“Community Guidelines Make this the Best Party on the Internet”: An In-Depth Study of Online Platforms’ Content Moderation Policies

Brennan Schaffner*
Arjun Nitin Bhagoji*
Siyuan Cheng
Jacqueline Mei
Jay L. Shen
University of Chicago
Chicago, Illinois, USA

Grace Wang
Marshini Chetty
Nick Feamster
Genevieve Lakier
Chenhao Tan
University of Chicago
Chicago, Illinois, USA

ABSTRACT

Moderating user-generated content on online platforms is crucial for balancing user safety and freedom of speech. Particularly in the United States, platforms are not subject to legal constraints prescribing permissible content. Each platform has thus developed bespoke content moderation policies, but there is little work towards a comparative understanding of these policies across platforms and topics. This paper presents the first systematic study of these policies from the 43 largest online platforms hosting user-generated content, focusing on policies around copyright infringement, harmful speech, and misleading content. We build a custom web-scraper to obtain policy text and develop a unified annotation scheme to analyze the text for the presence of critical components. We find significant structural and compositional variation in policies across topics and platforms, with some variation attributable to disparate legal groundings. We lay the groundwork for future studies of ever-evolving content moderation policies and their impact on users.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in collaborative and social computing**; • **Social and professional topics** → *Hate speech*; *Political speech*; *Governmental regulations*.

KEYWORDS

content moderation, dataset, qualitative analysis, quantitative analysis

ACM Reference Format:
Brennan Schaffner, Arjun Nitin Bhagoji, Siyuan Cheng, Jacqueline Mei, Jay L. Shen, Grace Wang, Marshini Chetty, Nick Feamster, Genevieve Lakier, and Chenhao Tan. 2024. “Community Guidelines Make this the Best Party on the Internet”: An In-Depth Study of Online Platforms’ Content Moderation Policies. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI ’24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3613904.3642333

*The first two authors contributed equally to this research.

CHI ’24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642333

1 INTRODUCTION

As much of the world’s discourse happens online, platforms that host and mediate online content play an increasingly influential role in moderating societal conversations. Platforms like Twitter, Threads, and TikTok have been compared to “global town squares” [58]. The practice of deciding whether to publish, remove, and flag content that is posted by third-party users is typically referred to as *content moderation* [34, 35]. This is what Goldman [24] refers to as “content regulation”. Since platform policies influence online behavior for millions of users and shape societal discourse, it is crucial to understand how content moderation policies are structured and what they contain.

Platforms that are faced with the prospect of moderating content face two primary challenges: (1) enforcing policies at scale; (2) ensuring that policies are applied consistently. First, as the *scale* of data to be monitored and moderated has increased, consistent content moderation has become extremely challenging. In part due to the scale of the problem, alongside regulatory pressures, platforms have increasingly attempted to rely on algorithmic automated content moderation [25]. Yet, in reality, automation has not been able to replace human decision-making: much content moderation is implemented by underpaid contractors, largely in the Global South [13], hired by large platforms [17]. A second goal of content moderation is *consistency* in how the policies are applied—across instances of content, users, geographies, and so forth. Greater consistency may lead to increased trust and improved discourse [35]. Unfortunately, public perception is that content moderation is inconsistently implemented [46, 49]. Keller et al. [34] also show it is often difficult to determine where, how and according to what rules content moderation is occurring.

In this paper, we seek a unified understanding of what these *rules* are in the first place. We look at how many online platforms specify their content moderation policies, with an in-depth study and structured analysis across a wide range of platforms. Fortunately, many platforms typically provide some level of transparency into their content moderation policies, allowing us to study them in some level of detail (including, for example, what is specified vs. unspecified, and how the structure and content of these policies vary across platforms). With the exception of copyright enforcement, content moderation largely lacks a prescriptive regulatory approach (particularly in the United States), leading to potential divergence of policies and inconsistency of application, even within a single platform. This state of affairs makes it paramount to study platforms' content moderation policies as they define public discourse, and being user-facing, influence the manner in which users interact and conduct themselves online.

We study content moderation policies for several topics, given that the nature of content moderation policies typically differs depending on the type of content. For example, copyright rules have a well-established legal regime, especially in the United States after the enactment of the Digital Millennium Copyright Act (DMCA), which makes it an insightful contrast to other types of content that might be moderated, such as misinformation and hate speech. To capture the ends of this spectrum in our work, we focus on three content moderation topics: (1) copyright infringement, (2) hate (or harmful) speech, and (3) misinformation (or misleading content). Each topic represents different intensities of legal grounding and has entered the public discourse about online content at different times. Moreover, the prominence of the three topics makes it likely that each studied platform will contain policies pertaining to each topic.

This paper performs the first in-depth collection, annotation, and analysis of content moderation policies, across the 43 largest online platforms (determined using Tranco [39]) hosting user-generated content. A better understanding of how content moderation policies are structured and what they contain may lead to improved alignment in platform policies, regulation, and user expectations. Towards this goal, we pose the following research questions:

(1) Collection: How do we systematically collect all text related to content moderation policies across a range of platforms?
(2) Annotation: How can we annotate content moderation policy text to reveal key components that are important for users and capture similarities and differences across topics and platforms?
(3) Analysis: How consistent are content moderation policies in structure and composition across different platforms, and how do they relate to existing legal frameworks?

The *collection* effort is substantial given the relatively unstructured nature of these policies, and the lack of standardized approaches to expressing them. Specifically, these policies are often scattered, imprecise and non-committal (past work has also demonstrated that at least in specific cases, enforcement may also be inconsistent [44]). Collecting policies across a wide range of platforms in a systematic manner to create a standardized, accessible dataset is thus an important first step to enabling both this paper and future work to reach more general conclusions concerning how platforms approach content moderation. Because data collection results in a large volume of text, having a clear *annotation* scheme to provide insights about the data at scale, particularly with regard to information provided to and actions for users of the platform to take. Finally, standardized and wide-ranging data collection allows for cross-platform and cross-topic *analysis*, which can provide insights about the role legal regimes, platform size, and other factors play in policy structure and composition. This paper presents the following contributions:

1. Open-source collection pipeline enabling continual collection (§4.1). We create an open-source pipeline[1] to collect and build a dataset of content moderation policies. The pipeline consists of a scraper and text extractor. Our scraper overcomes common hurdles for web-scrapers such as bot-blocking, dynamic page loading via javascript and rate limiting. The scraper includes an iterative process to obtain relevant policy text across a platform using parsing and keyword search.

2. Inductively-designed policy annotation scheme (§4.2) : We develop an annotation scheme (or codebook) that captures critical components of policies from a user perspective. The codebook, developed in consultation with legal experts, takes an intent-driven and user-centric approach to the text, aiming to highlight the purpose and the relevance of policy statements. It enables further mixed-methods analysis to bring out policy differences across platforms and topics, in terms of communication clarity and the impact of legal regimes.

3. Dataset of annotated policies (§5): The scraping and annotation processes result in a dataset we name OCMP-43 (Online Content Moderation Policies, set of 43 platforms) for convenience. The dataset consists of over 1000 annotated pages of policy text with tens of thousands of annotated segments.[2] We provide the open-source dataset, including annotated text, to enable further research into existing content moderation policies.[3] We show that our dataset contains policy text scattered across diverse areas of a platform's pages, further underscoring the value of our consolidation.

4. Mixed-methods analysis of policy content, structure and coverage (§6): We use our annotated dataset to analyze content moderation policies at scale. Our key questions concern platforms' stated intent and methods in content moderation policies, as well as possible impact on users. A few key findings are that (1) platforms rarely explicitly define what they intend to moderate; (2) platforms, depending on the topic, can rely heavily on users for moderation; and (3) except in the case of copyright infringement, users rarely have recourse after being moderated.

The results presented in this paper enable and encourage researchers to perform further detailed studies of content moderation policies. The custom data collection pipeline we have developed enables studies of the evolution of platform policies over time. Our annotated dataset can help highlight best practices and provide recommendations for future policies, as well as to find policies that may show platforms acting in bad faith. Finally, our dataset and preliminary findings lay the groundwork for future user studies to understand the interaction of users' with these policies, as well as large-scale audits to determine when policies are being consistently implemented.

[1]The repository can be found here: https://anonymous.4open.science/r/content-moderation-7AC0/README.md
[2]This paper's title is adapted from Imgur's policy pages.
[3]The dataset is made available at https://ocmp43.cs.uchicago.edu.

2 BACKGROUND

In this section, we first provide a brief history of online content moderation, with a focus on the legal and policy regime in the United States. This helps set the context for why analyzing the moderation policies of different platforms is meaningful, as they often serve as proxy regulators in the absence of focused government regulation. We then provide a brief survey of related academic work that has studied content moderation by online platforms.

2.1 History and overview of online content moderation

Since the development of the world wide web as a medium for information exchange, there has been a proliferation of online platforms where people gather to share information and ideas. Their evolution can be traced from the simplicity of early message boards, where users could linearly post text to respond to each other, to modern platforms like Twitch and Facebook where live video can be shared even as users comment on it. As these platforms have grown in scale and scope, the content being shared on them has begun to have real-world impact, including, in some cases, the incitement of violence [3, 41].

However, governments, particularly in the Global North, have been reluctant to hold platforms responsible for the content that is posted on them, out of fear that doing so will lead platforms to either take down too much valuable content, or refrain from moderating content altogether. In the United States, Section 230 of the Title 47 of the United States Code[4], explicitly provides platforms protection from liability for the third-party speech that they publish. This Section also protects platforms from liability for the "good faith" moderation of third-party content they deem objectionable. The result has been to make it difficult to impose legal liability on platforms for much of the speech that appears on their websites. In the absence of clear rules, detailing what they should do about harmful or controversial content, platforms have developed elaborate policies to guide their regulation of speech.

