critiques to the needs of the photographer: 'You need to phrase things using the same language as the person who is asking for feedback. Things such as camera angle, camera focal length, *etc.* You have to do a countercheck of the message before you comment.' P8 appreciated how the Assistant saved her time, saying 'I had a sentence but I didn't know how to explain it clearly, and the Assistant helps me explain it. 80% of time time I accept its suggestions.'

Author Agency and Edit Scope

Participants disliked when they perceived that the tools were changing the tone of their comments or putting words in their mouths. P5 said of the Coach that one time it 'made its own opinion and tried to push me into it.' Even if the suggested changes were accepted, participants still expressed discomfort with substantial edits, particularly those suggested by the Assistant. P1 said that the Assistant 'changed my comment very strongly, quite a bit' although he said he was happy with the edits that the Assistant suggested. P8 said that sometimes the Assistant made suggestions that 'lost the meaning that I intended, but that's when I can just reject the suggestion.'

4.2.5 Discussion & Implications

In this project, we developed and deployed an LLM-powered coaching tool to teach members of the /r/photocritique community to give better feedback to one another. Results from semi-structured qualitative interviews with five participants, as part of an ongoing large scale deployment, suggest that LLM-assisted coaching is both far more effective at teaching the skill of providing feedback, as well as far more effective

at increasing the quality of the feedback that is delivered, than a non-LLM control interface.

We evaluated two different pedagogical strategies: an Assistant, which provides easily-to-accept writing suggestions that *decrease* the amount effort required to write a comment, and a Coach, which uses a Socratic strategy to nudge authors to think more deeply and revise their comments without making actionable suggestions, *increasing* the amount of effort required. We found that the Coach interface, with its educational yet friction-increasing nudges, required more cognitive effort and was slower to use, yet participants reported that it taught them how to write high quality comments more effectively than the Assistant. Despite this, many participants preferred the Assistant, as it was faster to use and required less effort.

This work has important implications for the use of LLMs to teach humans new skills over the Internet. While many LLM applications are considered tools that their users have unrestricted access to, our results suggest that greater automation can be detrimental to the human's long-term ability to perform the task unassisted. We show that LLMs can instead be used to teach humans how to do the skills themselves, by encouraging the human to do additional reflection and iteration. This result prompts an important discussion over the role of LLMs in online communities.

4.3 Summary of Contributions to Thesis

In this chapter, we worked to deploy interventions and data-driven best practices to maximize the impact of this dissertation research. We deployed CritiquePoints, a system to gamify the delivery of high quality feedback in a large photography community on Reddit. In the first 3.5 years of an ongoing deployment, CritiquePoints have been awarded almost 10,000 times, enabling a valuable opportunity to study what makes for high quality content in a community specific fashion. We found that high quality feedback is polite and actionable, and addresses not only the questions posed by the critiquee but also goes beyond what was requested.

Next, we used these insights to design an LLM-powered coaching tool, integrated into a web-browser extension, which we deployed with members of a large photography community. We conducted semi-structured interviews with participants to evaluate two pedagogical strategies. We found that both strategies are effective at improving the quality of feedback delivered in the community, and a Socratic coaching strategy requires users to do more work, but results in more learning.

Chapter 5

Discussion and Conclusion

5.1 Summary of Thesis Contributions

In this dissertation research, I made online communities better through three research activities. In Chapter 2, I *characterized* what community members value for their own communities, and how these values vary between and within communities. In Chapter 3, I built upon this characterization by *assessing* the state of current governance practices and outcomes, in order to identify those that seem most promising. In Chapter 4, I partnered with a large community to *deploy* new tools and practices, and I disseminated my results from the previous chapters to the Reddit Moderator Council, to put findings into practice and maximize the impact of my research.

This dissertation includes several important contributions to the field of computational social science, content moderation, and online community governance.

1. I developed the first taxonomy characterizing the breadth of community members values for their own communities in their own words. This taxonomy consists of 9 broad values and 29 subcategories.

2. I conducted the largest-to-date survey of community members' values on Reddit, measuring how their values vary between and within communities, and I make anonymized responses public to support further research.

3. I developed a new method to measure community members' perceptions of their moderators at a massive scale, one of the first methods to scalably measure governance outcomes. I make the underlying

classification pipeline public, along with a dataset of over 2 million posts and comments discussing moderators.

4. I deployed two new systems to improve the quality of participation in a large online photography discussion community.

Together, these contributions demonstrate my thesis statement: Informed by online communities' individual values and needs, large scale analyses of online communities can leverage existing variance in governance and outcomes to improve communities through data-driven best practices, the development of novel tools, and real world collaborations.

5.2 Future Directions

While I am very proud of the work presented in this dissertation, there remains much more to do to improve online community governance. I am particularly excited about leveraging the new capabilities of large language models, which did not exist when I began my dissertation research, to make online communities healthier, and to build new tools that empower community members and moderators alike to make their own communities successful.

5.2.1 LLM Coaching to Improve Conversational Outcomes

My work with /r/photocritique demonstrates the significant potential of browser-based coaching systems to improve online community interactions (§4.2.2). The success of our easily installed browser extension in teaching community members how to write more effective comments suggests that similar interventions could have substantial impact when expanded to address broader communication challenges across diverse online spaces. This foundation points toward a particularly pressing application: helping people engage in productive disagreement online. In our increasingly polarized society, where much public discourse occurs on digital platforms, the ability to disagree constructively has become both more critical and more challenging [152, 191, 254]. Online communities often struggle with debates that devolve into unproductive arguments, personal attacks, or echo chambers that reinforce existing divisions [152, 41]. Yet as our society also becomes more geographically polarized, online communities have enormous potential to serve as

venues for meaningful dialogue across difference, if participants can be equipped with the right communication skills.

I propose developing an expanded coaching system designed to help community members disagree with one another in productive ways. This system would teach users to apply principles of effective communication during contentious discussions, including perspective-taking techniques that encourage consideration of others' viewpoints, strategies for identifying and utilizing shared moral frameworks, and methods for maintaining politeness and respect even when discussing deeply held beliefs. Large language models are ideally suited to quickly provide useful and personalized feedback, and by providing real-time guidance during actual conversations, such a system could help transform the quality of discourse in online communities. To evaluate this intervention's effectiveness, I would deploy the coaching system at scale in many diverse communities and conduct a randomized controlled trial. This study would robustly measure the system's impact on participants' ability to engage in meaningful conversations on challenging topics where disagreement is likely, and hopefully improve the quality of democratic and political discourse in online communities in a practical and scalable manner.

5.2.2 Expanding Community-Specific Participation Incentives

In the 3.5 years that the CritiquePoint system (§4.1) has been deployed, the moderators of several other communities have reached out, asking if it would be possible to deploy a similar system in their own communities. As we already demonstrated, systems such as CritiquePoints provide a valuable tool for communities to incentivize high quality participation, but also to scalably measure in quasi mealtime how well the community is functioning. Yet such systems are rarely used in practice, with only one other system currently deployed on Reddit, in /r/ChangeMyView [254, 246]. It is quite challenging for moderators and other community stakeholders to measure community-specific metrics, and the double-function of CritiquePoints (incentivizing quality participation while also measuring that quality) makes it a natural solution to this problem.

I propose developing an easily-to-adopt generalization of CritiquePoints that could be adopted in any community by moderators without any technical or coding skills. This system would have a self-service deployment process, so moderators would not need to wait for assistance to launch it in their community.

Simply granting moderator permissions to a universal version of /u/CritiquePointBot would be all that is necessary to launch the system in a new community. I envision a web-accessible settings page that moderators could use to customize their deployment, for example by change the name used for points, and the macro used to award them, as well as the ability to enable, disable, and customize automatic features such as real-time leaderboards, custom flair for awardees. Such a system could even-come with recommended prewritten announcement threads and documentation to make launching points in a new community as easy as possible.

I believe that this system would be an impactful research project on three different fronts. First, it would help incentivize and improve the quality of participation in a broad range of communities. Second, it would help moderators and other stakeholders to collect detailed information about the quality of participation in their communities,. Third, it would help establish connections between myself and the moderators of many communities. These connections could lead to future research collaborations and provide opportunities for additional data collection and to deploy future systems.

5.2.3 Community Dashboards to Enable Community Driven Experimentation

Online community governance is challenging for moderators to implement effectively, and moderators lack tools to measure the impact that their governance practices have on their communities. While the data-driven best practices I developed in my dissertation research are a valuable starting point for moderators (Chapter 3), it would be even more valuable if communities could do their own rigorous experimentation to understand what works best for their community's specific needs. While some simple systems for self-experimentation have already been deployed [182], these systems are rarely used, and only include a very narrow set of outcome data.

I propose the development of a 'community dashboard' that collects community outcome data from multiple streams (automated measures such as in §3.1, lightweight and repeated surveys of community members, CritiquePoint-style systems, and meta discussion threads) and presents it to moderators and community members in a convenient interface. Such a system could increase transparency and trust between moderators and non-moderator community members, would allow moderators to understand how community outcomes are trending, and enable A/B testing of different moderation strategies.

This is an ambitious research project that would likely consist of two phases: development and deployment. Development could begin with a simple needs-finding survey of a small number of moderators recruited through the Reddit Mod Council. The dashboard could publicly present present and past data to moderators and community members, incorporating data from a number of sources. The dashboard could leverage our moderator perceptions pipeline (§3.1) to identify comments which discuss the moderators, and label their sentiment. The dashboard could even link to or display the full text of these comments, although this raises some manageable privacy concerns. In addition, the dashboard could display aggregated responses from lightweight surveys (via Reddit Chat) of a random sample of community members conducted on a weekly basis, developed similar to the automatic voting and jury mechanisms in PolicyKit [291]. The dashboard could also include metrics such as the number of active community members, new members, retention rate, and amount of removed content. The dashboard could also potentially leverage a system similar to Pol.is,[1] a system which assists in crowd sourcing and distilling the thoughts of large groups, to provide additional insights into the community.

Deployment of the dashboard would ideally occur with a diverse set of partner communities. Working with the moderators, we could capture detailed quantitative data before and during the deployment, including moderator actions, users' contributions, and rule changes. We could interview moderators before, during, and after the deployment to understand how the system impacted their moderation strategy, and also hear perspectives from non-moderator community members. If we are able to recruit a large enough set of communities to make this feasible, we will also gather data from a randomly selected held out set of control communities, to provide a baseline to compare our intervention against in a RCT.

[1] `https://pol.is/home`

Bibliography

[1] Dagmar Abfalter, Melanie E Zaglia, and Julia Mueller. “Sense of virtual community: A follow up on its measurement”. In: *Computers in Human Behavior* 28.2 (2012), pp. 400–404.

[2] Roger S Ahlbrandt and James V Cunningham. *A new public policy for neighborhood preservation*. Praeger Publishers, 1979.

[3] Jennifer Allen et al. “Evaluating the fake news problem at the scale of the information ecosystem”. In: *Science Advances* 6.14 (2020).

[4] K. Allison and K. Bussey. “Communal Quirks and Circlejerks: A Taxonomy of Processes Contributing to Insularity in Online Communities”. In: *ICWSM* (2020).

[5] Hind Almerekhi, Haewoon Kwak, and Bernard Jim Jansen. “Statistical Modeling of Harassment against Reddit Moderators”. In: *WWW Companion* (2020).

[6] Gabriel A. Almond and Sidney Verba. *The civic culture: Political attitudes and democracy in Five nations*. Sage Publications, 1989.

[7] Tim Althoff, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. “How to ask for a favor: A case study on the success of altruistic requests”. In: *ICWSM*. Vol. 8. 1. 2014, pp. 12–21.

[8] Tim Althoff et al. “Quantifying dose response relationships between physical activity and health using propensity scores”. In: *NIPS ML4H* (2016).

[9] Tawfiq Ammari, Sarita Schoenebeck, and Daniel Romero. “Self-declared throwaway accounts on Reddit: How platform affordances and shared norms enable parenting disclosure and support”. In: *CSCW* (2019).

[10] Aris Anagnostopoulos et al. “Viral Misinformation: The Role of Homophily and Polarization”. In: *WWW Companion* (2015).

[11] Pablo Aragón, Vicenç Gómez, and Andreaks Kaltenbrunner. “To thread or not to thread: The impact of conversation threading on online discussion”. In: *ICWSM*. Vol. 11. 1. 2017, pp. 12–21.

[12] Ashwini Ashokkumar and James W Pennebaker. “Tracking group identity through natural language within groups”. In: *PNAS nexus* 1.2 (2022).

[13] Peter C. Austin and Elizabeth A. Stuart. “Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies”. In: *Statistics in Medicine* 34 (2015), pp. 3661–3679. URL: https://api.semanticscholar.org/CorpusID:14478957.

[14] Kenneth M Bachrach and Alex J Zautra. “Coping with a community stressor: The threat of a hazardous waste facility”. In: *Journal of health and social behavior* (1985), pp. 127–141.

[15] Lars Backstrom et al. “Characterizing and curating conversation threads: expansion, focus, volume, re-entry”. In: *WSDM*. 2013, pp. 13–22.

[16] James Baglin. “Improving your exploratory factor analysis for ordinal data: A demonstration using FACTOR”. In: *Practical Assessment, Research, and Evaluation* 19.1 (2014), p. 5.

[17] Ramy Baly et al. “We Can Detect Your Bias: Predicting the Political Ideology of News Articles”. In: *EMNLP*. ACL, 2020, pp. 4982–4991. URL: https://www.aclweb.org/anthology/2020.emnlp-main.404.

[18] Jiajun Bao et al. “Conversations Gone Alright: Quantifying and Predicting Prosocial Outcomes in Online Conversations”. In: *TheWebConf* (2021).

[19] Michael Barthiel et al. “Reddit news users more likely to be male, young and digital in their news preferences”. In: *Pew Research Center* (Feb. 2016). URL: https://www.pewresearch.org/journalism/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/.

[20] Jason Baumgartner et al. “The Pushshift Reddit Dataset”. In: *arXiv:2001.08435* (2020).

[21] Rashi Bhati et al. “Sustainable Computing for Future-Ready Engagement: Integrating Green Technologies and Gamification in Digital Ecosystems”. In: *International Conference on Education Technology Management* (2025).

[22] Anita Blanchard. “Blogs as virtual communities: Identifying a sense of community in the Julie/Julia project”. In: (2004).

[23] Anita L Blanchard. “Developing a sense of virtual community measure”. In: *CyberPsychology & Behavior* 10.6 (2007), pp. 827–830.

[24] Anita L Blanchard. “Testing a model of sense of virtual community”. In: *Computers in Human Behavior* 24.5 (2008), pp. 2107–2123.

[25] Anita L Blanchard and M Lynne Markus. “The experienced “sense”’ of a virtual community: Characteristics and processes”. In: *ACM Sigmis Database* (2004).

[26] Alexandre Bovet and Hernán A. Makse. “Influence of fake news in Twitter during the 2016 US presidential election”. In: *Nature Communications* (2019).

[27] Ryan L Boyd et al. “The development and psychometric properties of LIWC-22”. In: *Austin, TX: University of Texas at Austin* 10 (2022).

[28] Lia Bozarth, Aparajita Saraf, and C. Budak. “Higher Ground? How Groundtruth Labeling Impacts Our Understanding of Fake News about the 2016 U.S. Presidential Nominees”. In: *ICWSM*. 2020.

[29] Cody Buntain and Jennifer Golbeck. “Identifying social roles in reddit using network structure”. In: *WWW*. 2014, pp. 615–620.

[30] Sloane Burke Winkelman et al. “Exploring Cyber Harassment among Women Who Use Social Media”. In: *Univ. J. of Public Health* (2015).

[31] Susan M Burroughs and Lillian T Eby. “Psychological sense of community at work: A measurement system and explanatory framework”. In: *Journal of community psychology* 26.6 (1998), pp. 509–532.

[32] Brantly Callaway and Pedro H.C. Sant’Anna. “Difference-in-Differences with multiple time periods”. In: *Journal of Econometrics* 225.2 (2021), pp. 200–230.

[33] Justine Cassell and Dona Tversky. “The language of online intercultural community formation”. In: *JCMC* 10.2 (2005).

[34] Laura Ceci. *Reddit: quarterly number of DAU 2021-2025, by online status*. https://www.statista.com/statistics/1453133/reddit-quarterly-dau-by-online-status/. Accessed: 2025-09-15. May 2025.

[35] E. Chandrasekharan et al. “Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators”. In: *CHI* 3 (2019), pp. 1–30.

[36] Eshwar Chandrasekharan et al. “Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit”. In: *TOCHI* 29 (2020), pp. 1–26. URL: https://api.semanticscholar.org/CorpusID:221879027.

[37] Eshwar Chandrasekharan et al. “The Internet’s Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales”. In: *CSCW* (2018).

[38] Eshwar Chandrasekharan et al. “You Can’t Stay Here: The Efficacy of Reddit’s 2015 Ban Examined Through Hate Speech”. In: *CSCW* (2017).

[39] Janghee Cho and Rick Wash. “How Potential New Members Approach an Online Community”. In: *CSCW* (2021).

