"I Searched for a Religious Song in Amharic and Got Sexual Content Instead": Investigating Online Harm in Low-Resourced Languages on YouTube.

Hellina Hailu Nigatu
hellina_nigatu@berkeley.edu
UC Berkeley
USA

Inioluwa Deborah Raji
rajiinio@berkeley.edu
UC Berkeley
USA

ABSTRACT

Online social media platforms such as YouTube have a wide, global reach. However, little is known about the experience of low-resourced language speakers on such platforms; especially in how they experience and navigate harmful content. To better understand this, we (1) conducted semi-structured interviews (n=15) and (2) analyzed search results (n=9313), recommendations (n=3336), channels (n=120) and comments (n=406) of policy-violating sexual content on YouTube focusing on the Amharic language. Our findings reveal that – although Amharic-speaking YouTube users find the platform crucial for several aspects of their lives – participants reported unplanned exposure to policy-violating sexual content when searching for benign, popular queries. Furthermore, malicious content creators seem to exploit under-performing language technologies and content moderation to further target vulnerable groups of speakers, including migrant domestic workers, diaspora, and local Ethiopians. Overall, our study sheds light on how failures in low-resourced language technology may lead to exposure to harmful content and suggests implications for stakeholders in minimizing harm. **Content Warning**: This paper includes discussions of NSFW topics and harmful content (hate, abuse, sexual harassment, self-harm, misinformation). The authors do not support the creation or distribution of harmful content.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **HCI theory, concepts and models**; • **Social and professional topics** → *Cultural characteristics*; *Geographic characteristics.*

KEYWORDS

Online Harm, Recommendation Systems, Low-Resourced Languages, Community Guidelines, User Experience, Search Engines, Low-Resourced NLP, Policy

ACM Reference Format:
Hellina Hailu Nigatu and Inioluwa Deborah Raji. 2024. "I Searched for a Religious Song in Amharic and Got Sexual Content Instead": Investigating Online Harm in Low-Resourced Languages on YouTube.. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3630106.3658546

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658546

1 INTRODUCTION

As the dominant video platform in several countries [4, 53], YouTube is a global phenomenon. One study [21] indicates that around 67% of all content on YouTube in 2019 was posted in a language other than English. In addition to serving as a source of entertainment, YouTube provides a way to generate income for many, including low-income families in developing countries [53]. However, while online platforms such as YouTube bring benefits into the diverse contexts in which they operate, they also export their harms [53, 62, 77, 82].

Several government bodies around the world are trying to increase regulation over online content [6, 12]. However, across many legal contexts, "harmful content" includes both legal and illegal material [48], and so online platforms more reliably depend on their own "Community Guidelines" or other codes of conduct for moderating content on their platforms. Typically, platforms enforce their policies through a combination of human and automated content moderation pipelines [30, 71, 90]. Despite these guardrails, users can still be exposed to harmful content due to, for example, a lack of proper policy enforcement [17, 52] or content moderation avoidance by content creators [46]. Despite its global presence, and the platform's claim that "Community Guidelines are enforced consistently across the globe, regardless of where the content is uploaded"[13], previous work has already shown that "Non-English speakers are hit the hardest" when it comes to exposure to content the users "regret" [53]. Likely, the under-performance of language technologies in low-resourced languages [40, 65, 67][1]—i.e. languages that have limited digital data available [42, 61, 94]— combined with the lack of linguistic competence in these languages amongst human moderators [41], make content moderation in these contexts especially challenging.

While online content in low-resourced languages goes under-moderated, the experiences of users from these contexts are also under-studied[75, 84]. In this work, we study the experiences of low-resourced language speaking YouTube users. The YouTube transparency report [41] indicates that policy-violating sexual content is amongst the most moderated types of policy violation on the

[1]Following works from other disciplines [31, 86], we exclusively use the term *low-resourced* instead of *low-resource* to emphasize that the low-resourcedness comes from *choices* in design and development of language technologies that has left some languages—and by extension, the communities—behind, while these languages do not innately lack resource.

platform. While not all online sexual content is harmful [37, 79], online harm due to such content could occur in the form of non-consensual release of intimate videos and images [39], unintentional exposure to sexual content [87], and exposure of minors to sexual content [51]. Hence, we narrow our scope into studying *policy-violating sexual content on YouTube in Amharic*, one of the local languages in Ethiopia with over 25 million native speakers worldwide [16].

We set out to answer the following research questions:

> **RQ1**: What are the experiences and coping mechanisms of users exposed to policy-violating sexual content on YouTube in a low-resourced language context?
> **RQ2**: What are the characteristics of policy-violating sexual content on YouTube in Amharic?

To answer the above research questions, we conducted two studies. In *Study 1*, we conducted semi-structured interviews with low-resourced language speaking YouTube users to understand their experiences. In *Study 2*, we collected and analyzed search results, recommendations, and comments on policy-violating sexual videos in Amharic to understand the characteristics of the policy-violating content. To the best of our knowledge, this is the first study to directly and explicitly examine the experience of low-resourced language speakers with online harm on YouTube.

We describe our results in Sec. 4, and discuss the implications of our findings from both studies in Sec. 5. We hope that our work informs social media platforms like YouTube in forming strategies for better enforcing existing policies to protect these marginalized communities.

2 RELATED WORKS

2.1 Online Experiences and Harm

Global South communities are excluded in much of the design, development, and deployment of technologies [18, 20, 32, 35]. Often, such communities are left with technologies that are dysfunctional within their context or inadequate in addressing their needs. After surveying over 4000 participants from 14 countries, one study [74] found that the perception of online harm is highest in users outside the US. Another study [75] found that Bangladesh Facebook users perceived content moderation systems as methods that propagate historical power relations and perpetuate colonial viewpoints. Previous work [85] also found that Caribbean users had developed a perception of increased vulnerability and more severe harm than their US counterparts.

Previous work has also shown the prevalence of harms associated specifically with sexual content. For instance, past work reveals how searching for benign terms such as "black girls" on Google [63] or Google's Keyword planner [91] returns pornographic content. Further evidence reveals the toxicity in large, Internet-crawled datasets, including the existence of "sexual abuse, rape, and non-consensual explicit images" [25], indicating potential downstream impacts of online toxicity for machine learning model training and evaluation [68].

2.2 Challenges in Content Moderation

Despite efforts by social media platforms, policies sometimes fall short in protecting users. Some online platform users employ content moderation avoidance strategies to circumvent platform policy restrictions [29, 46, 57, 75]. A common example of such techniques is *lexical variation*, which involves modifying linguistic features (e.g phonetic) of constructs (e.g. words) in ways that do not alter their meaning, i.e. that are semantically equivalent [36]. For example, lexical variants could occur via intentional misspellings or use of emojis and special characters [29, 75] (eg. "seggs" instead of "sex"). While such tactics have been used to avoid over-moderation by non-malicious groups [29, 75] they have also been used to spread harmful content online [46].

Studies in multi-lingual Natural Language Processing (NLP) reveal that many of the automated functions of content moderation all fail to perform in the context of African languages, including Amharic – this involves tasks such as sentiment analysis [40, 80], detection of hate speech [27, 56, 65] or social manipulation [43], and even basic text classification tasks [67]. Large Language Model content filters degrade in low-resourced language settings, effectively "jail-breaking" such models to spew content that would otherwise be formally restricted [92]. Although there have been attempts to address this issue through novel benchmarks [58, 59] and datasets [22, 23, 47, 89], the current scope of interventions remains limited and fairly idiosyncratic – for instance, the efforts in the Amharic language so far have only largely focused on deterring politically motivated hate speech on Twitter [22].

