Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

Written for Lawyers or Users? Mapping the Complexity of Community Guidelines

Mia Nahrgang,[1] Nils B. Weidmann,[1] Friederike Quint,[2] Sebastian Nagel,[1] Yannis Theocharis,[2] Margaret E. Roberts[3]

[1]University of Konstanz
[2]Technical University of Munich
[3]University of California, San Diego
mia.nahrgang@uni-konstanz.de, nils.weidmann@uni-konstanz.de, friederike.quint@tum.de, sebastian.nagel@uni-konstanz.de, yannis.theocharis@tum.de, meroberts@ucsd.edu

Abstract

Recent regulatory efforts such as the EU's Digital Services Act aim to increase transparency in mostly opaque content moderation practices of social media platforms. They encourage platforms to post information about what content is prohibited on the platform. But what kinds of platforms follow these best practices, and how readable is the information posted for average users? This paper introduces the *Content Moderation Policies and Reports Dataset* (COMPARE), a new data collection comprising content moderation policies of 132 of the most popular social media platforms for comparative analysis. We then use COMPARE to measure the complexity of 89 community guidelines, focusing on length, readability, and semantic complexity. We find that while the largest platforms are most likely to make guidelines available, they also tend to have the longest and semantically most complex guidelines. In addition, community guidelines seem to grow in length with new regulation. Our results suggest that it is crucial to look beyond the mere provision of community guidelines and to study how complex guidelines are for users.

Introduction

Social media was once considered a vehicle for enhancing the democratic process by promising to empower users to participate in online discussions while bypassing traditional media gatekeepers. Yet, over time, their democratic potential has been overshadowed by the emergence of more troubling aspects (Tucker et al. 2017). Content such as hate speech and misinformation can pose challenges to democratic governance by decreasing trust in politicians, fostering polarization, and producing misinformed voters (Jerit and Zhao 2020; Tucker et al. 2018; Lorenz-Spreen et al. 2023). While the evidence on their effects is mixed, social media has also been weaponized by actors with malicious intentions through foreign influence campaigns (Eady et al. 2023; Pierri et al. 2023). To counter this and other potentially harmful content, platforms often aggressively moderate content on their platforms. By becoming "the arbitrators for what is and what is not allowed in the public sphere" (Maddox and Malson 2020, p. 3), platforms hold tremendous power. However, public trust in those platforms is low, with surveys revealing that three-quarters of Americans do not trust social media companies to make fair content moderation decisions (Kamp and Ekins 2021).

To maintain platform discretion over the moderation of content while also addressing the problem of public trust, scholars, public officials, and civil society activists have advocated for more transparency in content moderation. As part of this, many have argued that social media companies should provide community guidelines to users that outline the applicable rules of engagement on a platform (MacCarthy 2020; Suzor et al. 2019; European Union 2022; Access Now et al. 2018; Electronic Frontier Foundation et al. 2015). A community guidelines document is published and updated by the platform; it "lays out the platform's expectations of what is appropriate and what is not. It also announces the platform's principles, and lists prohibitions, with varying degrees of explanation and justification" (Gillespie 2019, p. 46). Proposed standards for these guidelines, such as those put forward by civil society organizations in the Santa Clara Principles, urge platforms to "publish clear and precise rules and policies relating to when action will be taken with respect to users' content or accounts" (Access Now et al. 2018). While the Digital Services Act (DSA) does not mention community guidelines directly, Article 14 requires platforms to provide "information on any restrictions that they impose in relation to the use of their service" in their terms and conditions. In addition, it mandates that the terms and conditions "shall be set out in clear, plain, intelligible, user-friendly and unambiguous language" (European Union 2022).

While posting community guidelines has become a common practice for most social media platforms, very little work has studied how accessible they are to users and their differential use across platforms. Achieving transparency through community guidelines requires that users can clearly understand what behavior is admissible and what is prohibited. However, there are vast differences in the way that community guidelines are presented across platforms. Our comparative analysis aims to study this variation empirically. Focusing on 132 social media platforms, we ask:

- **RQ1:** What are the characteristics of social media platforms that publish community guidelines compared to those that do not?
- **RQ2:** What is the complexity of the guidelines in terms of length, readability, and semantic complexity?

- **RQ3:** How does regulation impact the complexity and accessibility of community guidelines?

Focusing on the length, readability, and semantic complexity, our study is the first large-scale analysis of community guidelines. We find that very large companies are the most likely to provide community guidelines. However, these guidelines tend to be much longer, more semantically complex, and less readable than others. Moreover, our results suggest that regulation may drive the extension of guidelines' length, perhaps to the detriment of the usability. Our results imply that civil society organizations and policy officers should look beyond the mere availability and more into the complexity of community guidelines from a user perspective.

Related Work

Transparency in content moderation has been analyzed from various angles. Much conceptual work has focused on the benefits. For example, MacCarthy (2020) argues that transparency in content moderation is beneficial for both users and the public discourse at large because it facilitates holding platforms accountable. Empirical research by Suzor et al. (2019) surveys users showing that users are uncertain about the process of content moderation, including why and how their content was moderated. Juneja, Rama Subramanian, and Mitra (2020) analyze whether the content moderation practices of community moderators on Reddit are in line with transparency standards using moderation logs, and find that the rules are not sufficiently clear. Similarly, Urman and Makhortykh (2023) comparatively analyze transparency reports of ten major technical companies (e.g., YouTube, Apple, Google, Meta) using the recommended statistics that should be published by platforms according to the Santa Clara Principles as transparency reference categories. They find that in a pre-DSA era, none of the analyzed platforms fully comply with the principles. Moreover, Trujillo, Fagni, and Cresci (2025) examine self-reported moderation decisions of eight social media platforms through the first 100 days of the DSA transparency database. They also compare the platform-provided data with the transparency reports platforms are obliged to publish bi-annually under the DSA, revealing inconsistencies between the two reporting channels.

Other scholars have demonstrated the effects transparency can have on user education and online discourses. Jhaver et al. (2019) test whether being transparent about content moderation decisions has an effect on subsequent posting behavior of the user and find that providing moderation reasons reduces the odds of future post removals on Reddit, thus indicating that users post less problematic content. However, the same effect could not be confirmed for bystanders witnessing the moderation (Jhaver, Rathi, and Saha 2024).

Keeping the benefits in mind, scholars have suggested measures to increase transparency in content moderation. For example, MacCarthy (2020) recommends that platforms should inform users about the moderation of their content, including the reason, and put complaint and appeal systems in place. He suggests that platforms should publish aggregate statistics on how much content was moderated and provide insights into algorithms used for content curation and moderation. Finally, MacCarthy stresses that platforms should explain enforcement techniques and outline the content rules.

Another strand of the literature focuses on the actual documents outlining the rules, the community guidelines. In contrast to terms of services, community guidelines are what "users are more likely to read if they have a question about the proper use of the site, or find themselves facing content or users that offend them" (Gillespie 2019, p. 46). They are thus the main document establishing acceptable behavior and potential sanctions and hold important discursive power in disputes over content moderation. Kopf (2024) argues that vague guidelines can lead to uncertainty among users which could result in self-censorship and eventually to a decrease in content variety. Instead, platforms have incentives to formulate vague rules in order to avoid their interpretation being challenged.

Katzenbach et al. (2023b) show historical trends from 2004 to 2021, analyzing the number of characters and the frequency at which they changed, focusing on platform policies including community guidelines, privacy policies, and terms of services for Facebook, Instagram, YouTube, and X. They found that Facebook and X started to frequently update their community guidelines in the second half of 2017 with Facebook publishing nineteen updates in the second half of 2019.

Moreover, some scholars have directed their attention towards the content of community guidelines. For example, Jiang et al. (2020) identify 66 rules prohibiting content in community guidelines and code these across 11 social media platforms. They report substantial differences across platforms in prohibited content. Similarly, Singhal et al. (2023) provide a taxonomy of content moderation categories. Furthermore, they analyze the *comprehensibility* of community guidelines of 14 mainstream and fringe platforms. In this study, comprehensibility is conceptualized as the granularity of prohibited content categories, whether examples are provided, and whether these are complemented by images or video material.

Additionally, research has discussed reasons for community guidelines to expand. Barrett and Kreiss (2019, p. 2) argue that platform policies respond to "normative pressure from external stakeholders." Moreover, Gillespie (2019) describes how additions to community guidelines are similar to traffic lights in that they represent a precedent, where a previously unregulated situation has caused harm in some way and results in a new policy.

Research Gaps

As of yet, research analyzing content moderation policies has focused mostly on the major social media platforms in the U.S. and Europe. However, social media use is increasingly distributed across many different platforms, including many large social media platforms based outside the U.S. and Europe (Gorwa 2024). Therefore, there is a

need for a comparative analysis that goes beyond the top ten biggest platforms and Western centrism.

While publishing community guidelines is often seen as a benchmark for transparency in content moderation, there is little work that has studied the practical usability of the guidelines for users. While other scholars have looked at the readability of, for example, terms of services or privacy policies (Gyasi and Bangmarigu 2020; Derguech, Zainab, and D'Aquin 2018; Jensen and Potts 2004; McDonald et al. 2009), to our knowledge no research has looked at the complexity of community guidelines. In addition, very little work has investigated how platforms change their guidelines in reaction to regulation, and how this affects how accessible guidelines are to users.

COMPARE: The Content Moderation Policies and Reports Dataset

To address these gaps, we present a unique dataset called the Content Moderation Policies and Reports Dataset (COMPARE). COMPARE is a comprehensive data collection on social media platforms and their characteristics, which includes links to content moderation policies. To identify larger platforms, we started with an initial list of major social media platforms by combining the most popular social media platforms from ten global and regional (U.S., China, Germany) rankings (see Appendix A1 for more details). From this candidate list, we then selected those platforms for inclusion in the dataset if they fit our definition of "social media platforms", namely if they (i) host user-generated content that is (ii) at least to some extent public-facing and in principle visible to anyone upon registration and (iii) that stays consistently accessible over a longer period.

The COMPARE dataset consists of two parts: first, a set of metadata about each platform, and second, a list of references to websites where platforms detail their content moderation policies. In the following, we introduce each of these two parts.

Platform Information

For each platform, COMPARE contains information about the country, the size, the age, and the type of the platform as well as whether it is decentralized or an alt-tech platform.

Country We coded the platform country based on the country of the platform's headquarters. In order to retrieve information about the location of the headquarters, we consulted the platforms' self-descriptions on either the *About* sections on their website or their LinkedIn accounts. Additionally, we consulted websites like Crunchbase or Wikipedia.