[40] Daejin Choi et al. “Characterizing conversation patterns in reddit: From the perspectives of content properties and user participation behaviors”. In: *ACM COSN*. 2015.

[41] Daejin Choi et al. “Rumor propagation is amplified by echo chambers in social media”. In: *Scientific Reports* 10.1 (2020), pp. 1–10.

[42] Minje Choi et al. “Ten Social Dimensions of Conversations and Relationships”. In: *WWW* (2020).

[43] G. Ciampaglia and D. Taraborelli. “MoodBar: Increasing New User Retention in Wikipedia through Lightweight Socialization”. In: *CSCW* (2015).

[44] Lorenzo Cima et al. “The Great Ban: Efficacy and Unintended Consequences of a Massive Deplatforming Operation on Reddit”. In: *Companion Publication of the 16th ACM Web Science Conference* (2024). URL: https://api.semanticscholar.org/CorpusID:267069411.

[45] Mauro Coletto et al. “Automatic controversy detection in social media: A content-independent motif-based approach”. In: *Online Soc. Networks Media* (2017). URL: https://api.semanticscholar.org/CorpusID:54300115.

[46] Shaun E Cowman, Joseph R Ferrari, and Matthew Liao-Troth. “Mediating effects of social support on firefighters’ sense of community and perceptions of care”. In: *Journal of Community Psychology* (2004).

[47] Jonathon N Cummings, Lee Sproull, and Sara B Kiesler. “Beyond hearing: Where the real-world and online support meet.” In: *Group Dynamics: Theory, Research, and Practice* 6.1 (2002), p. 78.

[48] Cristian Danescu-Niculescu-Mizil et al. “No country for old members: user lifecycle and linguistic change in online communities”. In: *WWW* (2013).

[49] Kareem Darwish, Walid Magdy, and Tahar Zanouda. “Trump vs. Hillary: What Went Viral During the 2016 US Presidential Election”. In: *SocInfo*. 2017.

[50] Srayan Datta and Eytan Adar. “Extracting Inter-community Conflicts in Reddit”. In: *ICWSM*. 2019.

[51] Corinne David-Ferdon et al. “A comprehensive technical package for the prevention of youth violence and associated risk behaviors”. In: (2016).

[52] Jenny L Davis and Timothy Graham. “Emotional consequences and attention rewards: The social effects of ratings on Reddit”. In: *Information, Communication & Society* 24.5 (2021), pp. 649–666.

[53] Munmun De Choudhury and Sushovan De. “Mental health discourse on reddit: Self-disclosure, social support, and anonymity”. In: *ICWSM*. 2014.

[54] Dorottya Demszky et al. “Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings”. In: *NAACL-HLT*. 2019.

[55] Sebastian Deri et al. “Coloring in the Links: Capturing Social Ties as They are Perceived”. In: *CSCW* (2018).

[56] Tim Dettmers et al. “QLoRA: Efficient Finetuning of Quantized LLMs”. In: *ArXiv* abs/2305.14314 (2023).

[57] Yoan Dinkov et al. “Predicting the Leading Political Ideology of YouTube Channels Using Acoustic, Textual, and Metadata Information”. In: *INTERSPEECH*. 2019.

[58] Stacy Jo Dixon. *Daily time spent on social networking by internet users worldwide from 2012 to 2025*. 2025. URL: https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/.

[59] Stacy Jo Dixon. *Number of social media users worldwide from 2017 to 2027*. 2023. URL: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

[60] MeiXing Dong et al. “We Are in This Together: Quantifying Community Subjective Wellbeing and Resilience”. In: *ICWSM*. Vol. 17. 2023, pp. 185–196.

[61] Bryan Dosono and Bryan C. Semaan. “Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit”. In: *CHI* (2019).

[62] David L. DuBois et al. “How Effective Are Mentoring Programs for Youth? A Systematic Assessment of the Evidence”. In: *Psychological Science in the Public Interest* 12.2 (2011), pp. 57–91. DOI: 10.1177/1529100611414806.

[63] T. Erickson and W. Kellogg. “Social translucence: an approach to designing systems that support social processes”. In: *TOCHI* (2000).

[64] Facebook. *How Facebook uses artificial intelligence to moderate content*. 2025. URL: https://www.facebook.com/help/1584908458516247/?helpref=uf_share.

[65] Anna Fang, Wenjie Yang, and Haiyi Zhu. *Shaping Online Dialogue: Examining How Community Rules Affect Discussion Structures on Reddit*. 2025.

[66] Susan J Farrell, Tim Aubry, and Daniel Coulombe. “Neighborhoods and neighbors: Do they contribute to personal well-being?” In: *Journal of community psychology* 32.1 (2004), pp. 9–25.

[67] Michelle Faverio and Olivia Sidoti. “Teens, Social Media and Technology 2024”. In: *Pew Research Center* (2024). https://www.pewresearch.org/internet/2024/12/12/teens-social-media-and-technology-2024/.

[68] Emilio Ferrara. “Contagion dynamics of extremist propaganda in social networks”. In: *Information Sciences* 418 (2017), pp. 1–12.

[69] Joseph R Ferrari, Theresa Luhrs, and Victoria Lyman. “Eldercare volunteers and employees: Predicting caregiver experiences from service motives and sense of community”. In: *The Journal of Primary Prevention* (2007).

[70] Casey Fiesler and Brianna Dym. “Moving Across Lands: Online Platform Migration in Fandom Communities”. In: *CSCW* (2020).

[71] Casey Fiesler and Nicholas Proferes. ““Participant” Perceptions of Twitter Research Ethics”. In: *Social Media + Society* 4 (2018).

[72] Casey Fiesler et al. “Reddit Rules! Characterizing an Ecosystem of Governance”. In: *ICWSM* (2018).

[73] Ute Fischer, Lori McDonnell, and Judith Orasanu. “Linguistic correlates of team performance: Toward a tool for monitoring team functioning during space missions”. In: *Aviation, space, and environmental medicine* 78.5 (2007), B86–B95.

[74] Danyel Fisher, Marc Smith, and Howard T Welser. “You are who you talk to: Detecting roles in usenet newsgroups”. In: *HICSS’06*. IEEE. 2006.

[75] Sofie Flensted. “Exploring the Connection Between Newspaper Blogs and Sense of Community”. PhD thesis. 2011.

[76] Seth Frey and Nathan Schneider. *Effective Voice: Beyond Exit and Affect in Online Communities*. 2021.

[77] Seth Frey and Robert W. Sumner. “Emergence of integrated institutions in a large population of self-governing communities”. In: *PLOS ONE* 14.7 (July 2019), pp. 1–18. DOI: 10.1371/journal.pone.0216335. URL: https://doi.org/10.1371/journal.pone.0216335.

[78] Seth Frey et al. “Governing Online Goods: Maturity and Formalization in Minecraft, Reddit, and World of Warcraft Communities”. In: *CSCW* (2022). URL: https://api.semanticscholar.org/CorpusID:246485431.

[79] Batya Friedman, Peter H. Kahn, and Alan Borning. “Value Sensitive Design and Information Systems”. In: *Human-Computer Interaction and Management Information Systems* (2006).

[80] Soumen Ganguly et al. “Empirical Evaluation of Three Common Assumptions in Building Political Media Bias Datasets”. In: *ICWSM*. 2020.

[81] Abigail Geiger. *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. Sept. 2019.

[82] US Surgeon General. *Social Media and Youth Mental Health: The US Surgeon General's Advisory*. 2023.

[83] Jennifer L Gibbs, Heewon Kim, and Seol Ki. “Investigating the role of control and support mechanisms in members' sense of virtual community”. In: *Communication Research* 46.1 (2019), pp. 117–145.

[84] Eric Gilbert. “Designing Social Translucence over Social Networks”. In: *CHI* (2012).

[85] Sarah A. Gilbert. “"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”. In: *CSCW* (2020).

[86] Tarleton Gillespie. *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2021.

[87] B. Glaser, A. Strauss, and E. Strutzel. “The Discovery of Grounded Theory: Strategies for Qualitative Research”. In: *Nursing Research* (1968).

[88] M. Glenski, C. Pennycuff, and T. Weninger. “Consumers and Curators: Browsing and Voting Patterns on Reddit”. In: *IEEE TCSS* 4.4 (2017), pp. 196–206. DOI: 10.1109/TCSS.2017.2742242.

[89] Maria Glenski, Tim Weninger, and Svitlana Volkova. “Propagation from deceptive news sources who shares, how much, how evenly, and how quickly?” In: *IEEE TCSS* 5.4 (2018), pp. 1071–1082.

[90] Jeffrey Gottfried. “Americans' Social Media Use”. In: *Pew Research Center* (Jan. 2024). URL: https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/.

[91] Agam Goyal, Charlotte Lambert, and Eshwar Chandrasekharan. “Uncovering the Internet’s Hidden Values: An Empirical Study of Desirable Behavior Using Highly-Upvoted Content on Reddit”. In: *arXiv:2410.13036* (2024).

[92] Przemyslaw A. Grabowicz et al. “Distinguishing Topical and Social Groups Based on Common Identity and Bond Theory”. In: *WSDM ’13*. Rome, Italy, 2013, pp. 627–636.

[93] James Grimmelmann. “The Virtues of Moderation”. In: *Yale Journal of Law and Technology* (2015).

[94] Nir Grinberg et al. “Fake news on Twitter during the 2016 U.S. presidential election”. In: *Science* (2019).

[95] Lani Guinier. *The Tyranny of the Majority : Fundamental Fairness in Representative Democracy*. Free Press, 1994.

[96] Kilem L. Gwet. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Advanced Analytics, 2014.

[97] Hussam Habib et al. “To Act or React: Investigating Proactive Strategies For Online Community Moderation”. In: (2019). arXiv: 1906.11932 [cs.SI].

[98] Jonathan Haidt. *The Anxious Generation*. Penguin, Mar. 2024. ISBN: 9780593655030.

[99] Aaron Halfaker, Aniket Kittur, and John Riedl. “Don’t Bite the Newbies: How Reverts Affect the Quantity and Quality of Wikipedia Work”. In: *WikiSym* (2011).

[100] Aaron Halfaker et al. “The Rise and Decline of an Open Collaboration System: How Wikipedia’s Reaction to Popularity Is Causing Its Decline”. In: *American Behavioral Scientist* (2013).

[101] Aarash Heydari et al. *YouTube Chatter: Understanding Online Comments Discourse on Misinformative and Political YouTube Videos*. 2019. arXiv: 1907.00435 [cs.CY].

[102] Natalie C Hopkins-Best. “Psychological sense of community within mediated communities: the case of the news blog”. 2010.

[103] Sameera Horawalavithana et al. “Online discussion threads as conversation pools: predicting the growth of discussion threads on reddit”. In: *Computational and Mathematical Organization Theory* (2022), pp. 1–29.

[104] B. Horne, Jeppe Nørregaard, and Sibel Adali. “Different Spirals of Sameness: A Study of Content Sharing in Mainstream and Alternative Media”. In: *ICWSM*. 2019.

[105] Philip N. Howard et al. “Social Media, News and Political Information during the US Election: Was Polarizing Content Concentrated in Swing States?” In: *arXiv:1802.03573 [cs]* (Feb. 2018). URL: http://arxiv.org/abs/1802.03573.

[106] N.M. Hurd, M.A. Zimmerman, and Y Xue. “Negative Adult Influences and the Protective Effects of Role Models: A Study with Urban Adolescents”. In: *Journal of Youth and Adolescence* 38.6 (2009), p. 777.

[107] Clayton J. Hutto and Eric Gilbert. “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text”. In: *ICWSM* (2014). URL: https://api.semanticscholar.org/CorpusID:12233345.

[108] Sohyeon Hwang and Jeremy D. Foote. “Why Do People Participate in Small Online Communities?” In: *CSCW* (2021). URL: https://doi.org/10.1145/3479606.

[109] Sohyeon Hwang and Aaron Shaw. “Rules and Rule-Making in the Five Largest Wikipedias”. In: *ICWSM*. 2022. URL: https://api.semanticscholar.org/CorpusID:249892728.

[110] Jane Im et al. “Deliberation and Resolution on Wikipedia: A Case Study of Requests for Comments”. In: *CSCW* (2018).

[111] Mike Isaac and Kate Conger. *Reddit bans forum dedicated to supporting Trump*. https://www.nytimes.com/2021/01/08/us/politics/reddit-bans-forum-dedicated-to-supporting-trump-and-twitter-permanently-suspends-his-allies-who-spread-conspiracy-theories.html. Accessed: 2021-04-09. Jan. 2021.

[112] Abraham Israeli, Shani Cohen, and Oren Tsur. “Unsupervised discovery of non-trivial similarities between online communities”. In: *Expert Systems With Applications* (2022).

[113] Abraham Israeli and Oren Tsur. “With Flying Colors: Predicting Community Success in Large-scale Collaborative Campaigns”. In: *ICWSM*. 2024.

[114] Davy Janssen and Raphaël Kies. “Online forums and deliberative democracy”. In: *Acta política* 40 (2005), pp. 317–335.

[115] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. “Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit”. In: *CSCW* 3 (2019), pp. 1–27. URL: https://api.semanticscholar.org/CorpusID:203558216.

[116] Shagun Jhaver, Seth Frey, and Amy Zhang. *Decentralizing Platform Power: A Design Space of Multi-level Governance in Online Social Platforms*. 2023. arXiv: 2108.12529 [cs.HC].

[117] Shagun Jhaver, Seth Frey, and Amy Zhang. *Decentralizing Platform Power: A Design Space of Multi-level Governance in Online Social Platforms*. 2023. arXiv: 2108.12529 [cs.HC].

[118] Shagun Jhaver, Himanshu Rathi, and Koustuv Saha. *Bystanders of Online Moderation: Examining the Effects of Witnessing Post-Removal Explanations*. 2023. arXiv: 2309.08361 [cs.HC].

[119] Shagun Jhaver, P. Vora, and A. Bruckman. “Designing for Civil Conversations: Lessons Learned from ChangeMyView”. In: *GVU Technical Reports* (2017).

[120] Shagun Jhaver et al. “"Did You Suspect the Post Would be Removed?"”. In: *CSCW* 3 (2019), pp. 1–33.

[121] Shagun Jhaver et al. “Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator”. In: *TOCHI* (2019).

[122] Shagun Jhaver et al. “Online Harassment and Content Moderation: The Case of Blocklists”. In: *TOCHI* (2018).

[123] Jialun Aaron Jiang et al. “Understanding international perceptions of the severity of harmful content online”. In: *PLoS ONE* (2021).

[124] S. Jiang, Ronald E. Robertson, and Christo Wilson. “Bias Misperceived: The Role of Partisanship and Misinformation in YouTube Comment Moderation”. In: *ICWSM*. 2019.

[125] Shan Jiang, Ronald E. Robertson, and Christo Wilson. “Reasoning about Political Bias in Content Moderation”. In: *AAAI*. 2020.

[126] Ridley Jones et al. “R/Science: Challenges and Opportunities in Online Science Communication”. In: *CHI* (2019).

[127] Sanjay R Kairam and Jeremy Foote. “How founder motivations, goals, and actions influence early trajectories of online communities”. In: *CHI*. 2024, pp. 1–11.

[128] Sanjay R Kairam, Melissa C Mercado, and Steven A Sumner. “A social-ecological approach to modeling sense of virtual community (SOVC) in livestreaming communities”. In: *CSCW* 6 (2022), pp. 1–35.

[129] Sanjay Ram Kairam, Dan J Wang, and Jure Leskovec. “The life and death of online groups: Predicting group growth and longevity”. In: *WSDM*. 2012.

[130] Michael T Kalkbrenner. “Alpha, omega, and H internal consistency reliability estimates: Reviewing these options and when to use them”. In: *Counseling Outcome Research and Evaluation* (2023).

[131] Mostafa Kamalpour, J. Watson, and L. Buys. “How Can Online Communities Support Resilience Factors among Older Adults”. In: *Int. J. of Human–Computer Interaction* (2020).

[132] Maya Kavaliova et al. “Crowdsourcing innovation and product development: Gamification as a motivational driver”. In: *Cogent Business & Management* (2016).

[133] S. Kawanaka et al. “Gamified Participatory Sensing in Tourism: An Experimental Study of the Effects on Tourist Behavior and Satisfaction”. In: *Smart Cities* (2020).

[134] Brian Keegan and Casey Fiesler. “The Evolution and Consequences of Peer Producing Wikipedia’s Rules”. In: *ICWSM* (May 2017). DOI: 10.1609/icwsm.v11i1.14899. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14899.

[135] C. Kelty. “Too Much Democracy in All the Wrong Places: Toward a Grammar of Participation”. In: *Current Anthropology* 58 (2017).

[136] Cynthia Kennett and Malcolm Payne. “Understanding why palliative care patients ‘like day care’and ‘getting out’”. In: *Journal of Palliative Care* 21.4 (2005), pp. 292–298.

[137] Osama Khalid and Padmini Srinivasan. “Style matters! Investigating linguistic style in online communities”. In: *ICWSM*. 2020.