These rules are intended to guide users, and demonstrate to members of the general public that the platforms are exercising their power over speech responsibly. Despite these aims, it is often very difficult for either users or researchers to know how platforms are applying these policies, and whether they are doing so consistently. This is notwithstanding calls from civil society organizations for many years now for clear and consistent decision-making by the stewards of the digital public sphere [1]. Consistency and clarity are understood to be important goods in themselves—one of the rights that users are entitled to, when they speak on the internet—and also as a means of ensuring the legitimacy and thus the effectiveness of content moderation decision-making. The theory here is that consistently applied moderation criteria help regulate and make online communities safe for their participants, and also "increase the legitimacy and thus the effectiveness of moderation decisions" [35].

In this paper, we seek to understand what comprises moderation criteria for modern online platforms, as espoused by the platforms themselves. This is important for users, since in the absence of unified legal frameworks governing their behavior online, each platform's stated moderation criteria are all the user can use as a guide.

[4]This defined the structure and role of the Federal Communications Commission (FCC), the regulatory body that deals with communication via modern technologies such as radio and satellite.

2.2 Related Work

Content moderation in online communities: Content moderation has attracted substantial interest from computer science researchers because of its growing importance in online communities [35]. Content moderation can potentially filter "bad" content and "bad" actors and promote a healthy community, but may also infringe an individual's right to freedom of expression. Thus, there is a large body of research studying the effect of moderation on community behavior, including whether one should regulate online content at all [6, 9, 12, 14, 31, 43, 53, 56]. For instance, prior work has developed an observational method that leverages delayed feedback (*i.e.*, content moderation does not happen instantaneously) to understand the causal effect of comment removal on users' future behavior [56].

There is a also growing line of research on characterizing the rules on online platforms [11, 18, 20, 22, 33], all of which tend to focus on a single platform. For instance, some researchers analyzed wikipedia.com's publicly available records and editor discourse to uncover patterns in rules and rule-making communities [5, 27] and track rules over time [4, 33]. In another study, researchers [20] provided a characterization of different types of rules and show that community rules share common characteristics across subreddits. In comparison, other researchers [11] performed a large-scale study to understand content moderation through language used in removed comments on Reddit and identify norms that are universal (macro), shared across certain groups (meso), and specific to individuals (micro).

Our paper differs from this body of work in the scale of the cross-platform analysis we perform, enabled by a custom web-scraper and policy annotation scheme. By gathering and analyzing platforms' content moderation policies at scale, as well as releasing a fully annotated dataset, we set the stage for future user and audit studies that go beyond traditionally studied platforms such as reddit.com. The publication of our dataset takes some inspiration from Wilson et al. [62], who released an annotated dataset of privacy policies. However, our methodology and scope differ substantially. In particular, while they manually downloaded privacy policies, we diverged in our approach by developing a custom web-scraper to automatically retrieve and extract dispersed content moderation policies, showcasing a distinct approach to locating relevant policy text across the sites, and employing a novel annotation scheme. The closest related work to ours is Singhal et al. [54], which qualitatively surveys 14 different platforms' content moderation policies. Our work, in contrast, enables quantitative analysis due to our large annotated dataset and only 8 of the studied platforms overlap, with our work studying 35 additional platforms.

Legal research on content moderation: The importance of content moderation has spurred a robust legal discussion [23, 36, 57]. Three questions are outlined in previous work [24, 36]: (1) what content should be allowed online? This question is concerned with the First Amendment of the United States Constitution. As discussed in §1, the laws are much clearer for copyright infringement than for misleading content; (2) who should make the substantive rules of online content and activities? Klonick [36] further distinguishes standards (e.g., "don't drive too fast") from rules (e.g., "a speed limit set at sixty-five miles per hour") in the current practice of online platforms; (3) who should determine if a rule violation has occurred, and who should hear any appeals of those decisions? While existing research is mostly qualitative, our goal is to examine existing platform rules in detail and at scale.

User understanding of content moderation: There is a large and growing body of research around how users understand and react to content moderation [19, 21]. For instance, there are numerous studies that examine how reddit.com's users react to post-level and community-level moderation [7, 8, 10, 16, 28–32, 42]. Prior work has shown that while banning subreddits with hateful speech does decrease toxicity overall, however, toxicity within the banned subreddit may increase [10, 16]. In some cases, content moderation can increase toxicity and move users to more extreme behavior on other platforms [26]. Some studies also imply that users who know about community rules or who received explanations for their removals are often more accepting of moderation practices [28, 31]. Some user studies have also studied content moderators themselves [2, 47, 48].

Researchers have also examined how users react to algorithmic content moderation [60] and softer forms of content moderation such as and shadow banning [45, 63]. Most of this work focused on content moderation on a single platform, for a single topic. In contrast, our work aims to create an understanding of content moderation policies across multiple platforms by synthesizing what users see when trying to understand how they may be moderated.

3 RESEARCH SCOPE

In this section, we describe the research scope within which we analyze the content moderation policies of different websites. At a high level, we focus on three different topics across 43 of the 200 top websites from the Tranco [39] list[5] generated on 29 June 2022, which we determine to host user-generated content.

3.1 Topics of focus

Modern platforms host a wide range of content that may have to be regulated, either due to explicit legal frameworks that require it, or due to the platforms' own motivations, which could stem from ethical or economic considerations, or both. In this paper, we focus on three broad topics within which content that tends to get regulated often falls: *copyright infringment*, *harmful speech*, and *misinformation and misleading content*. Our choice of these three topics is motivated by three factors. First, we find policy pertaining to these three topics is highly prevalent across all the websites we consider, highlighting their importance and ubiquity. This is in contrast to niche or newly emerging topics that may not appear in all platform policies (*e.g.*, regulations concering deepfakes or Generative Artificial Intelligence). Second, these topics have been the focus of much recent debate and consideration, both in the public and academic spheres, due to their potential for direct harm to platform users [3, 40]. Finally, copyright infringement offers a meaningful contrast to the other two topics, due to the presence of strong copyright laws in many countries while the other content areas have fewer governmental regulations. We are aware that there are other areas such as child pornography where moderation is common, but we limit our scope in this paper for focus, and to avoid the additional complications of exploring these areas with complex ethical considerations.

3.1.1 Copyright Infringement. There are complex legal regimes that regulate intellectual property around the world, with a particularly prominent one being the Digital Millennium Copyright Act (DMCA) [15] in the United States. Although the DMCA does limit platforms' liability for copyright violations by content posted on them, they nonetheless tend to have well-defined policies that specify actions users can take with regard to copyrighted content.

3.1.2 Harmful speech. Naturally, platforms may restrict content that is directly harmful to other users of the platform, as this can reduce user engagement [35], leading to direct economic concerns. We refer to content of this form as *abuse*. In addition to content that directly targets other users, there is an increasing prevalence [25] of content that is offensive towards groups of people sharing characteristics, which is broadly referred to as *hate speech*. We group these two types of content under the area of harmful speech.

In many contexts, it is challenging to distinguish harmful speech from provocative or satirical speech, especially in light of cultural norms that vary both geographically and temporally. A lack of clear definitions of what constitutes harmful speech—and its increasing scale in recent years—has made it challenging to moderate. In the United States in particular, the First Amendment to the Constitution protects speech in a variety of contexts, leading to the absence of clear legal principles within which harmful speech can be regulated. This legal regime is in contrast to Germany, for example, where the Network Enforcement Act (NetzDG) [37] explicitly requires platforms to take down hate speech. There is thus significant variation in how platforms deal with the presence of harmful speech, making it an interesting area to study.

3.1.3 Misinformation and Misleading Content. As the reach of online platforms increases, the impact of and trust in conventional sources of information has eroded. This came to light in recent political campaigns in the United States and India, as well as during the course of the COVID-19 pandemic. However, the ease of spread of misinformation on online platforms has had debilitating consequences on public health and safety, making it a critical area for moderation for platforms [40]. The polarization regarding what is misinformation or not, particularly in light of mistrust of traditional institutions like universities and the government, makes it challenging for platforms to moderate without alienating portions of its user base. With rapidly changing conditions, it is often unclear what the consensus is on certain critical topics, which makes it difficult to moderate in real-time.

In addition to misinformation regarding public health and safety, we also categorize spam, fake products, and false advertising under misleading content, as these are all types of misleading content that platforms need to regulate in order to not erode user trust. Depending on the type of content hosted on the platform, there are subtle variations in how misleading content appears (*e.g.,* clothing

[5]Tranco provides a manipulation-resistant and publicly available list of the top websites for researchers. The list we use is accessible at https://tranco-list.eu/list/Q9X24/full.

Table 1: The 43 platforms in our dataset of online content moderation policies (OCMP-43), ordered by their Tranco ranking.

Platforms in OCMP-43
facebook.com, youtube.com, instagram.com, twitter.com, linkedin.com, wikipedia.org, amazon.com, pinterest.com, github.com, reddit.com, vimeo.com, wordpress.com, msn.com, tiktok.com, xvideos.com, tumblr.com, pornhub.com, nytimes.com, flickr.com, fandom.com, ebay.com, imdb.com, medium.com, soundcloud.com, aliexpress.com, twitch.tv, stackoverflow.com, archive.org, theguardian.com, bbc.co.uk, xhamster.com, quora.com, w3.org, sourceforge.net, indeed.com, etsy.com, sciencedirect.com, booking.com, imgur.com, spankbang.com, researchgate.net, washingtonpost.com, xnxx.com

with misinformation printed on it is found on Etsy). Throughout the remainder of the work, we use 'misleading content' to refer to all of these forms of misleading content and misinformation.