The failure of content moderation schemes in Global South communities has, on several occasions, materialized into physical harm, including acts of violence and escalation of war in countries like Mayanmar [17, 52], Nigeria [14, 55] and Ethiopia [22]. Additionally, the impact of content moderation on Global South communities is not just in how moderation schemes fail to protect Global South users from online harm. Several articles shed light on the impact of Big Tech hiring content moderators from Global South countries and how these content moderators then experience harm in the form of low wages, lack of access to mental health support, and lack of protection by law [5, 45, 49, 60, 72, 83].

2.3 YouTube Content Moderation Policies & Practices

On October 27, 2023, in compliance with the Digital Services Act (DSA), which came into effect November 16, 2022, Google released a Transparency Report,in conjunction with supplementary materials [28, 69, 93], on its products classified as a Very Large Online Search Engine (VLOSE) or a Very Large Online Platform (VLOP), including YouTube [41]. The report in particular provides concrete figures for Google's moderation efforts and resources during the period from August 28, 2023 to September 10, 2023. The report states, "…Community Guidelines play an important role in maintaining a positive experience for everyone on our platforms *no matter where they are in the world*" [41]. Notably, these guidelines apply to not just the videos themselves, but "all types of content on our platform, including videos, comments, links, and thumbnails" [93]. YouTube claims that, in 2023, "out of every 10,000 views on YouTube in Q2,

only 9-10 came from violative content", though it is unclear for which languages this evaluation was conducted [41].

The company also explains its approach to detection. 91.3% the content flagged for moderation is "self-detected" – i.e. perceived through pro-active platform features and processes, rather than reported by users, or flagged through some other external means [41][2]. They state that "the overwhelming majority of violative content is detected by automated systems", which is then re-directed to "human reviewers" [41]. These human content moderators "evaluate whether (the content) violates our policies" and "...content assigned for their review may have been posted in several different languages. In some cases and where appropriate, translation tools may be used to assist in the review process and allow us to moderate content 24/7 and at scale" [93]. YouTube reports that 89.2 % of their human moderators operate in English. Of the non-English moderators, only 363 moderators (2.1%) operate in a language other than the more highly resourced Spanish, Portuguese, French, German, or Italian [41][3]. Once evaluated by human reviewers, "we remove the content and use it to train our machines for better coverage in the future" [93].

Unless flagged for "a clear educational, documentary, scientific or artistic (EDSA) purpose" [93], policy-violating content faces a set of articulated consequences – most commonly "Restrictions of the visibility of content" (93.2% of flagged content) [41], such as restricting viewership or disabling its recommendation-enabled reach[4]. Over the reporting period, YouTube noted taking initiative on 6,320 cases related to "Nudity / Sexual " content – significantly more than the number of actions taken on many of the other categories of online harm [41][5]. This indicates that sexual content is amongst the most moderated types of policy violations on YouTube.

To understand YouTube's content moderation from the outside is much more difficult. There are several challenges in conducting audit studies on ever-changing, black-box platforms like YouTube. Prior works have used methods such as sock-puppet studies [24, 73, 77] or browser plug-ins [19, 53] to circumvent these challenges. Other studies use qualitative methods focused on understanding the nuanced beliefs and behaviors of the users and creators who engage with the platforms [88]. In our work, we combine both approaches, collecting platform data as well as conducting interview studies, to gain a wider perspective on the platform and user experiences.

3 METHODS

In this section, we outline the methodological details for both of our studies. Each study went through and passed a formal Institutional Review Board (IRB) process at an accredited U.S. institution and was approved with IRB Protocol ID: 2022-11-15801.

[2]Numbers calculated from Table 3.1.2 on "the number of measures taken on violative content, broken down by service and detection method", from Google Transparency Report [41]

[3]Numbers calculated from Table 3.3.1 which "reflects the human resources evaluating content across the official EU Member State languages, for each service", from Google Transparency Report [41]

[4]Numbers calculated from Table 3.1.3 on "Number of measures taken at Google's own initiative, by service and type of restriction applied", from Google Transparency Report [41]

[5]Numbers taken from Table 3.1.1.h on "own initiative actions taken on YouTube, by type of illegal content or violation of terms and conditions", from Google Transparency Report [41]

3.1 Study 1: Semi-Structured Interviews

Study 1 aims to answer **RQ1** through qualitative interviews discussing the *experiences* of Amharic-speaking YouTube users: how individuals become exposed to online harm, how they understand it, and what strategies, if any, they use to mitigate such harm. Further details can be found in Appendix A.

Participant Recruitment. We recruited participants through social media (Twitter, WhatsApp, Telegram), direct outreach to women's rights advocate groups in Ethiopia, and through the authors' personal networks. We used a screening survey to select participants. Following previous work [88], we exclusively recruited women as they are the population who would be most impacted by exposure to such content. Table 1 presents participants' details.

Consent and Compensation. Participants signed a consent form approved by our IRB. Depending on their location, participants were compensated with a 25 USD gift card or a 1000 ETB mobile card top-up per hour of participation.

Procedure. We conducted semi-structured interview sessions remotely over Zoom, Google Meet, and through end-to-end encrypted calls on Telegram. Sessions lasted from 45-75 minutes. At the beginning of each session, we requested and received verbal consent (in addition to written consent) to record all sessions for subsequent analysis. Appendix A presents our guiding questions. All interviews were conducted by the first author and were in a mix of Amharic and English, depending on participants' preferences.

Data Analysis. We analyzed 10 hours of footage from 15 sessions and 9 pages of written notes. We used inductive thematic analysis [26] to code the data. As the only author who speaks Amharic, the first author served as the interviewer, translator and transcriber as well as conducted direct open coding of each of the recordings with short, descriptive sentences using QualCoder [9]. All authors then discussed the open codes and synthesized higher-level themes, refining codes, and themes iteratively in weekly meetings – resulting in a total of 936 unique open codes, grouped into 26 first-level themes and further grouped into 6 second-level themes (see Appendix E).

3.2 Study 2: Investigating Characteristics of Policy-Violating Sexual Content

In order to characterize the nature of the policy-violating content identified in Study 1, we supplement our findings with an analysis of YouTube platform data to answer **RQ2**. Our language of study, Amharic, is primarily written in the Ge'ez script but sometimes also written in Latin characters online [81]—we conduct our analysis in both. Inspired by past work [70], we minimized personalization by (1) not signing in, (2) clearing all browser data before every query, and (3) running in a private, incognito window. To change locations, we used Virtual Private Network (VPN) services when using the YouTube platform, and use the coordinate parameter to specify the desired latitude and longitude when using the YouTube API. For the US, we collected data while physically being in the country. We collected all data using a Mozilla Firefox browser on a Linux machine during the period Dec 2022–April 2023.