[138] M. Kharratzadeh and Deniz Üstebay. “US Presidential Election: What Engaged People on Facebook”. In: *ICWSM*. 2017.

[139] Kiana Kheiri and Hamid Karimi. “SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning”. In: *ArXiv* abs/2307.10234 (2023). URL: https://api.semanticscholar.org/CorpusID:259991148.

[140] C. Kiene and Benjamin Mako Hill. “Who Uses Bots? A Statistical Analysis of Bot Usage in Moderation Teams”. In: *CHI EA* (2020).

[141] C. Kiene, A. Monroy-Hernández, and Benjamin Mako Hill. “Surviving an “Eternal September”: How an Online Community Managed a Surge of Newcomers”. In: *CHI ’16* (2016).

[142] Aniket Kittur and Robert E. Kraut. “Beyond Wikipedia: coordination and conflict in online production groups”. In: *CSCW* (2010).

[143] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. “The Challenge of Understanding What Users Want: Inconsistent Preferences and Engagement Optimization”. In: *Proceedings of the 23rd ACM Conference on Economics and Computation*. EC ’22. Boulder, CO, USA: Association for Computing Machinery, 2022, p. 29. URL: https://doi.org/10.1145/3490486.3538365.

[144] Joon Koh, Young-Gul Kim, and Young-Gul Kim. “Sense of virtual community: A conceptual framework and empirical validation”. In: *International journal of electronic commerce* 8.2 (2003), pp. 75–94.

[145] Piotr Konieczny. “Decision making in the self-evolved collegiate court: Wikipedia’s Arbitration Committee and its implications for self-governance and judiciary in cyberspace”. In: *Intl. Sociology* (2017).

[146] Vinay Koshy et al. “Measuring User-Moderator Alignment on r/ChangeMyView”. In: *CSCW* 7 (2023), pp. 1–36.

[147] Vinay Koshy et al. “Venire: A Machine Learning-Guided Panel Review System for Community Content Moderation”. In: *CSCW* (2025).

[148] Yubo Kou et al. “Understanding social roles in an online community of volatile practice: A study of user experience practitioners on reddit”. In: *ACM Transactions on Social Computing* 1.4 (2018), pp. 1–22.

[149] Ramez Kouzy et al. “Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter”. In: *Cureus* 12 (2020).

[150] Robert E Kraut and Paul Resnick. *Building successful online communities: Evidence-based social design*. Mit Press, 2012.

[151] Manfred Krifka, Silka Martens, and Florian Schwarz. “Group interaction in the cockpit: some linguistic factors”. In: *Linguistische Berichte* (2003).

[152] Emily Kubin and Christian von Sikorski. “The role of (social) media in political polarization: a systematic review”. In: *Annals of the International Communication Association* 45 (2021), pp. 188–206.

[153] Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. “Watch Your Language: Large Language Models and Content Moderation”. In: *ICWSM* (2023).

[154] Ravi Kumar, Mohammad Mahdian, and Mary McGlohon. “Dynamics of conversations”. In: *KDD*. 2010, pp. 553–562.

[155] Srijan Kumar et al. “Community Interaction and Conflict on the Web”. In: *WWW ’18* (Apr. 2018), pp. 933–943.

[156] Upendra Kumar et al. “Inducing personalities and values from language use in social network communities”. In: *Information Systems Frontiers* 20 (2018), pp. 1219–1240.

[157] Ema Kušen and Mark Strembeck. “Building blocks of communication networks in times of crises: Emotion-exchange motifs”. In: *Computers in Human Behavior* 123 (2021), p. 106883.

[158] Tom van Laer. “The Means to Justify the End: Combating Cyber Harassment in Social Media”. In: *Journal of Business Ethics* 123 (Aug. 2014), pp. 85–98. (Visited on 07/08/2019).

[159] Himabindu Lakkaraju, Julian McAuley, and J. Leskovec. “What’s in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media”. In: *ICWSM* (2013).

[160] Charlotte Lambert, Frederick Choi, and Eshwar Chandrasekharan. “"Positive reinforcement helps breed positive behavior": Moderator Perspectives on Encouraging Desirable Behavior”. In: *Proc. ACM Hum. Comput. Interact.* (2024).

[161] Charlotte Lambert, Koustuv Saha, and Eshwar Chandrasekharan. “Does Positive Reinforcement Work?: A Quasi-Experimental Study of the Effects of Positive Feedback on Reddit”. In: *CHI* (2025). DOI: 10.1145/3706598.3713830. URL: https://doi.org/10.1145/3706598.3713830.

[162] J. Richard Landis and Gary G. Koch. “The Measurement of Observer Agreement for Categorical Data”. In: *Biometrics* (1977).

[163] Leon Leibmann et al. “Reddit Rules and Rulers: Quantifying the Link Between Rules and Perceptions of Governance across Thousands of Communities”. In: *ICWSM* (2025). URL: https://api.semanticscholar.org/CorpusID:275906749.

[164] A. Lenhart et al. “Online Harassment, Digital Abuse, and Cyberstalking in America”. In: *Data & Society Research Institute* (2016).

[165] Hanlin Li, Brent J. Hecht, and Stevie Chancellor. “All That’s Happening behind the Scenes: Putting the Spotlight on Volunteer Moderator Labor in Reddit”. In: *ICWSM* (2022).

[166] Hanlin Li, Brent J. Hecht, and Stevie Chancellor. “Measuring the Monetary Value of Online Volunteer Work”. In: *ICWSM* (2022).

[167] Zhiyuan Jerry Lin et al. “Better When It Was Smaller? Community Content and Behavior After Massive Growth”. In: *ICWSM*. 2017. URL: https://api.semanticscholar.org/CorpusID:2524208.

[168] Darren L Linvill and Patrick L Warren. “Troll factories: Manufacturing specialized disinformation on Twitter”. In: *Political Communication* (2020), pp. 1–21.

[169] Yinhan Liu et al. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. In: *ArXiv* abs/1907.11692 (2019).

[170] Llama Team. *The Llama 3 Herd of Models*. 2024. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783.

[171] Travis Lloyd et al. “AI Rules? Characterizing Reddit Community Policies Towards AI-Generated Content”. In: *CHI* (2025).

[172] Haiwei Ma et al. “Write for life: Persisting in online health communities through expressive writing and social support”. In: *CSCW* (2017).

[173] Juan Manuel Machimbarrena et al. “Internet Risks: An Overview of Victimization in Cyberbullying, Cyber Dating Abuse, Sexting, Online Grooming and Problematic Internet Use”. In: *International Journal of Environmental Research and Public Health* 15 (2018).

[174] K. MacQueen et al. “Codebook Development for Team-Based Qualitative Analysis”. In: *Field Methods* (1998).

[175] Mahya Maftouni, Patrick Marcel Joseph Dubois, and Andrea Bunt. ““Thank you for being nice”: Investigating Perspectives Towards Social Feedback on Stack Overflow”. In: *Graphics Interface* (2022).

[176] Kaitlin Mahar, Amy X. Zhang, and David Karger. “Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation”. In: *CHI* (2018).

[177] Thomas J. Main. *The Rise of the Alt-Right*. Brookings Institution Press, July 2018.

[178] Stephann Makri and Sophie Turner. ““I can’t express my thanks enough”: The “gratitude cycle” in online communities”. In: *Journal of the Association for Information Science and Technology* 71.5 (2020), pp. 503–515.

[179] Trevor Martin. “community2vec: Vector representations of online communities encode semantic relationships”. In: *Proceedings of the Second Workshop on NLP and Computational Social Science*. Association for Computational Linguistics, Aug. 2017, pp. 27–31. URL: https://www.aclweb.org/anthology/W17-2904.

[180] Alice Marwick and Rebecca Lewis. “Media manipulation and disinformation online”. In: *Data & Society* (2017). https://datasociety.net/library/media-manipulation-and-disinfo-online/.

[181] J. Nathan Matias. “The Civic Labor of Volunteer Moderators Online”. In: *Social Media + Society* (2019).

[182] J. Nathan Matias and Merry Ember Mou. “CivilServant: Community-Led Experiments in Platform Governance”. In: *CHI* (2018).

[183] J. Nathan Matias, Tyler Simko, and Marianne Reddan. *Reducing the Silencing Role of Harassment in Online Feminism Discussions*. 2020.

[184] J.N. Matias. "Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016). URL: https://api.semanticscholar.org/CorpusID:9897274.

[185] Jorge Nathan Matias. "Preventing harassment and increasing group participation through social norms in 2,190 online science discussions". In: *PNAS* 116 (2019), pp. 9785–9789. URL: https://api.semanticscholar.org/CorpusID:140369949.

[186] Tara Matthews et al. "They said what? Exploring the relationship between language use and member satisfaction in communities". In: *CSCW*. 2015.

[187] Bree McEwan. "Communication of communities: Linguistic signals of online groups". In: *Information, Communication & Society* (2016).

[188] David W. McMillan and David M. Chavis. "Sense of community: A definition and theory". In: *Journal of Community Psychology* (1986).

[189] A. Medvedev, R. Lambiotte, and J. Delvenne. "The anatomy of Reddit: An overview of academic research". In: *ArXiv* abs/1810.10881 (2018).

[190] Meta. *Community Standards*. 2025. URL: https://transparency.meta.com/policies/community-standards/.

[191] Panagiotis Takis Metaxas and Eni Mustafaraj. "Social Media and the Elections". In: *Science* 338 (2012), pp. 472–473.

[192] MistralAI. *Mistral-7B-Instruct-v0.3*. May 2024. URL: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3.

[193] Amy Mitchell et al. "Many Americans Say Made-Up News Is a Critical Problem That Needs To Be Fixed". In: *Pew Research Center Science and Journalism* (2019). https://www.journalism.org/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/.

[194] Matti Nelimarkka, Salla-Maaria Laaksonen, and Bryan Semaan. “Social Media Is Polarized, Social Media Is Polarized: Towards a New Design Agenda for Mitigating Polarization”. English. In: *ACM DIS*. ACM conference on Designing Interactive Systems, DIS ; Conference date: 09-06-2018 Through 13-06-2018. United States: ACM, June 2018, pp. 957–970. DOI: 10.1145/3196709.3196764.

[195] Edward Newell et al. “User Migration in Online Social Networks: A Case Study on Reddit During a Period of Community Unrest”. In: *ICWSM* (2016).

[196] Nic Newman. *Digital News Report*. Reuters Institute, 2020.

[197] Matthew N. Nicholson, Brian C. Keegan, and Casey Fiesler. “Mastodon Rules: Characterizing Formal Rules on Popular Mastodon Instances”. In: *CSCW* (2023). URL: https://api.semanticscholar.org/CorpusID:264039192.

[198] Zach Nussbaum et al. *Nomic Embed: Training a Reproducible Long Context Text Embedder*. 2024. arXiv: 2402.01613 [cs.CL]. URL: https://arxiv.org/abs/2402.01613.

[199] Patricia Obst, Lucy Zinkiewicz, and Sandy G Smith. “Sense of community in science fiction fandom, Part 1: Understanding sense of community in an international community of interest”. In: *Journal of Community Psychology* 30.1 (2002), pp. 87–103.

[200] Patricia Obst, Lucy Zinkiewicz, and Sandy G Smith. “Sense of community in science fiction fandom, Part 2: Comparing neighborhood and interest group sense of community”. In: *Journal of Community Psychology* 30.1 (2002), pp. 105–117.

[201] Patricia L Obst and Katherine M White. “Revisiting the sense of community index: A confirmatory factor analysis”. In: *Journal of community psychology* 32.6 (2004), pp. 691–705.

[202] OpenAI et al. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL].

[203] OpenAI. *Hello GPT-4o*. May 2024. URL: https://openai.com/index/hello-gpt-4o/.

[204] Kyle Orland. *Reddit cashes in on AI gold rush with $203M in LLM training license fees*. 2024. URL: https://arstechnica.com/ai/2024/02/reddit-has-already-booked-203m-in-revenue-licensing-data-for-ai-training/.

[205] Maria Palacin-Silva et al. “The Role of Gamification in Participatory Environmental Sensing: A Study In the Wild”. In: *International Conference on Human Factors in Computing Systems* (2018).

[206] Katherine A. Panciera, Aaron L Halfaker, and Loren G. Terveen. “Wikipedians are born, not made: a study of power editors on Wikipedia”. In: *GROUP* (2009). URL: https://api.semanticscholar.org/CorpusID:6286454.

[207] S. Paoli, Nicolò De Uffici, and V. D’Andrea. “Designing badges for a civic media platform: reputation and named levels”. In: *British Computer Society Conference on Human-Computer Interaction* (2012).

[208] Ashwin Paranjape, Austin R Benson, and Jure Leskovec. “Motifs in temporal networks”. In: *WSDM*. 2017.

[209] Douglas D Perkins et al. “Participation and the social and physical environment of residential blocks: Crime and community context”. In: *American journal of community psychology* 18.1 (1990), pp. 83–115.

[210] Ben J. Plackett. *Unpaid and abused: Moderators speak out against Reddit*. https://www.engadget.com/2018-08-31-reddit-moderators-speak-out.html. Accessed: 2019-07-08. 2018.

[211] Deborah A. Prentice, Dale T. Miller, and Jenifer R. Lightdale. “Asymmetries in Attachments to Groups and to their Members: Distinguishing between Common-Identity and Common-Bond Groups”. In: *Personality and Social Psychology Bulletin* 20.5 (1994), pp. 484–493.

[212] Gale H. Prinster et al. “Community Archetypes: An Empirical Framework for Guiding Research Methodologies to Reflect User Experiences of Sense of Virtual Community on Reddit”. In: *CSCW* (2024).

[213] Nicholas Proferes et al. “Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics”. In: *Social Media + Society* 7 (2021).

[214] Vahed Qazvinian et al. “Rumor has it: Identifying misinformation in microblogs”. In: *EMNLP*. 2011, pp. 1589–1599.

[215] Erik X. Raj and Derek E. Daniels. “Psychosocial support for adults who stutter: Exploring the role of online communities”. In: *Speech, Language and Hearing* (2017).

[216] Ashwin Rajadesingan, P. Resnick, and C. Budak. “Quick, Community-Specific Learning: How Distinctive Toxicity Norms Are Maintained in Political Subreddits”. In: *ICWSM*. 2020.

[217] Sven Carsten Rasmusen et al. “Raising Consent Awareness With Gamification and Knowledge Graphs: An Automotive Use Case”. In: *International Journal on Semantic Web and Information Systems (IJSWIS)* (2022).

[218] Raquel Recuero, Felipe Bonow Soares, and Anatoliy A. Gruzd. “Hyperpartisanship, Disinformation and Political Conversations on Twitter: The Brazilian Presidential Election of 2018”. In: *ICWSM*. 2020.

[219] Reddit, Inc. *Community Funds Overview*. 2024. URL: https://support.reddithelp.com/hc/en-us/articles/15484345935508-Community-Funds-Overview.

[220] Reddit, Inc. *Moderator Code of Conduct*. 2025. URL: https://redditinc.com/policies/moderator-code-of-conduct.

[221] Reddit, Inc. *Moderator Reserves Overview*. 2025. URL: https://support.reddithelp.com/hc/en-us/articles/15484270707092-Moderator-Reserves-Overview.

[222] Reddit, Inc. *Reddit Rules*. 2025. URL: https://redditinc.com/policies/reddit-rules.

[223] Reddit, Inc. *Reddit, Inc. Home Page*. 2024. URL: https://redditinc.com/.

[224] Harita Reddy and Eshwar Chandrasekharan. “Evolution of Rules in Reddit Communities”. In: *CSCW* (2023).

[225] Francisco Regalado et al. “Gamifying Online News in a Senior Online Community: Insights from Designing and Assessing the Readers’ Experience”. In: *The social science* (2021).

[226] Yuqing Ren, Robert Kraut, and Sara Kiesler. “Applying Common Identity and Bond Theory to Design of Online Communities”. In: *Organization Studies* 28 (Mar. 2007), pp. 377–408.

[227] Michael D Resnick, Marjorie Ireland, and Iris Borowsky. “Youth violence perpetration: what protects? What predicts? Findings from the National Longitudinal Study of Adolescent Health”. In: *Journal of adolescent health* 35.5 (2004), 424–e1.

[228] Manoel Horta Ribeiro et al. “Do Platform Migrations Compromise Content Moderation? Evidence from r/The_Donald and r/Incels”. In: *CSCW* (2020). URL: https://api.semanticscholar.org/CorpusID:263873469.

[229] Manoel Horta Ribeiro et al. “Post Guidance for Online Communities”. In: *CSCW* (2025).

[230] Ganit Richter, D. Raban, and S. Rafaeli. “Tailoring a Points Scoring Mechanism for Crowd-Based Knowledge Pooling”. In: *Hawaii International Conference on System Sciences* (2018).

[231] Julian Risch and Ralf Krestel. “Top Comment or Flop Comment? Predicting and Explaining User Engagement in Online News Discussions”. In: *ICWSM*. 2020.