3.2 Platforms of focus

There is a proliferation of online platforms that host user-generated content. Our focus in this paper is to provide a method to collect content moderation policies from a wide set of platforms, as well as to collate a large dataset of policies. We considered the top 200 websites from the Tranco list [39] obtained on the 29th of June, 2022[6]. Of these, we filtered out those websites that do not host user-generated content, such as google.com, akamaiedge.net and baidu.com, as well as those which may host user-generated content, but are not primarily in English[7] such as bilibili.com. We also combined websites that correspond to the same platform, and thus use the same set of content moderation policies, such as wikipedia.org and wikimedia.org. The complete list of the resulting 43 platforms we consider is presented in Table 1.

To determine if a platform hosts user-generated content, and if it is primarily in English, we use manual analysis. We visited each website in turn, and checked if any portion of the platform contained user-generated content, which would in turn indicate the possible presence of policies regarding content moderation. For example, while booking.com largely contains links to stay and transport options for travelers, we include it in our dataset since it also hosts a comment and feedback section where users can interact with each other, governed by platform policies on moderation.

The resulting list of platforms host a diverse spectrum of user-generated content. For instance, large social media powerhouses like facebook.com, instagram.com, and twitter.com feature an array of multimedia content, including text, images, videos, and live streams. The variety of media is accompanied by a variety of intents, from consuming and debating news to friendly banter. On platforms such as linkedin.com and github.com, the focus shifts towards professional networking and collaborative coding, leading to distinct user generated content such as resumes and code documentation. The wealth of wikipedia.com's content comes from world-wide volunteer contributions to articles. Sites such as reddit.com and tumblr.com foster semi-siloed communities that generate content based on specific themes and, in some circumstances, are expected to self-moderate. Platforms like pornhub.com and xvideos.com cater to explicit adult content, posing unique challenges for content moderation. Meanwhile, news platforms must govern both their journalists' content and the community's comments, with heightened expectations of content accuracy. Further, e-commerce platforms like amazon.com and etsy.com take on additional responsibilities regulating both vendor content (e.g., product listings and FTC advertising regulations) and buyer content (e.g., reviews and ratings). Our dataset of content moderation policies captures the tailored content moderation policies necessitated by these variations in content types and user dynamics.

[6]The list can be accessed at https://tranco-list.eu/list/Q9X24/full

[7]This is a necessary limitation as English is the only language all authors of the paper are sufficiently proficient in. Regardless, our method for scraping can be adapted for platforms in other languages with some modification.

4 CREATING AND ANNOTATING THE DATASET

In this section, we outline our procedure for obtaining relevant content moderation policy text from the 43 websites and three policy areas we detailed in §3. We then describe the framework for discussing critical components of moderation policies and how we use this framework as a codebook to annotate the dataset. Figure 1 summarizes the pipeline that we describe in this section.

4.1 Dataset creation

This section describes the process for building the dataset of policy text, including the design and implementation of our custom web scraper.

4.1.1 Locating the policies. Modern web platforms, especially the most popular sites in our list, lack a consistent structure for distributing information regarding content moderation policies across a site's pages. Although almost all the platforms do feature a *Terms of Service (ToS)* page that outline key policies, including those regarding content moderation, many sites also have separate *Community Guidelines/Standards*, *Help Centers*, and/or official blog posts. Moreover, a particular platform's policies may be hosted on another domain entirely—often on parent platforms or customer service platforms. For instance, many of facebook.com's policy pages can be found on meta.com, and reddit.com's policies are on both redditinc.com and reddithelp.com. We first manually explored the 43 websites of focus and recorded URLs for each site's key policy pages. In most cases, we found the URLs corresponding to the *ToS*, as well as the *Community Guidelines/Standards* and *Help Center* if they existed—we refer to these as *canonical links.*

Informed by the manual exploration, we then curated a list of keywords that were closely associated with the three policy topics of interest (see Table 2), which we refer to as the *topic-wise keyword list* henceforth. We searched for these keywords paired with platform names on generic search engines such as google.com, as well as on the platforms themselves, if they had a *Help Center* or other searchable database of links. The aim of this process was to ensure we included URLs from diverse parts of each platform such that our subsequent scraping would be able to find all relevant policy text. The links obtained with this process, along with the canonical links, formed the set of *seed links* for the web scraper, which resulted in an average of 17.8 links per website, approximately evenly distributed

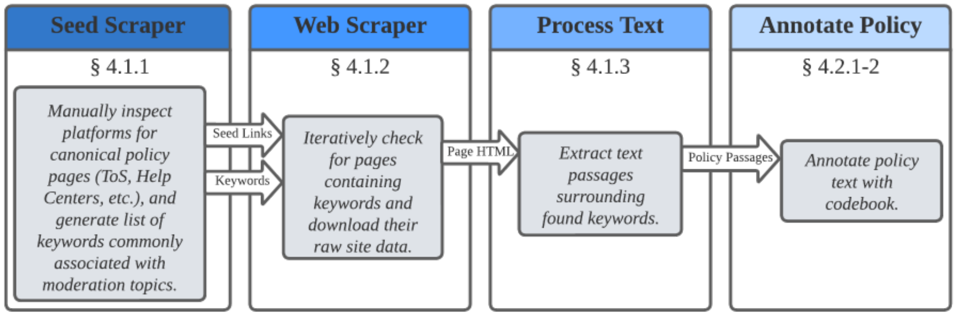

Figure 1: Pipeline for coding and annotating content moderation policy dataset OCMP-43.

Table 2: The keywords we associated with each content moderation topic.

Copyright Infringement	Misleading Content	Harmful Speech
['copyright', 'dmca']	['misinfo', 'mislead', 'disinfo', 'authentic', 'trust', 'integrity', 'misrepresent', 'impersonat', 'manipulat', 'decept', 'deceive', 'spam', 'fraud', 'fake', 'false']	['hate', 'abus', 'violen', 'discrimin']

across the three topics. For the full list of the seed links we used, see §A of the Supplementary Materials.

4.1.2 Scraping the policies. We use the set of seed links as the starting point to explore all web pages on a given platform that may contain topic-related content moderation policies for that platform. We designed and implemented a custom web scraper, as we found existing open-source solutions such as Scrapy [52] and MAXQDA's [55] built-in scraper often failed to retrieve text from pages of interest due to bot blocking, dynamic loading of webpages using JavaScript, as well as direct rate limiting for queries from a particular Internet Protocol (IP) address. Our custom scraper, which is open-sourced along with this paper[8], works as follows:

(1) We use a headless Selenium browser using the undetected-chromedriver (with edits[9] [59]) to download the HTML page source for each of the seed links. This driver helps bypass standard bot-blocking techniques.
(2) For each HTML page, we parse the HTML to identify embedded links. If the links lead to a non-empty page, we check if any of the text on the page matches a keyword in the topic-wise keyword list. If it does, we batch the page for further exploration, as well as download its HTML for analysis. Page sources are stored separated by which moderation topic they relate.
(3) For each batched page, we repeat steps (1) and (2) until we have visited every page two hops from the seed links, scraping those that have any text matching the topic-wise keyword list. Note that the dataset is entirely text, meaning we do not retain images, interactive layouts, bullet points, or other non-text elements.

The scraper can systematically locate and scrape more policy pages than is possible from manual scraping alone. In fact, early deployments of the scraper included more pages than necessary, often traversing URLs that were not on the original platform at all. To address this issue, we created an allow-list and block-list for each platform which matches URL patterns for desired web pages, and omits webpages that are clearly unrelated to the platform in question. In addition, due to the unrestricted nature of webpage design, and our specific requirements, we did not follow all links we found from a given page. For instance, we excluded mailto, fragment, and anchor links. We also had to join relative links, and follow redirection links. As per our earlier study design choice, we did not include pages that were not in English which we detected using langdetect's pretrained models [38]. A more detailed description of our iterative scraper design in response to issues we encountered is in §A of the Supplementary Materials.

We conducted measurements from May to July 2023, resulting in text from 8514 policy pages from all 43 platforms. In some cases, in spite of our best efforts at trying to bypass rate limiting, the scraper was unable to download the page source from relevant pages. We manually copied the text from these 115 pages, forming 1.4% of the total number of pages we scraped.

4.1.3 Extracting policy text. Even on the pages for which we identified as having content moderation policy keywords, the resulting data contains a wealth of text that is irrelevant to content moderation. This is byproduct of Step (2) in the method, which only requires a single word on the page to match the topic-wise keyword list for us to collect the corresponding text. There could thus be a host of irrelevant text on the page, adding overhead to the annotation process. For example, a *ToS* page may have large amounts of text reserved for subscriptions and fees, or text scraped from a *Help Center* post may include menu headers. To restrict the dataset to include only relevant policy text, we aim to extract only the passages from a page that contain the keywords. To this end, we parsed the raw text into sentences and included the 5 sentences before and after the sentence containing the keyword, merging

[8]Code Repository: https://anonymous.4open.science/r/content-moderation-7AC0/
[9]We had to locally edit this package's contents to work with our version of Chrome and to work with python multiprocessing.

Table 3: The policy annotation scheme. Subcodes (where applicable) are listed below top-level codes.

