AI Rules? Characterizing Reddit Community Policies Towards AI-Generated Content

TRAVIS LLOYD, Cornell Tech, USA
JENNAH GOSCIAK, Cornell University, USA
TUNG NGUYEN, Cornell Tech, USA
MOR NAAMAN, Cornell Tech, USA

How are Reddit communities responding to AI-generated content? We explored this question through a large-scale analysis of subreddit community rules and their change over time. We collected the metadata and community rules for over 300, 000 public subreddits and measured the prevalence of rules governing AI. We labeled subreddits and AI rules according to existing taxonomies from the HCI literature and a new taxonomy we developed specific to AI rules. While rules about AI are still relatively uncommon, the number of subreddits with these rules more than doubled over the course of a year. AI rules are more common in larger subreddits and communities focused on art or celebrity topics, and less common in those focused on social support. These rules often focus on AI images and evoke, as justification, concerns about quality and authenticity. Overall, our findings illustrate the emergence of varied concerns about AI, in different community contexts. Platform designers and HCI researchers should heed these concerns if they hope to encourage community self-determination in the age of generative AI. We make our datasets public to enable future large-scale studies of community self-governance.

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

Additional Key Words and Phrases: Online Communities, Reddit, Rules, AIGC, Generative AI, Governance, AI-Generated Content, Moderation

1 Introduction

The emergence of mainstream generative AI tools capable of producing compelling text, images, and videos—collectively referred to as AI-Generated Content (AIGC)—poses a unique challenge for online platforms that are dependent on User-Generated Content. Many such platforms have enacted top-down changes to respond to this new technology, with some deciding to ban it [43], while others are encouraging its use or even incorporating generative AI features into their user interfaces [42, 45]. Other platforms have declined to make top-down changes and have instead left it up to their users to come to their own positions on this technology. One such platform is Reddit, a social sharing and news aggregation site, where individual communities, known as subreddits, are free to enact their own rules and policies [17, 19, 58, 65]. This decentralized moderation approach allows communities to craft policies towards generative AI that are contextually-relevant: the policies align with how communities see the technology being used, and whether or not that use aligns with their values. Decentralized moderation also provides a rich source of data through which HCI researchers interested in helping online communities navigate sudden technological change can better understand the range of attitudes and stances that exist across communities who vary in size, purpose, and values.

Findings from past HCI studies have emphasized the important role that explicit and consistent community rules play in helping online communities navigate moments of intense change [35]. These rules interact with other forms of governance [40], such as norms [7] and technical affordances [31], to encourage certain behaviors and discourage others [36].

Authors' Contact Information: Travis Lloyd, tgl33@cornell.edu, Cornell Tech, New York, NY, USA; Jennah Gosciak, jrg377@cornell.edu, Cornell University, Ithaca, NY, USA; Tung Nguyen, tn375@cornell.edu, Cornell Tech, New York, NY, USA; Mor Naaman, mor.naaman@cornell.edu, Cornell Tech, New York, NY, USA.

Of course, generative AI and AIGC may impact the rules [43], dynamics [15, 26], and governance [6] strategies of online communities in complex ways. While some recent research has explored how generative AI might help online communities deal with problematic content [1, 39], other work has documented the ability of AIGC to pollute these same communities with deceptive information [52]. Recent work by Lloyd et al. [41] used interviews with Reddit moderators to explore the particular challenges of governing AI use in online communities and found that communities are enacting explicit rules about AIGC in order to clarify norms. Lloyd et al.'s work focuses on the perspectives and experiences of their interviewees and raises the question of how their findings generalize across the wide variety of communities on Reddit.

This paper addresses that question through a large-scale quantitative analysis of community rules on Reddit. Our approach allows us to paint a fuller picture for the HCI community by complementing Lloyd et al.'s qualitative findings with an analysis of platform-wide patterns and trends, while simultaneously providing deeper insight into the specific types of AI rules that communities are creating. We build on techniques developed by Fiesler et al. [17] to study the ecosystem of community rules on Reddit and adapt their methodology to specifically investigate the emergence of community rules governing AI use. We use Fiesler et al.'s community topic and rule taxonomies so that we can compare our analysis of AI rules to their analysis of general Reddit rules. We address the following research questions:

- **RQ1:** How common are subreddit rules governing the use of AI and how has their prevalence changed over time?
- **RQ2:** What types of communities are more likely to have these rules?
- **RQ3:** What types of rules exist for governing the use of AI?

To answer these questions we crawled over 300,000 public subreddits on two occasions, in July 2023 and November 2024, building a dataset of $N = 99,969$ English-language subreddits with published rules. Analyzing a longitudinal sample of subreddits that were present during both our crawls, our inquiry into **RQ1** finds that while only 1.2% of subreddits now have rules about AI, the number of subreddits with AI rules has more than doubled since July 2023. Larger subreddits are much more likely to have AI rules: for example, of the top 1% largest subreddits, 17% had AI rules in our most recent crawl. To explore **RQ2**, we perform subreddit-level analysis on all subreddits in our 2024 sample with rules. We apply a "Topic" label to these subreddits using a taxonomy from Fiesler et al. [17] and "Community Archetype" labels using a taxonomy from Prinster et al. [57]. We use quantitative methods to explore the relationship between these labels and the presence of AI rules and find these rules to be most common in communities with *Art* and *Celebrity* topics, as well as those that fit the *Content Generation* Community Archetype. To explore **RQ3**, we perform rule-level analysis on the rules in our sample that we identified as governing AI use. We apply an existing rule taxonomy from Fiesler et al. [17] and additionally produce our own set of emergent labels to describe meaningful dimensions of variation in both the requirements imposed by these rules and the language through which they frame AI usage. We find rules about AI images to be more common than rules about other types of AIGC and see quality and authenticity as the most frequently expressed concerns.

We compare these findings to the results of related work and discuss implications for platform designers. In particular, thoughtful design choices may help address the community concerns about quality and authenticity that our findings identify. Our findings also show, though, that communities of different types take varied approaches towards regulating AI use, and suggest that design solutions meant to help communities address these challenges must be sensitive to community context.

As an additional contribution, we make our community rules data and labels public. While we limit our current analysis to questions about AI rules, our data can be used to answer a wide range of questions about content moderation in self-moderating online communities. We hope that this will enable large-scale analysis by future HCI researchers interested in encouraging online community self-determination.

2 Related Work

Our study is motivated by two bodies of work from the HCI literature. The first set of related work includes empirical research into online communities' rules, norms, and values. This work demonstrates that online communities hold a wide range of values and emphasizes that rules and norms are always context-dependent. The second body of work concerns current and potential future impacts of generative AI on online communities. This research shows that communities are already grappling with the changes brought by generative AI and are adapting their policies and practices in response.

2.1 Online Communities' Rules, Norms, and Values

Online communities play an important role in social life [53], and occupy a complicated position as a site of nuanced interactions between people, technology, and policy [28]. This importance and complexity have made online communities a frequent topic of study in the HCI literature [38, 55]. HCI scholars have explored the variety of purposes that these communities can serve and how platform design choices may influence their success. For example, Preece introduced the idea of *Empathic Communities* [54], in which people come together to share and receive support, and suggested platform design guidelines to encourage sociality [56]. In addition to encouraging pro-social behavior, a frequent topic of study has been how to discourage anti-social behavior [36, 40]. Much of this work has focused on the benefits [22] and challenges [37] of effective and fair content moderation. The HCI community has been interested in top-down moderation [20], in which platforms set and enforce rules, but even more interested in bottom-up, or self-moderation, in which communities do [63]. A major finding of this research is the importance of context sensitivity in moderation decisions, which self-moderation is often more attuned to. Taken together, this research suggests that studying self-moderation decisions may provide insight for design solutions that can promote community self-determination at a time of sudden technological change.

Reddit is a fruitful platform for a study of community self-moderation because it is home to online communities of various size and purpose. This heterogeneity, combined with its largely public and text-based nature, have made Reddit a frequent topic of study in online communities research [58]. Recent work has explored the values [69], norms [7], and rules [17] of different communities across the platform, both via direct methods like surveys, and indirect methods such as analyzing digital trace data from these communities. Reddit's implementation of community self-moderation allows communities to write and make public their own community-enforced rules, which Fiesler et al. collect and characterize in a study that is the main methodological inspiration for our work [17]. In their study, Fiesler et al. find subreddit size to be the main predictor of having rules, but note that even among subreddits of similar sizes rules are context-dependent, not uniform, and often uniquely tailored to the topic of the subreddit (e.g., art subreddits have more rules about copyright). The authors also highlight that setting norms and developing rules is fundamentally a social process. Overall, Fiesler et al. show the importance of community rules and the feasibility of using them as a data source to study attitudes and stances across a platform. Given the potential impact of AI on online communities, we use community rules as a starting point to study context-specific attitudes towards AIGC as they emerge across Reddit.

2.2 Generative AI's Impact on Online Communities

A growing body of research in HCI, Social Computing, and adjacent behavioral science fields has documented ways in which generative AI may impact online communities through increased deception, decreased trust, and new forms of harassment. Studies have demonstrated that AI can effectively deceive humans [52] and have raised concerns that generated text could be weaponized for deception by either humans [21] or automated social media accounts (bots) [46]. Additionally, just the *suspicion* that other online community members are using AI may be enough to impact dynamics in online communities. Several studies in AI-Mediated Communication (AIMC) [24] provide evidence that the perception that another online community member is using AI causes members to regard each other more negatively [59] and trust each other less [29]. Finally, HCI research has long documented how online communities can be sites of harassment and abuse [8, 23, 32, 51]. The impact of AI-powered harassment on online communities, such as malicious synthetic, or deepfake, media [13] is much less studied [66], but there is concern that this extremely harmful phenomenon may grow in prevalence alongside the popularity of generative AI tools [62].