[232] Guido van van Rossum. *All Things Pythonic: Origin of BDFL*. 2008. URL: https://archive.ph/20120721081049/http://www.artima.com/weblogs/viewpost.jsp#selection-157.0-163.14.

[233] Haji Mohammad Saleem and D. Ruths. “The Aftermath of Disbanding an Online Hateful Community”. In: *ArXiv* abs/1804.07354 (2018).

[234] Mattia Samory, Vartan Kesiz Abnousi, and Tanushree Mitra. “Characterizing the Social Media News Sphere through User Co-Sharing Practices”. In: *ICWSM*. Vol. 14. 2020, pp. 602–613.

[235] Tiago Santos et al. “Can Badges Foster a More Welcoming Culture on Q&A Boards?” In: *ICWSM* (2020).

[236] Kai Sassenberg. “Common bond and common identity groups on the Internet: Attachment and normative behavior in on-topic and off-topic chats.” In: *Group Dynamics: Theory, Research, and Practice* (2002).

[237] Martin Saveski, Brandon Roy, and Deb Roy. “The structure of toxic conversations on Twitter”. In: *WebConf*. 2021.

[238] M. Scheuerman et al. “A Framework of Severity for Harmful Content Online”. In: *CSCW* (2021).

[239] Nathan Schneider. “Admins, mods, and benevolent dictators for life: The implicit feudalism of online communities:” in: *New Media & Society* (2021).

[240] Joseph Seering. “Reconsidering Self-Moderation: The Role of Research in Supporting Community-Based Models for Online Content Moderation”. In: *CSCW* (2020).

[241] Joseph Seering and Sanjay Ram Kairam. “Who Moderates on Twitch and What Do They Do?” In: *GROUP* 7 (2022), pp. 1–18. URL: https://api.semanticscholar.org/CorpusID:255226737.

[242] Joseph Seering et al. “Moderator engagement and community development in the age of algorithms”. In: *New Media & Society* 21 (2019), pp. 1417–1443. URL: https://api.semanticscholar.org/CorpusID:149757447.

[243] Elisa Shearer and Katerina Eva Matsa. *News Use Across Social Media Platforms 2018*. https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/. Accessed: 2021-04-09. Sept. 2018.

[244] C. Estelle Smith et al. “The Impact of Governance Bots on Sense of Virtual Community: Development and Validation of the GOV-BOTs Scale”. In: *CSCW* (2022).

[245] Olivia Solon. “Underpaid and Overburdened: the Life of a Facebook Moderator”. en-GB. In: *The Guardian* (May 2017). (Visited on 10/24/2019).

[246] Kumar Bhargav Srinivasan et al. “Content Removal as a Moderation Strategy: Compliance and Other Outcomes in the ChangeMyView Community”. In: *CSCW* (2019). URL: https://api.semanticscholar.org/CorpusID:203617040.

[247] Sameer B. Srivastava et al. “Enculturation Trajectories: Language, Cultural Adaptation, and Individual Outcomes in Organizations”. In: *Management Science* (2015). URL: https://pubsonline.informs.org/doi/10.1287/mnsc.2016.2671.

[248] Kate Starbird. “Examining the Alternative Media Ecosystem Through the Production of Alternative Narratives of Mass Shooting Events on Twitter”. In: *ICWSM*. 2017.

[249] Statista. *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. July 2021.

[250] Peter Stefanov et al. "Predicting the Topical Stance and Political Leaning of Media using Tweets". In: *ACL*. Online: ACL, July 2020.

[251] Igor Steinmacher et al. "Social Barriers Faced by Newcomers Placing Their First Contribution in Open Source Software Projects". In: *CSCW* (2015).

[252] Anselm L. Strauss. *Qualitative analysis for social scientists*. Cambridge University Press, 1987.

[253] Shima Sum et al. "Internet use as a predictor of sense of community in older people". In: *CyberPsychology & Behavior* 12.2 (2009), pp. 235–239.

[254] Chenhao Tan et al. "Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions". In: *WWW* (2016). URL: https://api.semanticscholar.org/CorpusID:8577096.

[255] Samia Tasnim, M. Hossain, and Hoimonty Mazumder. "Impact of Rumors and Misinformation on COVID-19 in Social Media". In: *Journal of Preventive Medicine and Public Health* 53 (2020), pp. 171–174.

[256] Yla R Tausczik and James W Pennebaker. "The psychological meaning of words: LIWC and computerized text analysis methods". In: *Journal of language and social psychology* 29.1 (2010), pp. 24–54.

[257] Nathan TeBlunthuis et al. "No Community Can Do Everything: Why People Participate in Similar Online Communities". In: *CSCW* (2022). URL: https://api.semanticscholar.org/CorpusID:245877572.

[258] "The Discourse of Online Content Moderation: Investigating Polarized User Responses to Changes in Reddit's Quarantine Policy". In: *ACL* (2019).

[259] The Internet Archive. *Wayback Machine*. URL: https://web.archive.org/.

[260] The Oversight Board. *Content Moderation in a New Era for AI and Automation*. 2025. URL: https://www.oversightboard.com/news/content-moderation-in-a-new-era-for-ai-and-automation/.

[261] Sarah-Kristin Thiel. "Let's play Urban Planner: The use of Game Elements in Public Participation Platforms". In: *plaNext–Next Generation Planning* (2017).

[262] Sarah-Kristin Thiel et al. "Inclusive Gamified Participation: Who are we inviting and who becomes engaged?" In: *Hawaii International Conference on System Sciences* (2019).

[263] Sarah-Kristin Thiel et al. "Playing (with) Democracy: A Review of Gamified Participation Approaches". In: *Journal of eDemocracy and Open Government (JeDEM)* (2016).

[264] Pamela Bilo Thomas et al. "Behavior Change in Response to Subreddit Bans and External Events". In: *IEEE TCSS* 8 (2021), pp. 809–818. URL: https://api.semanticscholar.org/CorpusID:230770317.

[265] TikTok. *Community Guidelines*. 2025. URL: https://www.tiktok.com/community-guidelines/en.

[266] Lisbeth Tonteri et al. "Antecedents of an experienced sense of virtual community". In: *Computers in Human Behavior* 27.6 (2011), pp. 2215–2223.

[267] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. arXiv: 2307.09288 [cs.CL].

[268] T. Tran et al. "An Investigation of Misinformation Harms Related to Social Media during Two Humanitarian Crises". In: *Information Systems Frontiers* (2020), pp. 1–9.

[269] Francesca Tripodi. "Ms. Categorized: Gender, notability, and inequality on Wikipedia". In: *New Media & Society* (June 2021).

[270] Selen Türkay and Sonam Adinolf. "Friending to flame: How social features affect player behaviours in an online collectible card game". In: *CHI*. 2019.

[271] Svitlana Volkova et al. "Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter". In: *ACL Short Papers*. July 2017, pp. 647–653.

[272] Soroush Vosoughi, Mostafa 'Neo' Mohsenvand, and Deb Roy. "Rumor gauge: Predicting the veracity of rumors on Twitter". In: *KDD* 11.4 (2017), pp. 1–36.

[273] Soroush Vosoughi, Deb Roy, and Sinan Aral. "The spread of true and false news online". In: *Science* (2018).

[274] David Wadden et al. "The Effect of Moderation on Online Mental Health Conversations". In: *ICWSM '21*. Vol. 15. 2021, pp. 751–763.

[275] Isaac Waller and Ashton Anderson. "Generalists and Specialists: Using Community Embeddings to Quantify Activity Diversity in Online Platforms". In: *TheWebConf* (2019).

[276] Isaac Waller and Ashton Anderson. "Quantifying social organization and political polarization in online platforms". In: *Nature* (2020).

[277] Yixue Wang and N. Diakopoulos. "Highlighting High-quality Content as a Moderation Strategy: The Role of New York Times Picks in Comment Quality and Engagement". In: *ACM Transactions on Social Computing* (2021).

[278] M McLure Wasko and Samer Faraj. ""It is what one does": why people participate and help others in electronic communities of practice". In: *The journal of strategic information systems* 9.2-3 (2000), pp. 155–173.

[279] Jennifer L Welbourne, Anita L Blanchard, and Marla D Boughton. "Supportive communication, sense of virtual community and health outcomes in online infertility groups". In: *C&T '09*. 2009, pp. 31–40.

[280] Galen Weld, Maria Glenski, and Tim Althoff. "Adjusting for Confounders with Text: Challenges and an Empirical Evaluation Framework for Causal Inference". In: *ICWSM* 16 (2021), pp. 1109–1120.

[281] Galen Weld et al. "Perceptions of Moderators as a Large-Scale Measure of Online Community Governance". In: *CSCW* (2025).

[282] Galen Cassebeer Weld, Maria Glenski, and Tim Althoff. "Political Bias and Factualness in News Sharing Across more then 100, 000 Online Communities". In: *ICWSM* (2021).

[283] Galen Cassebeer Weld, Amy X. Zhang, and Tim Althoff. “Making Online Communities ’Better’: A Taxonomy of Community Values on Reddit”. In: *ICWSM* (2024).

[284] Galen Cassebeer Weld, Amy X. Zhang, and Tim Althoff. “What Makes Online Communities ’Better’? Measuring Values, Consensus, and Conflict across Thousands of Subreddits”. In: *ICWSM* (2022).

[285] Galen Cassebeer Weld et al. “How Conversational Structure and Style Shape Online Community Experiences”. In: *ICWSM* abs/2508.08596 (2025). URL: https://api.semanticscholar.org/CorpusID:280635813.

[286] Wikipedia. *Wikipedia:Polling is not a substitute for discussion*. 2021.

[287] Wen Wu, Li Chen, and Qingchang Yang. “Inferring Students’ Sense of Community from Their Communication Behavior in Online Courses”. In: *UMAP*. 2017.

[288] X. *The X Rules*. 2025. URL: https://help.x.com/en/rules-and-policies/x-rules.

[289] Ramtin Yazdanian et al. “Eliciting New Wikipedia Users’ Interests via Automatically Mined Questionnaires: For a Warm Welcome, Not a Cold Start”. In: *ICWSM* (2019).

[290] Yulin Yu and Paramveer Dhillon. “Deconstructing the structure of online conversations on Reddit”. In: *CSCW* (2024).

[291] Amy X. Zhang, Grant Hugh, and Michael S. Bernstein. “PolicyKit: Building Governance in Online Communities”. In: *UIST* (2020).

[292] Justine Zhang et al. “Characterizing Online Public Discussions through Patterns of Participant Interactions”. In: *CSCW* (2018).

[293] Haiyi Zhu, Robert E. Kraut, and Aniket Kittur. “The Impact of Membership Overlap on the Survival of Online Communities”. In: *CHI* (2014), pp. 281–290. URL: https://doi.org/10.1145/2556288.2557213.

[294] Fabiana Zollo and Walter Quattrociocchi. “Misinformation Spreading on Facebook”. In: *Complex Spreading Phenomena in Social Systems: Influence and Contagion in Real-World Social Networks*. Ed. by Sune Lehmann and Yong-Yeol Ahn. Springer International Publishing, 2018, pp. 177–196. URL: https://doi.org/10.1007/978-3-319-77332-2_10.

Appendix A

Additional Materials for Taxonomy of Community Values

A.1 Additional Figures Describing Survey Interface and Recruiting

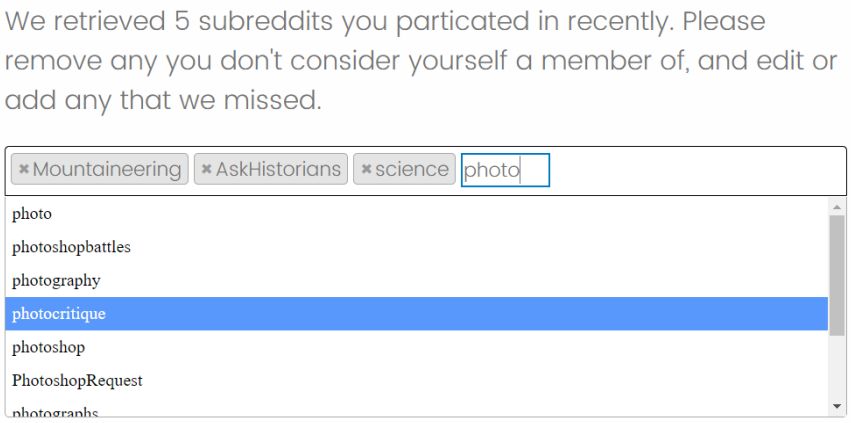


Figure A.1: A screenshot of the interface used by participants to enter the subreddits they consider themselves a member of. This search box queries the reddit API in real-time to populate the results and ensure that only valid subreddit names are entered.

Figure A.2: Reddit advertisements used to recruit participants.

A.2 Additional Recruiting and Incentive Details

Participants for this study were recruited primarily through the purchase of reddit advertisements. These advertisements display inline with other content in both individual subreddits and aggregated content views, and are shown on the reddit website as well as the official reddit mobile app. Figure A.2 shows the appearance of these advertisements.

Upon clicking on the advertisement, the user is taken to the first page of the survey which summarizes the aims of the study and requests informed consent to continue. Over the course of the study, 920,025 advertisement impressions were made, generating 2,084 clicks, for a click-through rate of 0.227%. Of these people who clicked through to the first page of the survey, 509 started to complete the survey, and 212 completed it. Eleven respondents did not consent to continuing the survey, and as a result were not shown any additional questions. To increase the diversity of recruiting, the survey was also distributed to relevant university mailing lists and Slack channels at two large American universities, as well as posted to /r/SampleSize, a subreddit for the distribution of surveys, recruiting an additional 41 participants.

To incentivize respondents to participate, a raffle was held, and winners were chosen at random from the pool of respondents who both completed the survey and supplied their reddit username, which was used to contact the winners. One ‘first place’ prize, a $100 Amazon gift card, and five ‘second place’ prizes of $20 Amazon gift cards were awarded (equivalent amounts in local currency were provided to participants outside of the United States).

A.3 Additional Details on Participant Demographics

	Our Survey ($n = 212$)		reddit Overall Barthiel et al. [19]
Age			
18-29	101	(71.1%)	59%
30-49	35	(24.6%)	33%
50-64	6	(4.2%)	7%
65 or older	0	(0.0%)	<1%
Gender			
Woman	49	(26.2%)	33%
Man	129	(69.0%)	67%
Non-binary	9	(4.8%)	Not reported
Additional Gender	5	(2.7%)	
Race and Ethnicity			
White (Hispanic, Latino, or Spanish)	32	(16.9%)	7% Hispanic
Non-white Hispanic, Latino, or Spanish	6	(3.2%)	
White (Not Hispanic, Latino, or Spanish)	89	(47.1%)	74% White
Black or African American	3	(1.6%)	8% Black
Middle Eastern or North African	8	(4.2%)	
Asian	46	(24.3%)	10% Other
Native Hawaiian or Pacific Islander	3	(1.6%)	
Indigenous American or Alaskan Native	1	(0.5%)	
Additional Race/Ethnicity	10	(5.3%)	

Table A.1: Demographics of survey respondents and overall reddit demographics. In general, our respondents' demographics are similar to the overall demographics of reddit. Participants could choose multiple gender and race/ethnicity options, so percentages may not sum to 100%. Overall reddit demographic data [19] uses different racial identity questions, so while an exact comparison is not possible, similar categories are provided here.

A.4 Survey Instrument

Informed Consent

This survey aims to learn more about the subreddits that reddit users participate in, and what values redditors have for those subreddits.

In this survey, we will ask you a few questions about your overall reddit usage, and then ask you a few questions each about the subreddits you consider yourself to be a member of. You may skip any questions you'd prefer not to answer. It should take less than 5 minutes to complete.

Only high-level data will be published as part of our research. Your responses are confidential, and will never be made public.

This study is run by researchers from the University of Washington, and has been determined to be exempt from IRB approval under University of Washington IRB ID STUDY00011457. For questions or concerns,

please contact gweld@cs.washington.edu, or /u/cyclistNerd on reddit.

Would you like to participate in this survey?

- ○ Yes, I would like to participate in this survey.
- ○ No, I would not like to participate in this survey.

Reddit Username and Compensation

As compensation for your participation in this study, you will be entered in a raffle for one of five Amazon gift cards, worth $20 each.

To contact you after the raffle drawing, we will send you a reddit private message. To do so, we ask for your reddit username.

We will also use your username to identify your posts and comments in the subreddits you participate in. Your username will be kept confidential, and we will never publish any of your reddit history.

Providing your username is entirely optional, but without it, we cannot enter you into the raffle.

What is your primary reddit username? Your answer will be kept confidential.

Please spell carefully, and do not include '/u/' or 'u/'.

[ ]

Do you have multiple reddit accounts?

○ No, I only have one reddit account. ○ Yes, I have multiple reddit accounts.

You entered that your primary username is /u/_____

If this looks correct, press next to start the rest of the survey. If this is incorrect, please press back to edit your response.

Multiple Accounts

Earlier, you indicated that you had multiple reddit accounts.

If you're comfortable doing so, please enter the names of your alternate reddit accounts, separated by commas.