Code	Memo
POLICY JUSTIFICATION	
Community, Trust, & Safety	References to community values or user trust safety as motivation for policy.
Legal	References to extant legal frameworks for motivation for policy.
MODERATION CRITERIA	
Definition	Definitions clarifying content that is not allowed.
Example	Examples of content that is not allowed; can also be broad description of content types.
Exception	Explains content that is allowed with aim to delineat border-line cases or explain special circumstances where otherwise violative content is allowed.
SAFEGUARDS	
Active User Role	When users play an active role in the content moderation, such as reporting and flagging content.
Platform Detection Methods / Prevention Initiatives	When platforms employ methods to safeguard against violative content, such as automated detection technology and moderator training initiatives.
PLATFORM RESPONSE	
User-Targeted Enforcement	Responses to becoming aware of violative content that focus on the user that posted the content.
Content-Targeted Enforcement	Responses to becoming aware of violative content that focus on the content itself.
Investigation / Review	Responding to potentially violative content by investigating context or gathering more information.
REDRESS / APPEAL	User pursuit of an enforcement being reconsidered/overturned.
BINDING LEGALESE	
Liability	Platform explains their lack of liability for actions related to content moderation policy (non-)enforcement.
User Rights Altered	Policy text that calls out the altering of user rights related to content moderation; often includes phrases such as "you warrant/agree."
SIGNPOST	When platform policy links to information on another page such as other policy pages or third-party resources.

overlapping passages as needed. Each passage is then labeled along two axes: which platform the text came from, and which topic of content moderation the text referred to—Copyright Infringement, Harmful Speech, or Misleading Content.

4.1.4 Ethical considerations around scraping. We argue here that our scraping method abides by research ethics, even though the Terms of Service of most the platforms we consider do not permit scraping. First, the ruling of the United States District Court for the District of Columbia in the case of *Sandvig vs. Sessions* established [61] that it is legally permissible for researchers to use automated tools to collect information from websites, in spite of the provision of the Computer Fraud and Abuse Act (CFAA). Second, we do not scrape text from any pages that contain the personal data of users. In fact, for most platforms, we considered only pages that were accessible without logging in to the platform, which is clearly publicly available information.

4.2 Annotation Scheme for Dataset Analysis

To provide further structure to the large volume of text gathered from the content moderation policies of the platforms we consider, we devised an annotation scheme (or "codebook") to label relevant sections of text. Such an approach is commonly used in the social sciences to both categorize and extract meaning from language-based data [50]. As is common, we combined both deductive and inductive approaches to codebook development. Using the deductive approach, we developed an initial list of codes that could help answer our hypotheses with regard to the composition of content moderation policies, and their link to existing legal regimes. We focused on codes that could capture the *purpose* of a portion of policy text, especially as it pertained to how a user would interact with it. We were guided by the principles outlined in Kiesler et al. [35] for healthy online communities, with emphasis on "moderation criteria, a chance to argue one's case, and appeal procedures" as key components of a comprehensive content moderation policy. The team, which includes a legal expert on free speech laws, iteratively refined the codebook over multiple cycles of coding subsets of the text corpus. The codebook thus both informs the analysis of the text, and is informed by the text itself. The codebook serves the dual purpose as both an annotation schema and a framework of critical components for content moderation policies. The complete codebook is presented in Table 3. The codebook makes use of both high-level categories for the codes (e.g., 'Policy Justification' and 'Platform Response') and more specific sub-codes (e.g., 'Policy Justification > Legal' and 'Platform Response > User-Targeted Enforcement').

Four members of the research team were involved in iterative codebook development and coding process. For each platform, one coder annotated all policy-page text with the final codebook in MAXQDA. The first coder annotated all the policy pages for the site (across all three topics of interest *copyright, harmful speech*, and *misleading content*) so that they could understand the required context of a platform's site-wide policy, such as self-referenced pages and guidelines. A secondary coder then checked through the annotated policies, indicating spots for further discussion, which we discussed and resolved in many recurring research meetings. Each coder performed both primary and secondary coding split across different platforms, dispersing perspective and accountability around the dataset. An example of a coded policy passage is shown in Figure 2.

The policy annotation enables powerful mixed-methods analysis (see §6). First, it allows us to effectively locate instances of policy text across the different codes, which we rely on as evidence in support of hypotheses regarding the policies. Second, code coverage can indicate platforms' focus on a given topic, as well as differences across platforms. Our final codebook serves as an initial taxonomy for the critical components of content moderation policy. For example, 'Moderation Criteria' refers to **What** is moderated (and what may not be); 'Policy Justification' refers to **Why** content is moderated; 'Safeguards', 'Platform Response', and 'Redress / Appeal' are different temporal aspects of **How** the process of content moderation manifests. Further, 'Binding Legalese' provides information on **Whom** liability and responsible falls for content, while 'Signposts' describes the structural components of the policy text.

4.3 Methodological Challenges and Limitations

With any data collection and annotation process of this scale, there are a number of challenges we faced, and limitations with respect to the final dataset. We document these in this section so users of the collection pipeline as well as the dataset can adjust future usage and analysis accordingly.

4.3.1 Data collection. Our scraper implementation introduces several limitations that may affect this work's conclusions. First, relying on keyword for corpus creation introduces the possibilities of a dataset not perfectly representative of extant content moderation policies. To combat this issue, we removed false positives (unrelated policy text) when applying qualitative codes, and we added the iterative scraper to reduce false negatives (uncaptured policy text). Still, there may be auxiliary policy text that was inaccessible to our scraper and therefore not represented in our dataset. Second, if a platform changed its policy pages during the study, we cannot guarantee that our seed links or allow/block-lists are up-to-date. Third, we discovered that a substantial fraction of the extracted text files contained no relevant policy due to the ubiquity of terms such "copyright" and "safety" in the headers and footers, which triggered our keyword matching. However, due to the unstructured nature of HTML and inconsistent choices across different platforms (and even pages within the same platform), we found no clear way to distinguish portions of a webpage that comprised the body. We thus left this filtering to the (manual) annotation step, which increased the load on the coders. Finally, we ended up with number of pages that had duplicated content but different URLs. We relied on the annotation process to exclude these from the analysis, but different design goals could change the trade-off between the reliance on the scraping versus the annotation process.

4.3.2 Data annotation and analysis. First, the annotation scheme was developed on a relatively small number of platforms in comparison to all possible of platforms that host user-generated content. However, these platforms still generated a large amount of policy text and taken together, represent a large portion of the Internet's user-generated content. We are thus confident that the annotation scheme will remain useful for the platforms that host user-generated content. Second, although we took care to disambiguate the labels in our annotation scheme as much as possible from each other, each coder still had to make choices about certain labels that may be close. In particular, deciding when policy text referencing moderation criteria was an EXAMPLE and when it functioned as a DEFINITION was challenging. We often saw the two interleaved, with definitions bleeding into examples and *vice versa*. We erred on the side of marking moderation criteria largely as examples unless they specifically included phrases such as "we define". Our choice here directly impacts Finding 3. Third, it is noteworthy that a platform's stated policies may differ from their actual moderation practices. This study focuses on what platforms communicate rather than on their operational conduct. While we acknowledge potential gaps between policy and action, we view this as a motivation for our work. Recognizing what platforms express they do is fundamental to grasping their operational reality.

Finally, the process of categorizing platforms for analysis poses significant challenges, as platforms often possess multiple facets that defy straightforward classification. Consequently, we resist analysis based on platform categorizations due to this inherent difficulty, apart from one exception where we analyze the completeness of policy components for select platform categories—specifically, those where we can clearly delineate closed, exhaustive groups from among the 43 platforms. We concentrate on three categories: (1) *Adult content*: pornhub.com, xhamster.com, xnxx.com, spankbang.com, and xvideos.com; (2) *News*: nytimes.com, theguardian.com, bbc.co.uk, and washingtonpost.com; (3) *E-commerce*: amazon.com, ebay.com, etsy.com, and aliexpress.com.

5 DATASET

In this section, we provide an overview of the annotated dataset (OCMP-43) generated using the method that we described in §4. We provide statistics about the composition of the annotated dataset, as well as where the policy text was located on respective pages.

5.1 Dataset descriptors

We create a repository of text files organized first by platform, and subsequently by content topic for each website from a scraper run at a specific point in time. Each text file contains policy text from a single page, organized by passages within. It also indicates the URL of the page to enable reproducibility. Each passage is labeled with the keyword from the topic-wise keyword list that led to that passage being included, which in turn led to the page being included. We note that each file can contain passages with policy text

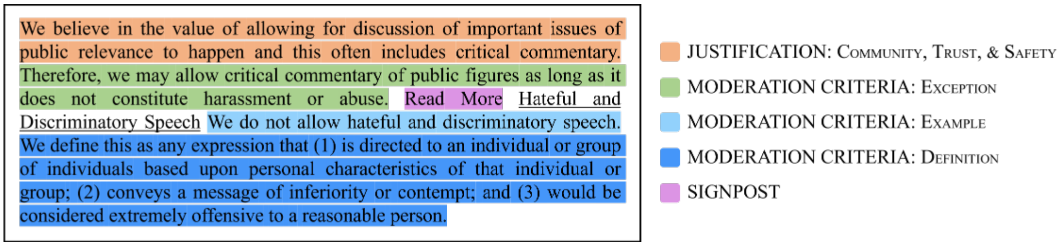

Figure 2: An example of an annotated policy passage.