The platforms in our dataset are headquartered in 22 different countries. Most platforms are based in the U.S. (n = 58), followed by China (n = 27), Japan (n = 7) and Germany (n = 7). Other countries include Russia, Taiwan, and the UK (n = 4), Canada and Iran (n = 3), and France and Poland (n = 2). Argentina, Brazil, India, Korea, Latvia, Luxembourg, New Zealand, Singapore, South Korea, Turkey, and the United Arab Emirates each host one platform in the dataset.

Platform Size To determine the platforms' size, we relied on the marketing portal Similarweb's monthly visits estimates (Similarweb 2024) as of February 2024. Platform size varies significantly, ranging from the smallest platforms in our sample with less than five thousand monthly visits to the biggest platforms including YouTube (33 billion monthly visits), Facebook (16 billion), and Instagram (seven billion). The mean across all platforms is 624.6 million monthly visits.[1]

Platform Age COMPARE also includes the year a platform was launched. The oldest platforms were launched in 1995 (Ptt and NewGrounds) and the newest in 2023 (Threads and Bluesky). The search strategy for the platform age was similar to the strategy employed for determining the country.

Platform Type We employ a platform typology that is inspired by Rajendra-Nicolucci and Zuckerman (2021). However, we chose to simplify their categories into four distinct types to minimize redundancies, eliminate overlaps, and ensure each type had a sufficient number of platforms. For example, we subsumed *forum* and *Q&A* platforms into one category. Moreover, we recoded platforms with a *local* focus, platforms used by a *subculture* or being considered as *alt-tech* platforms, since these are platform characteristics rather than constituent elements of different platform types. As a result, we use four platform types in COMPARE:

- *Chat* platforms revolve around private either one-on-one or small-group communication.
- *Creator* platforms "enable users to share a specific type of media (like video, live streams, blogs, or art), in a one-to-many fashion. They are home to 'creators,' people who consistently make content for the platform, often as a source of income, and to audiences who turn to these platforms for entertainment, information, and a sense of identity and community in fandom" (Rajendra-Nicolucci and Zuckerman 2021, p. 63).
- *Forum* platforms are focused on topics of common interest rather than on preexisting social relationships. Forums can be text- or image-based. Forums can follow a question-answer type, and users usually have an anonymous username.
- *Social network* platforms are general-purpose platforms that focus on connecting people who either already know each other or are looking for new connections (e.g., professional networking, dating).

Decentralized and Alt-Tech Platforms Additionally, COMPARE includes two different flags that identify (i) decentralized platforms and (ii) platforms considered as alt-tech platforms. Decentralized platforms such as Mastodon or Bluesky inherit their name from their decentralized technical set-up. Other than mainstream social-media platforms,

[1]To facilitate presentation later in this paper, we create five size categories as follows. *Very large:* platforms with 800,000,000 or more monthly visits; *large:* 100,000,000 to 800,000,000 visits; *medium*-sized: 1,000,000 to 100,000,000 visits; *small:* 100,000 to 1,000,000 visits; *very small:* less than 100,000 visits per month.

the ownership of their servers is distributed, thus circumventing central governance for example in regard to content moderation (Rozenshtein 2022). Alt-tech (*Alternative Technology*) platforms such as Gab, frequently used by but not exclusive to the far-right, provide an alternative to Silicon Valley-controlled mainstream platforms and are usually characterized by minimal content moderation and a strong emphasis on promoting protecting free speech rights (Freelon, Marwick, and Kreiss 2020; Siapera 2023; Balci, Sirivianos, and Blackburn 2024). Both of these characteristics were determined based on the literature (Rajendra-Nicolucci and Zuckerman 2021; Freelon, Marwick, and Kreiss 2020; Shaughnessy et al. 2024; Rozenshtein 2022) complemented by information from Wikipedia. The dataset includes 120 centralized and 12 decentralized platforms. Moreover, out of the 132 platforms, 17 were coded as alt-tech platforms.

Links to Content Moderation Policies For each platform, COMPARE includes links to six different types of platform policies: (1) the terms of services, (2) the privacy policies, (3) the community guidelines, (4) the transparency reports, and information about (5) a platform's content moderation enforcement options and (6) how the moderation process is structured (for definitions see Appendix A2). Those aspects were chosen because they cover the most relevant characteristics of a platform's content moderation strategy. The links were collected from October 2023 to January 2024.

Community Guidelines

For our analysis below, we focus on community guidelines, also sometimes referred to as community standards or codes of conduct. Unlike terms of services which are essentially legal documents, community guidelines are supposed to be written in an accessible manner because they are intended to be read by users (Singhal et al. 2023). Of the 132 platforms in the COMPARE dataset, 92 platforms have published such guidelines (see Appendix A3 for a platform overview).[2]

We scraped 76 community guidelines in August 2024 from Germany, using the default language settings. To do this, we visited the main community guidelines page and scraped information from links on the page if these links provided specific information about the rules and allowed behavior on the platform. We did not follow links to other policy documents, such as the terms of service and copyright policies. This results in 54 English and 22 non-English guidelines (see Figure 1).[3] To make the complexity measures (see below) applicable and comparable, we then translated the non-English texts into English using primarily the DeepL API (DeepL 2025). Because of language availability, we used the Google Cloud Translation API (Google 2025) for the two Persian texts.

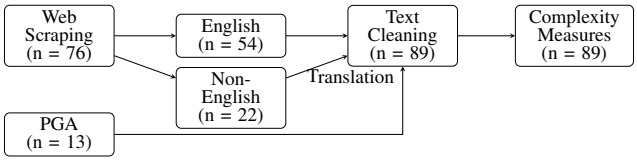

Figure 1: Text processing steps

The remaining 13 community guidelines were obtained from the Platform Governance Archive's (PGA) GitHub Repository with the versions from August 13, 2024. The PGA also makes historical versions available, capturing how the community guidelines and other platform policies have evolved since August 2022 (Katzenbach et al. 2023a). Moreover, for a small subset, the PGA traces the evolution of community guidelines since 2007 (Katzenbach et al. 2023c), which we used for our additional analysis of the evolution of community guidelines below. Finally, we applied text-cleaning steps to all the 89 community guidelines. More specifically, we replaced new line characters and semicolons with dots, removed special characters, and harmonized the numbering (for more information see Appendix A4).

Measuring the Complexity of Community Guidelines

We set out to investigate three different complexity dimensions of community guidelines: A first, simple measure is the overall *length*. While longer guidelines contain more information, they are also likely to be more time-consuming for users to navigate, since they need to process more information. Second, we consider the *readability* of the community guidelines, which takes into account the length and structure of sentences. Third, we focus on *semantic complexity*, captured by the number of categories that platforms explicitly mention in the guidelines as banned on their services. We now introduce each of these measures.

Length The length of community guidelines is measured as the number of tokens in the complete text of the guidelines (even if distributed across multiple web pages).

Readability Text readability refers to the comprehension difficulty of a given text. For determining the readability of community guidelines, we use an established measurement, the Flesch-Kincaid Grade Score (FKGS, Flesch 1948; Kincaid et al. 1975). This measurement relies on the average sentence length and the average word length in a given text, thus focusing on syntactic readability while disregarding content, semantic, or lexical aspects. To facilitate interpretation, the scale indicates the years of education (in the US school system) one needs to understand a given text.

[2]Our analysis only includes 89 community guidelines, because Instagram and Threads rely on Facebook's guidelines and were thus not included again. Moreover, the platform Caffeine went offline before we finalized the scraping.

[3]Of the latter, only 5 platforms published platform-provided translations to English. The translated versions provided by WeChat and Kakao are exact translations. Bilibili, Hatenablog, and Xiaohongshu provide summarized English translations, which are unlikely to be consulted by native speakers. To avoid introducing more noise, we decided to use machine translations for all the non-English community guidelines, regardless of whether the platform provides a translation, a summary, or no English version at all.

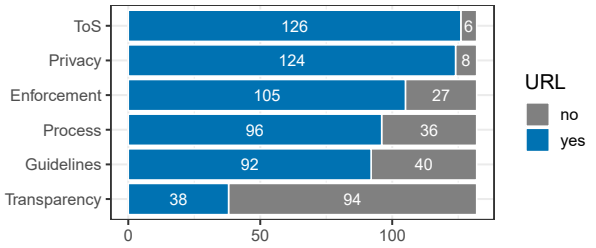

Figure 2: Distribution of content moderation variables in COMPARE in absolute numbers.

Therefore, higher scores indicate that texts are more complex when it comes to their readability.

Syntactic readability measurements are commonly used in medical research analyzing health literacy, i.e. patients' ability to understand the provided information and to make informed health decisions based upon this (Rooney et al. 2021; Sare et al. 2020). In political science, readability scores have been used in the context of speeches by members of parliament (Benoit, Munger, and Spirling 2019; Schoonvelde et al. 2019; Lin and Osnabrügge 2018), executive press releases (Rauh 2023) or election campaign messages (Bischof and Senninger 2018) and have recently been validated for an adult study population (Benoit, Munger, and Spirling 2019). FKGS is defined as

$$\text{FKGS} = 0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59$$

To make sure that our results do not depend on the FKGS measure, we also conduct a robustness test using Gunning's Fog Index (Gunning 1952), another syntactic readability measurement. The results are presented in Appendix A8.

Semantic Complexity Finally, we determine the semantic complexity of community guidelines, which captures the number of content categories platforms decide to explicitly prohibit in their community guidelines through human coding. In this process, we build on Jiang et al. (2020)'s compilation. We refined this initial list of categories through an iterative process, reviewing several guidelines and adding new categories until no further additions were needed. The final list consists of 80 categories in 16 groups (see Appendix A5). Importantly, if a behavior or content is coded as not prohibited, this only means that it is not specifically banned in the community guidelines, but does not indicate that the platform tolerates it since the content category could be prohibited somewhere else like in the terms of services or could be banned according to law. For example, child sexual exploitation materials are likely not allowed on any platform, however, not every platform explicitly mentions in their guidelines that this content is banned.