[ ]

Demographics

What is your age (in years)? [ ]

Please describe your gender (check all that apply)

- ☐ Woman
- ☐ Man
- ☐ Non-binary
- ☐ [ ] Prefer to self-describe

Please describe your race (check all that apply)

- ☐ White (Hispanic, Latino, or Spanish)
- ☐ White (Not Hispanic, Latino, or Spanish)
- ☐ Non-white Hispanic, Latino, or Spanish
- ☐ Black or African American
- ☐ Asian
- ☐ Middle Eastern or North African
- ☐ Native Hawaiian or Pacific Islander
- ☐ Indigenous American or Alaskan Native
- ☐ [ ] Prefer to self-describe

Overall Reddit Usage

In this section, we will ask you about how you use reddit.

Typically, how often do you use reddit?

- Every day.
- A few times a week.
- Once a week.

Typically, how long do you typically spend on reddit at one time?

- Less than five minutes.
- More than five minutes, but less than fifteen minutes.
- More than fifteen minutes, but less than an hour.
- More than an hour.

Typically, how often do you post or comment on reddit versus browsing what others have submitted (lurking)?

- Frequently, I frequently submit posts or comment on threads.
- Occasionally, I post or comment occasionally, but mostly browse what others have submitted.
- Rarely, I almost always just browse what others have submitted.
- Never, I only browse what others have submitted.

How often do you use aggregate subreddits (like your frontpage, /r/all, or multi-reddits) versus looking at individual subreddits?

- Always, I never look at individual subreddits.
- Frequently, I mostly use use aggregate subreddits, but sometimes I look at individual subreddits.
- Occasionally, I look at aggregate subreddits and individual subreddits about evenly.
- Rarely, I mostly look at individual subreddits.
- Never, I only look at individual subreddits.

Do you use reddit more from your computer or from your phone?

- ○ I only use reddit from my computer.
- ○ I mostly use my computer, but sometimes use my phone.
- ○ I use my phone and my computer about evenly.
- ○ I mostly use my phone, but sometimes use my computer.
- ○ I only use reddit from my phone.

Subreddit Selection

In this section, we'll ask you questions specific to the subreddits that you consider yourself a member of. These subreddits should be subreddits that are important to you,and that you are familiar enough with to feel comfortable commenting on different aspects of how they are run, and how their members interact with one another.

Thank you for selecting subreddits. The next section will ask you about your experiences with each subreddit, individually.

Open Ended Subreddit Value Questions In this section, we'll ask you questions specifically about your experience in /r/_____.

In your responses, please consider not only the content of /r/_____, but also how it is run, how its members treat one another, and anything else that impacts your experience.

As it exists right now, what are a few of the best aspects of the /r/_____ community?

If you could change anything, what are some aspects of the /r/_____ community you would like to improve upon?

Reflection Questions

These last two questions ask you to reflect on your experience across all the subreddits you've used.

Generally, what values do you think are important in an online community? What makes a community healthy?

Have you ever stopped participating in any subreddits? What are some signs that a community isn't worth your time, or isn't a community you want to participate in?

A.5 Codebook

Introduction

This taxonomy is intended to classify individual idea units, which are focused on a specific aspect of a community. As such, it is intended to be mutually exclusive - idea units should not be assigned to multiple categories. If a participants' response appears to belong to multiple categories, it likely should be subdivided into multiple idea units.

Furthermore, the taxonomy is hierarchical, with the categories being grouped into 9 high level groups. Idea units should be tagged with the most specific category possible - while the high level group is also available as a label itself, it should only be applied when the idea unit in question is either a) too short/vague to be assigned to a more specific category, or b) does not clearly fit into any of the subcategories.

Taxonomy Categories and Descriptions

1) Quality of Content Idea units that belong to this category should be commenting on the perceived value/utility of the content in the subreddit, or lack thereof, with exceptions for idea units regarding bullying/offensive content (these belong in category 5a) or idea units regarding the trustworthiness of the content (these belong in category 9b). As with all categories, idea units in the Quality of Content category should be assigned to as specific of a category as possible. Thus, many of the idea units that fall into category 1 (as opposed to subcategories 1a-d) are fairly vague/generic.

Examples

I like the content

Content is easy to understand

Counterexamples

The content is funny belongs in 1a due to the specificity of "funny" (it's entertaining).

There are not enough memes belongs in 1b because the desire for memes is a personal preference.

Some posts are degrading to women belongs in 5a because it refers specifically to content this offensive or abusive.

1a) Education, Entertainment This category should contain idea units that comment on the value of

the content, be it educational, entertaining, or some combination (such as "interesting"). It can sometimes be difficult to distinguish idea units in this category from those in 1b Personal Preferences. You should use your judgment to determine if the survey respondent's emphasis is more on the utility of the content (belonging in 1a) the the specific type of content they prefer (belonging in 1b).

Examples

The posts are funny

I learn a lot about different mushrooms

Counterexamples

I especially laugh at the pictures of silly cats belongs in 1b because the emphasis is on the personal preference towards a specific type of content (silly cat pictures).

1b) Personal Preferences This category should contain idea units that reflect the respondent's preferences towards (or against) particular types of content, frequently including idea units on memes, specific subject matter, etc.

Examples

Good memes

Custom gaming PCs

Less discussion about politics would be good

1c) Curation, Recency, Discovery This category focuses on how the content in the subreddit helps the community member discover new things that they may not have otherwise seen, or helps them make sense of large volumes of content (curation).

Examples

/r/seattle helps me keep up to date with news

/r/hiphopheads helps me find the best new music

1d) Spam, Reposts, Bots Idea units in this category focus on the presence of specific types of content such as spam, reposts, and content submitted by bots.

Examples

There are too many reposts in /r/pics

I don't like the repetitive questions

The community bot is helpful

Too much spam

2) Community Engagement

Idea units in this category focus not on the content, but on the community, either the community as a whole abstract entity, or on specific people or groups of people. One exception to this is idea units about the credentials or knowledge of people in the community, which belong in 9a Knowledgeable People. Most idea units in this category are likely to fall into the more specific idea units should be assigned to subcategories 2a and 2b, while very generic or high level idea units should be assigned to this top level category.

Examples

Community

Counterexamples

Friendly community belongs in 2a as it comments on a quality (friendliness) of the community as a whole.

2a) Quality of Interaction or Community as a Whole Idea units in this category comment on both good and bad aspects of the community or of specific interactions within the community. Note that some specific types of interactions belong in other categories (e.g. size in 4b, harassment in 5a, voting behavior in 8a, rule/norm-breaking in 8b).

Examples

My posts are replied to quickly

The community is nice

I get to discuss my favorite TV show [in /r/TwinPeaks]

2b) Connection, Universalization Idea units in this category should focus on the impact of the community on the survey respondent, such as feeling connected to community members, or aware that there are others out there who feel the same as they do (universalization).

Examples

/r/meow_irl is so relatable

/r/mentalhealth makes me feel connected to others

/r/raisedbynarcissists makes me know there are others who feel like me

3) Diversity This category focuses on the diversity of content and people within a community. Almost all idea units in this category should be able to be assigned to one of the two subcategories 3a and 3b.

3a) Variety of Content This category focuses on the variety of content. All idea units which praise or critique the diversity of content should fall within this category, with the exception of complaints about reposts (re-submission of the exact same content) which belong in 1d.

Examples

I get to see so many different kinds of snails [on /r/snails]

I don't like how there is a hivemind

3b) Diversity of People This category should contain idea units on the diversity of the people within the community or the community as a whole. Idea units that focus on the diversity (or lack thereof) peoples' opinions or ideas also belong in this category.

Examples

There are people of so many different backgrounds

I don't like the hivemind about some musicians

4) Size Idea units in the size category will most likely fall into either of the two subcategories unless they are very vague.

Counterexamples

Too big should be assigned to 4b as adjectives "big" and "small" generally refer to the size of the community unless explicitly stated otherwise.

4a) Volume of Content This category contains idea units that relate to the volume of content (including posts and comments) within a subreddit, or the rate at which this content is posted.

Examples

I wish there were more posts

Not enough people post

I like that there are always new posts for me to look at

Posts get lots of comments

Posts (or comments) are submitted so frequently

4b) Size of Community This category contains idea units that relate to the size of the community (i.e. the number of people who are in the community or who participate in the community)

Examples

I wish the community were bigger

The community is small and close-knit belongs in 4b as the emphasis is primarily on the size of the community. Idea units in the category can be similar to those in 2a Quality of Interaction or Community as a Whole. Use your judgment to determine if the primary emphasis is on the size of the community or on some other qualities

Counterexamples

The community is close-knit belongs in 2a as the primary emphasis of the idea unit is on the connection within the community, not the size.

5) Participation and Inclusion Idea units in this category should focus on who participates in the community, and actions and content that explicitly impact who participates and who is included. Idea units that mention aspects of a community that may have an impact on inclusion, but this impact is not explicitly stated (e.g. 'friendly people') should be categorized in 2a.

Examples

Everyone is included

Counterexamples

They are mean to new members belongs in subcategory 5a as it relates to behavior which is specifically abusive or harassing

5a) Offensive, Abusive, Harassing Content or Behaviors Idea units in this category relate to content and behaviors that are offensive, abusive, or harassing to specific people, groups of people, or in general. You should use your best judgment to interpret if the content is offensive to the respondent, in which case it belongs in this category, or if it is simply not in alignment with their personal preferences, in which case it belongs in 1b.

Examples

I wish there were less jokes that make fun of women is offensive

Bullying

[/r/pics] pictures of male genitalia belongs in this category because in the context of /r/pics, pictures of genitalia are unexpected and therefore likely offensive.

Counterexamples

[/r/nudes] pictures of male genitalia does not belong in this category because the most likely interpretation of "pictures of male genitalia" in the context of /r/nudes is that of a personal preference (given that such pictures would be expected on the /r/nudes subreddit) and therefore this idea unit is best categorized under 1b.

People are quick to criticize belongs in 5 because criticism isn't inherently offensive

5b) Outsiders and Demographics Idea units in this category should explicitly relate to who participates in the community, and their perceived out/in-group status.

Examples

[/r/wichita] too many people from outside the city post here

Posts are mostly submitted by people from outside the community

5c) Tools for Participation This category relates to tools (such as wikis, FAQs, and stickied posts) that help improve participation.

Examples

The bot that welcomes new members is really nice

There should be be an FAQ explaining how to post

6) Technical Features Idea units in this category and subcategories should relate to technical features offered (or not offered by reddit). As with all categories, the most specific subcategory should be used. As such, this top-level category should only be used for broad idea units that do not fit into 6a-c.

Examples

Videos don't load sometimes

I prefer old reddit to the redesign

The mobile app works well

6a) Flairs, Tags, NSFW Labels Idea units in this category should relate to flairs (small icons or text strings associated with usernames), tags applied to posts (this includes stickied/pinned posts or comments), and NSFW (Not Safe for Work) labels.

Examples

[/r/colonoscopy] NSFW tags are really well used

Flairs are helpful for knowing who is who

Post tags for categories are convenient

Sometimes NSFW content isn't marked as such

6b) Search, Filters Idea units in this category relate to searching for content or filtering for specific types of content.

Examples

I like that I can easily search for pictures

I wish I could filter by experience level

It would be nice to be hide NSFW content belongs in this category because it relates primarily to the ability to filter out NSFW content, not the NSFW label itself.

6c) Recommendation Systems Idea units in this category relate to recommendation systems for finding similar content.

Examples

I wish there were recommendations for seeing more posts like the ones I like

7) Moderation and Moderators This category contains idea units regarding who the moderators are, and how they perform their job (or fail to perform their job).

On reddit, moderators are responsible for:

- Setting rules
- Enforcing rules, including
- Removing posts
- Banning users
- Updating tags and flairs and NSFW labels
- Recruiting and selecting new moderators
- Communicating their work to the community

Note that this category should only contain idea units relating specifically to moderators and their actions. Idea units regarding rules themselves (e.g. requests for new rules) should go in 8 and its subcategories.

Examples

Moderators are very friendly and reply quickly to messages

The mods are corrupt and on a power-trip

I wish the community was involved in new rules being created

Rules aren't well enforced

Moderators enforce rules unfairly

Counterexamples

I wish content from 4chan was banned belongs in 8b because it is a request for a rule.

8) Norms and Rules Idea units in this category and its subcategories should relate to rules and norms, explicitly stated or not. Idea units relating to how rules are enforced, or other aspects of moderation, belong in 7 Moderators.

8a) Voting Behavior Idea units in this category should relate specifically and explicitly to voting behavior (upvotes and downvotes on reddit).

Examples

People shouldn't use downvote as a disagree button

Counterexamples

Upvotes help me find the best music belongs in 1c because it is primarily relating the the curation mechanism provided by votes, not how people vote.

8b) Adherence to Norms and Rules Idea units in this category should relate to the subreddits' rules and norms and how the community adheres to norms and rules. Many idea units in this category will be specific to the type of content within the subreddit, so checking the subreddit the idea unit corresponds to may be helpful. This category should also include requests for rules. Implicit norms can be particularly complex and context dependent, so consider these carefully and use your best judgment. Note that idea units relating to the enforcement of rules belong in 7 Moderators.

Examples

It is frustrating when newbs ask questions that are already answered in the FAQ

[/r/lgbtq] People sometimes ask us to basically pick a label for them - this is a deeply personal identity, not something we can pick for other people

I wish content from 9gag was banned

They should allow memes

[/r/birdspotting] People should be required to include the location the saw the bird in their request for ID

[/r/ElderScrolls] People should specify which version of the game they're referring to

Counterexamples

Rules aren't enforced strictly enough belongs in 7 because it primarily relates to enforcement, which is performed by moderators.

9) Trust Idea units in this category relate to the trustworthiness (or lack thereof) of the people and the content within subreddits. The top-level category should be used only when it is not possible to differentiate between idea units commenting on the people (9a) and the content (9b).

Examples

Students post their honest experiences and opinions so it's legit belongs in the top-level category because it is commenting both on the people (students are honest) and the content (posts are honest)

9a) Knowledgeable People Idea units in this category should relate to the knowledge, trustworthiness, or credentials of the people in the community.

Examples

There are lots of doctors so I know what they're saying is correct

People are mostly honest

9b) Trustworthy Content Idea units in this category should relate to the veracity or trustworthiness of content in the community.

Examples

There is no fake news

[/r/whatisthisplant] Multiple people often confirm the same ID so I know it's right

10) Exclude/Unintelligible/Off-topic

Notes for Raters

Responses are confidential. Please do not disclose them. Use your best judgment, and refer back to this codebook as necessary. Feel free to use the comments column to note any responses you were especially uncertain of. When labeling, mark the number and letter of the category in the label column. For example, an idea unit that you classify into "Tools for Participation" should be labeled "5c". Don't worry if the labels are deeply skewed - in our initial sample, categories 1a, 1b, and 2a were by far the most common. Leave blank idea units that you think are not categorizable, and add a comment explaining why.

Appendix B

Additional Materials for Large Scale Survey of Community Values

B.1 Categorizing Subreddit Topics

To operationalize higher-level notions of community topic and focus,[1] 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)

[1] We additionally experimented with pre-computed subreddit embeddings [155, 179, 275]. We found these representations were unable to differentiate between communities based on their values.

- **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.2 Power Analysis to Determine Validity of Responses

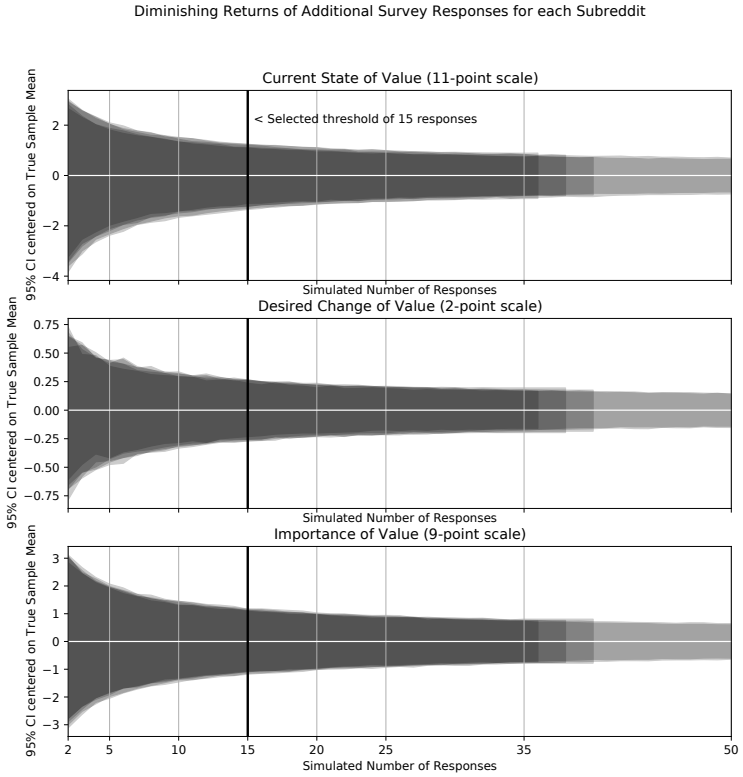


Figure B.1: Using responses from the 5 subreddits with more than 35 responses each, we randomly down-sampled (1,000-fold bootstrapping) responses to estimate the sample variance when collecting fewer responses. We found that beyond 15 responses per subreddit, sample variance does not decrease significantly,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

B.3 Participant Recruiting and Incentives

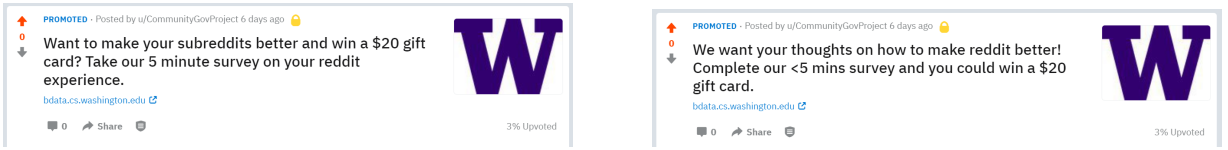


Figure B.2: Reddit advertisements used to recruit participants.