Many additional studies have started to document the impacts of AIGC on online communities "in the wild". Recent work has identified undisclosed AI use in several settings, such as LLM-powered social bots on Twitter [70], AIGC in fake online reviews [49], and spam content used for scams [12]. In addition to studies of deceptive use, other recent work has begun to explore how the disclosed use of AIGC may affect community dynamics, such as in art-sharing communities [68]. Generative AI can additionally impact online communities by drawing traffic away from communities and decreasing overall engagement: causal inference studies of Q&A sites in the Stack Exchange network, where users gather to ask and answer questions about programming and related topics, have demonstrated that the launch of ChatGPT caused an overall decline in website visits and question volume [4]. The authors of this study contrast this decline with a measured null-effect on similar Reddit communities, which they suggest may avoid a decline in traffic because of their emphasis on socialization, rather than pure information exchange. Finally, recent work by Lloyd et al. used interviews to qualitatively explore the attitudes and experiences of Reddit moderators dealing with AIGC [41]. This study interviewed moderators from subreddits of different sizes and topics and surfaces a variety of community objections to AIGC that range from practical to ideological. The work poses open questions about how prevalent these concerns are across the platform and suggests that larger trends could be explored quantitatively in future work. Our study fills that gap through a quantitative analysis that helps identify the most common rules and the most impacted communities, so that the HCI community can begin the task of addressing these issues with context-specific solutions.

3 Methods

In order to perform a large-scale analysis of community rules governing the use of AI, we collected the metadata of public, English-language subreddits. Building on a large-scale crawl in 2024, and an earlier one performed in 2023, we created five datasets of subreddits and their associated rules. We processed the more recent crawl to create four datasets: the **Broad Subreddit Set** ($N = 307,543$) consisting of the metadata of subreddits with and without rules; the **Rules Subreddit Set** ($N = 99,969$) of only subreddits with stated, published rules; the **AI Rules Subreddit Set** ($N = 4,251$) of subreddits with rules governing the use of AI; and the **AI Rules Set** ($N = 4,458$) of individual rules governing AI use. The process of constructing these datasets is summarized in Figure 1, and detailed below. We supplement the raw crawl data with additional information about subreddit and rule types through several rounds of labeling, using both human and LLM coders. We use the **AI Rules Set** and emergent coding to construct a novel taxonomy of AI Rule Types. In

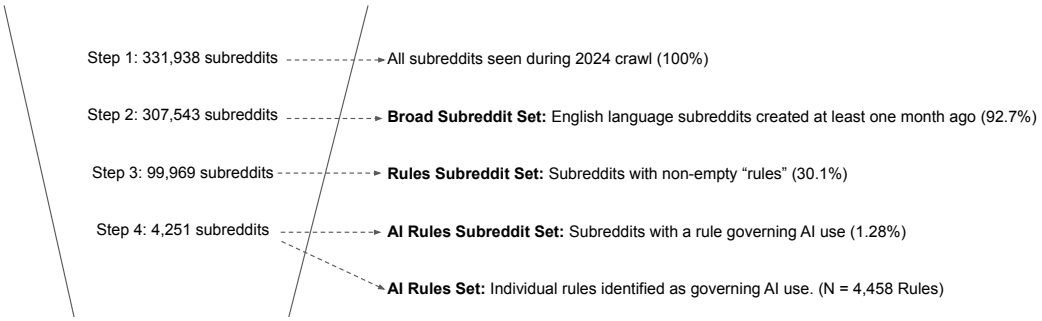

Fig. 1. The process for collecting subreddit metadata and building the datasets used in our analysis: the **Broad Subreddit Set** of English-language subreddits, the **Rules Subreddit Set** of subreddits with non-empty rules, the **AI Rules Subreddit Set** of subreddits with rules governing the use of AI, and the **AI Rules Set** of individual rules governing AI use.

addition, we created a longitudinal dataset of subreddits that were present in both the 2023 and 2024 crawls. We use this **Longitudinal Subreddit Set** ($N = 227,737$) to compare the prevalence of AI rules at different points in time for a consistent sample of subreddits. We will make our datasets and labels public upon completion of peer-review.

3.1 The Broad Subreddit and Rules Subreddit Sets

In order to better understand how online communities across Reddit are responding to the arrival of AIGC, we conducted a large-scale data collection and analysis. In November 2024 we performed an automated crawl of Reddit's "Top Communities" list[1], which includes public subreddits with more than ten subscribers. We used a web-scraping library[2] to paginate through this list and extract the IDs of all listed subreddits from the page HTML. We then used Reddit's Data API[3] to fetch metadata and rules for each subreddit ID.

From the $N = 331,938$ subreddits that we saw in our crawl, we used a process summarized in Figure 1 to create several datasets for analysis. We used an automated tool[4] to determine the language of each subreddit based on the text in its description and rules metadata. We removed subreddits with a primary language other than english or a creation date less than one month before the crawl date, in order to remove communities that had not yet had time to create rules. This yielded the **Broad Subreddit Set** of $N = 307,543$ public, English-language subreddits that were at least one month old, which we used to understand platform-wide patterns in the data and factors that contribute to whether subreddits have rules at all.

Using this broad dataset, we constructed a second, smaller dataset: the **Rules Subreddit Set** ($N = 99,969$). This dataset consists only of subreddits with non-empty `rules` metadata. We use this dataset to understand patterns specific to subreddits with rules. These two datasets are the focus of our descriptive analysis and were used to create two other datasets specific to *AI rules*, which are discussed in Sections 3.3.4 and 3.3.2.

[1]https://www.reddit.com/best/communities/1/
[2]https://docs.scrapy.org/en/latest/
[3]https://support.reddithelp.com/hc/en-us/articles/16160319875092-Reddit-Data-API-Wiki
[4]https://pypi.org/project/langdetect/

3.2 Longitudinal Subreddit Set

To address RQ1's concern with the shift in rule prevalence over time, we constructed a **Longitudinal Subreddit Set**, using a July 2023 data collection effort as a baseline, and identifying the subset of subreddits from the 2024 crawl that were seen in the earlier crawl. ChatGPT launched at the end of 2022, kicking off a period of increased public awareness and usage of AI. Given the dynamic nature of such a moment of flux, we decided to perform a longitudinal analysis on the same set of subreddits over a period of time, so that we could study the evolution of community norms and practices soon after this launch.

The 2023 crawl used the same technique as the 2024 crawl, capturing subreddits in Reddit's "Top Communities" list as described above. In the 2023 crawl, we scraped the subreddit metadata and rules data directly from the web page of each subreddit, using the same tool that we used to visit the "Top Communities" list—a simpler approach that was possible under Reddit's prior rate limit. We verified that the data that we gathered from a subreddit's web page is identical to the data returned by the API, which we used for the 2024 crawl. This 2023 crawl scraped $N = 337,399$ subreddits in total. From this 2023 dataset, we identified the subreddits whose primary language was English, and that were also present in the 2024 crawl. This process yielded the **Longitudinal Subreddit Set** ($N = 227,737$) which we used to study the change in prevalence of AI rules over time.

3.3 Data Labeling

To answer our research questions about rules governing the use of AI, we created multiple label categories for the AI rules and the communities that had rules. Through the process described below, we used a hybrid human-LLM coding technique to determine whether each of the subreddits in the **Rules Subreddit Set** and the **Longitudinal Subreddit Set** had an explicit rule about AI use, and to label them accordingly. We used these labels to identify the subset of subreddits with AI rules. We refer to the subset of the **Rules Subreddit Set** with AI rules as the **AI Rules Subreddit Set** ($N = 4,251$). The AI-related rules from this set became our **AI Rules Set** ($N = 4,458$). We used an iterative inductive coding technique to identify common types of AI rules. We then used a hybrid human-LLM coding technique to label the rules according to these categories.

Below, we describe the label categories we used to categorize subreddit AI rules (summarized in Table 1), as well as the human-LLM process that we used to label the data. We also describe how we determined and labeled the community type for each subreddit in the **Rules Subreddit Set**, according to several frameworks from existing literature (summarized in Table 2).

3.3.1 Detecting Community Rules Governing AI Use. Community rules are often very long and complex, which made our datasets too large to manually code for the presence of rules governing the use of AI. To get a rough sense of the prevalence of these rules, we first used a simple regex to search the subreddits' rules for AI-related keywords. This step found matches in the rules of less than 2% of the subreddits. We then devised a protocol, inspired by recent HCI work using LLMs for qualitative labeling tasks [2, 9, 60], to detect AI-related rules across the larger datasets with more precision. The approach is based on comparing LLM-produced labels with those produced by human experts and iterating on the LLM prompt until desirable inter-rater reliability (IRR) is achieved between the human and LLM labels.

In order to assess LLM performance on a sample with both positive and negative examples, the first two authors randomly selected 100 subreddits from the **Rules Subreddit Set** that matched the AI-keyword regex. The authors then independently applied a "Yes/No" label to each subreddit in response to the following question: *"Do any of this subreddit's rules explicitly govern the use of AI tools or the content that such tools produce?"* The two authors achieved strong agreement in their codes ($IRR > .8$) and only one human coder was used in future iterations.

Table 1. Labels applied to rules in the **AI Rules Set**

Label Category	Source	Mutually Exclusive?	Example Labels
Rule Type	Fiesler et al.	No	Advertising & Commercialization, Prescriptive, Restrictive
AI Rule: Rationale	New	No	Low Quality, Inaccurate, Inhuman
AI Rule: Stance	New	No	Unqualified Ban, Qualified Ban, Disclosure Required
AI Rule: Medium	New	No	Text, Images, Deepfakes
AI Rule: Framing	New	No	Generate/Create, Edit/Assist/Enhance, Unspecified

Table 2. Labels applied to subreddits in the **Rules Subreddit Set**

Label Category	Source	Mutually Exclusive?	Example Labels
Topic	Fiesler et al.	Yes	Art, Celebrity, Video Games
Community Archetype	Prinster et al.	No	Content Generation, Topical Q&A, Social Support
Has AI Rule?	New	Yes	Yes, No

We used these manually labeled rules as ground truth against which to evaluate the performance of an LLM coder, which we planned to use to label the rest of the data. We crafted a prompt that combined the human coder instructions with Chain-of-Thought prompting techniques [67], including the recommended eight few-shot, in-context examples, which we selected from the manually labeled sample. We produced LLM labels by using OpenAI's API to prompt *gpt-4o-mini-2024-07-18*, OpenAI's current flagship small model, using this prompt and the same input data that were given to the human coders (a subreddit's rules). We compared the LLM-produced labels with the human-produced ones and calculated an IRR of .7, which we deemed insufficient. Two human coders examined the discrepancies, discussed what they thought the correct label should be, and updated the prompt with a clarifying example. For example, one rule labeled *NO* by humans but *YES* by the LLM was about bots: *"No bots(No bots): No bots. All bots are banned unless the bot provides exceptional value to /r/ MealPrepSunday."* To better align the LLM label with our preferred definition of an AI rule, we added this rule as a negative example in the prompt, with the explanation:"This rule bans bots, not AI. Bots are not necessarily AI."