Table 4: Descriptive Metrics for the Dataset OCMP-43

Descriptive Metric	Copyright Infringement	Harmful Speech	Misformation	Total
No. Coded Segments	3,953	3,034	4,374	11,361
Coded/Total Pages	390/2,739	342/1,546	570/4,229	1,302/8,514
Coded/Total Characters	475,631/8,275,886	401,294/5,580,974	584,525/19,671,028	1,461,450/33,527,888

corresponding to different keywords. During the annotation process, we noticed that the same policy text was sometimes repeated across different pages. We used our discretion in flagging these to be excluded in any analysis.

In Table 4, we provide an overview of the annotated dataset. There is a clear difference in coding coverage across the three topics, with harmful speech having the largest fraction both in terms of pages coded (21.7%) as well as characters coded (7%). This occurs due to the frequent presence of terms like "copyright" and "spam" or "trust" in the headers and footers of webpages, which leads to our scraper acquiring pages that were not coded. §B of the Supplementary Materials contains a more detailed platform-wise breakdown of the metrics from Table 4.

Computation and memory requirements: The complete dataset comprises 8,514 text files, taking about 35 MB uncompressed. We ran the scraper and extractor on a Intel Xeon server equipped with a AMD EPYC 7502P 32-Core Processor. With our default settings, the process of scraping and extracting the complete dataset took about 86 compute hours (which we reduced with multiprocessing).

5.2 Where is Policy Text Located?

From a user's perspective, it is important to be able to locate policy concerning content that they think may be subject to moderation. They may be checking to determine if they are likely to face enforcement for content they wish to post, be looking for measures to flag posts from another user, or seeking recourse if they feel they have been unfairly moderated. In light of these user needs, we analyze where policy regarding content moderation is located on different platforms and how straightforward it is to navigate to the policy from the landing page.

Content moderation policies are typically found in three different places across different platforms: the *Terms of Service*, a *Help Center*, and an additional page (or set of pages) that broadly defines how users should conduct themselves on the platform. For brevity, we refer to the latter as *Community Guidelines*, although different platforms term these differently, with variations like 'community rules', 'code of conduct' and so forth.

As we expected to some degree, all but one of 43 platforms we consider has a ToS page, which in all cases is linked directly from the landing page of the platform. The World Wide Web (W3) Consortium's website is the only one that does not link to a ToS. The ToS contain broad instructions and information on the conditions under which users will use the platform and its services. It also contains what we call `Binding Legalese` in our codebook (see Table 3), which is text that fundamentally alters users' rights when using the platform. While a majority of platforms have ToS accessible directly from the landing page, this is not true for the *Community Guidelines*.

Location Observation 1: 79% (34/43) of platforms have page(s) dedicated to *Community Guidelines*, where platforms lay out their expectations of user behavior, especially as pertaining to inter-user interaction. However, only 35% (12/34) of these platforms link to the *Community Guidelines* from the landing page. During the process of collecting the dataset, we added the *Community Guidelines* as a canonical link by explicitly looking for them via external search engines. We were driven to do this to be exhaustive as motivated by our research questions, but this is not the approach most users have when interacting with a platform. We thus find it significant that only about a third of the platforms we consider link to the *Community Guidelines* directly from the landing page. This observation raises interesting questions for future user studies to determine the extent to which users are aware of the presence of *Community Guideline* and consciously hew to them when interacting on the platform.

Location Observation 2: 84% (36/43) of platforms have a *Help Center* that directs users to policy relevant to content moderation, with 97% (35/36) of these help centers linked from the landing page. In addition to the 36 platforms with *Help Centers* that contain text relevant to content moderation policy, three others have help centers but do not contain any relevant policy. Interestingly, one platform, quora.com, required users to be logged in to access both the *Community Guidelines* and *Help Center* from the landing page.

From a user perspective, it is encouraging that most large platforms have an easily accessible help center. However, these only lead to relevant policy when using terms from a topic-wise keyword list, which users may not think of when searching for these pages.

Location Observation 3: Content moderation policy is found scattered across pages including, but not limited to, *Transparency Centers*, privacy, advertising, developer, and seller policies, and blog posts. Platforms have a diversity of user types based on their purpose, such as advertisers, sellers, journalists, researchers, and property owners; with each type of user expected by the platform to conform to certain behavioral norms. We find that this diversity correlates with the profusion of locations on a platform where content moderation policy may be located. In addition, platforms can announce new policies and initiatives via internal newsrooms or blogs. Although the intent of the scattered policies is often similar, the language and organization differs based on the user type targeted.

6 FINDINGS

In this section, we present our findings from analysis of the content moderation policy corpus (§4.1). We use the codebook developed in §4.2 to analyze the text from the corpus both qualitatively and quantitatively. Our complete annotated dataset is large, containing thousands of pages and tens of millions of characters. We present the following findings by introducing a question followed by findings of statistical trends, for which we employ percentages of coding frequencies.[10] We then support the statistical trends with representative quotes from the dataset of policy text. Note all emphasis is our own.

6.1 Why do platforms say they moderate?

Our initial analysis of the policy text revealed that platforms largely cite two reasons for content moderation: meeting legal requirements (LEGAL) and maintaining community standards (COMMUNITY, TRUST, & SAFETY). We study the variation of these reasons across the three moderation topics.

Finding 1. Of all the policy text that cites meeting legal requirements as the justification for moderation, 73.5% is for copyright infringement, with only 13.8% for harmful speech and 12.7% for misinformation and misleading content.

We hypothesize that the presence of strong legal regimes for copyright infringement in the U.S. (e.g., DMCA) as well as other countries leads to legal requirements being cited more commonly under the topic of copyright infringement, as shown in this exemplary text from microsoft.com:

> *"However, **we are generally required by law** to disable access to copyrighted content (including videos, music, photographs, or other content you upload onto a Microsoft website) if the copyright holder claims that the use of the copyrighted work is infringing." (microsoft.com's Copyright FAQ)*

Laws such as "NetzDG" from Germany and the US Federal Trade Commission (FTC) regulations on false advertising likely account for the presence of legal justification for moderation in the other topics. The following policy text represents a typical example from the dataset that captures both major reasons platforms moderate (legal requirements and community values):

> *"We take a two-step approach to reviewing content that is reported through the NetzDG reporting form. First, **we review the reported content under our Community Guidelines.** [...] Second, if the reported content doesn't violate our Community Guidelines, [...] **we review it for legality** based on the report." (instagram.com's Help Center)*

Across all three topics, it is insightful to compare how often the justification is that the platform needs to maintain community standards.

Finding 2. For harmful speech and misleading content, platforms reference COMMUNITY, TRUST, & SAFETY as their motivation for moderation 83.2% and 82.0% of the time, respectively, compared to just 21.3% of the time for copyright infringement.

When platforms are not guided by comprehensive LEGAL structures (such as the case with most harmful speech and misleading content), they rely on their VALUES, TRUST, & SAFETY, exemplified by Tumblr's following pointed policy justification:

> *"Our users, as decent human beings, don't like that kind of thing." (Hate and Harassment section of tumblr.com's Global Advertising Policy)*

6.2 How do platforms describe what they moderate?

Platforms need to clearly specify what is likely to get moderated, and what is not, since this provides clarity to users, fostering healthy communities [35]. However, across all three topics, we found that platforms established what they were likely to moderate using EXAMPLES, instead of DEFINITIONS.

Finding 3. DEFINITIONS only comprise 7.5%, 6.6%, and 2.3% of moderation criteria in policy text regarding copyright infringement, harmful speech, and misinformation and misleading content, respectively. Describing moderation criteria by EXAMPLE accounts for the rest.

Even when platforms did choose to include definitions for the content they moderate, it was usually accompanied by a list of clarifying examples, such as:

> *"**Hate speech is defined as** a serious attack on a group or individual based on their race, ethnicity, gender, nationality, sexual orientation, sex, religion, caste, serious medical condition, or disability. [...] **Examples include:***

[10]We also analyzed trends using code coverage rather than code frequencies, calculating percents using length of coded text instead of number of occurrences of a code. Since all trends were consistent between both metrics, we opt to present code frequency. See §B of the Supplementary Materials for detailed tables and equations used for each statistic reported as well as alternative methods to calculate relevant statistics.

- Statements like '⟨insert race/ethnicity⟩ are not welcome in our country'." (quora.com's Platform Policies)

The specificity of what definitions we did find also varied by moderation topic, with harmful speech and copyright infringement tending to be more clearly defined as shown by the previous example for harmful speech and the following for copyright infringement:

> *"**Copyright infringement is** doing any of the following without permission from the copyright owner(s): making copies, distributing the work (such as uploading to SoundCloud), performing or displaying the work publicly, or making 'derivative works'." (soundcloud.com's Help Center)*

For misleading content we found most platforms adopted the following typical, rather vague tone and lack of specificity:

> *"We define misinformation as content with a claim that is determined to be false by an authoritative third party." (facebook.com's Community Guidelines)*

Examples are used either to expand on definitions in simpler, procedural terms as exemplified by this quote:

> *"If you do perform a cover song in a live Twitch stream, please make a good faith effort to perform the song as written by the songwriter(s), and **create all audio elements yourself, without incorporating instrumental tracks, music recordings, or any other recorded elements owned by others.**" (twitch.tv's Music Guidelines)*

Otherwise, in lieu of definitions entirely, platforms sometimes sought to establish concepts by example:

> *"Refrain from using broad and vague terms that have a potential to mislead your buyers (such as 'environmentally friendly' or 'eco-friendly')." (etsy.com's Recycled Content Policy)*

This dichotomy between definitions and examples both being used to guide users towards acceptable behavior on online platforms mirrors that between "rules" and "standards" when differentiating legal from illegal behavior as stated in [36, 51].