Participant ID	Profession	Languages	Location	Years Using YouTube
P0	Political Science	Amharic, Afaan Oromoo, Tigrinya, English	Ethiopia	8
P1	Software Engineer	Amharic, English	Ethiopia	6
P2	Management	Amharic, English	UAE	10
P3	Psychology	Amharic, Guragigna, Tigrinya, Arabic, English	Ethiopia	9
P4	Medical Student	Amharic, English, French	Ethiopia	11
P5	Communication and Media	Amharic, English	Ethiopia	9
P6	Software Engineering Student	Amharic, English, French	Ethiopia	8
P7	Computer Scientist	Amharic, English	USA	7
P8	Natural Science Student	Amharic, English	Ethiopia	4
P9	Clinical Pharmacy Student	Amharic, Tigrinya, English	Ethiopia	5
P10	Medical Student	Amharic, English	Ethiopia	8
P11	Political Science Student	Amharic, Arabic, English	Ethiopia	9
P12	Computer Scientist	Amharic, Arabic, English	USA	15
P13	Natural Science Student	Amharic, English	Ethiopia	5
P14	Management, Business Owner	Amharic, English	Ethiopia	7

Table 1: Table describing participant demographic for the semi-structured interview study (Study 2). Participants reported *age in the range of 18-30* and had a diverse set of professional backgrounds. Additionally, all participants spoke at least one Ethiopian language, with 3 participants speaking multiple Ethiopian languages. Our participant's minimum reported years of using YouTube was 5 while the maximum was 15. All participants consumed Amharic content on YouTube and 3 participants consumed content in Afan Oromo, Guragignya, or Tigrinya in addition to Amharic. All participants were women.

Gathering Search Queries. We assembled seed search queries from (1) the most common YouTube searches in Ethiopia [38] and (2) common search queries from previous work [64] auditing search engines, which we adapted to the Ethiopian context[6]. We then ran an initial search, where we observed that sexual videos were returned for several of the top benign queries, including the names of famous children's TV shows. Additional benign tags identified on policy-violating content was added to our initial list of queries until we reached theoretical saturation [78]. Fig. 8 presents our final query list [7]. The first author translated each English query into Amharic and rewrote Romanized Amharic words in Ge'ez; yielding a total of 61 queries.

Search results. We used the YouTube API [11] to collect the top 50 search results for each query we shortlisted. We collected the search results for each query in five different locations: Ethiopia, the United States (US), the United Kingdom (UK), the United Arab Emirates (UAE), and Saudi Arabia (SA), informed by common locations for local and diaspora Ethiopians. We collected a total of 9313 videos through YouTube search queries.

Video Recommendations. To understand what kinds of policy-violating sexual videos are diffused by YouTube's recommendation algorithm, we open each policy-violating sexual video identified in the benign search phase, and captured all the recommended videos for each opened video. We conducted our data collection using Tracking Exposed [10]. Similarly to our search data collection scheme, we collected recommendation results in five different locations. We collected 3336 videos in total through YouTube recommendations.

Policy-violating Channels. From the sexual videos returned in our data collection, we used the YouTube API to collect information about the channels that posted those videos. We collected information such as channel name, description, date of creation, location, number of views, number of subscribers, and number of videos. We collected data from 120 YouTube channels.

Comments. From the channels in our data that posted the policy-violating sexual videos, we collected comments for three of the most popular videos from the channel with the most views. In total, we collected 406 comments, including threads, to understand the experience of YouTube users exposed to sexual videos.

Data Labeling and Analysis. For the videos collected from search and recommendation results, the first author opened each of the videos in the collected data and labeled the videos based on their title, thumbnail, audio, and video. For comments, we used inductive thematic analysis [26]. The first author, as the only author who speaks Amharic, conducted line-by-line direct open coding of each of the comments in our collected data. Then, the first and second authors iterated over the codes and synthesized themes in frequent weekly meetings. Our analysis resulted in 163 unique codes which we iteratively grouped into 14 first-level themes which we further grouped into 3 second-level themes. For the channels, the first and second authors met weekly and categorized them based on the descriptions and list of videos. Throughout the labeling process, the

[6]For instance, we replaced 'Black Women' with 'Habesha Women'. "Habesha" is a term used to commonly refer to Semitic language-speaking communities in Ethiopia and Eritrea.

[7]We added a screenshot of the table as we could not add Ge'ez characters with the currently allowed TAPS packages.

first author provided translations for contents that were in Amharic. In total, the labeling and analysis lasted from Dec 2022 - July 2023.

3.3 Limitations

In this study, we do not make claims about the *prevalence* of policy-violating content in low-resourced languages. Rather, in Study 2, we focus on attempting to understand the nature of the content we identify on the platform, in order to ground the characterization of experiences described in Study 1 with real YouTube platform data. We did not compare our results to any other language, as we are not interested in the *degree* to which online harm occurs in low-resourced versus high-resourced languages. Instead, we focus exclusively on the chosen context (ie. of Amharic-speaking YouTube users) as a stand-alone contribution. This allows us to prioritize our primary research goal of understanding what online harm looks like in a non-Western community and how technological failures for low-resourced languages contribute to the speakers' exposure to online harm. Additionally, in Study 1, our participants are women who at the very least are attending undergraduate studies or have completed their Bachelor's degree. Additionally, Amharic is just one of over 80 languages spoken in Ethiopia. As such, we acknowledge our participant's sample is not representative of the realities of all women in the Ethiopian context and is biased towards educated women. Due to these constraints, we acknowledge the limits of our findings toward broader generalizability.

3.4 Positionality

All authors of this study are of African descent, with a primary affiliation at a US-based university. The first author is a native Amharic speaker and also traveled to Ethiopia for part of the duration of the study, which eased technological and geographic barriers to communication, such as timezone differences and choice of communication platforms. We acknowledge our relative position of power and took steps to minimize the effects on our participants. In Appendix B, we detail the steps we took to protect our participants from further harm, and how we employed the HarmCheck framework [33] to reduce perpetuating harm to our participants, ourselves, and readers of our work.

4 FINDINGS

In this section, we provide our findings from the two studies as they relate to our research questions. We summarize all the policy violations we observed from our two studies in Table 2.

4.1 Low-resourced language speakers' experience deteriorates when using YouTube in their languages.

All participants indicated using YouTube for educational and entertainment purposes. "*That is where we find tutorials and lecture videos for our course work*" said P1. Further, some participants indicated using YouTube for news and politics (P0, P4, P11) and religious content (P3, P7, P10, P11). Some participants (P6, P7, P11) said if YouTube no longer existed, they would use different platforms for different content types, and that there is currently no single platform that could satisfy all their needs like YouTube. Several participants (P2, P5, P8, P10, P14) indicated TikTok might come close but criticized it for its short content length.

Our participants (P0, P3, P4, P7, P8, P11, P12, P14) reported that their search experience for Amharic was "*horrible*" and "*requires a lot of scrolling*", "*contains lots of unrelated content*", and "*is full of click-baits and tabloids*". P0 and P3 further state that they have similarly bad experiences when searching in Afaan Oromoo and Guragigna[8], two other languages spoken in Ethiopia. Some participants (P5, P13, P14) reported having a bad search experience in Amharic unless "*you knew the exact title*" and the "*exact spelling the content creators used*". P5 further elaborates that search is fragile in Amharic, with something as small as a missing space offsetting the search experience.

> *Amharic [words] in Latin [letters] have [many] different spellings … sometimes you have to use 3 different [Latin] letters together for one Amharic letter. Usually what I do is use the suggestions on the auto-complete even if it is written in a way I would not normally write.* —P7

P8 and P14 reported that searching for Amharic content in Latin script was worse than searching in Ge'ez while P2 reported having better experience when searching in Latin. All but one participant reported getting results in other languages such as Spanish, Russian, Hindi, Turkish, and other South Asian languages when searching for Amharic, Afaan Oromoo, or Tigrinya content, especially in Ge'ez scripts.