To improve IRR we repeated this process of randomly sampling, manually coding, generating LLM labels, comparing the results, and updating the prompt using the examples that the model got wrong. We performed three iterations of this step, at which point the LLM labels had an IRR of .95 with the human labels. We then labeled the entirety of both the **Longitudinal Subreddit Set** and the **Rules Subreddit Set** using OpenAI's API and *gpt-4o-mini-2024-07-18*. The final prompt used can be found in Appendix A.1. This process identified the **AI Rules Subreddit Set** ($N = 4,251$) of subreddits with rules governing AI use.

3.3.2 Identifying Emergent AI Rule Types. In order to better understand the types of rules that subreddits use to govern AI, we identified these rules in our sample and used an iterative inductive coding process to develop a descriptive taxonomy. To do this coding, we slightly modified the prompt developed and validated in Section 3.3.1 so that it would flag individual rules rather than

a set of rules as being AI-related. Since the prompt required only a slight modification from the prior prompt, and since human coders would be manually reviewing the labels in the next step, we decided that we did not need to validate model performance before tagging the entire sample. We used this prompt to instruct *gpt-4o-mini-2024-07-18* to label each of the $N = 33,674$ individual rules in the **AI Rules Subreddit Set**, which produced the **AI Rules Dataset** ($N = 4,458$) of individual rules about AI.

Next, two of the authors collaborated to identify emergent labels for the rules. The authors randomly selected $N = 25$ rules from this dataset and independently performed affinity mapping [25] by grouping rules that appeared similar into emergent categories. The authors then compared results and consolidated their categories into labels that best captured the elements of interest. Each of the two authors then independently applied this set of labels to another random sample of $N = 25$ rules, once again grouping similar rules, including into new emergent groups where the existing labels did not suffice. The two authors repeated this process of randomly sampling, independently labeling, collaboratively comparing their results, and consolidating their labels, until no new labels emerged and they were confident that their taxonomy covered the relevant dimensions of variation in the dataset. The final label categories were *Rationale*, *Stance*, *Medium* and *Framing*, which are listed alongside some of their possible labels in Table 1 (see Table 6 for a full list of all labels and examples of tagged rules). We applied these labels to the entire **AI Rules Dataset**, as we describe next.

3.3.3 Applying Rule Classification Labels. We used LLMs to apply our emergent labels and labels from an additional *Rule Type* category from Fiesler et al. [17] to the entire **AI Rules Dataset**. The two first authors collaboratively developed a codebook with instructions for applying these labels. Fiesler et al. does not provide codebook instructions, so the authors interpreted the examples presented in the study and came up with instructions for how to apply each of the 24 possible labels[5].

The two authors then revisited a sample that they had used for emergent coding and used the instructions to apply *Rule Type* labels. Next, the authors used their codebook to prompt OpenAI's *gpt-4o-2024-08-06* model to generate the rule-level labels found in Table 1 for this same sample. The two authors considered the difference between LLM labels and their own and used the discrepancies to identify parts of their codebook that needed clarification. The two authors refined the unclear instructions in the codebook and repeated this process iteratively until the LLM labels matched their own with an $IRR > .8$. We then used *gpt-4o-2024-08-06* to generate labels for the entire **AI Rules Dataset**. The final prompt used can be found in Appendix A.2.

3.3.4 Applying Subreddit Classification Labels. We also sought to label the subreddits in the **Rules Subreddit Set** with details about the type of community that each subreddit was home to. To this end, we devised a process to apply labels from two label categories used in prior HCI literature: Topic labels from Fiesler et al. [17] and Community Archetype labels from Prinster et al. [57]. In applying labels, we considered a subreddit's `name`, `description`, and `rules` metadata fields.

In the first step of the process the first author created a codebook with instructions as to how to label the (1) Topic of the subreddit, using as potential labels the 28 topics provided by Fiesler et al., and (2) which of the five Community Archetypes defined by Prinster et al. apply to the community. While these prior studies provided useful taxonomies for classification, neither provided a formal codebook with instructions, so the first author used the examples in these studies to produce a codebook with detailed coding instructions. Two of the authors used this codebook to independently

[5]*"Rule Type"* labels are not mutually exclusive, so we represented them as a series of binary labels, one for each possible Rule Type.

Table 3. Prevalence of rules and rules governing the use of AI for subreddits in the **Longitudinal Subreddit Set**. We show the prevalence at the time of each crawl as well as the % change between the crawls. In addition to the prevalence in the entire dataset, we include the prevalence in the top 1% largest subreddits, ranked by subscriber count at the time of the 2024 crawl.

	Jul 2023		Nov 2024		
	N	%	N	%	**% Change**
All subreddits in dataset ($N = 227,737$)					
had rules at time of crawl	66,505	29.2%	71,862	31.6%	**+8.1%**
had AI rules at time of crawl	1,348	.6%	2,808	1.2%	**+108.3%**
Top 1% largest subreddits ($N = 2,278$)					
had rules at time of crawl	2139	93.9%	2197	96.4%	**+2.7%**
had AI rules at time of crawl	192	8.4%	390	17.1%	**+103.1%**

classify the same sample of 25 subreddits randomly selected from the **Rules Subreddit Set**. The two authors compared their results and calculated IRR, achieving an initial IRR of .56, which was deemed insufficient. To improve IRR, the two authors discussed discrepancies in their codes and updated the codebook instructions where clarity was needed to align their labels. The two authors repeated this process of randomly sampling 25 subreddits from the **Rules Subreddit Set**, independently coding them with the updated codebook, calculating IRR between the two coders, discussing discrepancies between their codes, and iterating on the codebook, until they achieved an IRR of .88. We then used the OpenAI API to generate labels for the most recently coded sample by passing the codebook and subreddit metadata as a prompt to the *gpt-4o-mini-2024-07-18* model. We compared these LLM-produced labels with the manual labels and found there to be excellent agreement ($IRR = .84$). We repeated this process on another randomly sampled 25 subreddits. On this new sample the manual labels again agreed with the LLM labels with excellent agreement ($IRR = .88$). We then used the OpenAI API to prompt *gpt-4o-mini-2024-07-18* to generate labels for each subreddit in the **Rules Subreddit Set**. The final prompt used can be found in Appendix A.3.

4 Findings

We group our findings by research question and present in sequence: (1) changes in the prevalence of rules governing AI use, (2) the types of communities that have these rules, and (3) the types of rules that we find for governing AI use.

4.1 RQ1: How Has the Prevalence of Subreddit Rules Governing the Use of AI Changed Over Time?

Table 3 shows the results of the detection task described in Section 3.3.1 on our **Longitudinal Subreddit Set**. We find that during the 2023 crawl, $N = 1,348$ subreddits had rules governing the use of AI (.6%). By the time of our 2024 crawl, this number increased by 108.3% to $N = 2,808$ (1.2%). Rules in general, and specifically rules governing the use of AI, are more common in larger subreddits. For example, 8.4% of the largest 1% of subreddits by subscriber count had AI rules during our first crawl. This percentage increased to 17.1% in our second crawl. Figure 2 provides a more detailed demonstration of the differences in both the prevalence of AI rules and the increase in such rules among subreddits with different numbers of subscribers. The figure plots the number of subreddits in the **Longitudinal Subreddit Set** with AI rules at the time of each crawl. In order to expose the variation across subreddits of different sizes, we bucket subreddits along the X-axis into deciles based on their number of subscribers at the time of the second crawl (the labels represent the lower edge of the decile). Each decile contains approximately 22,773 subreddits. For each decile

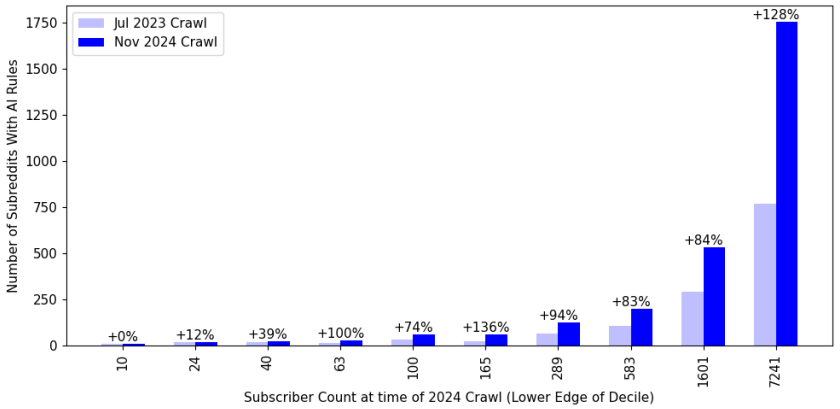

Fig. 2. Number of subreddits in the **Longitudinal Subreddit Set** with AI rules during each of our crawls. To demonstrate differences across subreddit size, subreddits are bucketed into deciles based on their subscriber count at the time of the second crawl. Each decile contains approximately 22,773 subreddits. The bars shows the number of subreddits with AI rules in each decile and the labels above the bars indicate the percent change between crawls.

we plot two bars representing the number of subreddits with AI rules: the left bar represents the number during the first crawl and the right bar represents the number during the second. Above the bars we include a numeric label representing the percent change between crawls. For example, the rightmost pair of bars shows that in the largest decile of subreddits, about 750 had AI rules during the first crawl, and that this number increased by 128% to about 1,750 at the second crawl. The figure shows that subreddits with more subscribers are more likely to have AI rules, and that communities of all sizes saw increases in the prevalence of AI rules between our two crawls, though that increase was more prominent for larger communities.

For comparison, we additionally analyzed the prevalence of AI rules in subreddits in the **Broad Subreddit Set** that were created *after* our first crawl and were thus excluded from our **Longitudinal Subreddit Set**. The **Broad Subreddit Set** contained $N = 30,962$ English language subreddits that were created after the first crawl (but were at least a month old during the second crawl). Of these newly created subreddits, $N = 467$ had AI rules (1.5%), higher than the percentage of subreddits with AI rules during the second crawl in the **Longitudinal Subreddit Set** (1.2%). While this difference can be due to the fact that new subreddits have different characteristics, our separate analysis in the following section also suggests that newer subreddits are more likely to have AI rules.

4.2 RQ2: What Types of Communities Have Rules Governing AI Use?

In this section, we examine the characteristics of communities that have rules governing the use of AI. To do that, we compare the characteristics of two groups of subreddits in the **Rules Subreddit Set**: subreddits with rules about the use of AI, and those without such rules. We find that larger subreddits and those associated with *Art* or *Celebrity* topics are more likely to have explicit rules regulating AI.