We also found surprisingly specific examples of disallowed harmful speech:

> *"Do not post [...] Dehumanizing speech or imagery in the form of comparisons, generalizations, or unqualified behavioral statements [...] (including but not limited to: **[censored explicit list of several harmful stereotypes]** [11])." (facebook.com's Community Standards)*

The positive impact of these type of examples are unclear [35] and could be triggering for some users.

Finally, platforms sometimes specified conditions under which content that would typically be moderated is allowed. For copyright infringement, EXCEPTIONS mostly referenced fair use, such as:

> *"**[U]sers are allowed to use copyright works** without the authorization of the copyright holder **for quotation, criticism, review and for the purpose of caricature, parody or pastiche provided that such use is fair**." (tiktok.com's Intellectual Property Policy)*

For EXCEPTIONS to harmful speech, platforms considered educational or historical value:

> *"**We may label rather than remove content** that evokes hateful rhetoric (including slurs) **in the context of counter speech, reclamation, or members' personal experiences** with racism, sexism, ableism, and other forms of prejudice or discrimination." (linkedin.com's Help Center)*

In some cases, platforms even considered newsworthiness as an EXCEPTION to their misleading content policies:

> *"Related to the elections: **We may not take action on violating content that is deemed newsworthy**." (tiktok.com's Election Integrity Policy)*

6.3 How do platforms find content that may need moderation?

In order to detect and then subsequently act upon content that may need to be moderated, we find platforms in general largely use three methods: automated detection, human moderators, and users flagging content, as summarized by this representative quote:

> *"Pornhub moderates user-uploaded content in three major ways: through the use of **automated detection technologies**, through **real-life human moderators**, and through **user-generated reports**. " (pornhub.com's Help Center)*

We categorize the use of human moderators and automated detection technologies together under PLATFORM DETECTION METHODS / PREVENTION INITIATIVES, and user-generated reports under ACTIVE USER ROLE to illustrate the potential imbalances between platform and user roles.

> **Finding 4.** Platforms rely heavily on an active user role when designing safeguards against content from all three topics that may need moderation. A large fraction of policy text that delineates safeguards for copyright (83.3%), harmful speech (61.5%), and misleading content (51.0%) references users taking an active role.

Some platforms boast successful automated approaches, as captured by this quote typifying what we saw:

> *"While this is an ongoing journey that we continue to refine, we're pleased that the metrics in our most recent Transparency Report show that, in the first half of 2021, close to 66.3 million violative pieces of content were removed from the site. Of these, **99.6% were removed through our automated defenses**." (LinkedIn's Data Science Manager's Blog Post)*

These automated approaches, platforms claim, are especially successful for blocking spam, a well-studied problem since the advent

[11] We redacted the list of stereotypes to minimize potential harm to readers. The original text can be found on the following page, which contains examples of stereotypes that some readers may find offensive: https://web.archive.org/web/20231120210834/https://transparency.fb.com/policies/community-standards/hate-speech/

of digital communication. In other domains, such as harmful speech and misleading content, automated detection methods are less viable when policy-violating language cannot be articulated and used for automatic detection. For instance:

> *"**Misinformation is different from other types of speech addressed in our Community Standards because there is no way to articulate a comprehensive list of what is prohibited.** With graphic violence or hate speech, for instance, our policies specify the speech we prohibit, and even persons who disagree with those policies can follow them. With misinformation, however, we cannot provide such a line. The world is changing constantly, and what is true one minute may not be true the next minute." (facebook.com's Community Standards on Misinformation)*

The varying types of user-generated content also prove troublesome for automated approaches. For example, twitch.tv, a platform featuring mostly *live* video streaming and *live* chat engagement, discusses their unique challenges for automation:

> *"**Content moderation solutions that work for uploaded, video-based services do not work, or work differently, on Twitch.** Through experimentation and investment, we have learned that for Twitch user safety is best protected, and most scalable, when we employ a range of tools and processes, and when we partner with, and empower, our community members. The result is a layered approach to safety—one that combines the efforts of both Twitch (through tooling and staffing) and members of the community, working together." (twitch.tv's NetzDG Transparency Report)*

Even with advanced automated approaches for flagging potentially problematic content, human reviewers are often needed to determine whether the flagged content does in fact violate platform policy. Platforms do address the somewhat heavy role that users are expected to play in content moderation, although it is unknown what user attitudes and seriousness are toward this task:

> *"Q: **It seems like you're asking users to do your job for you by reporting problems. Why should we?** A: We believe that when **all those involved in a community** - hosts and members, creators and contributors - **feel and take responsibility** for maintaining an appropriate and stimulating environment, the debate itself is improved and all those involved benefit. We work hard to make and keep the environment constructive and convivial but **we need your help to do so.**" (guardian.com's Frequently Asked Questions)*

The burden on users appears very pronounced for copyright infringement, with only 16.7% of all policy text relating to platforms' active role coming from copyright-related policy. This is likely due to the structure of reporting practices laid out in the DMCA, where platforms rely on user reports to take down violating content, and await counter-notices for possible reinstatement (see §6.5).

We also find evidence of heavy usage of human moderators, indicating that the promises of automated content moderation have perhaps not materialized [17] as demonstrated by this quote:

> *"24/7 Human Moderation Team: Our team of **moderators and support staff work 24 hours a day, 7 days a week to review all uploaded content for violations**, address user concerns, and remove all content that we identify or of which we are made aware of and deem as constituting hate speech." (pornhub.com's Hatespeech Policy)*

6.4 What happens to flagged content?

Once content has been flagged, platforms usually respond by targeting either the content itself, targeting the user who posted the content, or starting an investigation. Both USER- and CONTENT-TARGETED ENFORCEMENT are prevalent across all three content moderation topics. Having some form of both content each was also standard across platforms.

The following list distilled from twitter.com's enforcement options illustrates roughly all the possible enforcement actions we saw across platforms and topics:

> *"Below are some of the enforcement actions that we may take.*
> ***Tweet-level enforcement**: Limiting Tweet visibility, Excluding the Tweet from having ads adjacent to it, Requiring Tweet removal, Labeling a Tweet, Notice of public interest exception.*
> ***Account-level enforcement**: Suspend an account, Placing an account in read-only mode, Verifying account ownership" (twitter.com's Help Center page on range of enforcement options)*

In addition, we also saw platforms employ a 'strike policy', where first-time violators received strikes/warnings and repeat violators received escalated punishment like account termination. Strike policies were most commonly found with respect to copyright infringement, likely due to the DMCA legally requiring platforms to implement a repeat infringers policy in order to keep their copyright Safe Harbor status.

> *"Our '3 strikes' repeat infringement policy is implemented as follows: If one or more uploads occur after receipt of notice of a first infringement, uploaders are then given 2 chances to stop uploading videos or other content infringing any third party's copyrights. [...] **In the event that you accumulate three (3) such notices, your account will be terminated.**" (spankbang.com's 3 Strikes Policy Page)*

Still, platforms sometimes developed their own unique enforcement strategies. For instance, we found the following intriguing policy response by tumblr.com to disinformation campaigns by the Internet Research Agency (IRA) from Russia:

> *"What we're doing in response to the interference: First, we'll be emailing anyone who liked, reblogged, replied to, or followed an IRA-linked account with the list of usernames they engaged with. Second, **we're going to start keeping a public record of usernames we've linked to the IRA or other state-sponsored disinformation campaigns.**" (tumblr.com's Official Staff Blog)*

That is, tumblr.com pursued user-targeted enforcement by both contacting suspected users as well as public "naming-and-shaming".

For harmful speech and misinformation, we find a much greater prevalence of platforms intervening via investigation instead of immediate enforcement.

Finding 5. Of the policy text annotated as Platform Response: Investigation, only 14.9% is for copyright infringement, with 43.0% for harmful speech and 42.1% for misleading content.

We speculate this higher prevalence arises from two reasons. First, the DMCA requires platforms to take down content flagged as infringing copyright or face liability, so they typically only check if the claim itself meets legal requirements and don't necessarily review the potentially infringing content any further:

> *"GitHub Isn't The Judge.* ***GitHub exercises little discretion in this process other than determining whether the notices meet the minimum requirements of the DMCA.*** *It is up to the parties (and their lawyers) to evaluate the merit of their claims." (github.com's DMCA Guide)*

Second, what constitutes harmful speech and misleading content can sometimes be nebulous, leading platforms to hedge as encompassed by this Twitch quote:

> *"Twitch will consider a number of factors to determine the intent and context of any reported hateful conduct." (twitch.tv's Community Guidelines)*

6.5 Do users have any recourse after being moderated?

Specific instances of content moderation may be inconsistent with a platform's stated policies [44] and user expectations, leading users to seek rollback of moderation. However, we find that there is little users can concretely do unless they or their content was targeted on the basis of copyright infringement.

Finding 6. Only 9.4% and 15.3% of policy text labeled with Redress / Appeal falls under the topics of harmful speech and misleading content, respectively. Appeals regarding moderation related to copyright infringement dominate with 75.3%.

The appeals procedure for content removed due to copyright infringement is well-established, grounded in law, and clearly laid out for users to navigate on most platforms, with specialized copyright policy pages often in place as illustrated by a quote from Automattic (Wordpress's parent company):

> *"If you have received a Digital Millenium Copyright Act (DMCA) Infringement Notice and believe it was submitted in error,* ***you may submit a counter notice.*** *Counter notices must be submitted by the WordPress.com user who uploaded the material.* ***You may also use the form below to submit a DMCA counter notice:*** *" (automattic.com's Online DMCA Counter Notice Form)*

In fact, 8.1% of all policy text concerning copyright infringement deals with user appeals, but only 1.3% for harmful speech and 1.5% for misleading content does the same. This imbalance suggests that more work may need to be done to help users know what actions they can take if they feel their content has been unfairly moderated as harmful or misleading.