Two student research assistants were assigned to read the guidelines and to make an independent coding decision about whether a specific content or behavior is explicitly

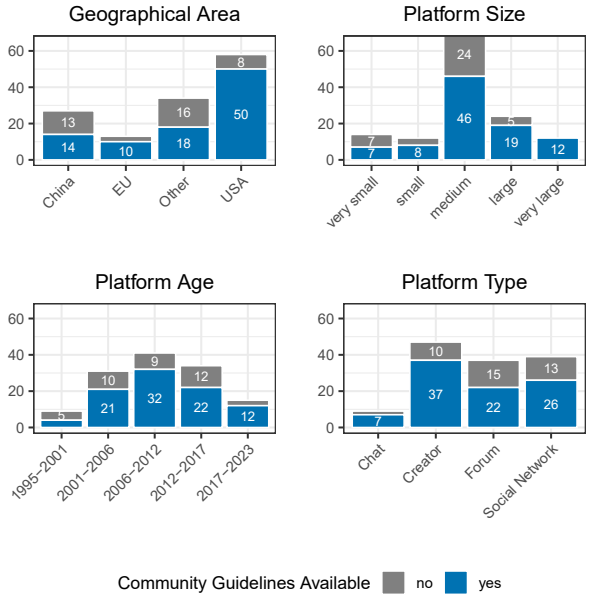

Figure 3: Publication of community guidelines by geographical area (top left), platform size (top right), age of platform (bottom left), and platform type (bottom right) in absolute numbers.

prohibited (coded as 1) or not (coded as 0) according to a given platform's community guidelines. Out of 7,120 coding decisions, coders agreed in 5,929 cases. Cohen's Kappa (Cohen 1960) between the two coders was 0.61, indicating substantial inter-coder agreement. In the remaining 1,191 cases where the coders disagreed, the lead author functioned as the tie-breaker.

Which Platforms Provide Community Guidelines?

Figure 2 shows how many platforms provide information on content moderation. Among the 132 platforms, just over two-thirds (n = 92) provide community guidelines. This is significantly fewer than the number that provide terms of services and privacy policies, or information on enforcement. However, community guidelines are more common than transparency reports, which only a quarter of platforms published, at least in the pre-DSA era.

While community guidelines are prevalent across all regions, platforms based in the U.S. and the EU are more likely to make community guidelines available – 86.2% and 76.9% respectively. In contrast, only slightly more than half of the platforms based in China (51.9%), and the remaining countries (52.9%) publish community guidelines (see Figure 3, top left panel). Platform size also matters, with larger platforms being more likely to provide community guidelines. While among the very large platforms, 100% provide community guidelines, this is only true for 50% of the very small platforms (see Figure 3, top right panel). In addition, there does not seem to be a clear trend in terms of platform

age. Platforms launched between 2017 and 2023 were most likely to provide community guidelines (80.0%), followed by platforms launched between 2006 and 2012 (78.0%; see Figure 3, bottom left panel). Looking at platform types (Figure 3, bottom right panel), creator platforms are most likely to provide community guidelines (78.7%) followed by chat platforms (77.8%), social networks (66.7%) and forums being the least likely (59.5%). Notably, both alt-tech platforms and decentralized platforms are less likely to provide community guidelines. However, the difference between mainstream (70.4%) and alt-tech platforms (64.7%) is less pronounced than the difference between centralized (70.8%) and decentralized platforms (58.3%). In Appendix A6, we present multivariate logistic regression results to explore which platform characteristics are the primary predictors of whether platforms publish community guidelines.

How Complex Are Community Guidelines?

As introduced above, we analyze the complexity of community guidelines along three dimensions: length, readability, and semantic complexity.

Length

Among platforms that do provide community guidelines, the length of the community guidelines varies considerably. While the community guidelines of Patriots.win (265 tokens), Mastodon (312 tokens),[4] Plurk (336 tokens), and Slug (408 tokens) would all fit on a single standard formatted Microsoft Word page, the longest guidelines can reach epic length. YouTube has the most extensive guidelines (28,954 tokens), followed by Facebook (23,456 tokens) and X (22,550 tokens). Assuming that the average adult reading speed for English non-fiction is approximately 238 words per minute (Brysbaert 2019), reading YouTube's guidelines in their entirety would require two full hours. The longest non-Western community guidelines are from Douyin (16,381 tokens), TikTok's mainland China equivalent. However, most guidelines are much shorter with an overall mean of 4,167 tokens and a median of 2,126 tokens.

Looking into differences in the length of community guidelines across platforms characteristics (see Figure 4), only platform size produces significantly different means, with the very large platforms also publishing by far the longest community guidelines (very large group is significantly different from other size groups with $p < 0.001$; see also Appendix A6 for OLS regression results). Possible explanations for this are that bigger platforms can invest more resources into content moderation and its specification, but also that these platforms are under heightened scrutiny due to their enhanced reach and social relevance.

Readability

In terms of readability, community guidelines range from requiring six years of education to needing 17 years of education in order to be able to understand the text. The

[4]These are the community guidelines of the original Mastodon operated instance Mastodon.Social.

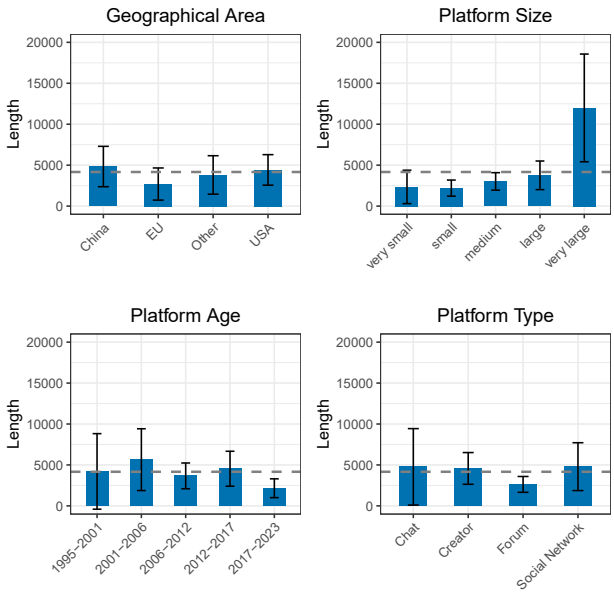

Figure 4: Average length in number of tokens by geographical area (top left), platform size (top right), age of platform (bottom left), and platform type (bottom right). The dashed line displays the mean across all guidelines.

readability score suggests that users would need on average 11.6 years of school education to understand the guidelines, which corresponds to high school education. The median is 11.9. While Foursquare (FKGS = 6.3), Xing (FKGS = 6.5), and 4Chan (FKGS = 7.0) communicate their rules in the most accessible way, the community guidelines of SlideShare (FKGS = 17.5), Douyin (FKGS = 16.3), and Kuaishou (FKGS = 15.4) were the most difficult to understand. In Appendix A7, we discuss how we ensure external validity of the readability scores using data from another study with participants in a laboratory environment.

Looking into group-based differences (see Figure 5), community guidelines of platforms based in the EU are significantly more readable than of those based in China ($p < 0.01$). Moreover, the community guidelines of forum platforms are on average significantly more readable than those of chat or creator platforms ($p < 0.05$, see Appendix A6 for the OLS regression results). In Appendix A8, we replace the Flesch-Kincaid Grade Score with the Gunning's Fog Index (Gunning 1952) and find that our results are robust to this change. In Appendix A9, we also explore whether translation is an issue and show that (i) translated and non-translated community guidelines follow a similar trend; (ii) translation does not substantially reduce the readability; (iii) changing the translation software does not alter the results; and that (iv) translating all guidelines to Chinese and applying a Chinese Readability Index (Xu, Yao, and Chen 2019) produces similar results.

Readability Compared to Other Types of Text Benchmarking the readability of community guidelines against

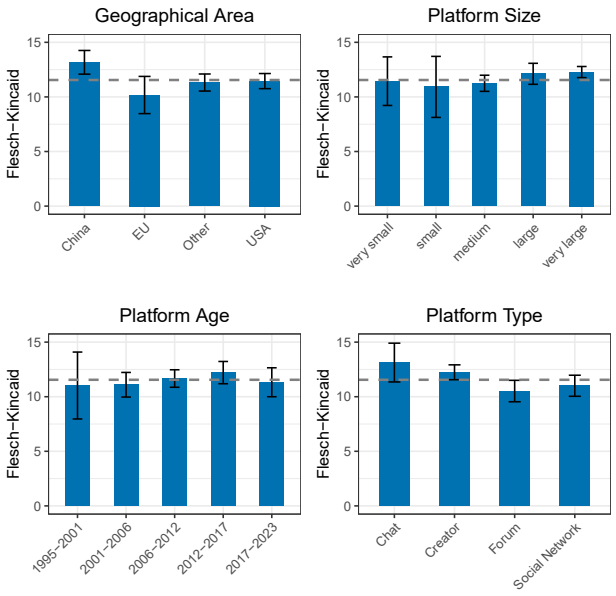

Figure 5: Average readability (Flesch-Kincaid Grade Score) by geographical area (top left), platform size (top right), age of platform (bottom left), and platform type (bottom right). The dashed line displays the mean across all guidelines.

other text types (see Figure 6) shows that these guidelines have an intermediate level of readability. On average, community guidelines require a high school level education to understand them (FKGS = 11.6). As one would expect, they are more readable than political science abstracts (FKGS = 17), which are scientific texts written for an expert audience. Community guidelines also are more readable than privacy policies (FKGS = 14), essentially legal documents, and broadsheet newspapers (FKGS = 12.6). At the same time, community guidelines are less readable than tabloid newspaper articles (FKGS = 8.4) and social media posts (FKGS = 9.9).[5] This is particularly interesting because there seems to be an one and a half year readability gap between the text produced by social media users and platforms' community guidelines written for these users.

Semantic Complexity

Finally, we measure the semantic complexity, defined as the number of content categories platforms explicitly ban in their guidelines. On average, platforms include 26.3 content categories (median = 22) in their community guidelines. Out of all the 80 possible content categories, Facebook's community guidelines prohibit the highest number of categories (n = 69), followed by TikTok (n = 67), Douyin (n = 61), WeChat (n = 60) and YouTube (n = 59). Foursquare (n = 2), Xing, Slug, and Hacker News (n = 3) include the fewest content categories in their community guidelines.

[5]The readability scores for privacy policies are extracted from Jensen and Potts (2004). The scores for the abstracts and newspaper articles are replicated based on data and scripts from Rauh (2023) and for social media posts from Özdemir and Rauh (2022).

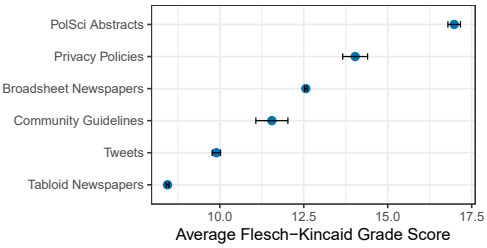

Figure 6: Average readability of different text types with 95% confidence intervals.

Figure 7 shows what proportion of categories per group are explicitly banned in the community guidelines on each platform (see Appendix A5 for the definitions of categories and groups). Darker shading in the heatmap indicates that a higher proportion of categories from the respective group is explicitly mentioned in the guidelines. The heatmap also reveals clusters among platforms. For example, only roughly a fifth of the platforms (n = 17) ban a category subsumed under political content.

Comparing platforms according to group characteristics (see Figure 8), there is a clear trend in regard to platform size where with growing platform size, more categories are banned within the community guidelines. Very large platforms include significantly more categories than very small, small, medium-sized, and large platforms (all group comparisons $p < 0.01$, see also Appendix A6 for OLS regression results).