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. B.2 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.

B.4 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 B.1: 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 §2.2.3) topic of the subreddit.

Table B.2: Descriptions of features used in the prediction tasks (§2.2.7).

Appendix C

Additional Materials for Sense of Virtual Community Surveys

C.1 Survey Instrument

The following survey instrument was presented to respondents using Qualtrics.

Intro and Screener Questions

The information collected in this study will be used as part of a research study conducted by Reddit, Inc. Your survey responses will be de-identified. Any data collected will be used and shared in accordance with our Privacy Policy.

By agreeing to participate in this study, you consent to our collection and processing of your survey responses. You represent you are 18 years of age or older. You acknowledge that any information about the study, including any study details and any information provided to you through the study, are considered Reddit Confidential Information. You acknowledge Reddit may use your de-identified survey answers for marketing purposes.

By clicking the link to the survey, you acknowledge that you have read and agree to this consent.

○ I acknowledge that I have read the consent language above and agree to participate in this survey study.

What is your age?

- Under 15 [terminate]
- 16-17 [terminate]
- 18-24
- 25-34
- 35-44
- 45-54
- 55-64
- 65+

Reddit Usage

How long have you been active on Reddit? [Choose the answer that best applies]

- Less than 1 week
- 1-4 weeks
- 1-3 months
- 4-6 months
- 7 months - 1 year
- Over 1 year

How often do you typically visit Reddit? [Choose the answer that best applies]

- Multiple times per day
- Once per day
- A few times per week

- A few times per month
- Once a month or less often

Subreddit-Specific Questions

Please think about the reasons that you visit r/_____ and rate your level of agreement or disagreement with each of the following statements [5-point Likert Agree/Disagree]

- I use this subreddit to get information
- I use this subreddit to learn how to do things
- I use this subreddit to provide others with information
- I use this subreddit to contribute to the pool of information
- I use this subreddit to meet people with my interests
- I use this subreddit to build relationships with others
- I use this subreddit to learn about myself
- I use this subreddit to gain insight into myself
- I use this subreddit to gain prestige
- I use this subreddit to feel important
- I use this subreddit to receive entertaining content
- I use this subreddit to have fun
- I use this subreddit for relaxation or stress relief
- I use this subreddit as a way to pass time when bored

Please rate your level of agreement or disagreement with each of the statements regarding r/_____. Choose the option that best describes how you personally feel [5-point Likert Agree/Disagree]

- This subreddit has quality content.
- The content in this subreddit has variety.
- The people in this subreddit are diverse.
- This subreddit has trustworthy people and information.
- Members of this subreddit engage with one another.
- Members of this subreddit are included and able to contribute.
- This community is an appropriate size.
- Members of this community have input into moderator decisions.
- This community is free of offensive or harassing behavior.

Think about how important each of the following is to your experience in r/_____, and rank them in terms of importance.

- Quality of the content
- Variety in/of the content
- Diversity of the people
- Trustworthiness of the people and information
- Members' engagement with one another
- Members' inclusion and ability to contribute
- Size of the community
- Community input into moderator decisions
- Absence of offensive or harassing behavior
- Other (please explain) [ ]

Please rate your level of agreement or disagreement with each of the statements regarding the community within r/_____. Choose the option that best describes how you personally feel [5-point Likert Agree/Disagree]

- I expect to be a part of this community for a long time.
- I think this community is a good thing for me to be a part of.
- It is important to me to be a part of this community.
- I feel at home in this community.
- I recognize the screen names of most participants in this community.
- If there is a problem in this community, members can get it solved.
- Members of this community can be counted on to help others.
- I want the same things from this community as other members.
- Members of this community share the same values.
- I have friends in this community that I can depend on.
- If I have a personal problem, I can turn to members of this community.
- I care about what other community members think of me.
- Most members of this community know me.
- I feel like I have influence over what this community is like

What other, existing subreddits would you recommend to a redditor who enjoyed and participated in r/_____, and why? (Optional)

As it exists right now, what are a few of the best aspects of r/_____? (Optional)

If you could change anything about r/_____, what are some aspects that could be improved upon? (Optional)

Demographics and Conclusion

How would you describe your gender identity? Please mark any answers that apply:

- ○ Man
- ○ Woman
- ○ Non-binary
- ○ Prefer to self-describe (please specify)
- ○ Prefer not to answer

If you'd like to receive a gratuity for completing this survey, please provide a valid email address, where you would prefer to receive your Amazon gift card. (optional)

Is there anything else that you'd like to tell us about your experience using Reddit, in general? (optional)

C.2 Conversational Patterns

We computed the number of occurrences of eight unique conversational patterns in each target subreddit. While not all patterns were included in our final model (§2.4.1), we include them here for completeness.

Pattern Abbreviation	Description of Pattern
ApBc	OP Posts, Receives Reply
ApBcAr	OP Posts, Receives Reply, Replies Back
ApBcAu	OP Posts, Receives Reply, Upvotes
ApBcAd	OP Posts, Receives Reply, Downvotes
AcBr	OP Comments, Receives Reply
AcBrAc	OP Comments, Receives Reply, Replies Back
AcBrAu	OP Comments, Receives Reply, Upvotes
AcBrAd	OP Comments, Receives Reply, Downvotes

Table C.1: Descriptions of the eight conversational patterns that we identified occurrences of in target subreddits. Not all patterns were included in our final model.

C.3 SOVC Scores by Topic

To assess how SOVC varies between communities of different topics, we made use of Reddit's publicly available topic listings, which consist of 29 top level topic categories viewable at `www.reddit.com/explore`. For communities assigned multiple topics, we used only the topic assigned the highest relevance to that community. We excluded 6 topics (Nature & Outdoors, Sciences, Adult Content, Wellness, Q&As & Stories, and Spooky) for which we had fewer than three participating communities assigned that topic.

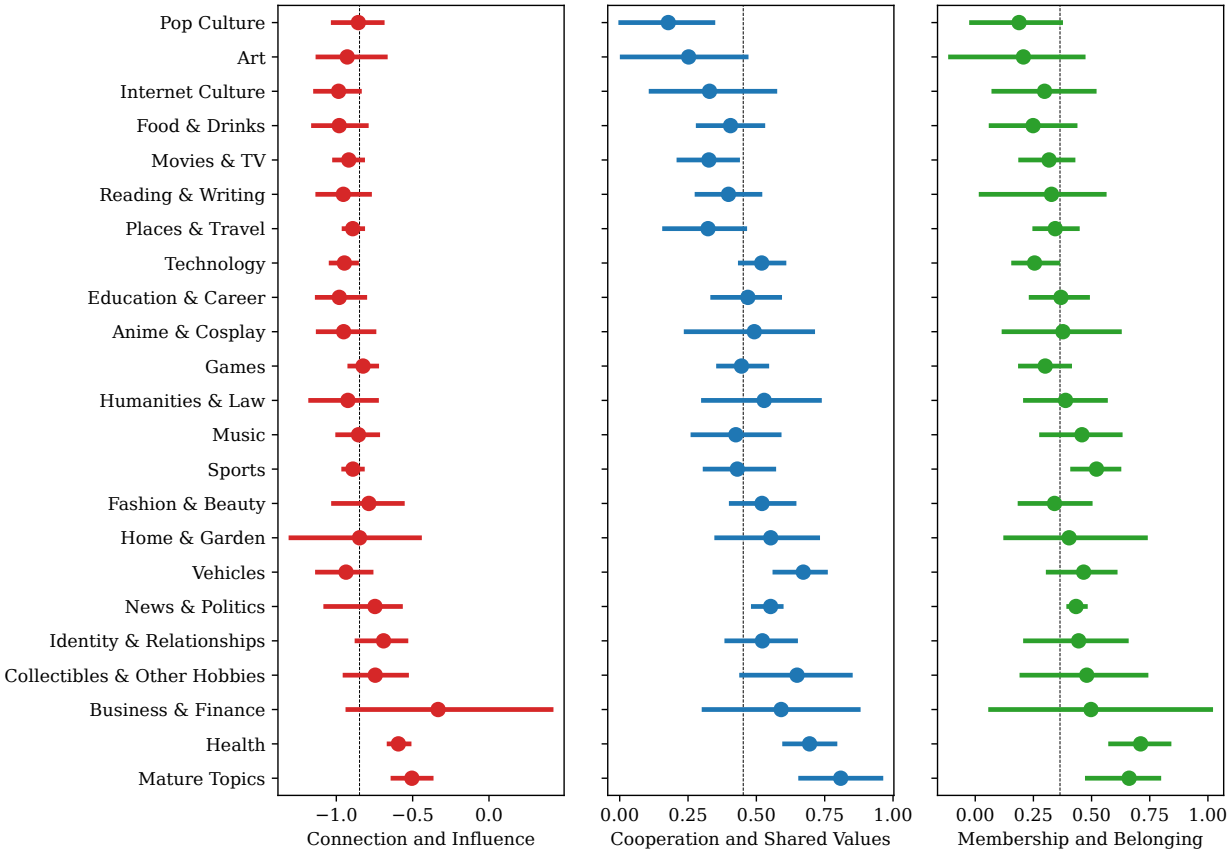


Figure C.1: Communities with different topics differ somewhat in their senses of virtual community. The dotted line shows the average SOVC score across all included in our study, while points show the average SOVC score for communities of that type, along with bootstrapped 95% confidence intervals. Sports communities have fairly typical scores for Connection and Influence and Cooperation and Shared Values, yet above average scores for Membership and Belonging. Health communities have above average scores for all three SOVC factors.

Appendix D

Additional Materials for Measuring Perceptions of Moderators

D.1 Moderator Sentiment Codebook

D.1.1 Positive Sentiment

This label should be used for comments expressing a positive sentiment towards the moderator or moderator team.

Examples

"This subreddit is so lucky to have such a great mod team"

"Make the life of our hard-working mods here easier"

"The mods are always so helpful, but this thread got a bit messy" — this is a tricky judgment call, but I'd say that the overall sentiment is positive with this thread being an exception.

Counterexample

"This subreddit used to be well-run, but in the past year or so the moderation has really gone to shit" — a judgment call similar to earlier, but here I would say *negative*.

D.1.2 Negative Sentiment

This label should be used for comments expressing a negative sentiment towards the moderator or moderator team.

Examples

“The mods here suck”

“The mods made a mistake” — everyone makes mistakes, but it’s still better if they don’t.

“I’m so tired of mods not removing crap like this”

D.1.3 Neutral Sentiment

This label should be used when there isn’t enough context for you to make a judgment about the sentiment of the comment or post, or the sentiment seems neutral.

Examples

“I didn’t delete the post, maybe the mods did?”

“Edit: reworded a slur after getting a warning from the mods”

“Mods please ban this person” — not enough explicitly stated sentiment to know what is meant with certainty.

“Why don’t you go and complain to the mods like you usually do?” — negative sentiment, but not directed towards the moderators.

D.2 Prompts for Topic and Sentiment Classification

D.2.1

The following few-shot prompt was used with GPT-4 [202] to classify the topic of subreddits in our analyses (§3.1.3), based on the name of the subreddit.

```
Given a subreddit, classify its topic into exactly one
of the following categories:

Hobby communities, which focus on a specific hobby,
sports related topic, or pastime.

Discussion communities, which are for discussion and
text-based content.

Media communities, which are for sharing videos,
and pictures.

News communities, which are for sharing news and
similar content.

Meme communites, which are for sharing memes or low
effort content.

Identity communities, which are for groups of a
spcecific identity or background.

/r/india: Identity
```

/r/bicycling ,: Hobby

/r/CrappyDesign: Media

/r/me_irl: Memes

/r/worldnews: News

/r/AskReddit: Discussion

/r/nba: Hobby

/r/relationship_advice: Discussion

/r/science: News

/r/teenagers: Identity

/r/dankmemes: Memes

/r/pics: Media

/r/{subreddit}:

D.2.2 Sentiment Classification Step Prompt

To classify the sentiment with regards to the moderators of posts and comments discussing mods (§3.1.3), we used the following prompt for our LLaMA 2 model, fine tuned with QLoRA [267, 56].

```
<|im_start|>system

Given a comment from Reddit which discusses moderators (or mods),
and its parent, identify how the author of the comment feels about
the moderators of the subreddit the comment was made in.

If the author feels that the moderators are doing a good job, mark
the sentiment as positive. If the author feels the moderators are
doing a bad job, mark the sentiment as negative. If it's not
possible to tell, mark the sentiment as neutral.


If the comment does not discuss moderators, but instead discusses
video game mods, or other types of modifications, mark the exclude
field as true and the sentiment as undefined.

If the comment is discussing moderators, but in a different
subreddit or different community, mark the other community
field as true.

The parent might be able to help your identification by providing
additional context.

Your answer should follow the format given in the examples.
<|im_end|>
```

```
<|im_start|>user
/r/{subreddit}
parent : {parent}
comment : {body}
<|im_end|>
```

D.3 Supplementary Figures

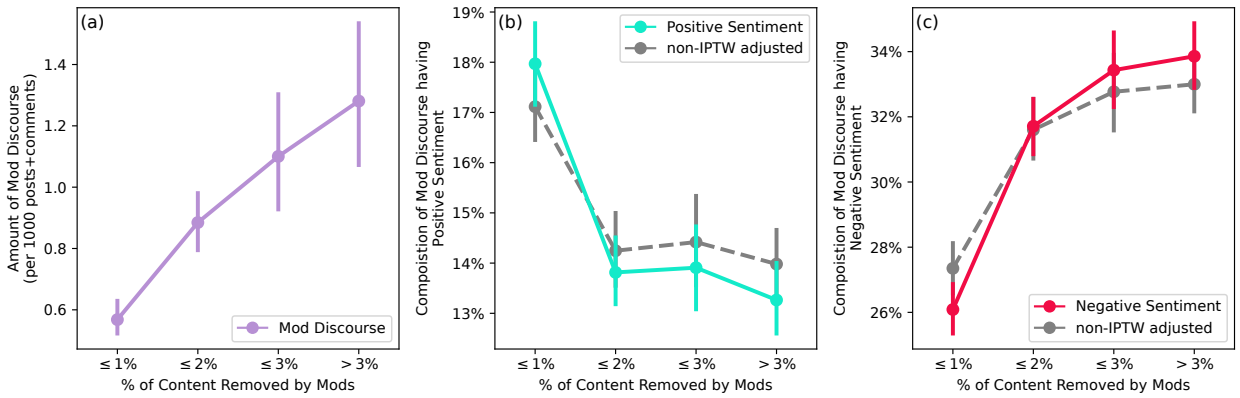


Figure D.1: On average, communities with a larger fraction of their content removed by mods tend to have more a smaller fraction of their mod discourse have positive sentiment (b), and a larger fraction with negative sentiment (c). Communities with more content removed by mods also tend to have more total mod discourse (a).