To identify subreddit characteristics that suggest the existence of *AI rules*, we first perform a difference-in-means test between subreddits with and without AI rules. Table 4 presents the results of the test, showing various subreddit characteristics that differ between these two subreddit types. For numerical subreddit characteristics like Subscriber Count and Age the table shows the average

Table 4. Differences between subreddits in the **Rules Subreddit Set** ($N = 99,969$) with and without AI rules. For numerical subreddit characteristics, we show the average value across the subreddits in each group. For binary characteristics, we show the percentage of subreddits with this characteristic in each group and denote these percentages with (%). While we computed differences for all 28 topics, we show here only the ten topics with the largest absolute differences.

		No AI Rule	AI Rule	Difference
Metadata	Subscriber count	37,865.69	27,3571.19	235705.5***
	Subreddit Age (Years)	4.94	6.32	1.38***
	Subreddit created after ChatGPT (%)	24.78	31.33	6.56***
Community Archetype	Content generation (%)	51.68	73.75	22.07***
	Affiliation with an entity (%)	57.47	63.3	5.83***
	Topical question and answer (%)	7.29	6.52	-0.78
	Learning and perspective broadening (%)	20.52	16.61	-3.91***
	Social support (%)	10.29	4.38	-5.91***
Topic	Celebrity (%)	3.87	18.33	14.45***
	Art (%)	2.36	12.99	10.63***
	Entertainment (%)	19.01	22.96	3.95***
	Politics (%)	2.21	0.54	-1.67***
	Business and Organizations (%)	2.71	0.99	-1.73***
	Games (%)	6.97	4.68	-2.29***
	Local (%)	4.22	1.11	-3.11***
	Hobby (%)	8.07	4.7	-3.36***
	Humor (%)	9.75	3.32	-6.43***
	Support (%)	8.98	2.52	-6.46***

Note: $^{*}p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$

value across these two groups of subreddits. For binary characteristics, such as whether a subeddit was created after the launch of ChatGPT or if it has a specific Community Archetype or Topic label, we show the percentage of subreddits in each group with this characteristic. To test for significance we use chi-squared tests for the binary characteristics and the non-parametric Mann-Whitney U test for numerical characteristics. We implement Bonferroni correction to adjust for multiple comparisons.

Table 4 exposes several key differences between subreddits with and without AI rules. Subreddits with AI rules tend to have more subscribers. These subreddits also tend to be newer; correspondingly, a higher percentage of subreddits with AI rules were created after ChatGPT was launched. We observe differences between the groups with respect to their association with the different Community Archetype labels. For example, there is a large and significant difference related to the *Content Generation* label; subreddits with this label are about 22 percentage points more likely to have AI rules. In contrast, subreddits of the *Social Support* Community Archetype are about 6 percentage points *less likely* to have AI rules. Lastly, the *Art* and *Celebrity* topics are more common in subreddits with AI rules, encompassing about a third of all subreddits with AI rules compared to only 6% of subreddits with no AI rules.

We run several logistic regression (LR) models to validate our results. As a baseline, we repeat the analysis done in Fiesler et al. [17] and use a LR model to predict the presence of *any* set of rules for subeddits in the **Broad Subreddit Set**, using as input variables the following metadata gathered in our second crawl: *Subscriber Count*[6], *Subreddit Age* (in years), and a binary indicator of whether or not the subreddit was created after the launch of ChatGPT. We compare these results with a

[6]We log transform *Subscriber Count* to account for the variable's skewed distribution.

LR model that specifically predicts the presence of *AI* rules, using as input all of the subreddit characteristics shown in Table 4. Comparing the two regression results isolates the subreddit characteristics that are specifically predictive of AI rules, not rules in general. We find that the *Art* and *Celebrity* Topic labels are both predictive of AI rules ($\beta = 1.47, p < 0.01$ and $\beta = 1.41, p < 0.01$), supporting the analysis above. In terms of Community Archetype labels, the *Content Generation* label is predictive of AI rules ($0.83, p < 0.01$). These regression results are in line with the findings from our difference-in-means test.

One possible explanation for why *Content Generation* communities are more likely to enact rules about AI could be because this type of community is generally less personal, with looser ties between members, and thus may need to rely on formal rules to maintain content quality and standards. Communities that prioritize relationships among users, on the other hand, may be able to use more personal methods to shape user behavior, such as informal norms. For example, consider a Wes Anderson subreddit whose purpose is to discuss "photos that accidentally share Wes Anderson's unique style." We labeled this subreddit with both the *Art* topic and the *Content Generation* community archetype. The language of this subreddit's AI rules reflects a distinct concern with the veracity of posted *images*, rather than concerns about authentic interactions. The subreddit's rules state that "Your photos don't have to be images taken by you, but they must be real photos." Similarly, subreddits labeled with the *Celebrity* topic tend to group AI rules with bans of photoshop, white borders, or low-resolution images, all of which are concerned with the format and quality of the subreddit's content. Some *Celebrity* subreddits seem to encourage AI, but only in cases where it may produce enhanced images. In contrast, subreddits that we labeled with the *Social Support* community archetype like r/introverts – a subreddit that does *not* have an AI rule – emphasize in their rules that they care about "authentic connections." We further explore the implications of these observations in Section 5.

4.3 RQ3: What Types of Rules Exist for Governing AI?

While our previous analyses explored what types of communities are likely to explicitly regulate AI, we also consider differences in *how* communities regulate AI. Drawing on the **AI Rules Set** ($N = 4,458$), we produce frequencies of the different Rule Type labels described in Sections 3.3.2 and 3.3.3. Table 5 shows the Rule Type labels used by Fiesler et al. [17], while Table 6 presents the new labels that we produced via an iterative, inductive coding process. The tables show the percentage of rules in the set that were tagged with each label.

Comparing our results to those of Fiesler et al. [17] illuminates some interesting differences between AI rules and general subreddit rules. Fiesler et al.'s results found that with general rules, *Prescriptive* rules, or rules that tell users what to do, occurred in about equal numbers as *Restrictive* rules that tell users what *not* to do[7]. However, in our **AI Rules Set**, 96.84% of rules were tagged as *Restrictive* while only 67.7% were tagged as *Prescriptive*. A *Restrictive* rule might include statements like "No AI (chatGPT etc.)" or "No AI generated content" while a *Prescriptive rule* might instead tell users whether to cite if AI is used. This pattern further bears out in the data for our *AI Rule: Stance* label category, which shows *Unqualified Bans* (e.g., "No AI-Generated Content") as the most common label (55.23% of AI rules). It seems that many subreddits may find it easier to simply ban the use of AI outright, rather than trying to develop policies that can accommodate it under certain circumstances (e.g., requiring the disclosure of AI use). Still, we do see evidence of some communities adopting stances other than outright bans, like the 24.23% of AI rules labeled *Stance:*

[7]These labels are not mutually exclusive: a rule could be tagged as both *Prescriptive* and *Restrictive* if it contains elements of both style.

Qualified Ban and the 17.95% labeled *Stance: Disclosure*. Perhaps this implies that some communities will be flexible towards figuring out how to work with, rather than against, this new technology.

Overall, we observe that a high number of AI rules are related to images. Table 5 shows that 46.55% of rules were tagged with the *Images* label from Fiesler et al. [17], which according to our codebook, indicates that the rule "is about posting visual content like images, photos, or paintings". For comparison, Fiesler et al. [17] tagged less than 8% of the rules in their dataset with this label. We see slightly smaller numbers with our inductively coded labels, which used a codebook that looked for more specific language: Table 6 indicates that 31.4% of AI rules are labeled as *Medium: Images* and 30.26% are labeled *Medium: Art*. This observation further supports the finding in Section 4.2 that subreddits with *Art* and *Celebrity* topics are more likely to have AI rules, as these are usually image-based subreddits.

Tables 5 and 6 also demonstrate that many AI rules on Reddit focus on concerns about content *quality*. Around 28% of rules were labeled as relating to *Low quality content*; 13% and 11% respectively were labeled as *Rationale: Low quality* and *Rationale: Low effort*, meaning that they explicitly stated these concerns as their rationale for regulating AI use. Rules about *Low quality content* emphasize the importance of "creative merit" and content that takes more than "only a few minutes of effort." These rationales are more prevalent than rationales based on spam, ethics, or accuracy. The focus on quality may suggest that explicit AI regulation is driven by practical concerns about content and reputation, not ideological objections over the use of AI more broadly. This observation echoes the "content quality concerns" shared by moderators in Lloyd et al. [41].

Since the *Art* and *Celebrity* Topics featured prominently in Section 4.2, we also separately examined patterns among Rule Types within each of these Topics. Among subreddits in the *Art* Topic, we find a higher share of AI rules with the *Copyright and Piracy* label (25.54%). These *Copyright and Piracy* rules primarily focus on proper attribution and posting original content. Among *Celebrity* subreddits, we see that the *Low Quality Content* and *Format* Rule Type labels are much more common (58.32% and 55.32%). The *Format* label means that references to AI often co-occur with regulations about the format of posts — such as statements about how posts should be titled or specifications about minimum word counts. In terms of the explicit rationales given for the AI rules, almost half of all AI rules in *Celebrity* subreddits were labeled *Rationale: Inauthentic* and 34.79% were labeled *Rationale: Low Quality*. Additionally, 77% of AI rules from *Celebrity* subreddits explicitly govern the use of "Fakes" (labeled with *Medium: Fakes*), which upon closer examination appears to be referring to heavily-edited photos (these rules separately make references to "Deepfakes", which we distinguish with a different label).

5 Discussion

Our quantitative study offers three main research contributions: (1) a large-scale analysis of rules about AI use in Reddit communities, (2) a taxonomy for classifying these rules, and (3) a dataset of subreddits, their metadata, and public community rules at two distinct points in time. Our study is an effort to support community self-moderation inspired by Seering's second "guiding question" for the HCI research community: *"What are the processes of context-sensitivity in online community moderation, and how might they be better supported?"* [63]. The sudden arrival of generative AI poses unique challenges for online communities [41] that HCI research is uniquely positioned to help address. Our study attempts to measure at scale the prevalence and variety of rules that online communities have enacted in response to this new technology. We hope that our data and analysis can add additional context to qualitative studies into online communities' responses to AIGC, as well as to general knowledge about online community self-moderation practices. Below we put our findings in dialogue with prior work and discuss design implications for social platforms.