Are The Complexity Dimensions Related?

Overall, the platforms that include more categories of banned content in their community guidelines also tend to have longer guidelines, which makes sense because they cover more ground. However, these platforms also tend to have less readable guidelines, suggesting that with length does not come increased usability (see Figure 9).

Length and semantic complexity are highly correlated ($r = 0.71$). This suggests that platforms have longer guidelines if they include more categories of prohibited content in the guidelines. While the correlation is fairly strong, it is not a perfect correlation, thus indicating that some platforms may make more efficient use of space.

Interestingly, however, guidelines highlighting a higher number of prohibited content categories is also correlated with less readable text and a higher Flesch-Kincaid Grade Score (r = 0.60). This indicates that the more categories covered, the more difficult the text is to read. Similarly, the length and readability scores of community guidelines are slightly positively correlated (r = 0.35), indicating that longer texts are less readable and require more years of education to be fully accessible.

However, there are also platforms defying this trend. For example, while YouTube's community guidelines are by far the longest, the guidelines are more readable (FKGS = 10.5)

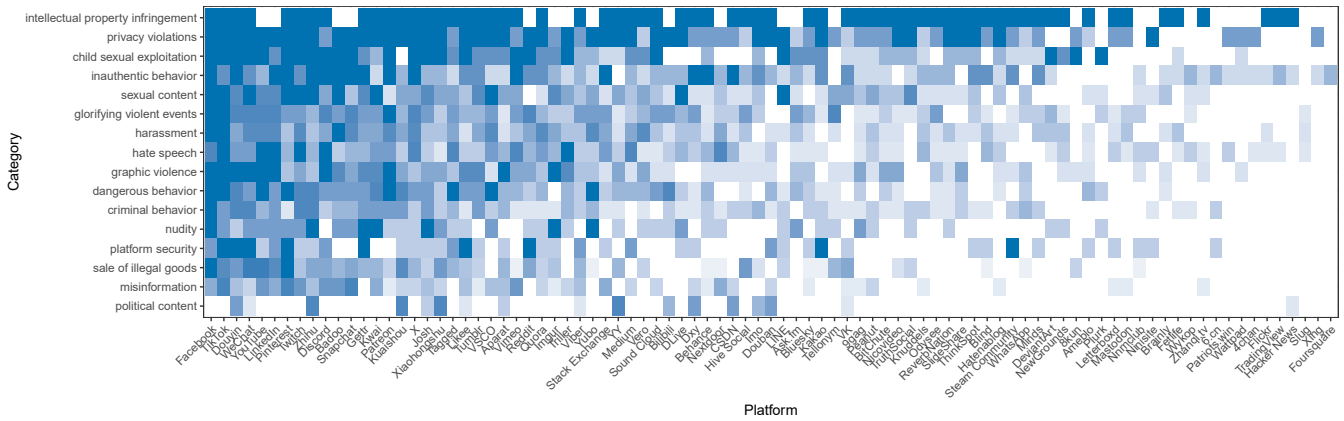

Figure 7: The heatmap shows what proportion of the content categories from each group are highlighted in the guidelines. A darker shading indicates that a higher proportion of categories are mentioned. White panels suggest that no category in this group is mentioned in the guidelines. The groups on the y-axis are ordered such that content groups with higher proportions appear at the top. The platforms on the x-axis are ordered according to the absolute number of categories mentioned per platform.

than the average. On the other side, there are also platforms that provide very short but still relatively difficult-to-read guidelines such as Letterboxd (FKGS = 12.8), Mastodon (FKGS = 12.6), and Plurk (FKGS = 12.3). This indicates that there may be more and less optimal ways of organizing guidelines that platforms should explore.

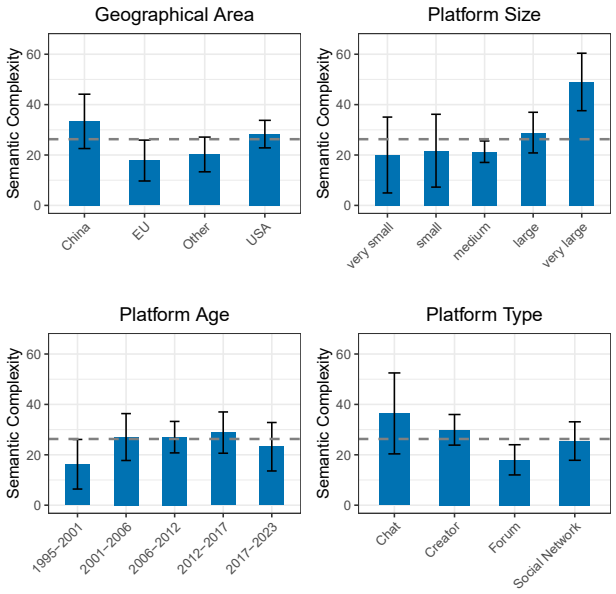

Figure 8: Average number of in the guidelines explicitly prohibited content categories by geographical area (top left), platform size (top right), age of platform (bottom left), and platform type (bottom right). The dashed line displays the mean across all guidelines.

How Complexity Evolves Over Time

With data provided by the Platform Governance Archive (Katzenbach et al. 2023a,c), it is possible to trace the evolution of the length and the readability of community guidelines for a subset of platforms over time. The first recorded community guidelines are available from August 2007 (Facebook) and January 2009 (X/Twitter).[6] From 2022 onwards, the PGA's data generation method changed and moved away from internet archive sources to tracking real-time updates. At the same time, the platform scope increasingly expanded and covers community guidelines of 13 of the platforms that are part of our sample at the time of analysis.

Overall, we can see that the length of community guidelines follows an increasing trend (see Figure 10)[7], as already noted by Katzenbach et al. (2023b). While the increase in length can be attributed to various factors, such as for example the pandemic or changing world affairs, it appears that sudden expansions are accompanied also by policy initiatives trying to improve transparency in content moderation. Particularly, the development of the length of Facebook's community guidelines aligns with this, revealing bigger jumps at the time of the publication of the *Manila Principles on Intermediary Liability* (Electronic Frontier Foundation et al. 2015), the publication of the *Santa Clara Princi-*

[6] As above, we drop Instagram's guidelines because they are essentially identical to Facebook's guidelines.

[7] These findings should be treated with a grain of salt. Some outliers could be due to the data generation procedure of the Platform Governance Archive, based on volunteers adding or removing subpages, or irregularities in technical nature. Therefore, we removed one outlier from Facebook's community guidelines (20/03/2019) because the text length momentarily doubled, and logically impossible references to COVID-19 were made. The temporary peak in X's community guidelines from November 11, 2022 to October 12, 2023 is due to volatility in the inclusion of sub-links and can be disregarded as well.

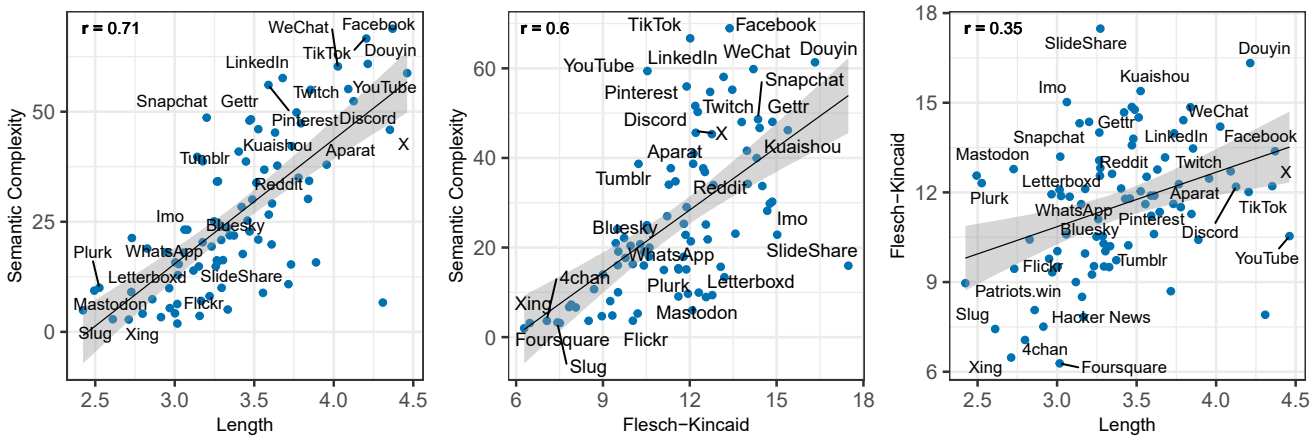

Figure 9: Scatterplots showing correlations of complexity dimensions across social media platforms. Higher Flesch-Kincaid Grade Score indicates less readable text. Length is measured using the base-10 logarithm of the number of tokens.

ples On Transparency and Accountability in Content Moderation (which were later endorsed by many major social media platforms including Facebook and Twitter; Access Now et al. 2018) and the adoption of the *Council of the European Union's position on the DSA* (Council of the European Union 2021). Since 2022 (right panel), particularly YouTube's community guidelines may have changed due to the new DSA regulations introduced in August 2023.

In terms of readability, there is no clear trend and for most platforms readability scores remain stable over time.

Discussion

Using the new COMPARE dataset, we have investigated the complexity of community guidelines of social media platforms. With over two-thirds of the platforms now publishing community guidelines, we show that there is tremendous variation in their complexity as measured by three dimensions – length, readability, and semantic complexity. Shorter guidelines can fit on a single standard Microsoft Word page, while the longest guidelines would take an average reader over two hours to read. Community guidelines, on average, require a high school level education in order to understand them. Interestingly, guidelines are also less readable than social media posts produced by those users that these guidelines are intended for.

Platforms use community guidelines to ban content categories to varying degrees, ranging from including only two categories in community guidelines to explicitly banning 69 categories of content. Moreover, we find that longer and semantically more complex guidelines suffer from lower readability, thus revealing trade-offs in platform's design choices. When considering the evolution of community guidelines, some platforms seem to expand their guidelines when there is increased attention due to the publication of transparency initiatives. This aligns with findings from the literature about the evolution of community guidelines over time.

Contributions

The COMPARE dataset provides information on platforms as well as links to content moderation policies for 132 platforms. We make this dataset publicly available, to facilitate further comparative research on platforms and their policies. Moreover, our analysis studies 89 community guidelines, including those of 14 platforms based in China. In doing so, we significantly expand the scope of previous research and assume a more global focus. This is important at a time in which the vast majority of research, but also the heated debates about content moderation and free speech, are based almost exclusively on Western platforms and the Western legal frameworks that underpin them (most importantly, the American and the European ones). Additionally, we contribute conceptually to the literature on transparency in content moderation by drawing attention to an overlooked aspect of transparency. We argue that beyond the mere provision of community guidelines, it is also important to consider how information is communicated and how community guidelines are formulated for consumption by the users of these platforms. After all, these users should be the main audience that these guidelines speak to, which is a perspective that has been neglected in previous research.