> *When I search in Amharic, I get Amharic and languages I do not know. I am sometimes surprised and ask myself 'I am still on YouTube, right?'* —P1

From our interviews, we found that our participants reported being exposed to sexual videos when searching for "*Amharic Movies*" (P1, P7), "*Habesha romance movies*" (P11), when searching for Amharic music (P10), looking up Ethiopian artists (P5, P12), searching for benign general terms like "*Ethiopian girls*" (P2), searching for literature work (P3), or searching for religious songs:

> *It was in the morning and I was about to pray … I searched for a religious song in Amharic and got sexual content instead.* —P3

Participants (P3, P6) also reported encountering sexual videos as "Up Next" recommendations after watching non-sexual videos such as entertainment talk shows in Amharic. Some participants (P6, P9, P12) also reported getting sexual and violent videos on YouTube Shorts. P12 further elaborates that she saw these videos while "*mindlessly scrolling*". Several participants (P1, P2, P4, P5, P8, P14) reported getting location-based recommendations which are usually related to politics and include hate speech or violent, graphic content. Additionally, participants had to use VPN services when YouTube was blocked in Ethiopia from February 2023 till July 2023 [76]. Some participants (P3, P8, P14) reported getting content in other languages although they did not report seeing a difference in the level or nature of harmful content.

Participants (P0, P1, P4, P7, P9, P11, P12) reported seeing explicit words used in Amharic; "*in English it is censored … in Amharic it is*

[8]Afaan Oromoo is written in Latin characters with the Qubee alphabet while Guragigna is primarily written with the Ge'ez alphabet.

out in the air and there are no guardrails." (P4). Participants (P0, P1) further elaborate that in some cases, content creators make efforts to blur explicit images while they have explicit Amharic writing. Others commented on how content creators use the conservative culture of Ethiopia to justify the release of such types of videos; by framing the videos as shaming women and putting women in their right place:

> *They use phrases like 'she deserves to get a beating' and 'the government should lock her up' to get clout and acceptance from the community for their content. What business does the government have over how an individual dresses? The culture is already conservative and content like this could translate to physical harm for these women.* —P0

4.2 Once the Damage is Done: How low-resourced language speakers cope when policies fail.

Our participants stated that they had previously reported videos with sexually explicit content (P0, P1, P5, P8, P12, P13, P14), religious and ethnic hate speech (P10, P11, P14), and graphic and violent content (P8). Two participants (P2, P10) said they had a successful reporting experience, where they got notifications the video was taken down. Meanwhile, others admitted that they either did not check after they reported (P9, P14), were told the video was not harmful (P6), or did not get feedback (P13, P9, P14). Participants (P0, P5) noted that they "*lost hope after repeated failed attempts*". Some participants (P5, P6, P10, P14) also reported a sense of not feeling prioritized. One participant, whom we will avoid disclosing their participant ID to avoid any risk of de-identification, shared that they used to work for one of the big social media companies as a content moderator and that they had diminished hope and interest in using such platforms or relying on the reporting mechanisms. They further elaborated how online harm in low-resourced languages might have less priority: "*Our content might have small priority because ... it could be automated failures ... or lack of reporting by users.*" Participants also expressed feeling like existing polices do not account for their cultural context. In fact, many (P3, P8, P11, P13, P14) inquired about who gets to define fairness or what is harmful:

> *My country's and another country's definition of right and wrong is not the same. So I think they only take their own way of defining what is right. I do not think what I feel matters.* —P14

In addition to reporting, participants also indicated using the "Do Not Recommend" feature on YouTube (P1, P6, P9, P12) or just skipping the content (P3, P5, P7). "*... even if you hover for a while, it will add it to your 'watched' list and start recommending similar stuff ... I just skip things as fast as I can.*"(P7). One participant told us she used three different Google accounts for different content types:

> *I created separate accounts: one for religious content, one for educational material, and one for entertainment and casual content. I didn't want to be hit with disturbing content when I was watching a religious sermon or looking at a lecture. My [channel] subscriptions are also different for each of the accounts. My search experience is still not ideal; sometimes I get Muslim content... even though I exclusively search and consume Orthodox Christian content on that account ... but that is fine.* —P10

4.3 Policy Violations In Search and Recommendation

CW: Includes discussion of rape. Using the snowball sampling approach described in Section 3.2, we found that 19 out of 31 search queries returned at least 2 and at most 10 sexual videos in the top-50 results for Latin-based queries, with the highest number of sexual videos returned for the query 'Doctor' (see Fig. 9b). For Ge'ez-based queries, we found that 17 out of 31 queries returned at least 1 and at most 12 sexual videos in the top-50 search results, with the most number of sexual videos returned for a name of a children's show, and the second highest number of sexual videos returned for the query 'Habesha Doctor' (see Fig. 9a) The impact of the identified policy violations is also influenced by other factors such as the placement of the video in the list of results and the nature of video thumbnails and titles. In Fig 1, we show examples of how searching for two benign queries results in sexual videos returned as the fifth and fourth search results respectively. Both results contain a neutral image of a woman, accompanied by provocative and explicit Amharic writing on the thumbnail and title.

Next, we assess if policy-violating content is further diffused by YouTube's recommendation system. In the five locations we studied, we found that if a policy-violating sexual video is opened, a proportion of the recommended videos will also include policy-violating sexual content. In Appendix C, we give visual examples: Fig.3 shows how, for a sexual video posted by a verified channel, all the recommendations are sexual videos from the same channel. Fig. 9, shows examples of recommendations for sexual videos opened on YouTube. In some cases, these recommended videos included serious offenses, such as depictions of sexual violence. In one case, a migrant worker records the sexual abuse she experiences by her employer with a hidden camera. In another case, there is a narrative recording of sexual assault on public transportation, accompanied by explicit imagery. Recommendations also included videos of animals mating in parks or in zoos and sexual scenes cut from movies. Table 5 in Appendix D shows the percentage of policy-violating content in the recommendations across the five locations we studied.

4.4 On Strategies Employed by Policy-Violating Channels

In total, there were 131 unique channels identified from our search (n=30) and recommendation (n=110) data analysis that posted videos we labeled as containing policy-violating sexual content. 25 out of 30 channels (83.33%) identified through search exclusively posted Ethiopian content in Amharic while for the 53 of the 110 channels (48.18%) identified through recommendations posted Ethiopian content in Amharic. We observed that from the channels that disclosed their location (n=90), around half of them(n=46) had their locations set to the US(n=35) or the UK(n=11), while none had their location