Table 5. Prevalence of Fiesler et al.'s *Rule Type* labels in the **AI Rules Set** ($N = 4,458$). We show the percentage of rules in the set that were tagged with each label. Rules may be tagged with multiple labels and each subreddit may have multiple rules in this set.

Rule Type	**%**
Content and behavior	99.73
Restrictive	96.84
Prescriptive	67.7
Images	46.55
Consequences, moderation, and enforcement	32.48
Low quality content	27.77
Post format	19.25
Copyright and piracy	16.49
Links and outside content	10.54
Spam	9.02
Off topic	7.11
Reposting	5.88
NSFW	5.59
Advertising and commercialization	3.93
Reddiquette and sitewide rules	1.05
Politics	1.03
Harassment	0.96
Spoilers	0.7
Hate speech	0.61
Doxxing and personal information	0.54
Trolling	0.43
Voting	0.18
Personal army	0.13
Personality	0.04

5.1 The Role of Community Rules

While we focus our analysis on rules about AI use, one of our primary contributions is a dataset of community rules that the HCI research community can use to answer a wider range of questions about community self-moderation. As far as we know, this is the first dataset of its kind, as prior studies of self-governing online communities' rules [5, 17, 27, 48, 61] have not made their rules datasets public. Large-scale, quantitative analysis of these rules offers a view into platform-wide patterns that complement past qualitative HCI research into online community self-moderation. Our AIGC-focused analysis finds that subreddit communities, especially those with more subscribers, are changing their rules to respond to generative AI. But what do these rules tell us about the experience of participating in or moderating these communities?

Past research into community self-moderation has found that community rule changes are most often enacted, "in response to unexpected incidents" that impact a community [64]. Qualitative work on the experience of moderating AIGC [41] provides evidence that AI rules are created for the same reason, as many moderators report problematic AI use in their communities as a precursor to their formal rules. This mechanism of rule creation suggests that a community's rules are not a perfect proxy for its values, as community values do not need to be codified into rules unless there is some incident that surfaces community misalignment. Seen in this light, the communities with AI rules in our dataset are not necessarily the communities with the strongest stances on AI, but those where some misalignment about AI use emerged between community members. Viewing

Table 6. Prevalence of AI Rule Type labels in the **AI Rules Set** ($N = 4,458$). We show the percentage of rules in the set that were tagged with each label. Rules may be tagged with multiple labels and each subreddit may have multiple rules included in this set. Examples are drawn from the data but have been shortened for brevity. See the tagging prompt in Appendix A.2 for label definitions.

Rule Category	Rule Type	% Of Rules	Example
Rationale	Low quality	13.19	*Low quality content will be removed. This includes AI generated images and other AI media.*
	Inauthentic	11.93	*NO FAKES OR OVERLY SHOPPED PHOTOS - may result in a ban, serious offense. Images must be posted as published except for Quality AI-enhanced images. We ask that those titles include "Enhanced."*
	Low effort	11.06	*AI generated images or Chatgpt text content will be removed as "low effort" content. Please do not post AI generated content on this subreddit.*
	Original content	7.0	*AI generated text does not qualify as an original source.*
	Off topic	3.1	*Content generated by a computer algorithm is off topic and unwanted.*
	Spam	3.25	*Any form of spam / blogspam is not allowed. This includes links to book reviews, AI generated content, bots, social media channels and any affiliate links.*
	Ethics	2.98	*Given Miyazaki's vehement opposition to AI and the ethical issues with AI generated art, posts that feature content generated by an AI model will be removed.*
	Policy	1.37	*Deep fakes violate reddit TOS. You will be banned if you post one.*
	Misleading	1.7	*AI altered videos are misleading and will be removed.*
	Inaccurate	0.83	*AI is awesome, but when used to answer technical questions it often gives incorrect or sometimes dangerous recommendations.*
	Inhuman	0.9	*We believe Hip hop is human made, and for that reason we've decided to ban AI generated content.*
Stance	Unqualified ban	55.23	*AI-generated content of any kind is forbidden.*
	Qualified ban	24.23	*AI Generated Fan Art is also not allowed, UNLESS it has been substantially edited by a human.*
	Disclosure	17.95	*If posting AI-generated art, specify in the title that it is AI-created, include what service used.*
	Enabling	18.51	*Recipes written by AI are ok but you must: 1) make the salsa, and 2) post a real photo of the completed salsa.*
	Requiring	1.1	*Only post images generated by Ai, that's it*
Medium	Content	52.74	*On r/NFT, content that is fully generated by AI and not modified in any way shape or form is prohibited and will result in a ban from the community.*
	Images	31.4	*Please no DALL·E mini or other AI generated images.*
	Art	30.26	*No AI art allowed, as per our content policy.*
	Fakes	15.79	*No fake images of Leah Willamson are permitted, either photoshopped or AI generated*
	Unspecified	5.36	*No AI - https://www.gofundme.com/f/protecting-artists-from-ai-technologies*
	Text	6.59	*This includes both AI generated images/artwork and text generated by tools such as ChatGPT.*
	Deepfake	4.6	*Deepfakes or other types of fakes are not allowed here, you will be banned for posting them.*
	Bots	3.34	*Karma-farming with ChatGPT bots, Malware, phishing, spam or scam will result in your account being banned without warning.*
	Videos	3.95	*No low effort AI content or Videos.*
Framing	Generate	53.88	*AI-generated content of any kind is forbidden.*
	Assist	16.08	*Traditional paintings and drawings only. No videos, screenshots, photography, audio, multi-panel comics, content generated or AI-assisted submissions.*
	Unspecified	12.02	*The removal of low quality and, AI submissions, are subject to remove at mod discretion.*

rules as evidence of past misalignment could explain why *Social Support* subreddits were negatively associated with AI rules in our analysis: perhaps these communities are more aligned in general, so that even if they have strong stances on AIGC, they do not need rules to enforce them.

It is worth noting that while rules may indicate some past community misalignment, they do not necessarily tell us how rules shape behavior in practice. For example, prior work has shown that humans cannot reliably detect AIGC [30], which raises the question of how enforceable AIGC rules are. At the same time, it is unclear whether or not AIGC rules even *need* to be enforced: past work on community rules suggests that they can shape behavior by clarifying norms [41, 44], even without an effective enforcement mechanism. Still, it is reasonable to assume that some communities with AI rules are attempting to enforce them. We see the prevalence of AI rules, especially in the largest subreddits, as evidence that HCI research should explore design solutions that can help with their enforcement.

One way that HCI research can empower communities to effectively enforce their AIGC policies is by offering tools for policy design. While some online communities are enacting top-down AI policies [50], Reddit has so far left this decision to individual communities. We like this approach, as past research has shown that community moderators learn by doing and interacting closely with their community [10], resulting in a quick feedback loop that better positions them to craft context-sensitive policies [63]. Still, opportunities exist for HCI research to help communities enact effective policies, both by offering empirical evidence of the effects of certain types of rules, as well as new systems for policy development [71]. Future work could use our rules dataset to find similar communities that differ in their policy choices, which could be used as a natural experiment to study the effects of different rules. Studying the effects of policies may uncover general knowledge about which kinds of rules (such as "prescriptive" vs "restrictive") encourage community engagement and which cause members to leave for other communities. Platforms could use this knowledge to offer moderators a reasonable set of "default" rules at community creation time, based on the type of community being created. Some alternative systems for policy development from past HCI research have explored ways of engaging more members of a community, beyond just moderators, in the policy creation process [71]. We see these more democratic systems of rule development as promising at a time when norms around AI use are still in flux.

5.2 Supporting Community Self-Governance in the Age of Generative AI

Comparing our findings to those from an earlier large-scale analysis of Reddit rules conducted by Fiesler et al. [17] provides insight into what is different about AI rules. We found that rules governing AI usage are more likely to be about images than the general rules analyzed in Fiesler et al. Together with the finding that images are the medium most referenced in AI rules, this finding suggests that the use of AI images may be the most common cause of community misalignment about AI norms. Our findings that *Restrictive* AI rules are more common than *Prescriptive* ones seems to suggest that, at least at this point in time, communities appear more interested in clarifying how AI *cannot* be used than how it can. Still, there is some evidence of flexibility in these stances, and it will be interesting to observe if community stances towards AI grow more or less flexible over time. It may be difficult to generalize about when communities require AI disclosure vs when they ban it outright, but digging into the motivations for these different approaches could be a topic of future HCI research.

These rules—and norms—will continue to change over time. So far our data shows that AI rules have become more prevalent, but it is possible that these rules disappear if AI use becomes widespread to the point that it is unremarkable, or where bans have negative impacts on community engagement. Two of the top three most common rationales in the AI rules in our data were *"Inauthentic"* and *"Low effort"*, which seems to suggest that some communities perceive that AI use

cheapens the social interactions that occur in online communities. These rationales for AI rules echo the findings from prior studies into the use of automated tools in interpersonal communication [47, 59]. We suspect that even if communities cannot effectively ban AIGC, those that care about gatekeeping [33] may seek out other ways for users to demonstrate their authenticity, similar to those that have developed in some anonymous online communities [3]. Platform designers can help by offering communication tools that encourage authenticity, such as those explored in recent work on *effortful communication* [34, 72]. Additionally, past HCI studies of online communities have shown that close-knit communities with strong social identity are able to navigate the challenging task of developing new norms around controversial technology use [16]. In the age of generative AI, HCI research could develop novel ways to encourage strong social identity formation within online communities.

Our results complement Lloyd et al. [41]'s qualitative investigation into moderators' experience of enforcing AIGC rules by documenting the prevalence of these rules across Reddit. Their study identifies three main categories of moderator concern about how AIGC will impact their communities: concerns about content quality, social dynamics, and governance processes. Of these, content quality concerns are the most represented in our data. What a community considers to be "low quality" is extremely context-dependent, but this could be one area where UI features that clearly communicate norms and expectations to community members, such as rules that display at post creation time, could go a long way to positively shape user behavior [44]. We suspect that rules about social dynamics are less prevalent in our data because of the phenomenon that we discuss in Section 5.1, where desired social dynamics may not need to be codified into rules because communities that care about them may be naturally better aligned. For example, in our data we do find rules about authenticity, but it tends to be more about the authenticity of content, not of interpersonal communication. Additionally, our findings provide further evidence of Lloyd et al.'s finding that moderators' concerns about AIGC are based on their communities' values. While rules usually do not explicitly state a community's values, sometimes we can infer them, e.g. it is not a coincidence that *Copyright and Piracy* AI rules are more common in *Art* subreddits. Finally, we note that rules about images are very common in our data, while the majority of concerns raised in Lloyd et al. pertain primarily to text. This is a significant difference for the HCI research community to note, as helping communities respond to AI-generated text as opposed to AI-generated images is a different challenge with different potential solutions. For example, images may be more likely to be robustly watermarked [11] or to have provenance metadata [14], which platforms could clearly communicate to viewers.