Limitations

Despite all its contributions, our comparative study suffers from a few limitations. First of all, the text of the community guidelines may not be the only way that platform policies are communicated. Our analysis ignores other types of content such as images (e.g., Xing) or videos (e.g., Yubo), thus potentially leaving out additional content. However, it seems that visual additions generally only repeat the content already contained in the text. Also, the text of community guidelines does not follow traditional sentence structures (e.g., by using subheadings or bullet point lists) and thus poses a challenge to applying syntactic readability measurements. We addressed this with extensive text cleaning. A second challenge is the comparison of guidelines in dif-

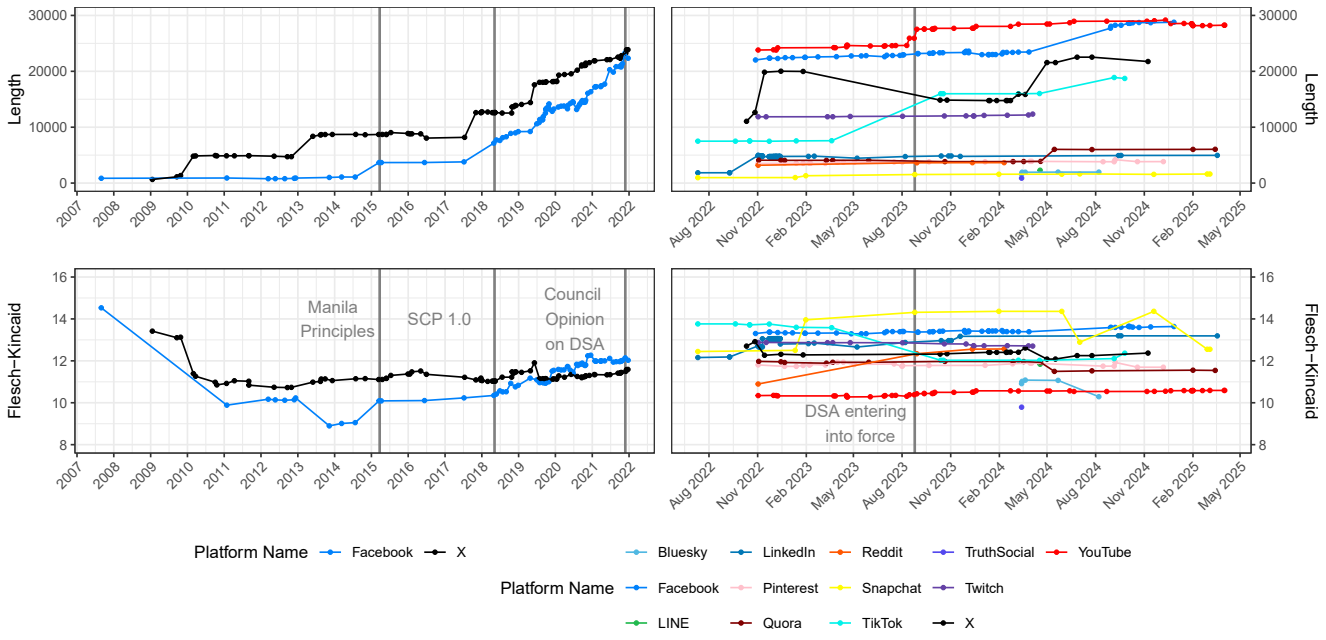

Figure 10: Evolution of length and readability from 2007 to 2021 (left panel) and from 2022 onwards (right panel).

ferent languages (Licht and Lind 2023). Similar to much other comparative work across different languages, our decision to translate the guidelines and apply our measurements to the English texts raises potential issues. Differences between guidelines in different languages could be driven by the different levels of complexity between languages, not the texts themselves. We have attempted to address this issue in different ways: we show that both English and non-English texts tend to follow the same trend and that the platform-provided English guidelines (in cases where they exist) lead to the same conclusions. We have also tested the robustness of our findings using different translation tools. In a final (and demanding) robustness test, we translate all the non-Chinese texts to Chinese and compute complexity measures designed for this language. A third challenge for our analysis is the time-varying nature of guidelines. Social media and their community guidelines operate in a dynamic environment. Guidelines change and platforms go offline (e.g., Caffeine). Therefore, much of our work (with the exception of the time trends shown in Figure 10) only captures a snapshot in time.

Recommendations

As we show in our analysis, the push for more transparency in content moderation has resulted in increased complexity of the community guidelines, the main documents that are supposed to communicate reasons for content moderation to end users. We want to caution against this problematic trend. While platforms certainly have reasons to publish community guidelines of epic length, including legal certainty and protection from criticism, the question remains whether this is helpful to social media users. There are two ways in which social media platforms could address the current situation. First, they can attempt to keep guidelines shorter. As all modern humans, social media users operate under time and attention constraints. Publishing lengthy community guidelines makes it thus less likely for users to find the relevant information about the rules for using a platform. Second, part of the problem also results from our finding that guideline length goes along with decreased readability. In other words, it is not the case that longer guidelines use less complex text, but rather the opposite. This issue could be addressed by providing guidelines in simple language, which is what other organizations such as certain news outlets already do. This would make guidelines more accessible to larger audiences and could thus help to increase transparency and public acceptance in the contested area of content moderation.

Data Availability

The COMPARE dataset, as well as other data and materials for our study, are available on GitHub at https://github.com/transparency-in-content-moderation/COMPARE.

Acknowledgments

This work is supported by a Max Planck/Alexander von Humboldt Research Award to Professor Margaret E. Roberts. The authors would like to thank Christian Rauh for providing the data for benchmarking the readability scores, and Xiaolin Chen and Leonard Tiedemann for their excellent research assistance.

References

Access Now; ACLU Foundation of Northern California; ACLU Foundation of Southern California; ARTICLE 19; Brennan Center for Justice; Center for Democracy & Technology; Electronic Frontier Foundation; Global Partners Digital; InternetLab; National Coalition Against Censorship; New America's Open Technology Institute; Ranking Digital Rights; Red en Defensa de kos Derechos Digitales; and WITNESS. 2018. The Santa Clara Principles on Transparency and Accountability in Content Moderation.

Balci, U.; Sirivianos, M.; and Blackburn, J. 2024. Exploring Left-Wing Extremism on the Decentralized Web: An Analysis of Lemmygrad.ml. *Workshop Proceedings of the 18th International AAAI Conference on Web and Social Media - Workshop: DeWeb 2024: 1st International Workshop on Decentralizing the Web*, 1–6.

Barrett, B.; and Kreiss, D. 2019. Platform transience: changes in Facebook's policies, procedures, and affordances in global electoral politics. *Internet Policy Review*, 8(4): 1–22.

Benoit, K.; Munger, K.; and Spirling, A. 2019. Measuring and Explaining Political Sophistication through Textual Complexity. *American Journal of Political Science*, 63(2): 491–508.

Bischof, D.; and Senninger, R. 2018. Simple politics for the people? Complexity in campaign messages and political knowledge. *European Journal of Political Research*, 57(2): 473–495.

Brysbaert, M. 2019. How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*, 109: 1–30.

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1): 37–46.

Council of the European Union. 2021. Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC - General approach.

DeepL. 2025. DeepL API.

Derguech, W.; Zainab, S. S. E.; and D'Aquin, M. 2018. Assessing the Readability of Policy Documents: The Case of Terms of Use of Online Services. In *Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance*, 247–256. Galway Ireland: ACM.

Eady, G.; Paskhalis, T.; Zilinsky, J.; Bonneau, R.; Nagler, J.; and Tucker, J. A. 2023. Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. *Nature Communications*, 14(1): 1–11.

Electronic Frontier Foundation; Access Now; Article 19; Center for Internet and Society; and Red en Defensa de kos Derechos Digitales. 2015. Manila Principles on Intermediary Liability.

European Union. 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).

Flesch, R. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3): 221–233.

Freelon, D.; Marwick, A.; and Kreiss, D. 2020. False equivalencies: Online activism from left to right. *Science*, 369(6508): 1197–1201.

Gillespie, T. 2019. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press. ISBN 978-0-300-23502-9.

Google. 2025. Cloud Translation API.

Gorwa, R. 2024. *The Politics of Platform Regulation: How Governments Shape Online Content Moderation*. Oxford University PressNew York, 1 edition. ISBN 978-0-19-769285-1 978-0-19-769289-9.

Gunning, R. 1952. *The technique of clear writing*. McGraw-Hill.

Gyasi, W. K.; and Bangmarigu, M. J. 2020. Exploring the Readability of Terms of Service of Social Networking Sites. *Covenant Journal of Informatics & Communication Technology*, 8(1): 16–33.

Jensen, C.; and Potts, C. 2004. Privacy policies as decision-making tools: an evaluation of online privacy notices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, 471–478. New York, NY, USA: Association for Computing Machinery.

Jerit, J.; and Zhao, Y. 2020. Political Misinformation. *Annual Review of Political Science*, 23(1): 77–94.

Jhaver, S.; Appling, D. S.; Gilbert, E.; and Bruckman, A. 2019. "Did you suspect the post would be removed?": understanding user reactions to content removals on reddit. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–33.

Jhaver, S.; Rathi, H.; and Saha, K. 2024. Bystanders of Online Moderation: Examining the Effects of Witnessing Post-Removal Explanations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–9. Honolulu HI USA: ACM.

Jiang, J. A.; Middler, S.; Brubaker, J. R.; and Fiesler, C. 2020. Characterizing Community Guidelines on Social Media Platforms. In *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing*, 287–291. Virtual Event USA: ACM.

Juneja, P.; Rama Subramanian, D.; and Mitra, T. 2020. Through the Looking Glass: Study of Transparency in Reddit's Moderation Practices. *Proceedings of the ACM on Human-Computer Interaction*, 4(GROUP): 1–35.

Kamp, D.; and Ekins, E. 2021. Poll: 75% Don't Trust Social Media to Make Fair Content Moderation Decisions, 60% Want More Control over Posts They See.

Katzenbach, C.; Dergacheva, D.; Fischer, A.; Kopps, A.; Kolesnikov, S.; Redeker, D.; and Viejo Otero, P. 2023a. Platform Governance Archive (PGA) v2.

Katzenbach, C.; Kopps, A.; Magalhães, J. C.; Redeker, D.; Sühr, T.; and Wunderlich, L. 2023b. The Platform Governance Archive v1 – A longitudinal dataset to study the governance of communication and interactions by platforms and the historical evolution of platform policies [Data Paper].