Policy	YouTube Violation Definition	Violations from Study 1	Violations from Study 2
Spam, deceptive practices, and scam	Misleading Metadata or Thumbnails: Using the title, thumbnails, and description to trick users into believing the content is something it is not	Thumbnails that have nothing to do with the content (P4, P7, P10, P11)	Thumbnails that have nothing to do with the content (Search, Recommendation)
Playlists	Playlists with thumbnails, titles, or descriptions that violate our community guidelines, such as those that are pornographic, or that consist of images that are intended to shock or disgust.	Playlists with sexual narrations (P5)	Full sexual playlists (Search, Recommendation, Channels)
Child Safety	Sexualization of minors: Sexually explicit content featuring minors and content that sexually exploits minors.	Sex tape of a minor; exposure to sexual content while a minor (P8, P2)	Sexual video involving a minor; comment from minor exposed to sexual video (Recommendation, Comments)
Thumbnails	Pornographic imagery. Imagery that depicts unwanted sexualization. Violent imagery that intends to shock or disgust. Graphic or disturbing imagery with blood or gore Vulgar or lewd language. A thumbnail that misleads viewers to think they're about to view something that's not in the video.	Graphic and violent content; sexual imagery; explicit, vulgar language (P0, P1, P2, P3, P6, P7, P8, P9 P13)	violent content, sexual imagery; explicit vulgar language (Search, Recommendation, Comment)
Nudity and Sexual Content	Explicit content meant to be sexually gratifying. Clips extracted from non-pornographic films, shows, or other content in order to isolate sexual content. Groping, kissing, public masturbation, "upskirting", voyeurism, predatory exhibitionism, or any other content that depicts someone in a sexualized manner without their consent. Content that depicts sexual acts, behaviors, or sex toys that's meant for sexual gratification	Sex Tapes, Expose videos, Narrations, Rape, Sexual Acts, Sexual scenes from movies, Sexualization of women on streets, Nude leaks (P0, P1, P2, P3, P4, P5, P6, P7, P8, P9, P11, P10, P12, P13, P14)	Sex Tapes, Expose videos, Narrations, Rape, Sexual Acts, Phone Sex Recordings, Sexual scenes from movies, Nude leaks (Search, Recommendation)
Suicide, self-harm, and eating disorders	Videos showing the lead-up to a suicide, or suicide attempts and suicide rescue footage without sufficient context.	Live suicide video (P7)	None
Vulgar Language	Use of sexually explicit language or narratives. Use of excessive profanity in the content. Use of heavy profanity or sexually suggestive terms in the content's title, thumbnail, or associated metadata. Use of excessive sexual sounds.	Sexually explicit titles in Amharic; Vulgar words in thumbnails; Vulgar words in video content; Narrations of sexual acts (P0, P1, P3, P4, P7, P9, P11, P12)	sexually explicit titles in Amharic; vulgar words in thumbnails; vulgar words in video content; narrations of sexual acts, Explicit Comments (Search, Recommendation, Comments)
Harassment and Cyberbullying	Content featuring non-consensual sex acts, unwanted sexualization or anything that graphically sexualizes or degrades an individual.	Harassment of women on streets; (P0, P1, P2, P6, P8)	Harassing Comments, Reactionary Videos Harassing Women (Recommendation, Comments)
Hate speech	Encourage violence and incite hatred against individuals or groups based on any of the attributes noted above.	Hate speech videos, songs with open call for violence, ethnic hate, religious hate (P0, P4, P3, P10, P14)	Ethnic Slurs (Comments)
Violent or graphic content	Inciting others to commit violent acts against individuals or a defined group of people. Footage or imagery showing bodily fluids, such as blood or vomit, with the intent to shock or disgust viewers. Footage of corpses with massive injuries, such as severed limbs. Graphic content that features animals and intends to shock or disgust. Violent physical sexual assaults (video, still imagery, or audio).	Graphic videos of massacres, attacks; video showing bodily fluid (P0, P1, P3, P7, P8, P11, P13)	Rape caught on Camera (Search)
Misinformation	Misattributed content. Content that may pose a serious risk of egregious harm by falsely claiming that old footage from a past event is from a current event.	False news; Misattributed Content (P5, P6)	Misattributed Content (Search)

Table 2: CW: Discussion of sexual content, rape, self-harm and suicide, hate, abuse, and violence. YouTube policy violations we observed from Study 1 and Study 2. For Study 1, we present the ID of the participant who reported encountering online harm in low-resourced languages when using YouTube in Ethiopian languages. For Study 2, we indicate on which type of data collection and analysis we observed the content that violates the YouTube policy.

set to Ethiopia. The year the channels were created ranged from 2006-2022 and of the 131 channels, 34 were *verified*. Total number of views for the channels ranged from 558 views to 28 billion, while subscriber number ranged from 0 to 36 million. In Table 3, we show examples of policy-violating channels, along with their channel descriptions, and sample video titles to demonstrate the strategies used by the channels to promote their content. Narrowing in on the channels that post sexual videos in Amharic, we found that the content creators used the following content moderation avoidance strategies:

- **Search Engine Optimization (SEO) Manipulation**: Placing hashtags with famous TV shows and celebrity names in the descriptions of their sexual videos.
- **Presented credentials and appeals to authority**: Using channel names like "Dr. [popular Ethiopian name]", and having misleading descriptions that suggested they were providing "health and lifestyle advice."
- **Explicit Amharic Thumbnail Writings**: Posting content with thumbnails with explicit writing in Amharic next to stock images or neutral images of Ethiopian female celebrities.
- **Innocuous Visuals with Explicit Audio**: Posting videos with explicit, sexual Amharic audio while the accompanying visuals were something innocuous (e.g demo of how to use an online website-making service).
- **English Content as Disguise**: Manipulating the description box by putting benign text in English (e.g saying the video is about learning about different types of microbes) but explicitly sexual descriptions in Amharic.
- **Lexical variation**: Mixing phonetically identical Ge'ez and Latin letters when writing sexually explicit words.
- **Cross-Referencing**: Referring to similar channels as hashtags or in their channel descriptions.
- **Sharing Contact Information**: Posting phone numbers, usernames for other platforms, and email addresses.

4.5 Disparate Experiences in the Comment Section

Harmful content was not limited to the videos; we observed comments that spread hate, used slurs, used vulgar and sexually explicit language, and advocated for violence. In one thread, we observed an exchange where users over-sexualized a specific ethnicity and used ethnic slurs in the exchanges. There were also several instances of users employing the credentials presented by the content creators. For instance, several referred to the content creator as a "doctor", although there was no clear evidence of such a qualification beyond the channel's own description. Even those who disagreed with the content would say "*Doctors should not act this way.*" or "*You should focus on giving medical advice.*" Further, there was one comment that gave details about a real medical condition and asked for medical advice. We also found comments that support and further the demeaning of women, with some including borderline rape content, indicating violent acts, and disregard for consent. In Table 4, we give some samples of the comments we analyzed along with the themes and open codes associated with them.

We observed several users indiscriminately share personal information such as age, marital status, and location. One commenter claimed to be 16 years old and expressed how the posted video impacted the way they think about intimacy. Of the comments that disclosed location information (7.06%), about half (51.6%) indicated being located in Middle Eastern countries, and some had indicators in their usernames and comments that they were migrant domestic workers (e.g using phrases migrant domestic workers have used to self-identify on online platforms). Some users would expose their location, and request others to pray for them to make it back to their country. In terms of marital status, we observed in the data that single people or people soon to be married would engage with the videos saying they "*will try it when I get married.*" One comment had a phone number with a mix of words and numerals (e.g five5five0one0five [9]) and indicated "*women can contact me on [PLATFORM] if you are located in [CITY] and want abortion pills.*".

Some comments (5.17%) explicitly opposed the content. Critical commenters noted, "Such content is against our cultural beliefs", and "This is religiously unacceptable", expressing fear that "There are young kids exposed to this type of content". One comment opposing a video that had advice for violent sexual acts indicated how the suggestions are damaging to women's health:

> *What type of demonized lesson are you giving to men? You have never experienced what it is like to be a woman and yet you are talking as if you are an expert... You are going to ruin people's marriages with this content.* —Comment from a YouTube user, translated from Amharic

5 DISCUSSION

Our findings indicate that Amharic-speaking users report being regularly exposed to policy-violating sexual content on the platform (Sec. 4.1 and Sec. 4.3). This could be due to malicious content creators relying on the under-performance of content moderation in low-resource languages (Sec. 4.4), or a set of other related factors. Regardless, our work suggests that while low-resourced language speakers find the platform genuinely enriching, they feel as though existing policies are being inadequately enforced. As a result, these users develop their own mitigation strategies beyond platform reporting in order to minimize the personal impact of online harms (Sec. 4.2).