Our study is limited in that it is based on samples of subreddits that were publicly listed on Reddit's web interface in November 2024 and July 2023. Our sample is thus missing subreddits that are public but unlisted, which makes it likely that certain types of subreddit are underrepresented, such as NSFW subreddits. Our sample also excluded subreddits with a primary language other than English. Future research could explore whether the trends that we observed still hold across the types of communities excluded from our sample. Our study also limits our analysis to text that is contained within a subreddit's "rules" metadata field. There are various other ways that online communities communicate rules and norms, such as through public discussions, wikis, direct messages, or moderation actions. Future work interested in studying communities' attitudes towards AI use could consider these other sources of data, as well as qualitative approaches that would elicit direct input from community members and moderators.

Ethics Statement. Our study was not required to undergo IRB review as we worked only with publicly available data. Still, we recognize the importance of conducting ethical social media research and have taken steps to follow the best practices of the HCI community [18]. We did not collect any information about individual users or conversations. Additionally, we did not include in

our study any subreddits who were not listed in Reddit's public index of communities. We plan to make our datasets public after this publication completes the peer-review process, but will take additional steps to ensure no sensitive information is contained before doing so.

6 Conclusion

Communities across Reddit have responded to the arrival of generative AI by updating their public community rules with explicit positions on the use of this new technology. A specific community's attitude towards the use of AIGC is nuanced and context-dependent, but our findings suggest a general trend of opposition from content generation communities that value quality and authenticity. It remains to be seen how these rules will be enforced, but HCI researchers should stay attuned to the changing needs and norms of these online communities as they continue to adapt to this new technology. Especially as social platforms begin incorporating generative AI features, it is important for designers to consider how to encourage the self-determination of communities with nuanced and varied stances on the use of AIGC.

Acknowledgments

This material is based upon work partially supported by the National Science Foundation under Grant No. CHS 1901151/1901329.

References

[1] Dhruv Agarwal, Farhana Shahid, and Aditya Vashistha. 2024. Conversational Agents to Facilitate Deliberation on Harmful Content in WhatsApp Groups. https://doi.org/10.1145/3687030

[2] Marianne Aubin Le Quéré, Hope Schroeder, Casey Randazzo, Jie Gao, Ziv Epstein, Simon Tangi Perrault, David Mimno, Louise Barkhuus, and Hanlin Li. 2024. LLMs as Research Tools: Applications and Evaluations in HCI Data Work. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3613905.3636301

[3] Michael Bernstein, Andrés Monroy-Hernández, Drew Harry, Paul André, Katrina Panovich, and Greg Vargas. 2011. 4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community. *Proceedings of the International AAAI Conference on Web and Social Media* 5, 1 (2011), 50–57. https://doi.org/10.1609/icwsm.v5i1.14134 Number: 1.

[4] Gordon Burtch, Dokyun Lee, and Zhichen Chen. 2024. The consequences of generative AI for online knowledge communities. *Scientific Reports* 14, 1 (May 2024), 10413. https://doi.org/10.1038/s41598-024-61221-0 Publisher: Nature Publishing Group.

[5] Jie Cai, Cameron Guanlao, and Donghee Yvette Wohn. 2021. Understanding Rules in Live Streaming Micro Communities on Twitch. In *Proceedings of the 2021 ACM International Conference on Interactive Media Experiences* (Virtual Event, USA) *(IMX '21)*. Association for Computing Machinery, New York, NY, USA, 290–295. https://doi.org/10.1145/3452918.3465491

[6] Yang Trista Cao, Lovely-Frances Domingo, Sarah Ann Gilbert, Michelle Mazurek, Katie Shilton, and Hal Daumé III. 2023. Toxicity Detection is NOT all you Need: Measuring the Gaps to Supporting Volunteer Content Moderators. https://arxiv.org/abs/2311.07879v2

[7] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 32 (11 2018), 25 pages. https://doi.org/10.1145/3274301

[8] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1217–1230. https://doi.org/10.1145/2998181.2998213

[9] Madiha Zahrah Choksi, Marianne Aubin Le Quéré, Travis Lloyd, Ruojia Tao, James Grimmelmann, and Mor Naaman. 2024. Under the (neighbor)hood: Hyperlocal Surveillance on Nextdoor. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–22. https://doi.org/10.1145/3613904.3641967

[10] Amanda L. L. Cullen and Sanjay R. Kairam. 2022. Practicing Moderation: Community Moderation as Reflective Practice. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1 (April 2022), 111:1–111:32. https://doi.org/10.1145/3512958

[11] Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, Jamie Hayes, Nidhi Vyas, Majd Al Merey, Jonah Brown-Cohen, Rudy Bunel, Borja Balle, Taylan Cemgil, Zahra Ahmed, Kitty Stacpoole, Ilia Shumailov, Ciprian Baetu, Sven Gowal, Demis Hassabis, and Pushmeet Kohli. 2024. Scalable watermarking for identifying large language model outputs. *Nature* 634, 8035 (Oct. 2024), 818–823. https://doi.org/10.1038/s41586-024-08025-4 Publisher: Nature Publishing Group.

[12] Renee DiResta and Josh A. Goldstein. 2024. How Spammers and Scammers Leverage AI-Generated Images on Facebook for Audience Growth. arXiv:2403.12838 [cs.CY]

[13] Hany Farid. 2022. Creating, Using, Misusing, and Detecting Deep Fakes. *Journal of Online Trust and Safety* 1, 4 (Sept. 2022). https://doi.org/10.54501/jots.v1i4.56 Number: 4.

[14] K. J. Kevin Feng, Nick Ritchie, Pia Blumenthal, Andy Parsons, and Amy X. Zhang. 2023. Examining the Impact of Provenance-Enabled Media on Trust and Accuracy Perceptions. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 270 (Oct. 2023), 42 pages. https://doi.org/10.1145/3610061

[15] Casey Fiesler. 2024. AI chatbots are intruding into online communities where people are trying to connect with other humans. http://theconversation.com/ai-chatbots-are-intruding-into-online-communities-where-people-are-trying-to-connect-with-other-humans-229473

[16] Casey Fiesler and Amy S. Bruckman. 2019. Creativity, Copyright, and Close-Knit Communities: A Case Study of Social Norm Formation and Enforcement. *Proc. ACM Hum.-Comput. Interact.* 3, GROUP, Article 241 (Dec. 2019), 24 pages. https://doi.org/10.1145/3361122

[17] Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. 2018. Reddit rules! characterizing an ecosystem of governance. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[18] Casey Fiesler, Michael Zimmer, Nicholas Proferes, Sarah Gilbert, and Naiyan Jones. 2024. Remember the Human: A Systematic Review of Ethical Considerations in Reddit Research. *Proc. ACM Hum.-Comput. Interact.* 8, GROUP (Feb. 2024), 5:1–5:33. https://doi.org/10.1145/3633070

[19] Eric Gilbert. 2013. Widespread underprovision on Reddit. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (San Antonio, Texas, USA) *(CSCW '13)*. Association for Computing Machinery, New York, NY, USA, 803–808. https://doi.org/10.1145/2441776.2441866

[20] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.

[21] Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. arXiv:2301.04246 [cs.CY]

[22] James Grimmelmann. 2015. The Virtues of Moderation. *Yale Journal of Law and Technology* 17 (2015), 42–109. https://heinonline.org/HOL/P?h=hein.journals/yjolt17&i=42

[23] Catherine Han, Joseph Seering, Deepak Kumar, Jeffrey T. Hancock, and Zakir Durumeric. 2023. Hate Raids on Twitch: Echoes of the Past, New Modalities, and Implications for Platform Governance. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 133:1–133:28. https://doi.org/10.1145/3579609

[24] Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication* 25, 1 (March 2020), 89–100. https://doi.org/10.1093/jcmc/zmz022

[25] Gunnar Harboe and Elaine M. Huang. 2015. Real-World Affinity Diagramming Practices: Bridging the Paper-Digital Gap. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 95–104. https://doi.org/10.1145/2702123.2702561

[26] Yiqing Hua, Shuo Niu, Jie Cai, Lydia B Chilton, Hendrik Heuer, and Donghee Yvette Wohn. 2024. Generative AI in User-Generated Content. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 471, 7 pages. https://doi.org/10.1145/3613905.3636315

[27] Sohyeon Hwang and Aaron Shaw. 2022. Rules and Rule-Making in the Five Largest Wikipedias. *Proceedings of the International AAAI Conference on Web and Social Media* 16 (May 2022), 347–357. https://doi.org/10.1609/icwsm.v16i1.19297

[28] Steven J. Jackson, Tarleton Gillespie, and Sandy Payette. 2014. The policy knot: re-integrating policy, practice and design in cscw studies of social computing. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 588–602. https://doi.org/10.1145/2531602.2531674

[29] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T. Hancock, and Mor Naaman. 2019. AI-Mediated Communication: How the Perception That Profile Text Was Written by AI Affects Trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300469

[30] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences* 120, 11 (March 2023), e2208839120. https://doi.org/10.1073/pnas.2208839120 Publisher: Proceedings of the National Academy of Sciences.

[31] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.* 26, 5, Article 31 (7 2019), 35 pages. https://doi.org/10.1145/3338243

[32] Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. 2019. Moderation Challenges in Voice-based Online Communities on Discord. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 55 (nov 2019), 23 pages. https://doi.org/10.1145/3359157

[33] Brian Keegan and Darren Gergle. 2010. Egalitarians at the gate: one-sided gatekeeping practices in social media. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work (CSCW '10)*. Association for Computing Machinery, New York, NY, USA, 131–134. https://doi.org/10.1145/1718918.1718943

[34] Ryan Kelly, Daniel Gooch, Bhagyashree Patil, and Leon Watts. 2017. Demanding by Design: Supporting Effortful Communication Practices in Close Personal Relationships. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 70–83. https://doi.org/10.1145/2998181.2998184

[35] Charles Kiene, Andrés Monroy-Hernández, and Benjamin Mako Hill. 2016. Surviving an "Eternal September": How an Online Community Managed a Surge of Newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1152–1156. https://doi.org/10.1145/2858036.2858356

[36] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. *Building successful online communities: Evidence-based social design* 1 (2012), 4–2. Publisher: MIT Press, Cambridge, MA, USA.