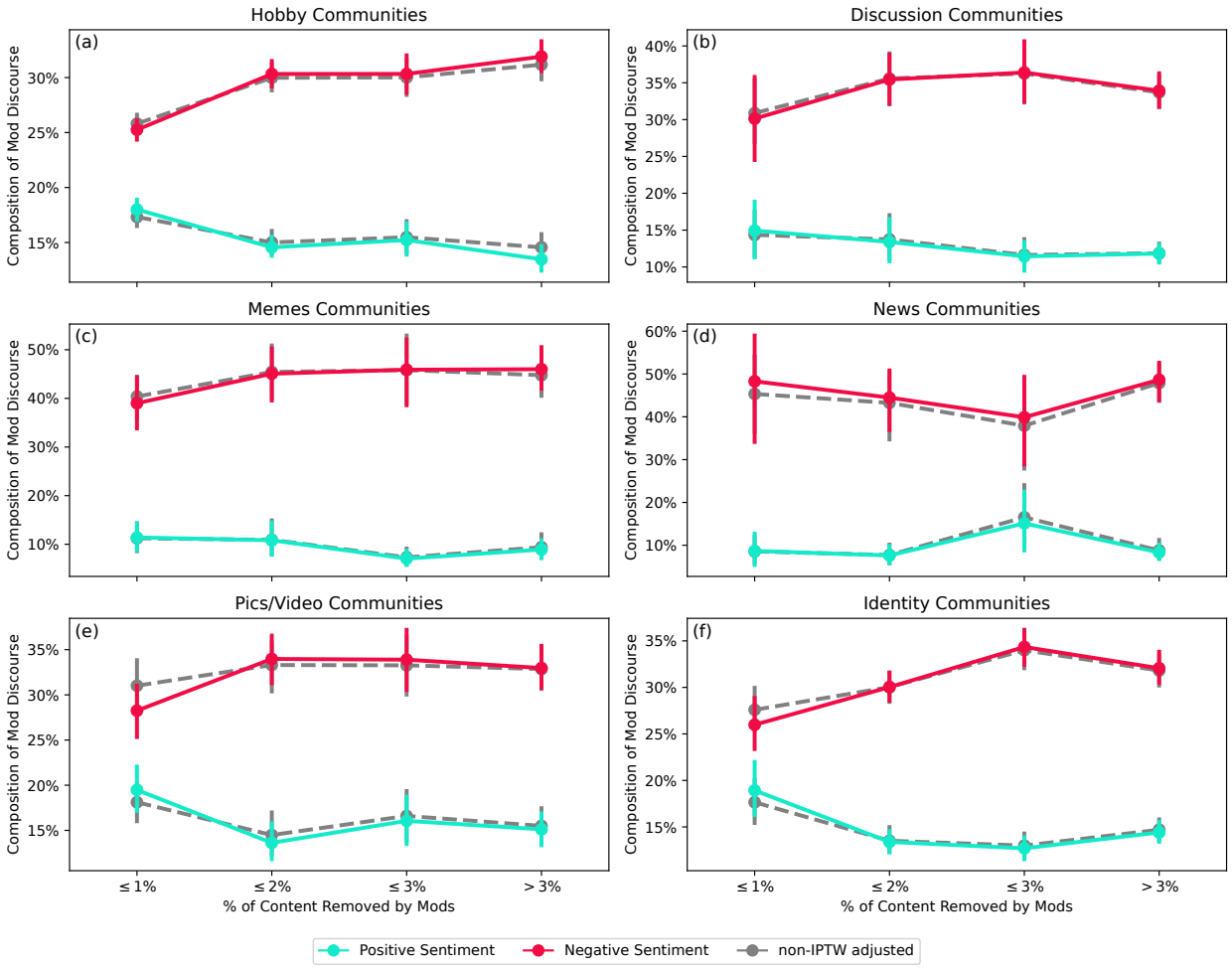


Figure D.2: Generally, the fraction of mod discourse with negative sentiment is higher in communities with more removed content. However, these trends vary depending on the topic of the community.

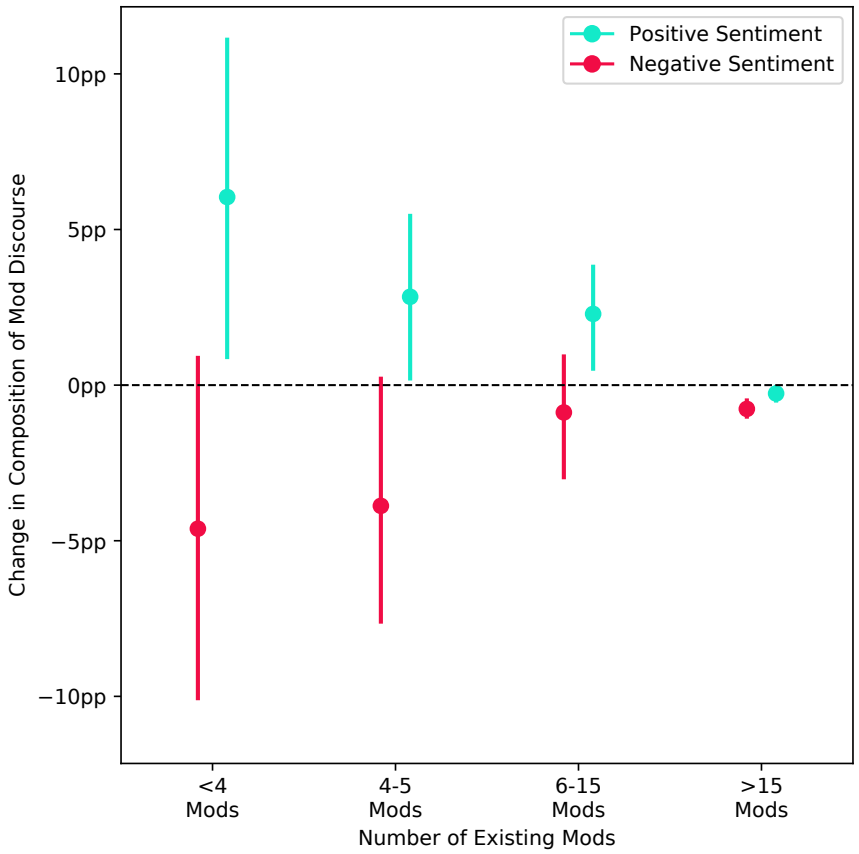


Figure D.3: On average, adding a new moderator to a subreddit results in an increase in the fraction of mod discourse which has positive sentiment, and a corresponding decrease in the fraction that has negative sentiment. However, the magnitude of the impact varies with the size of the community's moderator team; adding a single new mod to a community with fewer than 4 mods has a much larger impact than adding a single mod to a community with more mods. Adding a single moderator to a community with more than 15 mods has an impact which is not significantly different from 0.

D.4 Covariates Used in Propensity Score Modeling for IPTW

D.4.1 Covariates used for Mod Workload Analyses

The following table shows the covariates that were used in a logistic regression propensity score model for the moderator workload analyses (Figure 3.4), along with the resulting Standardized Mean Differences (SMDs) after reweighting, and the associated propensity score model (P.S.) coefficients used in the logistic regression.

	Moderator Workload Treatment Bin (Posts+Comments per Mod per Day)							
	0-5		5-10		10-100		>100	
Covariate	SMD	P.S. Coefficient	SMD	P.S. Coefficient	SMD	P.S. Coefficient	SMD	P.S. Coefficient
`total_items`	-0.18	-0.55	-0.17	-0.45	-0.10	+0.07	+0.48	+0.93
`frac_deleted`	+0.05	+0.01	+0.02	+0.02	-0.07	-0.07	+0.20	+0.04
`frac_removed`	+0.19	+0.04	+0.07	+0.01	-0.06	-0.05	-0.03	+0.00
`num_mods`	+0.04	+0.29	-0.02	+0.23	-0.03	-0.08	-0.01	-0.44
`category_hobby`	-0.22	-0.02	+0.01	+0.01	+0.06	+0.02	-0.07	-0.01
`category_discussion`	+0.18	+0.01	-0.02	-0.01	-0.05	-0.01	+0.04	+0.01
`category_memes`	+0.18	+0.03	+0.01	+0.00	-0.04	-0.05	+0.00	+0.01
`category_news`	+0.03	-0.01	+0.01	+0.00	-0.01	+0.01	+0.00	-0.01
`category_media`	+0.34	+0.02	+0.10	+0.00	-0.08	-0.04	-0.00	+0.01
`category_identity`	-0.24	-0.04	-0.10	-0.01	+0.05	+0.07	+0.07	-0.02

D.4.2 Covariates used for Strictness of Rule Enforcement Analyses

The following table shows the covariates that were used in a logistic regression propensity score model for the rule enforcement analyses (Figure 3.5), along with the resulting Standardized Mean Differences (SMDs) after reweighting, and the associated propensity score model (P.S.) coefficients used in the logistic regression.

	Amount of Removed Content Treatment Bin (Percentage of All Content)							
	0%-1%		1%-2%		2%-3%		>3%	
Covariate	SMD	P.S. Coefficient	SMD	P.S. Coefficient	SMD	P.S. Coefficient	SMD	P.S. Coefficient
total_items	-0.10	-0.12	+0.03	-0.10	+0.12	+0.01	+0.12	+0.21
frac_deleted	-0.58	-0.51	+0.02	-0.01	+0.30	+0.10	+0.67	+0.42
mod_workload	-0.08	-0.00	+0.04	+0.14	+0.15	+0.02	+0.04	-0.16
num_mods	-0.06	-0.21	-0.01	+0.12	-0.01	+0.01	+0.16	+0.08
category_hobby	+0.53	+0.10	-0.17	-0.01	-0.29	-0.01	-0.46	-0.08
category_discussion	-0.26	-0.09	+0.00	+0.01	+0.02	-0.01	+0.40	+0.09
category_memes	-0.04	+0.08	-0.01	-0.00	+0.04	+0.01	-0.02	-0.08
category_news	-0.11	-0.12	-0.06	-0.06	-0.00	-0.00	+0.28	+0.19
category_media	-0.12	+0.03	-0.08	-0.01	+0.02	-0.00	+0.21	-0.02
category_identity	-0.33	+0.00	+0.32	+0.08	+0.31	+0.02	+0.02	-0.10

Appendix E

Additional Materials for Analyses of Rules

E.1 Codebook

We provide a copy of our codebook, with further details and examples, as shared among authors for study reproduction. This codebook was created iteratively until saturation, with further detail found in §3.2.3.

E.1.1 Rule Tone

Prescriptive. A rule expressing general guidelines or desires for a community.

- Ex. 6. Include [REQUEST] in text posts for gif requests.
- Ex. Remember your reddiquette.

Restrictive. A rule which explicitly limits or forbids certain actions.

- Ex. Don't be an ass.
- Ex. Low effort & low quality content.

E.1.2 Rule Target

Rules generally target one or more of the following.

Post Content. Any rule explicitly stating desired or undesired content within the subreddit.

- Ex. 2. Text body must include context.
- Ex. OC is welcome.
- Ex. 1. No advertising, marketing research and spam.
- Ex. 5. Low effort/Blog spam/Repost - mods, please review.

User-Related. Any rule related to users, or which would cause unequal enforcement if two different users posted the same content. This includes verification and prior approval rules.

- Ex. Submissions from new accounts or from accounts with fewer than 10 karma will be removed.
- Ex. IamAs/AMAs must be approved by mods.

Post Format. Any rule prescribing post structure, formatting, titling, tagging or referencing a location to post (other subreddits, specific threads).

- Ex. 3. All News And Tweets Should Follow Standard Format.
- Ex. 2. R2: BootTooBig format.
- Ex. 10. Posts must begin with "ELI5:"
- Ex. 8. Use the stickied Megathreads If there are stickied Megathreads provided about certain topics they must be used when discussing that topic. Other posts will be removed.

Not a Rule. Sidebar content that is not necessarily a rule.

- Ex. Welcome to r/guitar, a community devoted to the exchange of guitar related information...
- Ex. Discuss all the Real Housewives franchises by Bravo TV with us...
- Ex. 9. Have fun!

E.1.3 Rule Topic

A rule can have any number of specific topics of focus.

Post Tagging & Flairing. Any rule pertaining to labeling/flairing, marking nsfw, using tags, etc.

- Ex. If your post is NSFW, please label it as such.
- Ex. Please add flair to your submissions.
- Ex. Use spoiler tags appropriately.

Peer Engagement. Any rule encouraging or discouraging the quantity of on-site peer engagement, including reporting, commenting, voting, and general activity.

- Ex. Please report any rule-breaking posts, as well as abusive comments or harassment.
- Ex. The downvote button is NOT a disagree button.
- Ex. Upvote Begging.
- Ex. If posting about weight loss, please provide details of your current plan and respond to questions.

Brigading. Any rule regulating large group actions, including raiding or mass-voting.

- Ex. This subreddit will not participate in or be the source of brigades or raids on other subreddits.
- Ex. Vote brigading.
- Ex. Use NP Links When Posting A Link To Other Subs.

Links & External Content. Any rule regulating links or the content of external resources. These rules do not regulate the content of images, but can regulate the host thereof.

- Ex. Do not submit a shortened link using a URL shortener like tinyurl.
- Ex. No promoting other subs/discords/crews.
- Ex. Do not mention or ask for usernames for other services.

Images. Any rule pertaining to the content or quality of images or videos, often including memes.

- Ex. While the places posted should be aesthetically unappealing, it is recommended that the photo quality is good. Artsy shots are more than welcome.
- Ex. Picture quality.
- Ex. Sourcing fan-art.
- Ex. Rule 4 - No memes, macros, low-effort, rant posts.

Commercialization. Any rule regulating advertisement, self-promotion, referrals and referral links, or buying/selling.

- Ex. Limit self-promotion.
- Ex. Do not advertise products or groups without permission.
- Ex. No VPN or Crypto-Currency Discussions Due to the commercial nature of VPNs...
- Ex. Rule 18 - No product/store/page/app review/rating.

Illegal Content. Rules about illegal content, including copyright infringement and piracy. Does not include harassment and hate speech.

- Ex. No talking about where to buy or get hold of MDMA or other illegal drugs.
- Ex. No links to pirated materials.
- Ex. No discussion of theft whatsoever.

Divisive Content. Rules regulating content which is inherently divisive, such as politics, current events, or community-specific hot topics. This only includes specific topics, i.e, the what, not the how.

- Ex. Ukraine Conflict.
- Ex. Religious preaching.

- Ex. Politics / Current Events.
- Ex. Do Not Editorialize.

Respect for Others. Rules explicitly discouraging hate speech, antagonization, harassment or encouraging respect for others. Also covers rules about swearing, which is often disrespectful.

- Ex. Any comments that include hate speech will be deleted.
- Ex. No hate speech.
- Ex. No profane/vulgar/undignified language.
- Ex. Social jerking.

Spam, Low Quality, Off-Topic, and Reposts. Any rule covering low-quality, off-topic, spam, reposted, or otherwise generally-undesired content.

- Ex. No advertising, marketing research and spam.
- Ex. Low effort/Blog spam/Repost - mods, please review!
- Ex. Common reposts and overdone topics will be removed.
- Ex. Embargoed Topics.
- Ex. Your post must be an unpopular opinion.

Ban Mentioned. Any rule that explicitly mentions banning.

- Ex. Malicious attempts to spoil other users will result in a ban.

Karma/Score/Voting Mentioned. Any rule that explicitly mentions karma, user scores, or voting.

- Ex. No Pandering for Upvotes.

E.2 Classification Prompt

The prompt was created collaboratively and iterated on to maximize classification performance across a manually-labeled validation set of 100 examples. Few-shot examples were chosen from a train-set of 200 randomly-sampled rules. All large-language models evaluated use the same prompt and sample selection technique. See §3.2.3 for details.

```
Given a rule in a specific subreddit, identify topics and qualities about
the rule.

If the rule explicitly limits or forbids certain actions, mark it as
"Restrictive".
If a rule expresses general guidelines or desires for a community, mark
it as "Prescriptive".
If a rule explicitly states desired or undesired content within the
subreddit, mark it as "Post Content".
If a rule is related to users or would cause unequal enforcement if two
different users posted the same content (including verification and prior
approval rules), mark it as "User-Related".
If a rule prescribes post structure, formatting, titling, or references a
location to post (such as other subreddits or specific threads), mark it
as "Post Format".
If the content is sidebar information and not necessarily a rule, mark it
as "Not a Rule".
If a rule pertains to labeling/flairing, marking nsfw, using tags, etc.,
mark it as "Post Tagging & Flairing".
If a rule encourages or discourages the quantity of on-site peer
engagement, including reporting, commenting, voting, and general
activity, mark it as "Peer Engagement".
If a rule regulates large group actions, including raiding or
mass-voting, mark it as "Brigading".
If a rule is about links or the external, off-reddit content (excluding
```

image content), mark it as "Links & External Content".
If a rule pertains to the content or quality of images or videos, often including memes, mark it as "Images".
If a rule regulates advertisement, self-promotion, referrals and referral links, or buying/selling, mark it as "Commercialization".
If a rule explicitly mentions illegal content, including copyright infringement and piracy (but excluding harassment and hate speech), mark it as "Illegal Content".
If a rule regulates content that is inherently divisive, such as politics, current events, or community-specific hot topics. , mark it as "Divisive Content". Only mark if the rule covers specific topics, i.e, the what being said, not how it was said.
If a rule discourages hate speech, antagonization, harassment, or encourages respect for others (including swearing), mark it as "Respect for Others".
If a rule covers low-quality, off-topic, spam, reposted, or otherwise generally-undesired content, mark it as "Spam, Low Quality, Off-Topic, and Reposts".
If a rule explicitly mentions banning users, mark it as "Ban Mentioned".
If a rule explicitly mentions karma or user scores, mark it as "Karma/Score Mentioned".
Your answer should follow the format given in the examples:

subreddit : AskReddit
rule : 3. Rule 3 - Open ended questions only
rule_description :
Prescriptive : False
Restrictive : True
Post Content : False
Post Format : True
User-Related : False
Not a Rule : False

```
Spam, Low Quality, Off-Topic, and Reposts : False
Post Tagging & Flairing : False
Peer Engagement : False
Links & External Content : False
Images : False
Commercialization : False
Illegal Content : False
Divisive Content : False
Respect for Others : False
Brigading : False
Ban Mentioned : False
Karma/Score Mentioned : False
```

... 5 more examples formatted as above

E.3 Additional Details on IPTW, Covariates, & Balance

As described in §3.2.3, we use Inverse Probability of Treatment Weighing (IPTW) [13] to adjust for two key confounding factors, community topic and size, in our analyses in §3.2.5. We use IPTW to reweight observations in the treatment (having a rule of a given type) and control (lacking a rule of a given type) groups to make their weighted distributions of covariates more similar to the reference distribution, consisting of the entire population [8]. For more details, see §3.2.3.