Our work shows how, even when policies exist, policy-violating content in low-resourced languages harms users due to (1) lack of proper enforcement by platforms and (2) due to under-performance of language technologies for low-resourced settings (Sec. 2.2). While recent efforts by community-led research groups [15, 34, 50, 54] have contributed works to move the NLP community in a less Anglo-centric [1] direction, we show how the exclusion of low-resourced languages in mainstream NLP, along with exclusion in policy enforcement by platforms, has real-world, immediate downstream harm.

Recommendations for Social Media Platforms with A Global Reach. Our results indicate weak moderation efforts on these platforms in the low-resourced language context. Failure to detect harmful

[9] The example is a fictional phone number. https://en.wikipedia.org/wiki/555_(telephone_number)

Channel Info	Channel Description	Video Title	Video Description	Strategies used
Name: Dr.[NAME_1] **Verified**: Yes **Location**: United Kingdom **Subscribers**: 169,000 **Video Type**: Narrations of sexual acts	Dr [NAME_1] is Similar Channel to Dr [NAME_3] and Dr [NAME_2]. You can freely contact me with https://t.me/[TELEGRAM USERNAME]	Dr [NAME_1] How to make the lady [EXPLICIT WORD] repeatedly in one go. dr [NAME_1] insight habesha dr [NAME_4] #Ethiopia	Dr [NAME_5] - Dr [NAME_7] - Dr [NAME_2] -[NAME] TUBE - DR [NAME_3]	Cross-Referencing; Providing contact information; Explicit Amharic writing in Thumbnail; Presented credentials and appeals to authority; SEO Manipulation
Name: Dr.[NAME_6] **Verified**: No **Location**: UK **Subscribers**: 52,400 **Video Type**: Narrations of sexual acts	Thank you for subscribe! My name is dr [NAME_6], in this channel you will be getting information about health care and skin routine tips and you [can] ask any [questions] about health care! If you [want] consulting you can call +[PHONE NUMBER] https://www.youtube.com /channel/ [CHANNEL ID]	Dr.[NAME_6] If you want to [EXPLICIT WORD] her, touch her at these places[.] You will be amazed!	Thank you for subscribing! things you need to know. for any questions related to [EXPLICIT WORD] you will find answers. All you need to do is subscribe. Become a member of this channel by subscribing. Disclaimer: This Channel Does Not Promote or encourage Any Illegal Activities, all contents provided by this Channel is meant for EDUCATIONAL PURPOSE Only. - Telegram Address - @[TELEGRAM USERNAME]. - Music from non-copyright music store … #dr[NAME_1] #dr[NAME_3] #ethiopiannews	Presented credentials and appeals to authority; English as a Disguise; Cross-Referencing; Providing contact information; Innocuous Visuals with Explicit Audio; SEO Manipulation; Explicit Amharic writing in Thumbnail;

Table 3: CW: contains discussion of sexual acts. Examples of channels that post policy-violating sexual content in Amharic. Basic information about the channel like name and number of subscribers is provided. We also list out which of the common strategies we identified the channel employs. All text translated from Amharic is highlighted in color. For instance with channel Dr.[NAME_6] We see that non-highlighted text (originally in English) mentions educational use and shows compliance with copyright while the highlighted text (Amharic) has explicit words. It also cross-references the first channel Dr.[NAME_1] in the video description to feign legitimacy.

content in low-resourced languages (Sec. 4.3) allows malicious content creators to surpass guardrails put in place by the platforms and leaves their users vulnerable to harmful content (Sec. 4.4). Additionally, search and recommendation features of YouTube further the reach of policy-violating content on the platform (Sec. 4.3). As Fig. 3 shows, a channel that posts policy-violating content in Amharic is (1) verified by the platform and (2) recommendations to videos from the channel are exclusively other policy-violating videos from the same channel. Furthermore, users feel a sense of not being prioritized and their cultural context not being accounted for in content moderation (Sec. 4.1). Hence, (1) even when policies exist, they are not properly enforced in low-resourced language contexts and (2) users feel moderation strategies and current policies are not reflective of the diverse cultural contexts and realities of global users. Platforms should be aware of and actively consider their limitations within the low-resourced language setting, and potentially explore new directions for culturally aware and context-specific moderation strategies. This echoes previous findings [74] advocating for the inclusion of diverse perspectives in content moderation. While warning against the *over-penalization* of marginalized communities [44], we argue that the effectiveness of online platform policy enforcement cannot be defined independently of cultural appropriateness or the cautious identification of policy violators.

Recommendations for NGOs focused on Protecting Marginalized Groups. By exclusively studying the experiences of women who speak low-resourced languages, our findings show how these marginalized groups can be exposed to and genuinely harmed by policy-violating sexual content on YouTube. We observed that vulnerable groups such as migrant domestic workers, who may not have their own access to healthcare services [2], might instead engage with videos by malicious content creators claiming to be medical doctors (Sec. 4.4). This puts them at risk of, for example, exposing their PII and medical information in the comments (Sec. 4.5). Our findings can help support NGOs working with marginalized communities (ie. women, children, migrant workers) in providing training regimes or awareness campaigns for effective online navigation tactics.

Recommendations for Government Bodies Hoping to Protect their Citizens. Content moderation failures have exposed citizens of Global South countries to physical and psychological harm–both as users of the platforms and employees for content moderation (Sec. 2.2). Efforts by individual governments [66] and collective agencies like the AU [3, 7, 8] should emphasize the culturally and contextually appropriate protection of their citizens from online harms. Government bodies could, for instance, set up oversight agencies that require platforms to demonstrate and disclose how

Second-Level Theme	First-Level Theme	Example Open-Code	Example Comment
Interactions among Commenters (N = 187, C = 68)	Peer judgement	Using religion to scold user engagement	you are an impostor. You really listen to the Quran?
	Sharing PII and personal details	Disclosing location	I will [EXPLICIT WORD] you. If you are in [CITY] let me give you my address.
	Vulgarity	Threats of sexual violence	I want to [EXPLICIT WORD] you. I will coercively tie your hands and legs tightly and [EXPLICIT WORD]
Direct Engagement with Content (N = 187, C = 65)	Emotive expression	Discussing arousal due to content exposure	Wow. I am like you too. But now I am [EXPLICIT WORD]
	Advice giving	Sharing personal experiences	My husband really likes it when I touch his [EXPLICIT WORD]
	Expressing encouraging, positive sentiments	Expressing gratitude	Thank you, doctor, for the wisdom you share with us
	Engaging in medical discourse	Sharing health information	below my belly button, my uterus.
	Expressing positionality	Explaining personal context or mode of engagement with the content	I am now married. I am listening to you attentively for my husband's sake.
	Asking for more	Asking the content creator to add visuals	Why don't you show us with video?
Opinion about Content (N = 75, C = 30)	Expressing concerns	Arguing that content is inappropriate	Please use words that are acceptable in Ethiopian culture. Please don't ever use vulgar words. Seems like you are forgetting this is social media.
	Questioning channel legitimacy	Questioning the creator's medical knowledge and credentials	are you really a doctor? first, treat yourself. There are plenty of medical topics that need discussion. You are just doing this to get more subscribers.
	Expressing approval	Defending content creators from critique	:laughing: this is something God created. Where is the inappropriateness?
	Fear of societal consequences	Concerns for impact to younger generation	I'll report your channel for inappropriate sexual content, there are young kids exposed to your filth

Table 4: CW: contains discussion of sexual violence. Analysis of Comments on Policy-Violating Sexual Content. Here, we present the number of codes assigned to each theme (C) and the number of comments coded with each theme (N). In total, we analyzed 406 comments – 371 were legible. Legible comments were open-coded with a total of 163 unique codes, grouped into 14 first-level themes, and 3 second-level themes. Examples codes are provided in the table. Note that some comments have multiple codes.

their policies account for the languages of the countries in which they operate.