Katzenbach, C.; Magalhães, J. C.; Kopps, A.; Sühr, T.; and Wunderlich, L. 2023c. Platform Governance Archive (PGA) v1.

Kincaid, J. P.; Fishburne, J.; Robert P., R.; Richard L., C.; and Brad S. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel:. Technical report, Defense Technical Information Center, Fort Belvoir, VA.

Kopf, S. 2024. Corporate censorship online: Vagueness and discursive imprecision in YouTube's advertiser-friendly content guidelines. *New Media & Society*, 26(4): 1756–1774.

Licht, H.; and Lind, F. 2023. Going cross-lingual: A guide to multilingual text analysis. *Computational Communication Research*, 5(2): 1–31.

Lin, N.; and Osnabrügge, M. 2018. Making comprehensible speeches when your constituents need it. *Research & Politics*, 5(3): 1–8.

Lorenz-Spreen, P.; Oswald, L.; Lewandowsky, S.; and Hertwig, R. 2023. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour*, 7(1): 74–101.

MacCarthy, M. 2020. Transparency requirements for digital social media platforms: recommendations for policy makers and industry. *SSRN Electronic Journal*, 1–36.

Maddox, J.; and Malson, J. 2020. Guidelines Without Lines, Communities Without Borders: The Marketplace of Ideas and Digital Manifest Destiny in Social Media Platform Policies. *Social Media + Society*, 6(2): 1–10.

Marchisio, K.; Guo, J.; Lai, C.-I.; and Koehn, P. 2019. Controlling the Reading Level of Machine Translation Output. *Proceedings of MT Summit XVII*, 1: 193–203.

McDonald, A. M.; Reeder, R. W.; Kelley, P. G.; and Cranor, L. F. 2009. A Comparative Study of Online Privacy Policies and Formats. In Goldberg, I.; and Atallah, M. J., eds., *Privacy Enhancing Technologies*, 37–55. Berlin, Heidelberg: Springer.

Pierri, F.; Luceri, L.; Chen, E.; and Ferrara, E. 2023. How does Twitter account moderation work? Dynamics of account creation and suspension on Twitter during major geopolitical events. *EPJ Data Science*, 12(1): 1–21.

Quint, F.; Theocharis, Y.; Nahrgang, M.; Weidmann, N. B.; and Roberts, M. E. 2025. Does the Community Understand the Community Guidelines? Working paper.

Rajendra-Nicolucci, E. C.; and Zuckerman, E. 2021. *An Illustrated Field Guide to Social Media. Knight First Amendment Institute*.

Rauh, C. 2023. Clear messages to the European public? The language of European Commission press releases 1985–2020. *Journal of European Integration*, 45(4): 683–701.

Rooney, M. K.; Santiago, G.; Perni, S.; Horowitz, D. P.; McCall, A. R.; Einstein, A. J.; Jagsi, R.; and Golden, D. W. 2021. Readability of Patient Education Materials From High-Impact Medical Journals: A 20-Year Analysis. *Journal of Patient Experience*, 8: 1–9.

Rozenshtein, A. Z. 2022. Moderating the Fediverse: Content Moderation on Distributed Social Media. *SSRN Electronic Journal*, 218–236.

Sare, A.; Patel, A.; Kothari, P.; Kumar, A.; Patel, N.; and Shukla, P. A. 2020. Readability Assessment of Internet-based Patient Education Materials Related to Treatment Options for Benign Prostatic Hyperplasia. *Academic Radiology*, 27(11): 1549–1554.

Schoonvelde, M.; Brosius, A.; Schumacher, G.; and Bakker, B. N. 2019. Liberals lecture, conservatives communicate: Analyzing complexity and ideology in 381,609 political speeches. *PLOS ONE*, 14(2): 1–15.

Shaughnessy, B.; DuBosar, E.; Hutchens, M. J.; and Mann, I. 2024. An attack on free speech? Examining content moderation, (de-), and (re-) platforming on American right-wing alternative social media. *New Media & Society*, 1–19.

Siapera, E. 2023. Alt Tech and the public sphere: Exploring Bitchute as a political media infrastructure. *European Journal of Communication*, 38(5): 446–465.

Similarweb. 2024. Monthly Visits.

Singhal, M.; Ling, C.; Paudel, P.; Thota, P.; Kumarswamy, N.; Stringhini, G.; and Nilizadeh, S. 2023. SoK: content moderation in social media, from guidelines to enforcement, and research to practice. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, 868–895.

Suzor, N. P.; West, S. M.; Quodling, A.; and York, J. 2019. What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication*, 13: 1526–1543.

Trujillo, A.; Fagni, T.; and Cresci, S. 2025. The DSA Transparency Database: Auditing Self-reported Moderation Actions by Social Media. ArXiv: 2312.10269 [cs].

Tucker, J.; Guess, A.; Barbera, P.; Vaccari, C.; Siegel, A.; Sanovich, S.; Stukal, D.; and Nyhan, B. 2018. Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. *SSRN Electronic Journal*.

Tucker, J. A.; Theocharis, Y.; Roberts, M. E.; and Barberá, P. 2017. From liberation to turmoil: Social media and democracy. *Journal of Democracy*, 28(4): 46–59.

Urman, A.; and Makhortykh, M. 2023. How transparent are transparency reports? Comparative analysis of transparency reporting across online platforms. *Telecommunications Policy*, 47(3): 1–15.

Wubben, S. 2012. Sentence Simplification by Monolingual Machine Translation. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 1015–1024.

Xu, W.; Yao, Z.; and Chen, D. 2019. Chinese annual report readability: measurement and test. *China Journal of Accounting Studies*, 7(3): 407–437.

Özdemir, S.; and Rauh, C. 2022. A Bird's Eye View: Supranational EU Actors on Twitter. *Politics and Governance*, 10(1): 133–145.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, answering the research questions posed in this paper will help advance science with novel data that in no way violates social contracts. Community Guidelines are not sensitive data and are publicly available.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, the abstract accurately and briefly reflects the contributions and scope of the paper.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, we thoroughly explain the operationalization of the key measurements and our methodological approach.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, possible artifacts are addressed by discussing country- and language-specific differences in the data collection and processing.
   (e) Did you describe the limitations of your work? Yes, see the Limitations section.
   (f) Did you discuss any potential negative societal impacts of your work? Not applicable. No negative societal impacts are involved.
   (g) Did you discuss any potential misuse of your work? Not applicable. There is no potential misuse of our work.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, all steps taken are transparently described to ensure reproducibility and limitations are described.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, the paper conforms to the ethics review guidelines.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Not applicable. No hypothesis testing is involved.
   (b) Have you provided justifications for all theoretical results? Not applicable. No hypothesis testing is involved.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Not applicable. No hypothesis testing is involved.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Not applicable. No hypothesis testing is involved.
   (e) Did you address potential biases or limitations in your theoretical framework? Not applicable. No hypothesis testing is involved.
   (f) Have you related your theoretical results to the existing literature in social science? Not applicable. No hypothesis testing is involved.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Not applicable. No hypothesis testing is involved.
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? Not applicable. No theoretical proofs are involved.
   (b) Did you include complete proofs of all theoretical results? Not applicable. No theoretical proofs are involved.
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Not applicable. No machine learning experiments were involved.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Not applicable. No machine learning experiments were involved.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Not applicable. No machine learning experiments were involved.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Not applicable. No machine learning experiments were involved.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Not applicable. No machine learning experiments were involved.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Not applicable. No machine learning experiments were involved.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, all original creators are cited.
   (b) Did you mention the license of the assets? Not applicable. No license is needed for data access.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, we are planning to make our dataset available on GitHub.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Not applicable. The data used from other sources is freely accessible and correctly cited.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Not applicable. The data used does not contain personal information nor offensive content.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes, we will make sure that our data is findable, accessible, interoperable, and reusable.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? Yes, aspects regarding dataset composition and data collection are thoroughly discussed in the data section as well as the appendix.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Not applicable. No human subjects were involved.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Not applicable. No human subjects were involved.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Not applicable. No human subjects were involved.

(d) Did you discuss how data is stored, shared, and deidentified? Not applicable. No human subjects were involved.

Appendix

A1 Sources for Social Media List Compilation

COMPARE is a dataset of social media platforms with a worldwide user base. To identify larger platforms, we started with an initial list of major social media platforms by combining the most popular social media platforms from ten global and regional (U.S., China, Germany) rankings. This resulted in 156 platforms. Seven platforms were not considered further because they became inactive. From this candidate list, we then selected those platforms for inclusion in the dataset if they fit our definition of "social media platforms" by (i) hosting user-generated content that is (ii) at least to some extent public-facing and in principle visible to anyone upon registration and (iii) that stays consistently accessible over a longer period. We followed an inclusive approach and decided that if any feature of a social media platform fulfills our three criteria, it should be included in our list. Eight platforms each did not meet the first or second inclusion criterion. Five platforms did not meet the third inclusion criterion.

- Rajendra-Nicolucci, C. & E. Zuckerman. 2021. Top 100: The most popular social media platforms and what they can teach us. https://knightcolumbia.org/blog/top-100-the-most-popular-social-media-platforms-and-what-they-can-teach-us.
- Similarweb. 2023. Top Websites Ranking. Most Visited Social Media Networks Websites. https://www.similarweb.com/top-websites/computers-electronics-and-technology/social-networks-and-online-communities.
- Wikipedia. 2023. List of social platforms with at least 100 million active users. https://en.wikipedia.org/w/index.php?title=List_of_social_platforms_with_at_least_100_million_active_users&oldid=1173223873.
- Statista. 2023. Most popular social networks worldwide as of January 2023, ranked by number of monthly active users. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users.
- Statusbrew. 2023. 100+ Social Media Statistics You Need To Know in 2023 [All Networks]. https://statusbrew.com/insights/social-media-statistics/#most-popular-social-networks.
- Lua., A. 2023. 23 Top Social Media Sites to Consider for Your Brand in 2023. Buffer. https://buffer.com/library/social-media-site.
- Gaasly. 2023. Social media trends in Germany 2023. https://www.gaasly.com/blog/social-media-trends-in-germany.
- Kemp, S. 2023. Digital 2023: The United States of America. Datareportal. https://datareportal.com/reports/digital-2023-united-states-of-america.
- Azoya. 2023. Top Chinese Social Media Platforms & Apps in 2023 You Need to Know. https://www.azoyagroup.com/page/view/top-chinese-social-media-platform-you-need-to-know.
- Singh, S. 2023. Top 100+ Social Media Sites & Platforms. Moneymint. https://moneymint.com/top-social-media-sites.