[37] Vinay Koshy, Tanvi Bajpai, Eshwar Chandrasekharan, Hari Sundaram, and Karrie Karahalios. 2023. Measuring User-Moderator Alignment on r/ChangeMyView. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 286 (10 2023), 36 pages. https://doi.org/10.1145/3610077

[38] Robert E. Kraut and Paul Resnick. 2012. *Building Successful Online Communities: Evidence-Based Social Design*. MIT Press. Google-Books-ID: lIvBMYVxWJYC.

[39] Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. 2024. Watch Your Language: Investigating Content Moderation with Large Language Models. *Proceedings of the International AAAI Conference on Web and Social Media* 18 (May 2024), 865–878. https://doi.org/10.1609/icwsm.v18i1.31358

[40] Lawrence Lessig. [n. d.]. *Code: And Other Laws of Cyberspace*. ReadHowYouWant. com. https://books.google.com/books?hl=en&lr=&id=tmE-pvNIX38C&oi=fnd&pg=PR2&dq=info:45fl71S8FGgJ:scholar.google.com&ots=Gd1wpKqwJa&sig=yeKlLD68eX5DRNkm19Bp5ecbRfs

[41] Travis Lloyd, Joseph Reagle, and Mor Naaman. 2023. "There Has To Be a Lot That We're Missing": Moderating AI-Generated Content on Reddit. https://arxiv.org/abs/2311.12702v4

[42] Ingrid Lunden. 2023. LinkedIn, now at 1B users, turns on OpenAI-powered reading and writing tools. https://techcrunch.com/2023/11/01/linkedin-now-at-1b-users-turns-on-openai-powered-reading-and-writing-tools/

[43] Makyen. 2023. Temporary policy: Generative AI (e.g., ChatGPT) is banned. https://meta.stackoverflow.com/q/421831/9737437

[44] J Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences* 116, 20 (2019), 9785–9789.

[45] Ivan Mehta. 2023. Instagram introduces GenAI powered background editing tool. https://techcrunch.com/2023/12/14/instagram-introduces-gen-ai-powered-background-editing-tool/

[46] Filippo Menczer, David Crandall, Yong-Yeol Ahn, and Apu Kapadia. 2023. Addressing the harms of AI-generated inauthentic content. *Nature Machine Intelligence* 5, 7 (July 2023), 679–680. https://doi.org/10.1038/s42256-023-00690-w Number: 7 Publisher: Nature Publishing Group.

[47] Pegah Moradi and Karen Levy. 2024. Sociotechnical Change: Tracing Flows, Languages, and Stakes Across Diverse Cases| "A Fountain Pen Come to Life": The Anxieties of the Autopen. *International Journal of Communication* 18 (2024), 9–9. https://ojs3.ijoc.org/index.php/ijoc/article/view/21842

[48] Matthew N. Nicholson, Brian C Keegan, and Casey Fiesler. 2023. Mastodon Rules: Characterizing Formal Rules on Popular Mastodon Instances. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (Minneapolis, MN, USA) *(CSCW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 86–90. https://doi.org/10.1145/3584931.3606970

[49] Rajvardhan Oak and Zubair Shafiq. 2024. Understanding Underground Incentivized Review Services. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 950, 18 pages. https://doi.org/10.1145/3613904.3642342

[50] Stack Overflow. [n. d.]. Generative AI Policy. https://stackoverflow.com/help/gen-ai-policy

[51] Joon Sung Park, Joseph Seering, and Michael S. Bernstein. 2022. Measuring the Prevalence of Anti-Social Behavior in Online Communities. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 451 (nov 2022), 29 pages. https://doi.org/10.1145/3555552

[52] Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2024. AI deception: A survey of examples, risks, and potential solutions. *Patterns* 5, 5 (May 2024), 100988. https://doi.org/10.1016/j.patter.2024.100988

[53] Jenny Preece. 1998. Empathic communities: reaching out across the Web. *interactions* 5, 2 (March 1998), 32–43. https://doi.org/10.1145/274430.274435

[54] Jenny Preece. 1999. Empathic communities: balancing emotional and factual communication. *Interacting with Computers* 12, 1 (Sept. 1999), 63–77. https://doi.org/10.1016/S0953-5438(98)00056-3

[55] Jenny Preece. 2000. *Online communities: Designing usability and supporting socialbilty*. John Wiley & Sons, Inc.

[56] Jenny Preece. 2001. Sociability and usability in online communities: Determining and measuring success. *Behaviour & Information Technology* 20, 5 (Jan. 2001), 347–356. https://doi.org/10.1080/01449290110084683 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01449290110084683.

[57] Gale H Prinster, C Estelle Smith, Chenhao Tan, and Brian C Keegan. 2024. Community Archetypes: An Empirical Framework for Guiding Research Methodologies to Reflect User Experiences of Sense of Virtual Community on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–33.

[58] Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society* 7, 2 (April 2021), 20563051211019004. https://doi.org/10.1177/20563051211019004 Publisher: SAGE Publications Ltd.

[59] Irene Rae. 2024. The Effects of Perceived AI Use On Content Perceptions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3613904.3642076

[60] Varun Nagaraj Rao, Eesha Agarwal, Samantha Dalal, Dan Calacci, and Andrés Monroy-Hernández. 2024. QuaLLM: An LLM-based Framework to Extract Quantitative Insights from Online Forums. https://doi.org/10.48550/arXiv.2405.05345 arXiv:2405.05345 [cs].

[61] Harita Reddy and Eshwar Chandrasekharan. 2023. Evolution of Rules in Reddit Communities. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (Minneapolis, MN, USA) *(CSCW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 278–282. https://doi.org/10.1145/3584931.3606973

[62] Regina Rini and Leah Cohen. 2022. Deepfakes, Deep Harms. *Journal of Ethics and Social Philosophy* 22, 2 (2022), 143–161. https://heinonline.org/HOL/P?h=hein.journals/jetshy22&i=150

[63] Joseph Seering. 2020. Reconsidering self-moderation: the role of research in supporting community-based models for online content moderation. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2 (Oct. 2020). https://doi.org/10.1145/3415178 Number of pages: 28 Place: New York, NY, USA Publisher: Association for Computing Machinery tex.articleno: 107 tex.issue_date: October 2020.

[64] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* 21, 7 (July 2019), 1417–1443. https://doi.org/10.1177/1461444818821316 Publisher: SAGE Publications.

[65] Philipp Singer, Fabian Flöck, Clemens Meinhart, Elias Zeitfogel, and Markus Strohmaier. 2014. Evolution of reddit: from the front page of the internet to a self-referential community?. In *Proceedings of the 23rd International Conference on World Wide Web* (Seoul, Korea) *(WWW '14 Companion)*. Association for Computing Machinery, New York, NY, USA, 517–522. https://doi.org/10.1145/2567948.2576943

[66] Rebecca Umbach, Nicola Henry, Gemma Faye Beard, and Colleen M. Berryessa. 2024. Non-consensual synthetic intimate imagery: Prevalence, attitudes, and knowledge in 10 countries. In *Proceedings of the CHI conference on human factors in computing systems (Chi '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3613904.3642382 Number of pages: 20 Place: Honolulu, HI, USA tex.articleno: 779.

[67] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. http://arxiv.org/abs/2201.11903 arXiv:2201.11903 [cs].

[68] Yiluo Wei and Gareth Tyson. 2024. Understanding the Impact of AI Generated Content on Social Media: The Pixiv Case. arXiv:2402.18463 [cs.CY] https://arxiv.org/abs/2402.18463

[69] Galen Weld, Amy X. Zhang, and Tim Althoff. 2024. Making Online Communities 'Better': A Taxonomy of Community Values on Reddit. *Proceedings of the International AAAI Conference on Web and Social Media* 18 (May 2024), 1611–1633.

https://doi.org/10.1609/icwsm.v18i1.31413
[70] Kai-Cheng Yang and Filippo Menczer. 2024. Anatomy of an AI-powered malicious social botnet. *Journal of Quantitative Description: Digital Media* 4 (May 2024). https://doi.org/10.51685/jqd.2024.icwsm.7
[71] Amy X. Zhang, Grant Hugh, and Michael S. Bernstein. 2020. PolicyKit: Building Governance in Online Communities. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 365–378. https://doi.org/10.1145/3379337.3415858
[72] Lei Zhang, Tianying Chen, Olivia Seow, Tim Chong, Sven Kratz, Yu Jiang Tham, Andrés Monroy-Hernández, Rajan Vaish, and Fannie Liu. 2022. Auggie: Encouraging Effortful Communication through Handcrafted Digital Experiences. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 427 (Nov. 2022), 25 pages. https://doi.org/10.1145/3555152