6 CONCLUSION

Our paper investigates the experience of Amharic-speaking users with policy-violating sexual content on YouTube. Our findings shed light on a small number of individual experiences within one of many low-resourced language speaking communities – in fact, Amharic is just one of the many low-resourced languages spoken in Ethiopia alone. More work is needed to further investigate the role of language in characterizing online experiences in other settings – i.e. multiple geographic regions, more languages, more dialects, involving more extensive platform data collection on a wider array of online harms. While conclusions may emerge that are common to low-resourced languages in general, the intersections of culture, language, and identity unique to each community or even each individual mean that we should be hesitant to generalize findings too quickly, and should instead prioritize in-depth, particular research on specific groups and individuals over broader theories. Overall, we hope our study will help inform policy-making, digital literacy research, and technological design on online platforms to properly protect and serve users operating in low-resourced language contexts.

ACKNOWLEDGMENTS

Hellina Hailu Nigatuis a SIGHPC Computational and Data Science Fellow. Inioluwa Deborah Raji is a Senior Fellow at Mozilla. We would like to thank members and friends of PLAIT lab for their feedback on this work. Additionally, we thank Daricia Wilkerson and Niloufar Salahi for their consistent feedback and guidance on this work. We thank Abeba Birhane for reviewing and giving feedback on versions of the manuscript. We thank Liza Gak and Sijia Xiao for the early conversations regarding this work. We also want to thank our colleagues from Addis Powerhouse. Finally, to our participants, we say Enamesegnalen.

REFERENCES

[1] 2019. The #BenderRule: On Naming the Languages We Study and Why It Matters. https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/

[2] 2019. Lebanon: 'Their house is my prison': Exploitation of migrant domestic workers in Lebanon. https://www.amnesty.org/en/documents/mde18/0022/2019/en/

[3] 2022. The African Union Artificial Intelligence Continental Strategy For Africa | AUDA-NEPAD. https://www.nepad.org/news/african-union-artificial-intelligence-continental-strategy-africa

[4] 2022. Essential YouTube Statistics. https://datareportal.com/essential-youtube-stats

[5] 2022. Inside Facebook's African Sweatshop. https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/

[6] 2022. World-first online safety laws introduced in Parliament. *GOV* (March 2022). https://www.gov.uk/government/news/world-first-online-safety-laws-introduced-in-parliament

[7] 2023. Artificial Intelligence is at the core of discussions in Rwanda as the AU High-Level Panel on Emerging Technologies convenes experts to draft the AU-AI Continental Strategy | AUDA-NEPAD. https://www.nepad.org/news/artificial-intelligence-core-of-discussions-rwanda-au-high-level-panel-emerging

[8] 2023. Pioneering Africa's AI Future: Convening of African AI Experts to Finalise the AU-AI Continental Strategy | AUDA-NEPAD. https://www.nepad.org/news/pioneering-africas-ai-future-convening-of-african-ai-experts-finalise-au-ai-continental

[9] 2023. QualCoder. https://qualcoder.wordpress.com/

[10] 2023. Tracking Exposed. https://tracking.exposed/

[11] 2023. YouTube Data API. https://developers.google.com/youtube/v3

[12] 2024. Principles and background. https://www.esafety.gov.au/industry/safety-by-design/principles-and-background

[13] 2024. YouTube Community Guidelines enforcement – Google Transparency Report. https://transparencyreport.google.com/youtube-policy/removals [Online; accessed 20. Jan. 2024].

[14] Clement Ola Adekoya. 2021. Information and Misinformation during the# End-SARS Protest in Nigeria: An Assessment of the Role of Social Media. *COVENANT JOURNAL OF LIBRARY AND INFORMATION SCIENCE* (2021).

[15] David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 3053–3070. https://doi.org/10.18653/v1/2022.naacl-main.223

[16] Gabe Adugna. 2022. Research: Language Learning - Amharic: Home. https://library.bu.edu/amharic/Home

[17] Emmanuel Akinwotu. 2021. Facebook's role in Myanmar and Ethiopia under new scrutiny. *the Guardian* (Oct. 2021). https://www.theguardian.com/technology/2021/oct/07/facebooks-role-in-myanmar-and-ethiopia-under-new-scrutiny

[18] Syed Mustafa Ali. 2016. A brief introduction to decolonial computing. *XRDS: Crossroads, The ACM Magazine for Students* 22, 4 (June 2016), 16–21. https://doi.org/10.1145/2930886

[19] Nina Altmaier, Davide Beraldo, Maria Castaldo, Daniel Jurg, Salvatore Romano, Matteo Renoldi, Tatiana Smirnova, Natacha Seweryn, and Luukas Veivo. 2020. YouTube Tracking Exposed: Investigating Brexit Polarization. https://youtube.tracking.exposed/trexit/

[20] Ahmed Ansari. 2019. Decolonizing design through the perspectives of cosmological others: Arguing for an ontological turn in design research and practice. *XRDS: Crossroads, The ACM Magazine for Students* 26, 2 (Nov. 2019), 16–19. https://doi.org/10.1145/3368048

[21] Sara Atske. 2019. 1. Popular YouTube channels produced a vast amount of content, much of it in languages other than English. https://www.pewresearch.org/internet/2019/07/25/popular-youtube-channels-produced-a-vast-amount-of-content-much-of-it-in-languages-other-than-english/

[22] Abinew Ali Ayele, Skadi Dinter, Tadesse Destaw Belay, Tesfa Tegegne Asfaw, Seid Muhie Yimam, and Chris Biemann. 2022. The 5js in ethiopia: Amharic hate speech data annotation using toloka crowdsourcing platform. In *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*. IEEE, 114–120.

[23] Abinew Ali Ayele, Seid Muhie Yimam, Tadesse Destaw Belay, Tesfa Asfaw, and Chris Biemann. 2023. Exploring Amharic Hate Speech Data Collection and Classification Approaches. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*. 49–59.

[24] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 74:1–74:34. https://doi.org/10.1145/3449148

[25] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. http://arxiv.org/abs/2110.01963 arXiv:2110.01963 [cs].

[26] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa.

[27] Galo Castillo-López, Arij Riabi, and Djamé Seddah. 2023. Analyzing Zero-Shot transfer Scenarios across Spanish variants for Hate Speech Detection. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*. 1–13.

[28] Transparency Report Help Center. 2023. *YouTube Community Guidelines Enforcement FAQs*. Technical Report. Google. https://support.google.com/transparencyreport/answer/9209072#zippy=%2Chow-is-violative-view-rate-vvr-calculated

[29] 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 the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1201–1213. https://doi.org/10.1145/2818048.2819963

[30] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3175–3187. https://doi.org/10.1145/3025453.3026018

[31] Jim Cummins. 2017. Teaching Minoritized Students: Are Additive Approaches Legitimate? *Harvard Educational Review* 87, 3 (Sept. 2017), 404–425. https://doi.org/10.17763/1943-5045-87.3.404

[32] Dipto Das. 2023. Decolonization through Technology and Decolonization of Technology. In *Companion Proceedings of the 2023 ACM International Conference on Supporting Group Work (GROUP '23)*. Association for Computing Machinery, New York, NY, USA, 51–53. https://doi.org/10.1145/3565967.3571754

[33] Leon Derczynski, Hannah Rose Kirk, Abeba Birhane, and Bertie Vidgen. 2022. Handling and Presenting Harmful Text. *arXiv:2204.14256 [cs]* (April 2022). http://arxiv.org/abs/2204.14256 arXiv: 2204.14256.