A2 Definition of COMPARE Content Moderation Links

Table A2 contains definitions of the six content moderation links we collected for COMPARE. For each platform, we collected links to the (1) privacy policies, (2) terms of service, (3) community guidelines, (4) transparency reports, and information about (5) the platform's content moderation enforcement options and (6) how the moderation process is structured. Those aspects were chosen because they cover the most relevant characteristics of a platform's content moderation strategy. The links were collected from October 2023 to January 2024. As we were interested in how transparent platforms themselves are, these links in general refer to statements of the platforms on their websites except if a platform has incorporated the policies of an affiliated company. For example, as YouTube is part of Google, it relies on Google's privacy policy. Moreover, links could

Platform	Guidelines	Platform	Guidelines	Platform	Guidelines
4chan	yes	5ch	no	6.cn	yes
8kun	yes	9gag	yes	Ameblo	yes
Aparat	yes	Ask.fm	yes	Babytree	no
Badoo	yes	Baidu Tieba	no	Behance	yes
Bilibili	yes	BitChute	yes	BizSugar	no
Blind	yes	Bluesky	yes	Brainly	yes
Caffeine	yes	CloutHub	no	Cnblogs.com	no
Computerbase.de	no	CSDN	yes	Dcard	no
DeSo	no	DeviantArt	yes	Diaspora	no
Discord	yes	DLive	yes	Douban	yes
Douyin	yes	DTube	no	Dxy	yes
Exblog	no	Eyny	no	Facebook	yes
Fetlife	yes	Flickr	yes	Foursquare	yes
Gab	no	Gettr	yes	Hacker News	yes
Hatenablog	yes	Hive Social	yes	Imgur	yes
Imo	yes	Instagram	yes	Instiz	no
Josh	yes	Kakao	yes	KizlarSoruyor	no
Knuddels	yes	Kuaishou	yes	Kwai	yes
Letterboxd	yes	Lihkg	no	Likee	yes
LINE	yes	LinkedIn	yes	LiveJournal	no
Mastodon	yes	Medium	yes	MeWe	no
Miaopai	no	Minds	yes	Mixi	no
Mydigit	no	Namasha	no	NewGrounds	yes
Nextdoor	yes	Nicovideo	yes	Ninisite	yes
Nnmclub	yes	Odysee	yes	OK	no
Patreon	yes	Patriots.win	yes	Peanut	yes
Pinterest	yes	Pixnet	no	Plurk	yes
Ptt	no	QQ	no	Quora	yes
QZone	no	Reddit	yes	ReverbNation	yes
Rumble	no	Sina Weibo	no	Skoob	no
SlideShare	yes	Slug	yes	Snapchat	yes
Sound Cloud	yes	Stack Exchange	yes	Steam Community	yes
Steemit	no	Tagged	yes	Taringa	no
Telegram	no	Tellonym	yes	The Dots	no
ThinkSpot	yes	Threads	yes	TikTok	yes
Toutiao	no	Trading View	yes	Triller	yes
TruthSocial	yes	Tumblr	yes	Twitch	yes
Valence	no	Vero	yes	Viber	yes
Vimeo	yes	VK	yes	VSCO	yes
Wattpad	yes	WeChat	yes	WhatsApp	yes
WrongThink	no	Wykop	yes	X	yes
Xiaohongshu	yes	Xing	yes	Yizhibo	no
Youku	no	YouTube	yes	Yubo	yes
YY	yes	Zhanqi.tv	yes	Zhihu	yes

Table A1: 132 Platforms in the COMPARE dataset and whether they provide community guidelines.

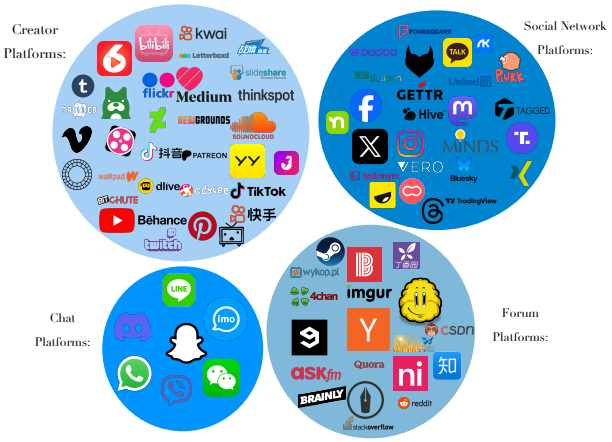

Figure A1: Visualization of platform logos of those platforms that provide community guidelines, grouped by platform type.

be collected and reused for multiple variables if they contain the corresponding information. For example, platforms might decide to address how they enforce content moderation and how the content moderation process is structured under the same link.

Variable	Definition
Privacy Policies (privacy)	URL to a platform's privacy policy or similar document detailing data protection rules.
Terms of Services (tos)	URL to a platform's terms of service/use or similar document containing legal rules of the platform's usage.
Community Guidelines (comguide)	URL to a platform's community guidelines or similar document describing the rules of behavior for the platform.
Transparency (trarep)	URL to a platform's transparency report where the platform reports about the scope of content moderation in a given timeframe.
Enforcement (enfopt)	URL to a platform's enforcement options of content moderation. These can be for example the removal, downgrading, or labeling of content or users.
Process (cmpro)	URL to a platform's website where the platform describes how the content moderation process is organized for example if it relies on automated means, users, or professional reviewers.

Table A2: Definitions of the content moderation variables in the COMPARE data set

A3 COMPARE Platforms

Table A1 provides an overview of the social media platforms contained in the COMPARE dataset and whether they published community guidelines. This is complemented by Figure A1 that visualizes platform logos grouped by platform type.

A4 Text Cleaning Steps of Community Guidelines

All community guidelines were scraped as markdown documents. Therefore, a markdown parser solved the main formatting issues by, for example, removing links. Additionally, we replaced newline characters and semicolons with dots. This was a strategic decision in order to address the particularities of the sentence structure of community guidelines (e.g., headlines, bullet points, etc.). Moreover, we removed special characters or translation-related combinations such as '#39;'. Finally, we harmonized the numbering from a wide range of options such as 1), (1), I., A., a., a) or (a) to Arabic numbers preceded and followed by a dot (as for example in "The following behavior is prohibited. 1. Hate Speech. 2. Harassment."). This was done in order to ensure that the average sentence length is not affected by different numbering systems.

A5 Prohibited Content Categories

This list displays the 80 categories in 16 groups which were used for determining semantic complexity.

- Child Sexual Exploitation
  - Minors Sexualization
  - Child Nudity
  - Child Exploitation Imagery
- Criminal Behavior
  - Vandalism
  - Scams
  - Human Trafficking
  - Celebrating Own Crime
  - Theft
- Depiction of Dangerous Behavior Risking Imitation
  - Dangerous Challenges
  - Suicide Depiction
  - Self-injury Depiction
  - Incitement to Dangerous Behavior
  - Eating Disorder Depiction
- Glorifying Violent Groups/Events
  - Incitement to Violence
  - Terrorist Propaganda
  - Mass Murder Support
  - Hate Group Propaganda
  - Criminal Group Propaganda
- Graphic Violence
  - Sexual Violence

  - Violence against Humans
  - Animal Abuse
  - Child Abuse
- Harassment
  - Bullying
  - Non-Consensual Intimate Imagery Threat
  - Non-Consensual Sexual Touching
  - Repeated Unwanted Advances
- Hate Speech
  - Slurs
  - Inferiority
  - Hateful Images and Symbols
  - Exclusion/Segregation
  - Dehumanization
- Inauthentic Behavior
  - Fake Profiles
  - Spam
  - Engagement Abuse
- Intellectual Property Infringement
- Misinformation
  - Fake News
  - Interference with Elections
  - Propagating Conspiracy Theories
  - Health Related Misinformation
  - Denying well Documented Historical Events
  - Denying Climate Change
- Nudity
  - Adult Non-Sexual Nudity
  - Adult Non-Consensual Intimate Imagery
- Platform Security
  - Sharing Malicious Software
  - Interrupting Platform Services
- Political Content
  - National Security
  - National Unity
  - National Interests
  - Criticizing the Government/Authorities
  - Distorting Historical Narratives
- Privacy Violations
  - Impersonation
  - Exposure of Personal Information
- Sale of Illegal or Regulated Goods
  - Pharmaceutical Sales
  - Non-medical Drug Sale
  - Marijuana Sales
  - Live Animal Sale
  - Endangered Species Sale
  - Human Organ Sale

	Dependent variable:
	Availability of Community Guidelines
EU	2.251^{**}
	(0.914)
Other	-0.042
	(0.599)
USA	2.657^{***}
	(0.779)
Monthly Visits	0.295^{***}
	(0.096)
Year	0.086^{*}
	(0.046)
Creator	0.404
	(1.080)
Forum	-0.659
	(1.068)
Social Network	-0.299
	(1.108)
Decentralized	-1.828^{**}
	(0.838)
Alt-Tech	-1.002
	(0.796)
Constant	-176.416^{*}
	(92.810)
Observations	130
Log Likelihood	-58.879
Akaike Inf. Crit.	139.758
Note:	*p<0.1; **p<0.05; ***p<0.01

Table A3: Logistic Regression Results

  - Firearm Sales
  - Counterfeit Products or Documents Sale
  - Alcohol and Tobacco Sale
- Sexual Content
  - Sexual Activity
  - Prostitution
  - Sexual Solicitation
  - Sexually Explicit Language

A6 Regression Results

To complement the descriptive graphs in the main part, we performed two regressions. Table A3 shows the results of a logistic regression used to determine which platform characteristics are more predictive of platforms publishing community guidelines. This confirms that platforms based in the U.S. and the EU as well as larger and centralized platforms are most likely to provide community guidelines. Moreover, Table A4 shows the results of an OLS regression with the three complexity dimensions as dependent variables.