6.6 How comprehensive are policies across platforms?

We find that all 43 platforms contain at least some policy text from each of the 3 topics. While most platforms have policy related to enforcement and expect users to play an active role in moderation, only about half (21/43) have definitions of content moderation criteria. To demonstrate a lacking of policy components, we *define a platform to be complete in terms of policy composition* if it contains policy text relating to all sub-codes under Policy Justification, Moderation Criteria, Safeguards, Platform Response and Redress/Appeal (see Table 3). Binding Legalese and Signposts are often present in policies, but we do not deem them necessary for completeness.

Finding 7. 39.5%(17/43) of the platforms we consider are complete in terms of policy composition.

As expected, many of the major platforms such as facebook.com and youtube.com are complete, but we also find (perhaps surprisingly) inclusions of sites such as pornhub.com. When looking at policy completeness by platform type, we find that no specific category of platforms has consistent completeness. For example etsy.com was the only complete platform out of the four e-commerce sites; pornhub.com was the only complete platform out of the five adult content sites; and none of the four traditional news platforms were complete. A detailed table of code-wise distribution of policy text is in §B of the Supplementary Materials.

7 DISCUSSION AND FUTURE DIRECTIONS

In this section, we outline the implications of our paper for different stakeholders, as well as clear directions in which our dataset can be used and our research methodology expanded upon.

7.1 Implications for regulators

The difficulty in locating, gathering, and analyzing text pertaining to content moderation underscores a significant challenge currently facing regulators. Specifically, although some areas of content are not subject to moderation, certain categories (e.g., violent speech, terrorist speech, copyrighted content) are subject to regulation, particularly in certain jurisdictions. An important challenge regulators face is thus determining both whether a platform's stated policy complies with laws and regulations, as well as, ultimately, whether the platform's enforcement complies with these stated policies.

Standardizing policy formats and language: The process of gathering these policies across a large number of websites was painstaking. We ended up finding relevant policy text across a large number of pages for each platform, as observed in §5.2. These pages often had different names on different platforms, making it difficult to determine what the intent of the page was compared to a similar one on another platforms. This difficulty in even locating policy text, let alone interpreting it, point to a need for dialogue between platforms and regulators to agree upon standardization around content moderation policy. The current difficulties we face in even locating these policies—regardless of the type of content—underscores the difficulty of regulatory enforcement in this area, particularly at scale. Specifically, while it may be feasible for a regulator to investigate a specific site (perhaps in response to an incident), large-scale compliance checking would require a more systematic approach. Standard locations for content moderation policies, as well as standard (possibly even machine-readable) policies, could make it easier for regulators to check that stated policies comply with the laws and regulations of a particular region. Such standardization could also ultimately help users locate relevant content moderation policies and information on what avenues they have for recourse on a moderation decision, e.g., in the event of what they may deem as unfair moderation.

Protecting consumer rights to their own content: In addition, we find very strongly worded text in platforms' policies governing the right users waive when using platforms. For example, the following excerpt is typical of the platforms' ToS. It appears extremely invasive with respect to users' right to their own content and how it may be modified and used:

> *"Specifically, you provide us with a royalty-free, irrevocable, perpetual, worldwide, exclusive, and fully sublicensable license to use, reproduce, modify, adapt, publish, translate, create derivative works from, incorporate into other works, distribute, perform, display, and otherwise exploit such content, in whole or in part in any form, media or technology now known or later developed." (washingtonpost.com's Terms of Service)*

We find an alarming prevalence of text that excludes platforms from any liability for exposing users to objectionable content even as users' rights are signed away. In addition, given the absence of widespread redressal mechanisms—as found in §6.5—users that disagree with any platform provisions have limited options. For instance:

> *"**[Y]our only remedy with respect to any dissatisfaction** with (i) the Services, (ii) any term of these Terms of Use, (iii) any policy or practice of Fandom in operating the Services, or (iv) any content or information transmitted through the Services, **is to terminate your account and to discontinue use** of any and all parts of the Services." (fandom.com's Terms of Service)*

That is, often a user's only recourse upon disagreement with any platform provisions is to stop using the platform entirely, which can result in users migrating to other services.

7.2 Implications for platforms

As difficult as it is for regulators to enforce laws concerning content moderation, the platforms themselves might ultimately prefer to arrive at solutions that both protect consumers while at the same time limiting the possibility of facing potentially costly and cumbersome prescriptive content moderation strategies. Various areas, particularly concerning terrorist speech and child pornography, have seen healthy industry-wide collaboration, particularly around the sharing of datasets containing objectionable content.

Platform providers who read our paper may find the opportunity to examine their own policies, in light of our industry-wide analysis. Our critical analysis of existing policies and the lens we have provided into other platforms' policy language and practice concerning different content types provides content providers the opportunity to improve their own policy content and structure—and perhaps even come to a collective agreement about both the way these policies should (or could) be structured, as well as where these policies could be placed on their respective websites.

Furthermore, in response to finding that most platforms in our dataset are *incomplete* in terms of expressing crucial policy components (see §6.6), we advocate for increased transparency in moderation policies as it directly correlates with enhancing user experience. We argue that shedding light on the intricacies of moderation strategies, algorithmic processes, and decision-making criteria not only builds trust with users but also contributes to a more informed and empowered user base. That said, we recognize the inherent tensions surrounding the implementation of transparency. For instance, platforms may be reluctant to disclose moderation mechanisms for fear of revealing trade secrets or aiding future moderation evasion. This tension is apparent in the current popularity of Transparency Reports which often include only high-level statistics on content removals and user bans. This brings forth a complex discussion on the extent and nature of transparency that platforms should adopt, one ripe for HCI researchers and industry representatives to engage in.

Ultimately, such a collaborative consensus requires not only the data and analysis that we have offered in this area, but also a deeper understanding of consumer expectations and understanding of these policies, and how to interpret them. Such a line of inquiry presents many opportunities for future work in human-computer interaction research, as we discuss in the next section.

7.3 Implications for HCI researchers

We suggest several avenues for future research in user and audit studies, automation of ongoing data collection, and improvements to collection and annotation of policy text.

Further analysis and usage of OCMP-43: Although our current findings in §6 provide a comprehensive high-level overview of why, when, and how platforms moderate, more nuanced insights can be extracted from the dataset. For instance, future studies can further contrast moderation policies for different types of users for the same platforms. Platforms such as the nytimes.com have policy for journalists, advertisers, and regular users, all of whom use the platform in different ways, with some types of users being governed by legal regimes other than the United States. Further, the policy text in OCMP-43 provides a useful starting point for both user and audit studies in the future. User studies could investigate how users perceive the policies of different platforms, and as well as understand how they manage the occasionally onerous responsibilities placed on them to moderate content themselves. Audit studies can help address the tensions around platform transparency discussed above by looking for instances of policy text within OCMP-43 where platforms commit to certain actions to moderate content and verifying if those conditions are met. Evidence of discrepancies between a platform's stated policy and practice revealed through audits can strengthen demands for greater transparency. Our dataset greatly eases the process of finding these instances cross-platform and by topic.

Using our data collection pipeline for extending OCMP-43: Our data collection pipeline itself (§4) can be effectively used to collect additional policy text. The fact that our scraper is built specifically to extract policy text greatly reduces the technical burden for inter-disciplinary research. First, simply adding more platforms and their corresponding seed links will provide researchers access to even a larger set of online content moderation policy text. Second, a longitudinal study across platforms can also be performed by running our scraper at regular intervals and annotating the text that has changed from the previous run. This could provide an insightful study of the impact of political and cultural shifts on platform policies. For instance, there has been a large shift in the content moderation policies of X.com, formerly known as Twitter at the time we collected our data. Finally, by adjusting the topic-wise keyword list, researchers can find policy around other topics of interest, such as data collection for training machine learning models, or platform policies on AI generated content being posted by users.

Improving the collection and annotation of policy text: As discussed in §4.3, there are several limitations to our collection and annotation pipelines. In the future, we would like to add further automated checks to reduce the load on annotators having to go through irrelevant policy pages by better parsing of HTML or even training natural language models to identify pages that are unlikely to have relevant policy. Since annotation can be laborious, future studies could ascertain how well a natural language model can perform the annotation task.

8 CONCLUSION

The moderation of user-generated content has emerged as a paramount issue that must balance the needs of both user safety and, in some countries, the right to free expression. As the proliferation of third-party content on these platforms continues, content moderation has become increasingly complex due to the sheer scale of data involved, as well as the variety of types of content that can be posted (and, ultimately, must be moderated). The absence of a prescriptive approach for many of these types of content has led to challenges in consistency of content moderation policies across different types of content, and across platforms. Lack of consistency is apparent in both the language of the policies, and structure of the articulated policies. This paper seeks to address these concerns with the first systematic study of content moderation policies across 43 major online platforms, examining their variation, and providing valuable insights for future research and policy alignment. Quantitative and qualitative analysis of the annotated dataset of policies created in this paper reveals immense variation across both topic and platform. Our data acquisition pipeline, dataset itself, and the corresponding analysis points to many future avenues for researchers and policymakers to continue to explore, especially in terms of how these polices might be better structured and articulated. We hope this paper acts as a catalyst for future research aimed at improving content moderation policies to enable healthier online communities.

ACKNOWLEDGMENTS

This work was funded by the Data & Democracy research initiative, a collaboration between the Data Science Institute and Center for Effective Government at the University of Chicago.