A LLM Prompts

A.1 Detecting Community Rules Governing AI Use

```
You will be given the public community rules for a subreddit. You will determine if any of the subreddit`s rules explicitly govern the use of AI tools or the content that such tools produce. Please return a JSON object with your determination, either YES or NO, and your explanation.

Examples:

1. Input: - AI generated(AI generated ): No AI (chatGPT etc.) generated messages/comments. No questions about chatGPT/AI generated code.
Output: {"determination": "YES", "explanation":"This rule bans code generated with ChatGPT or other AI tools"}

2. Input: - No bots(No bots): No bots. All bots are banned unless the bot provides exceptional value to /r/ MealPrepSunday.
Output: {"determination": "NO", "explanation":"This rule bans bots, not AI. Bots are not necessarily AI."}

3. Input: - Tributes, Split Edit, Deepfakes, roleplay, and DM Request are not allowed(Tributes, Split Edit, Deepfakes, roleplay, and DM Request are not allowed): The mods will take strict action against members who didn't follow this rule. The users who post Tributes will suffer a temporary ban.
Output: {"determination": "YES", "explanation":"This rule states that Deepfakes, which are a form of AI-generated content, are not allowed."}

4. Input: Rule 2 - Future Focus(Rule 2 - Future Focus): Submissions must be *future focused*. All posts must have an initial comment, a **Submission Statement**, that suggests a line of future-focused discussion for the topic posted. We want this submission statement to elaborate on the topic being posted and suggest how it might be discussed in relation to the future.
AI-focused posts are only allowed on the weekend.
Output: {"determination": "NO", "explanation":"This rule governs the use of posts about AI, not those made using it."}

5. Input: - Restricted accounts(New, Low Karma, bot, and underage accounts are restricted): * New or low karma accounts - may not post, can participate in the comments
* Negative karma accounts  - may not post or comment
* Under 18 users - may not post or seek diet advice
* No bot / AI accounts
Output: {"determination": "YES", "explanation":"This rule bans AI accounts."}

6. Input: - Rule 1: Must be relevant to language models(Rule 1: Must be relevant to language models): All language models are welcome here.
```

```
25
26 Dall-E, GPT-3, GPT-Neo, GPT-J, ...
27 Output: {"determination": "NO", "explanation":"This rule says that posts must be
       about AI but says nothing about their use."}
28
29 7. Input: - No fakes or overly shopped photos, images must be posted as published.
        No memes.(No fakes or overly shopped photos, images must be posted as
       published. No memes.): NO FAKES OR OVERLY SHOPPED PHOTOS - may result in a ban
       , serious offense. Images must be posted as published except for Quality AI-
       enhanced images. We ask that those titles include ""Enhanced."" Cartoonish/
       amateurish attempts will be removed solely at the mod's discretion. Don't
       complain. Do better. ! No X-rays or ""bubbling."" No watermarks. No emojis. No
        weird borders on the images (i.e., white or black bars). No text on images.
30 Output: {"determination": "YES", "explanation":"This rule specifies that all AI-
       enhanced images must be titled 'Enhanced'"}
31
32 8. Input: Properly Source Fan Content or Concept Art(This fan/concept art isn't
       properly sourced): We like to support the community and creators of the Marvel
        Cinematic Universe as much as possible and properly crediting them for their
       work is the least we can do. When submitting fan art or a piece of concept art
       , we ask that you properly link to the original source and mention the name of
        the artist in the title.
33 Output: {"determination": "NO", "explanation":"This rule does not mention art
       produced by AI."}
```

A.2 Applying Rule Classification Labels

```
1  You will be given one of a subreddit's public community rules and must determine
       which labels apply.
2  Follow the instructions for each label and determine if the label applies to the
       rule.
3  Return a JSON object where the keys are the label names and the values are TRUE if
        the label applies and FALSE if not.
4
5  Category: rule types
6
7  Label name: Advertising and Commercialization
8  Instructions: Is this rule about advertising and commercialization in the
       community?
9
10 Label name: Consequences/Moderation/Enforcement
11 Instructions: Does this rule specify what will happen if the rule is broken?
12
13 Label name: Content/Behavior
14 Instructions: Does this rule govern the types of behavior or content that are
       allowed or prohibited in the subreddit?
15
16 Label name: Copyright/Piracy
17 Instructions: Does this rule have to do with copyright, piracy, or giving credit
       to content's original creator?
18
19 Label name: Doxxing/Personal Info
20 Instructions: Is this rule about doxxing or revealing someone's personal
       information?
21
22 Label name: Format
```

```
Instructions: Does this rule specify that posts must follow a specific format?

Label name: Harassment
Instructions: Is this rule about harassment?

Label name: Hate Speech
Instructions: Is this rule about hate speech?

Label name: Images
Instructions: Is this rule about posting visual content like images, photos, or
    paintings?

Label name: Links and Outside Content
Instructions: Is this rule about posting links or outside content?

Label name: Low-Quality Content
Instructions: Is this rule about posting low-quality content?

Label name: NSFW
Instructions: Is this rule about posting NSFW content?

Label name: Off-topic
Instructions: Is this rule about posting off-topic content?

Label name: Personal Army
Instructions: Is this rule about using a group of other members to argue on your
    behalf?

Label name: Personality
Instructions: Does this rule govern what personality traits belong in the
    community?

Label name: Politics
Instructions: Does this rule govern posts about politics?

Label name: Prescriptive
Instructions: Does this rule include any positive statements about what users must
     do, or what is allowed?

Label name: Reddiquette/Sitewide
Instructions: Does this rule reference sitewide policy or norms?

Label name: Reposting
Instructions: Is this rule about reposting content that has already been posted?

Label name: Restrictive
Instructions: Does this rule include any negative statements about what users must
     not, or are not allowed to, do?

Label name: Spam
Instructions: Is this rule about posting spam, such as karma farming, or posting a
     large volume of content?

Label name: Spoilers
Instructions: Is this rule about posting spoilers?

Label name: Trolling
```

```
Instructions: Is this rule about trolling?

Label name: Voting
Instructions: Is this rule about voting in the community?

Category: Usage
Label name: usage--generate
Instructions: Does this rule use variations of the words "create" or "generate" to
     discuss content made with AI, such as AI-generated content?

Label name: usage--assist
Instructions: Does this rule reference using AI assistance, for example to enhance
    , alter, or edit images?

Label name: usage--general
Instructions: Does this rule mention AI, but does not mention a specific type of
    use?

Category: Rationale
Label name: rationale--lowquality
Instructions: Does this rule explicitly mention that content made with AI is low
    quality?

Label name: rationale--loweffort
Instructions: Does this rule mention that content made with AI is low effort?

Label name: rationale--spam
Instructions: Does this rule mention that content made with AI is spam?

Label name: rationale--misleading
Instructions: Does this rule mention that content made with AI is misleading?

Label name: rationale--inaccurate
Instructions: Does this rule mention that content made with AI is inaccurate?

Label name: rationale--inauthentic
Instructions: Does this rule mention that content made with AI is fake or
    inauthentic?

Label name: rationale--inhuman
Instructions: Does this rule mention that content made with AI is inhuman?

Label name: rationale--offtopic
Instructions: Does this rule mention that content made with AI is off topic?

Label name: rationale--policy
Instructions: Does this rule mention laws or site policies?

Label name: rationale--ethics
Instructions: Does this rule mention an ethical or moral objection to AI?

Label name: rationale--original_content
Instructions: Does this rule mention that content made with AI is not original
    content?

Category: Medium
Label name: form--text
```

```
Instructions: Does this rule explicitly mention text made with AI?

Label name: form--images
Instructions: Does this rule explicitly mention images generated or enhanced by AI?

Label name: form--videos
Instructions: Does this rule explicitly mention videos made with AI?

Label name: form--art
Instructions: Does this rule explicitly mention art made with AI?

Label name: form--deepfake
Instructions: Does this rule explicitly mention deepfakes?

Label name: form--fakes
Instructions: Does this rule explicitly mention fakes?

Label name: form--bots
Instructions: Does this rule explicitly mention AI bots?

Label name: form--content
Instructions: Does this rule explicitly mention AI content, for example AI-generated content?

Label name: form--general
Instructions: Does this rule govern the use of AI without explicitly referencing a specific type of use?

Category: Stance
Label name: stance--unqualified_ban
Instructions: Does this rule ban or disallow AI or some sort og content made using AI without providing exceptions?

Label name: stance--qualified_ban
Instructions: Does this rule ban some forms of AI use, but allow it in certain circumstances or if users label it?

Label name: stance--disclosure
Instructions: Does this rule require that users identify when the content that they post is made to some extent using AI, such as with a label, or flair, or certain text in the title?

Label name: stance--enabling
Instructions: Does this rule explictly say that some content made with AI is allowed under certain circumstances, for example if the fact that AI was used is disclosed?

Label name: stance--requiring
Instructions: Does this rule explictly require that posted content must be made with AI?
```

A.3 Applying Subreddit Classification Labels

```
You are to assist with a task to qualitatively code Reddit communities, which are known as subreddits.
```

```
You will be given a JSON string containing metadata about a subreddit that will contains its name, public description, description and rules.
For each category below, follow the instructions to determine which labels apply to the subreddit.
Return a JSON object where they keys are "Topical Question and Answer", "Learning and Perspective Broadening", "Social Support", "Content Generation", "Affiliation with an Entity", "Topic" and the values are the most appropriate label values.

Category: Community Archetype
Instructions: An archetype describes an abstract set of qualities that are likely to co-occur in individual instances of an object. A Community Archetype applies this concept to subreddits and classifies them into types of communities based on identifiable subreddit features and common user behaviors. The following labels each correspond to a Community Archetype. Consider each label and the characteristics associated with it and determine if it describes the given subreddit. For each label, choose the value Y if it best describes the subreddit and N if it does not.
Labels:
- Name: Topical Question and Answer
  Characteristics: Subreddits that are explicitly set up with strict rules to create a Question and Answer format around specific topics. Most top-level posts are questions, while most comments are answers, or discussions and enrichments of others` answers.
- Name: Learning and Perspective Broadening
  Characteristics: Subreddits that users visit primarily to learn or broaden their perspective. Top-level posts are often pointers to current events, publications, or relevant news, or relatable personal stories, experiences, and questions. Comments tend to elaborate upon the ideas raised, for example by celebrating or disagreeing, making jokes, providing contrasting or similar personal stories and experiences, or adding new ideas and references into the mix.
- Name: Social Support
  Characteristics: Subreddits where socially supportive behaviors are the main purpose and organizing principle for gathering. Top-level posts are often specific personal experiences or sensitive disclosures, questions about a health issue, announcements of milestones, reflections or venting, or sharing of resources, artwork, and encouraging thoughts and memes/jokes. Comments generally offer support, reflection, commiseration, resources, and validation.
- Name: Content Generation
  Characteristics: Subreddits that are devoted to content with a specific format, sense of humor, viewpoint, purpose, or type of creative expression. Top-level posts exemplify a subreddit`s specific content style and comments often include peoples` opinions on the content, extra information about the content, or sometimes commiseration with the original poster.
- Affiliation with an Entity
  Characteristics: Subreddits that have explicit affiliations with particular entities, such as geographical places or organizations (cities, universities), popular media (book or fan series, TV shows), sports teams, etc. Posts often feature local or breaking news and events related to the entity. In the comments, users generally express their feelings about news or events, and answer questions or offer advice and opinions about the entity.

Category: Topic
```

```
Instructions: Pick one of the following labels that best describes the topic of
    the subreddit. If multiple labels apply, choose the one that is most
    descriptive.
Labels:
- Video Games
- Image Sharing
- Entertainment
- Personal
- Technology and Science
- Celebrity
- Hobby
- Local
- Sports
- Support
- Music
- Art
- Humor
- Meta
- Business and Organizations
- Animals
- Work
- Writing
- Learning
- Politics
- Games
- News
- Health
- Food
- Parody
- Culture
- Fashion
- Drugs
```

Received 11 December 2024