	Dependent variable:		
	Length (log10)	Flesch-Kincaid	Semantic Complexity
	(1)	(2)	(3)
EU	−0.640	−2.475***	−10.765*
	(0.409)	(0.860)	(6.405)
Other	−0.515	−1.809**	−14.993***
	(0.344)	(0.723)	(5.384)
USA	−0.344	−1.287**	−4.870
	(0.305)	(0.641)	(4.772)
Monthly Visits	0.112***	0.037	2.424***
	(0.038)	(0.080)	(0.598)
Year	0.013	0.102**	0.910***
	(0.021)	(0.045)	(0.335)
Creator	0.072	−0.838	−3.248
	(0.397)	(0.835)	(6.220)
Forum	−0.242	−2.000**	−12.527*
	(0.422)	(0.888)	(6.614)
Social Network	0.166	−1.666*	−2.559
	(0.419)	(0.882)	(6.571)
Decentralized	−0.451	−0.338	−7.356
	(0.433)	(0.912)	(6.789)
Alt-Tech	−0.325	−1.643*	−6.801
	(0.411)	(0.865)	(6.443)
Constant	−20.682	−191.024**	−1,831.363***
	(43.144)	(90.760)	(675.927)
Observations	88	88	88
R^2	0.217	0.301	0.367
Adjusted R^2	0.115	0.210	0.285
Residual Std. Error (df = 77)	0.944	1.986	14.791
F Statistic (df = 10; 77)	2.133**	3.314***	4.466***
Note:			*p<0.1; **p<0.05; ***p<0.01

Table A4: OLS Regression Results

A7 External Validation of Readability Measurements

We obtained access to data from a related project (Quint et al. 2025), where in a laboratory setting, 337 participants were assigned one of ten randomly assigned social media platforms' community guidelines, and were asked to adjudicate whether a set of given content examples violate these guidelines. Following this, respondents were asked to report (i) the difficulty and the (ii) level of uncertainty of making these decisions based on the guidelines, on a 7-point Likert scale. Notably, the subjective difficulty and uncertainty measures correlate highly with the Flesch-Kincaid Grade Scores for the respective community guidelines ($r_{\text{difficulty}} = 0.75$ and $r_{\text{uncertainty}} = 0.70$, see Figure A2). This provides further evidence that readability (as measured by the FKGS) indeed does capture how accessible guidelines are to users.

A8 Internal Validation of Readability Measurements

In addition to our main results, we applied a different syntactic readability measure to test the robustness of our results. We chose the Gunning's Fog Index (Gunning 1952), which takes into account the average sentence length and the number of complex words as measured with words with more than three syllables.

$$\text{Gunning's Fog Index} = 0.4 \left(\frac{\text{total words}}{\text{total sentences}} \right) + 100 \left(\frac{\text{number of words with 3-syllables or more}}{\text{total words}} \right)$$

With this test we want to ensure that our findings are not specific to the Flesch-Kincaid Grade Score. The Gunning's Fog Index can be interpreted similarly to the Flesch-Kincaid Grade Score in that higher scores indicate less readable text. Figure A3 confirms that community guidelines of platforms based in the EU are significantly more readable than those of platforms based in China ($p < 0.01$). Using the Gunning's Fog Index readability metric, the differences between U.S.-based and China-based platforms ($p < 0.05$) also becomes significant. Our results are thus robust to changing the readability metric.

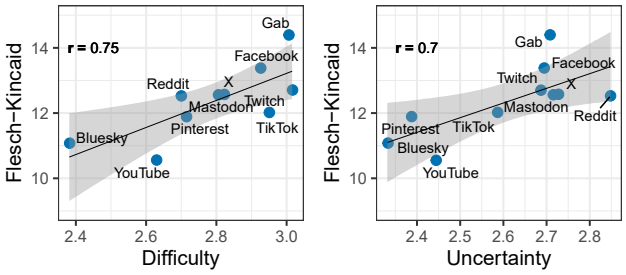

Figure A2: Scatterplots with average perceived difficulty/ uncertainty and average readability (Flesch-Kincaid Grade Score) per platform.

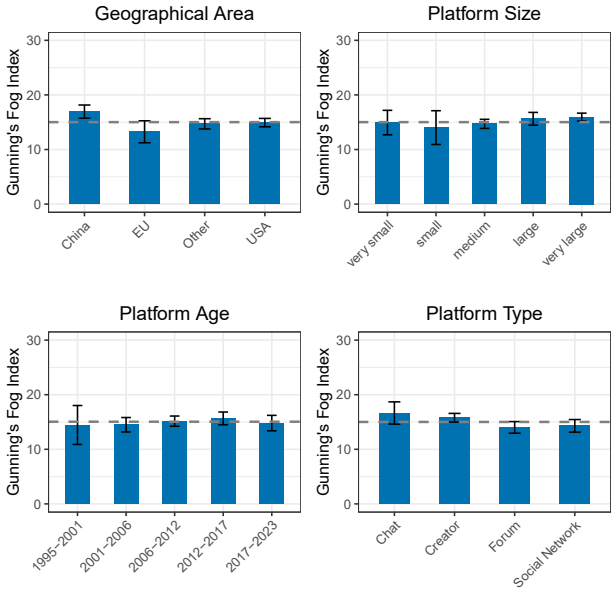

Figure A3: Average readability (Gunning's Fog Index) by geographical area (top left), platform size (top right), age of platform (bottom left), and platform type (bottom right). The dashed line displays the mean across all guidelines.

A9 Readability Measurements Applied to a Multilingual Corpus

To test the robustness of our findings, especially concerning the challenges of measuring readability in a multilingual corpus, we conducted a series of robustness checks.

First, we compared the translated texts (n = 22) with the non-translated texts (n = 67). We can see that both subsets follow essentially the same trend (see Figure A4) with longer guidelines being on average slightly less readable ($r_{\text{English}} = 0.28$, $r_{\text{Translations}} = 0.56$).

Moreover, in order to assess translation effects, we made use of the fact that some platforms provide their own English translations of the community guidelines. We compared the translations provided by the platform with the translations generated by us with DeepL for Kakao and WeChat. We scraped both versions of Kakao's Community Guidelines. The readability scores of both versions were almost identical with the platform-provided translation being slightly more readable ($\text{FKGS}_{\text{Platform Translation}} = 11.6$ and $\text{FKGS}_{\text{DeepL Translation}} = 11.8$). For WeChat, we used the platform-translated English community guidelines as provided by the PGA (Katzenbach et al. 2023a) and compared it to our DeepL translation. Again, both readability scores were fairly similar with the machine-translated text being slightly less readable (FKGS = 13.9) compared to the platform-translated text (FKGS = 13.4). In these cases, it thus seems that the machine translation made the text slightly more difficult to understand. However, differences are marginal, suggesting that automated translation is likely working for our application.

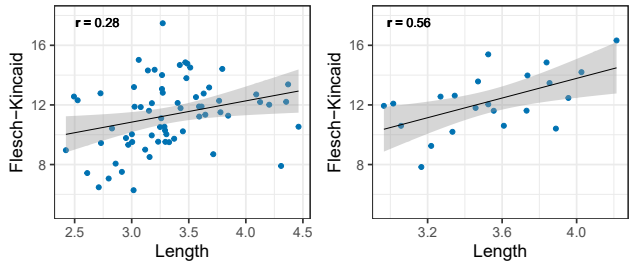

Figure A4: Scatterplots showing length and readability of non-translated (left) and translated (right) community guidelines. Length is measured using the base-10 logarithm of the number of tokens.

Furthermore, we used an alternative translation tool to generate the English translations, to make sure that our results do not depend on the choice of a particular translation service (Marchisio et al. 2019; Wubben 2012). For this, we translated the 22 community guidelines where the default language was not English using the Google Cloud Translation API (Google 2025) instead of the DeepL API (DeepL 2025). Figure A5 again confirms that EU-based platforms produce more readable guidelines than China-based platforms ($p < 0.05$). Moreover, we see that creator platforms have significantly less readable guidelines than forum platforms ($p < 0.05$).

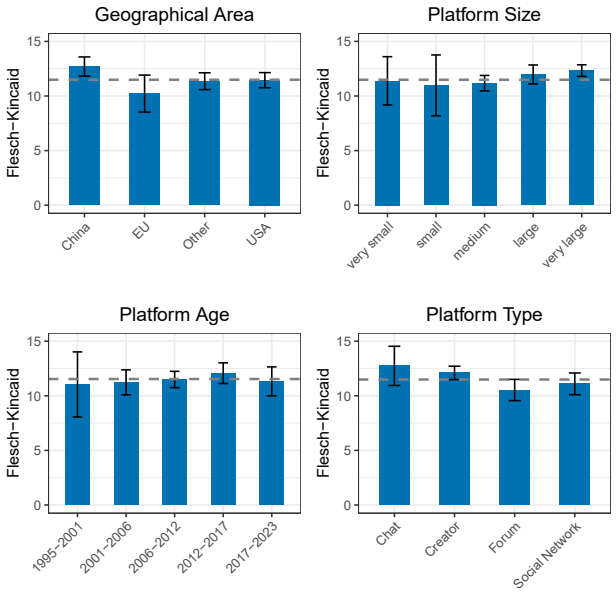

Figure A5: Average readability (Flesch-Kincaid Grade Score) calculated on translations generated with the Google Cloud Translation API by geographical area (top left), platform size (top right), age of platform (bottom left), and platform type (bottom right). The dashed line displays the mean across all guidelines.

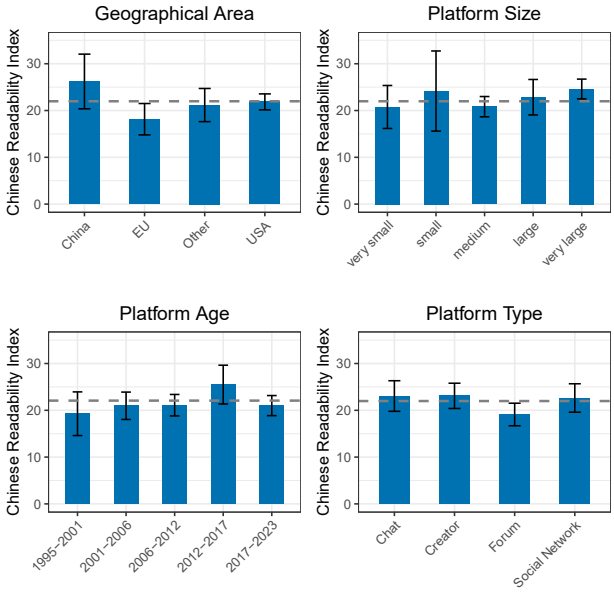

Figure A6: Average readability (Chinese Readability Index) calculated on community guidelines translated to Chinese by geographical area (top left), platform size (top right), age of platform (bottom left), and platform type (bottom right). The dashed line displays the mean across all guidelines.

Finally, we conducted an even more difficult robustness test. We translated all 89 community guidelines to Chinese with the DeepL API and applied a Chinese readability measurement (Xu, Yao, and Chen 2019). This metric is inspired by the Gunning's Fog Index but was developed on Chinese text, and combines the average word number in each clause and the proportion of adverbs and conjunctions in each sentence. Figure A6 again confirms our findings and shows that EU-based platforms have community guidelines that are more readable than China-based platforms ($p < 0.05$).

To summarize, our robustness tests confirm that (i) translated and non-translated community guidelines follow a similar trend; (ii) translation does not substantially reduce the readability; (iii) changing the translation software does not alter the results; and (iv) that translating all guidelines to Chinese and applying a Chinese Readability Index produces similar results. We are thus confident in our results despite the fact that the sample includes source texts in different languages.