To understand the degree to which IPTW changes our experimental results from those computed without any confounding adjustment (non-adjusted/unweighted), the following figure shows both IPTW results (in blue and red) and non-adjusted results (in gray). In these analyses, IPTW mostly slightly decreases the estimated effect size while not zeroing it out. This suggests that while some observed differences between communities' perceptions of governance can be partially, but only partially, be explained by community size and topic, and therefore they suggest that rules also play an important role.

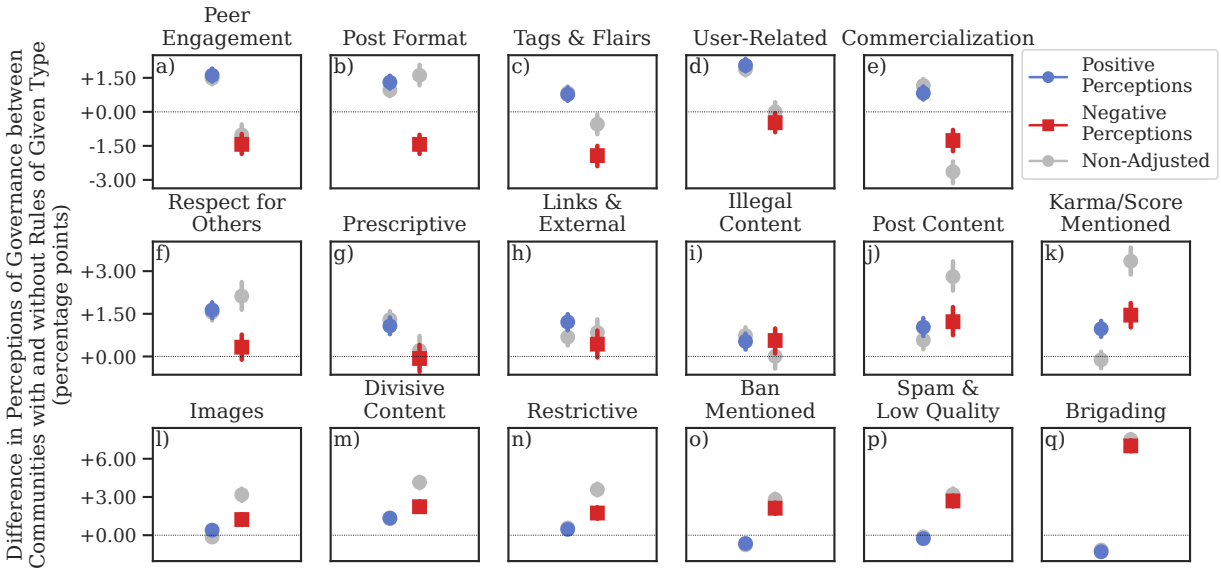


Figure E.1: For most rule types, adjusting for community topic and size (blue and red markers) slightly reduces the difference between communities with and without different rules, compared to no adjustment for confounding (gray markers). This suggests that topic size partially, but only partially, explain some difference in perceptions of governance, and that rules play an important rule. Additional discussion of these results is provided in §3.2.5.

For IPTW computation, we measure Community Size by computing the average number of posts and comments uploaded to a given community each day during the study period (April 2018 to December 2023).

Community Topic is classified into one of six different topical classes provided by an existing dataset [283] and are one-hot encoded for our logistic regression propensity score model. As such, the covariate vector $\mathbf{X}_i$ is of length 7.

We assess the efficacy of IPTW by examining the standard mean difference (SMD) between each weighted treatment/control group and the reference distribution, consisting of the entire population. Two groups are often considered 'balanced' or 'indistinguishable' if all covariates are within an SMD of 0.25 from the reference distribution [8]. Across 238 condition-covariate-rule type pairs (17 experiments $\times$ 2 treatment/control conditions $\times$ 7 covariates), our method achieves balance in 235 cases (98.74%). In the three cases where our method fails to achieve balance, no experiment has more than a single unbalanced covariate (out of seven), and no SMD exceeds 1.00. A complete list of covariates and SMDs for each experiment is given the in the tables below, with SMDs greater than 0.25 indicated with bold text.

E.3.1 Covariate Balance for Prescriptive Rules

Covariate	Mean of Reference Population	Mean of Control Group (no rule)	Mean of Treated Group (has rule)	SMD for Control	SMD for Treated
Community Size	1661.26	1289.02	1781.88	**0.46**	0.00
Discussion	0.11	0.11	0.11	0.02	0.01
Hobby	0.51	0.48	0.52	-0.02	0.00
Identity	0.17	0.17	0.17	0.00	0.00
Media	0.14	0.16	0.13	0.02	-0.01
Memes	0.05	0.05	0.05	-0.04	0.01
News	0.02	0.03	0.02	0.03	-0.01

E.3.2 Covariate Balance for Restrictive Rules

Covariate	Mean of Reference Population	Mean of Control Group (no rule)	Mean of Treated Group (has rule)	SMD for Control	SMD for Treated
Community Size	1661.26	379.22	1798.68	-0.04	0.00
Discussion	0.11	0.11	0.11	0.10	0.00
Hobby	0.51	0.53	0.51	0.00	0.00
Identity	0.17	0.14	0.17	-0.04	0.00
Media	0.14	0.15	0.14	0.00	0.00
Memes	0.05	0.04	0.05	-0.08	0.00
News	0.02	0.03	0.02	-0.01	0.00

E.3.3 Covariate Balance for Post Content Rules

Covariate	Mean of Reference Population	Mean of Control Group (no rule)	Mean of Treated Group (has rule)	SMD for Control	SMD for Treated
Community Size	1661.26	258.60	1790.65	-0.12	0.00
Discussion	0.11	0.10	0.11	0.03	0.00
Hobby	0.51	0.52	0.51	0.04	0.00
Identity	0.17	0.15	0.17	-0.01	0.00
Media	0.14	0.16	0.14	-0.02	0.00
Memes	0.05	0.04	0.05	-0.11	0.00
News	0.02	0.03	0.02	0.04	0.00

E.3.4 Covariate Balance for Post Format Rules

Covariate	Mean of Reference Population	Mean of Control Group (no rule)	Mean of Treated Group (has rule)	SMD for Control	SMD for Treated
Community Size	1661.26	707.31	2659.29	0.14	0.01
Discussion	0.11	0.11	0.12	0.05	0.01
Hobby	0.51	0.51	0.50	-0.05	-0.01
Identity	0.17	0.18	0.16	-0.02	-0.01
Media	0.14	0.13	0.15	0.03	0.01
Memes	0.05	0.04	0.05	0.05	0.00
News	0.02	0.02	0.03	-0.02	0.00

E.3.5 Covariate Balance for User-Related Rules

Covariate	Mean of Reference Population	Mean of Control Group (no rule)	Mean of Treated Group (has rule)	SMD for Control	SMD for Treated
Community Size	1661.26	1540.82	2028.53	0.04	0.03
Discussion	0.11	0.10	0.15	0.00	0.01
Hobby	0.51	0.52	0.48	-0.01	0.01
Identity	0.17	0.16	0.19	0.00	-0.01
Media	0.14	0.14	0.12	0.00	-0.01
Memes	0.05	0.05	0.03	0.00	0.00
News	0.02	0.02	0.02	0.00	-0.01

E.3.6 Covariate Balance for Spam, Low Quality, Off-Topic, and Reposts Rules

Covariate	Mean of Reference Population	Mean of Control Group (no rule)	Mean of Treated Group (has rule)	SMD for Control	SMD for Treated
Community Size	1661.26	502.99	1992.24	0.01	0.00
Discussion	0.11	0.14	0.10	0.13	0.01
Hobby	0.51	0.51	0.51	-0.05	-0.01
Identity	0.17	0.17	0.17	0.01	0.00
Media	0.14	0.13	0.14	-0.03	0.00
Memes	0.05	0.03	0.05	-0.02	0.00
News	0.02	0.02	0.03	-0.04	0.01

E.3.7 Covariate Balance for Post Tagging & Flairing Rules

Covariate	Mean of Reference Population	Mean of Control Group (no rule)	Mean of Treated Group (has rule)	SMD for Control	SMD for Treated
Community Size	1661.26	1201.43	3035.81	0.02	0.01
Discussion	0.11	0.13	0.07	0.00	0.00
Hobby	0.51	0.50	0.55	0.00	0.00
Identity	0.17	0.19	0.10	0.00	0.00
Media	0.14	0.12	0.20	0.01	0.00
Memes	0.05	0.04	0.06	-0.01	0.00
News	0.02	0.03	0.02	0.03	-0.01

E.3.8 Covariate Balance for Peer Engagement Rules

Covariate	Mean of Reference Population	Mean of Control Group (no rule)	Mean of Treated Group (has rule)	SMD for Control	SMD for Treated
Community Size	1661.26	1328.24	2373.23	-0.02	0.00
Discussion	0.11	0.10	0.14	0.00	0.00
Hobby	0.51	0.50	0.53	0.00	0.00
Identity	0.17	0.17	0.17	0.00	-0.01
Media	0.14	0.16	0.10	0.00	0.01
Memes	0.05	0.05	0.04	0.00	0.00
News	0.02	0.03	0.02	0.01	0.00

E.3.9 Covariate Balance for Links & External Content Rules

Covariate	Mean of Reference Population	Mean of Control Group (no rule)	Mean of Treated Group (has rule)	SMD for Control	SMD for Treated
Community Size	1661.26	1414.87	2026.90	0.21	0.01
Discussion	0.11	0.13	0.08	0.03	0.00
Hobby	0.51	0.50	0.52	-0.01	0.01
Identity	0.17	0.17	0.17	0.00	-0.01
Media	0.14	0.14	0.14	-0.01	0.00
Memes	0.05	0.05	0.05	0.00	-0.01
News	0.02	0.02	0.04	0.00	0.00

E.3.10 Covariate Balance for Images Rules

Covariate	Mean of Reference Population	Mean of Control Group (no rule)	Mean of Treated Group (has rule)	SMD for Control	SMD for Treated
Community Size	1661.26	1043.11	3110.61	0.19	0.01
Discussion	0.11	0.14	0.05	-0.01	0.04
Hobby	0.51	0.52	0.49	-0.02	-0.01
Identity	0.17	0.19	0.11	-0.01	-0.01
Media	0.14	0.10	0.22	-0.01	0.00
Memes	0.05	0.03	0.10	-0.03	0.00
News	0.02	0.02	0.03	0.17	-0.01

E.3.11 Covariate Balance for Commercialization Rules

Covariate	Mean of Reference Population	Mean of Control Group (no rule)	Mean of Treated Group (has rule)	SMD for Control	SMD for Treated
Community Size	1661.26	1384.51	1954.49	-0.01	0.00
Discussion	0.11	0.15	0.08	0.00	0.00
Hobby	0.51	0.38	0.65	0.00	0.01
Identity	0.17	0.16	0.18	0.00	0.01
Media	0.14	0.20	0.07	0.00	0.01
Memes	0.05	0.08	0.01	0.01	-0.05
News	0.02	0.03	0.01	0.00	0.00

E.3.12 Covariate Balance for Illegal Content Rules

Covariate	Mean of Reference Population	Mean of Control Group (no rule)	Mean of Treated Group (has rule)	SMD for Control	SMD for Treated
Community Size	1661.26	1596.03	1856.06	0.02	0.02
Discussion	0.11	0.12	0.10	0.00	0.03
Hobby	0.51	0.48	0.59	0.00	-0.01
Identity	0.17	0.17	0.17	0.00	0.01
Media	0.14	0.15	0.11	0.00	0.02
Memes	0.05	0.06	0.03	0.01	-0.04
News	0.02	0.03	0.01	0.01	-0.05

E.3.13 Covariate Balance for Divisive Content Rules

Covariate	Mean of Reference Population	Mean of Control Group (no rule)	Mean of Treated Group (has rule)	SMD for Control	SMD for Treated
Community Size	1661.26	1316.22	2772.01	**0.96**	0.04
Discussion	0.11	0.10	0.16	0.09	0.01
Hobby	0.51	0.56	0.34	-0.03	0.01
Identity	0.17	0.14	0.25	-0.01	-0.01
Media	0.14	0.14	0.12	-0.01	-0.02
Memes	0.05	0.04	0.08	-0.02	0.01
News	0.02	0.02	0.04	-0.02	0.03

E.3.14 Covariate Balance for Respect for Others Rules

Covariate	Mean of Reference Population	Mean of Control Group (no rule)	Mean of Treated Group (has rule)	SMD for Control	SMD for Treated
Community Size	1661.26	824.14	1934.97	**0.28**	0.00
Discussion	0.11	0.10	0.12	0.15	0.00
Hobby	0.51	0.53	0.50	-0.07	0.01
Identity	0.17	0.10	0.19	-0.06	0.00
Media	0.14	0.18	0.12	0.01	-0.02
Memes	0.05	0.06	0.05	0.03	-0.01
News	0.02	0.02	0.02	-0.02	0.00

E.3.15 Covariate Balance for Brigading Rules

Covariate	Mean of Reference Population	Mean of Control Group (no rule)	Mean of Treated Group (has rule)	SMD for Control	SMD for Treated
Community Size	1661.26	1461.99	3498.79	0.07	0.11
Discussion	0.11	0.11	0.16	0.01	0.02
Hobby	0.51	0.52	0.43	0.00	-0.01
Identity	0.17	0.16	0.20	0.00	-0.01
Media	0.14	0.14	0.08	0.00	0.01
Memes	0.05	0.04	0.10	-0.01	0.00
News	0.02	0.02	0.02	0.00	-0.01

E.3.16 Covariate Balance for Ban Mentioned Rules

Covariate	Mean of Reference Population	Mean of Control Group (no rule)	Mean of Treated Group (has rule)	SMD for Control	SMD for Treated
Community Size	1661.26	1556.12	3432.80	0.01	0.08
Discussion	0.11	0.11	0.16	0.00	0.01
Hobby	0.51	0.51	0.43	0.00	0.05
Identity	0.17	0.17	0.15	0.00	-0.04
Media	0.14	0.14	0.15	0.00	-0.02
Memes	0.05	0.05	0.08	0.00	-0.02
News	0.02	0.02	0.03	0.00	-0.01

E.3.17 Covariate Balance for Karma/Score Mentioned Rules

Covariate	Mean of Reference Population	Mean of Control Group (no rule)	Mean of Treated Group (has rule)	SMD for Control	SMD for Treated
Community Size	1661.26	1519.41	4628.85	0.00	0.03
Discussion	0.11	0.11	0.13	0.00	0.05
Hobby	0.51	0.51	0.45	0.00	0.01
Identity	0.17	0.17	0.06	0.00	-0.06
Media	0.14	0.14	0.14	0.00	0.00
Memes	0.05	0.04	0.20	-0.01	0.02
News	0.02	0.02	0.02	0.00	-0.01

Bias	# of Links	# of News Sources
Extreme Left	15,157	51
Left	3,023,382	364
Center Left	17,648,711	544
Center	4,494,687	442
Center Right	4,254,705	263
Right	3,226,828	352
Extreme Right	997,703	423
Unlabeled	*525,443,378*	

Table F.1: Numbers of links and unique news sources in our dataset, by the political bias of the link.

Factualness	# of Links	# of News Sources
Very Low	609,229	72
Low	749,202	369
Mixed	7,116,130	677
Mostly	2,217,719	110
High	22,055,943	1,313
Very High	2,263,604	134
Unlabeled	*524,092,724*	

Table F.2: Numbers of links and unique news sources in our dataset, by the factualness of the link.

Appendix F

Additional Materials for News Sharing Analyses

F.1 Dataset Summary

Our dataset was created from all public Reddit submissions posted between January 2016 and August 2019, the most recent data available at the time of this study. These submissions were downloaded using the Pushshift archives [20], and consist of 580 million submissions, 35 million unique authors, and 3.4 million unique subreddits. As each submission may consist of 0 or more links, the dataset includes a total of 559 million links. These links are to 5.1 million unique domains, of which we are able to label 2,801 unique domains with annotations from MBFC.

Table F.1 shows the number of links in the dataset, as well as the number of unique news sources, for each bias category.

Table F.2 shows the number of links in the dataset, as well as the number of unique news sources, for each factualness category.

The dataset may be downloaded from our website at `https://behavioral-data.github.io/news_labeling_reddit/`