REFERENCES

[1] Access Now, ACLU Foundation of Northern California. 2018. The Santa Clara Principles on Transparency and Accountability in Content Moderation. https://santaclaraprinciples.org/.

[2] Sabrina Ahmad. 2019. *"It's Just the Job": Investigating the Influence of Culture in India's Commercial Content Moderation Industry.* preprint. SocArXiv. https://doi.org/10.31235/osf.io/hjcv2

[3] Amnesty. 2022. Myanmar: Facebook's systems promoted violence against Rohingya; Meta owes reparations. https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/.

[4] Ivan Beschastnikh, Travis Kriplean, and David McDonald. 2008. Wikipedian self-governance in action: Motivating the policy lens. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 2. 27–35. https://doi.org/10.1609/icwsm.v2i1.18611

[5] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. 2008. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. In *Proceedings of the SIGCHI conference on human factors in computing systems*. https://doi.org/10.1145/1357054.1357227

[6] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. #Thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of CSCW* (San Francisco, California, USA) *(CSCW '16)*. ACM, New York, NY, USA, 1201–1213.

[7] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–30.

[8] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! Examining the effects of a community-wide moderation intervention on Reddit. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29, 4 (2022), 1–26.

[9] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *In Proceedings of ACM Human Computer Interaction* 1, Computer-Supported Cooperative Work and Social Computing, Article 31 (2017), 22 pages.

[10] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (Dec. 2017), 1–22. https://doi.org/10.1145/3134666

[11] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–25. https://doi.org/10.1145/3274301

[12] Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trajectories of Blocked Community Members: Redemption, Recidivism and Departure. In *Proceedings of WWW*. 184–195.

[13] Adrian Chen. 2023. The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed. https://www.wired.com/2014/10/content-moderation/

[14] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2014. How community feedback shapes user behavior. In *Proceedings of ICWSM*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8066

[15] US Congress. 1998. Digital Millennium Copyright Act. *Public Law* 105, 304 (1998), 112.

[16] G. Copeland, T. Keller, R. Krishnamurthy, and M. Smith. 1989. The Case For Safe RAM. In *Proceedings of the 15th International Conference on Very Large Data Bases* (Amsterdam, The Netherlands) *(VLDB '89)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 327–335. http://portal.acm.org/citation.cfm?id=88830.88887

[17] Vittoria Elliott. 2023. Meta's Gruesome Content Broke Him. Now He Wants It to Pay. https://www.wired.com/story/meta-kenya-lawsuit-outsourcing-content-moderation/

[18] Casey Fiesler, Michaelanne Dye, Jessica L. Feuston, Chaya Hiruncharoenvate, C.J. Hutto, Shannon Morrison, Parisa Khanipour Roshan, Umashanthi Pavalanathan,

Amy S. Bruckman, Munmun De Choudhury, and Eric Gilbert. 2017. What (or Who) Is Public?: Privacy Settings and Social Media Content Sharing. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Portland Oregon USA, 567–580. https://doi.org/10.1145/2998181.2998223
[19] Casey Fiesler, Jessica L. Feuston, and Amy S. Bruckman. 2015. Understanding Copyright Law in Online Creative Communities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, Vancouver BC Canada, 116–129. https://doi.org/10.1145/2675133.2675234
[20] 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.
[21] Casey Fiesler, Cliff Lampe, and Amy S. Bruckman. 2016. Reality and Perception of Copyright Terms of Service for Online Content Creation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, San Francisco California USA, 1450–1461. https://doi.org/10.1145/2818048.2819931
[22] Casey Fiesler, Joshua McCann, Kyle Frye, and Jed R Brubaker. [n. d.]. Reddit Rules! Characterizing an Ecosystem of Governance. ([n. d.]), 10.
[23] Tarleton Gillespie. 2018. *Custodians of the Internet.* Yale University Press.
[24] Eric Goldman. 2021. Content Moderation Remedies. *Michigan Technology Law Review, Forthcoming* (2021).
[25] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020), 2053951719897945.
[26] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. 2021. Do Platform Migrations Compromise Content Moderation? Evidence from r/The_Donald and r/Incels. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–24. https://doi.org/10.1145/3476057
[27] Sohyeon Hwang and Aaron Shaw. 2022. Rules and Rule-Making in the Five Largest Wikipedias. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 347–357. https://doi.org/10.1609/icwsm.v16i1.19297
[28] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did You Suspect the Post Would be Removed?": Understanding User Reactions to Content Removals on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–33. https://doi.org/10.1145/3359294
[29] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction* 26, 5 (Sept. 2019), 1–35. https://doi.org/10.1145/3338243
[30] Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. 2021. Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–30. https://doi.org/10.1145/3479525
[31] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter? User behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–27.
[32] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Transactions on Computer-Human Interaction* 25, 2 (April 2018), 1–33. https://doi.org/10.1145/3185593
[33] Brian Keegan and Casey Fiesler. 2017. The Evolution and Consequences of Peer Producing Wikipedia's Rules. In *Proceedings of ICWSM*. https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15698
[34] Daphne Keller, Paddy Leerssen, et al. 2020. Facts and where to find them: Empirical research on internet platforms and content moderation. *Social media and democracy: The state of the field and prospects for reform* 220 (2020), 224.
[35] Sara Kiesler, Robert E. Kraut, Paul Resnick, Aniket Kittur, Moira Burke, Yan Chen, Niki Kittur, Joseph Konstan, Yuqing Ren, and John Riedl. 2011. *Regulating Behavior in Online Communities.* The MIT Press, 125–178. http://www.jstor.org/stable/j.ctt5hhgvw.7
[36] Kate Klonick. 2017. The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev.* 131 (2017), 1598.
[37] Knight, Ben. 2018. Germany implements new internet hate speech crackdown. https://www.dw.com/en/germany-implements-new-internet-hate-speech-crackdown/a-41991590.
[38] langdetect 2023. https://github.com/Mimino666/langdetect
[39] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS 2019)*. https://doi.org/10.14722/ndss.2019.23386
[40] Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature human behaviour* 5, 3 (2021), 337–348.
[41] Mackintosh, Eliza. 2021. Facebook knew it was being used to incite violence in Ethiopia. It did little to stop the spread, documents show. https://www.cnn.com/2021/10/25/business/ethiopia-violence-facebook-papers-cmd-intl/index.html.
[42] J. Nathan Matias. 2016. Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 1138–1151. https://doi.org/10.1145/2858036.2858391
[43] 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.
[44] J Nathan Matias, Austin Hounsel, and Nick Feamster. 2022. Software-supported audits of decision-making systems: Testing google and facebook's political advertising policies. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–19.
[45] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (Nov. 2018), 4366–4383. https://doi.org/10.1177/1461444818773059
[46] Marie Ozanne, Aparajita Bhandari, Natalya N Bazarova, and Dominic DiFranzo. 2022. Shall AI moderators be made visible? Perception of accountability and trust in moderation systems on social media platforms. *Big Data & Society* 9, 2 (2022), 20539517221115666.
[47] Sarah T Roberts. 2014. *Behind the screen: The hidden digital labor of commercial content moderation.* University of Illinois at Urbana-Champaign.
[48] Sarah T Roberts. 2016. Commercial content moderation: Digital laborers' dirty work. (2016).
[49] Sabin, Sam. 2019. On Policing Content, Social Media Companies Face a Trust Gap With Users. https://pro.morningconsult.com/articles/on-policing-content-social-media-companies-face-a-trust-gap-with-users.
[50] Johnny Saldaña. 2021. *The coding manual for qualitative researchers.* sage.
[51] Hans-Bernd Schäfer. 2002. Legal rules and standards. In *The Encyclopedia of Public Choice.* Springer, 671–674.
[52] Scrapy | A Fast and Powerful Scraping and Web Crawling Framework 2023. https://scrapy.org/
[53] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping pro and anti-social behavior on Twitch through moderation and example-setting. In *Proceedings of CSCW.* ACM, New York, NY, USA.
[54] Mohit Singhal, Chen Ling, Nihal Kumarswamy, Gianluca Stringhini, and Shirin Nilizadeh. 2023. SoK: Content moderation in social media, from guidelines to enforcement, and research to practice. In *8th IEEE European Symposium on Security and Privacy (EuroSP 2023).*
[55] VERBI Software. 2021. MAXQDA. https://www.maxqda.com/
[56] Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content removal as a moderation strategy: Compliance and other outcomes in the ChangeMyView community. In *Proceedings of CSCW.*
[57] Nicolas P Suzor. 2019. *Lawless: The secret rules that govern our digital lives.* Cambridge University Press.
[58] Lorenz Taylor. 2023. How Twitter lost its place as the global town square. https://www.washingtonpost.com/technology/2023/07/07/twitter-dead-musk-tiktok-public-square/
[59] undetected-chromedriver 2023. https://github.com/ultrafunkamsterdam/undetected-chromedriver
[60] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "At the End of the Day Facebook Does What ItWants": How Users Experience Contesting Algorithmic Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–22. https://doi.org/10.1145/3415238
[61] Williams, Jamie. 2018. D.C. Court: Accessing Public Information is Not a Computer Crime. https://www.eff.org/deeplinks/2018/04/dc-court-accessing-public-information-not-computer-crime.
[62] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, et al. 2016. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* 1330–1340.
[63] Savvas Zannettou. 2021. "I Won the Election!": An Empirical Analysis of Soft Moderation Interventions on Twitter. *arXiv:2101.07183 [cs]* (April 2021). http://arxiv.org/abs/2101.07183 arXiv: 2101.07183.