[34] Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Chinenye Emezue. 2022. AfroLM: A Self-Active Learning-based Multilingual Pretrained Language Model for 23 African Languages. http://arxiv.org/abs/2211.03263 arXiv:2211.03263 [cs].

[35] Paul Dourish and Scott D. Mainwaring. 2012. Ubicomp's colonial impulse. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp '12)*. Association for Computing Machinery, New York, NY, USA, 133–142. https://doi.org/10.1145/2370216.2370238

[36] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric Xing. 2010. A Latent Variable Model for Geographic Lexical Variation. *ACL Anthology* (Oct. 2010), 1277–1287. https://aclanthology.org/D10-1124

[37] Shawna Ferris and Danielle Allard. 2016. Tagging for activist ends and strategic ephemerality: creating the Sex Work Database as an activist digital archive. *Feminist Media Studies* 16, 2 (March 2016), 189–204. https://doi.org/10.1080/14680777.2015.1118396 Publisher: Routledge _eprint: https://doi.org/10.1080/14680777.2015.1118396.

[38] Saifaddin Galal. 2023. Queries with the highest volume of YouTube search activity in Ethiopia in 2022. https://www.statista.com/statistics/1307177/most-popular-youtube-searches-in-ethiopia/

[39] Aina M. Gassó, Katrin Mueller-Johnson, and Esperanza L. Gómez-Durán. 2021. Victimization as a Result of Non-Consensual Dissemination of Sexting and Psychopathology Correlates: An Exploratory Analysis. *International Journal of Environmental Research and Public Health* 18, 12 (June 2021), 6564. https://doi.org/10.3390/ijerph18126564

[40] Seraphina Goldfarb-Tarrant, Björn Ross, and Adam Lopez. 2023. Cross-lingual Transfer Can Worsen Bias in Sentiment Analysis. *arXiv preprint arXiv:2305.12709* (2023).

[41] Google. 2023. *U Digital Services Act (EU DSA) Biannual VLOSE/VLOP Transparency Report*. Technical Report. Google. https://storage.googleapis.com/transparencyreport/report-downloads/pdf-report-27_2023-8-28_2023-9-10_en_v1.pdf

[42] Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal Neural Machine Translation for Extremely Low Resource Languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 344–354. https://doi.org/10.18653/v1/N18-1032

[43] Samar Haider, Luca Luceri, Ashok Deb, Adam Badawy, Nanyun Peng, and Emilio Ferrara. 2023. Detecting Social Media Manipulation in Low-Resource Languages. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 1358–1364.

https://doi.org/10.1145/3543873.3587615

[44] Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–35. https://doi.org/10.1145/3479610

[45] Karen Hao and Andrea Paola Hernández. 2022. How the AI industry profits from catastrophe. *MIT Technology Review* (July 2022). https://www.technologyreview.com/2022/04/20/1050392/ai-industry-appen-scale-data-labels

[46] Álvaro Huertas-García, Alejandro Martín, Javier Huertas-Tato, and David Camacho. 2023. Countering malicious content moderation evasion in online social networks: Simulation and detection of word camouflage. *Applied Soft Computing* 145 (Sept. 2023), 110552. https://doi.org/10.1016/j.asoc.2023.110552

[47] Abhik Jana, Gopalakrishnan Venkatesh, Seid Muhie Yimam, and Chris Biemann. 2022. Hypernymy Detection for Low-resource Languages: A Study for Hindi, Bengali, and Amharic. *ACM Transactions on Asian and Low-Resource Language Information Processing* 21, 4 (March 2022), 67:1–67:21. https://doi.org/10.1145/3490389

[48] Daphne Keller. 2022. Lawful but awful? Control over legal speech by platforms, governments, and internet users. *U. Chi. L. Rev. Online* (2022), 1.

[49] Kate Linebaugh. 2023. The Hidden Workforce That Helped Filter Violence and Abuse Out of ChatGPT - The Journal. - WSJ Podcasts. https://www.wsj.com/podcasts/the-journal/the-hidden-workforce-that-helped-filter-violence-and-abuse-out-of-chatgpt/ffc2427f-bdd8-47b7-9a4b-27e7267cf413

[50] Gabriela Litre, Fabrice Hirsch, Patrick Caron, Alexander Andrason, Nathalie Bonnardel, Valerie Fointiat, Wilhelmina Onyothi Nekoto, Jade Abbott, Cristiana Dobre, Juliana Dalboni, Agnès Steuckardt, Giancarlo Luxardo, and Hervé Bohbot. 2022. Participatory Detection of Language Barriers towards Multilingual Sustainability(ies) in Africa. *Sustainability* 14, 13 (Jan. 2022), 8133. https://doi.org/10.3390/su14138133 Number: 13 Publisher: Multidisciplinary Digital Publishing Institute.

[51] Sonia Livingstone and Peter K. Smith. 2014. Annual Research Review: Harms experienced by child users of online and mobile technologies: the nature, prevalence and management of sexual and aggressive risks in the digital age. *Journal of Child Psychology and Psychiatry* 55, 6 (2014), 635–654. https://doi.org/10.1111/jcpp.12197 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcpp.12197.

[52] Eliza Mackintosh. 2021. Facebook knew it was being used to incite violence in Ethiopia. It did little to stop the spread, documents show | CNN Business. *CNN* (Oct. 2021). https://www.cnn.com/2021/10/25/business/ethiopia-violence-facebook-papers-cmd-intl/index.html

[53] Jesse McCrosky, Brandi Geurkink, Kevin Zawacki, Anna Jay, Carys Afoko, Maximilian Gahntz, and Owen Bennett. 2021. YouTube Regrets. https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf

[54] Josh Meyer, David Ifeoluwa Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack Julian Weber, Salomon Kabongo, Elizabeth Salesky, Iroro Orife, Colin Leong, Perez Ogayo, Chris Emezue, Jonathan Mukiibi, Salomey Osei, Apelete Agbolo, Victor Akinode, Bernard Opoku, Samuel Olanrewaju, Jesujoba Alabi, and Shamsuddeen Muhammad. 2022. BibleTTS: a large, high-fidelity, multilingual, and uniquely African speech corpus. http://arxiv.org/abs/2207.03546 arXiv:2207.03546 [cs, eess].

[55] T Mhaka. 2020. How social media regulations are silencing dissent in Africa. Aljazeera.

[56] Syrielle Montariol, Arij Riabi, and Djamé Seddah. 2022. Multilingual Auxiliary Tasks Training: Bridging the Gap between Languages for Zero-Shot Transfer of Hate Speech Detection Models. *arXiv preprint arXiv:2210.13029* (2022).

[57] Rachel E. Moran, Izzi Grasso, and Kolina Koltai. 2022. Folk Theories of Avoiding Content Moderation: How Vaccine-Opposed Influencers Amplify Vaccine Opposition on Instagram. *Social Media + Society* 8, 4 (Oct. 2022), 20563051221144252. https://doi.org/10.1177/20563051221144252 Publisher: SAGE Publications Ltd.

[58] Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma OUSIDHOUM, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Said Ahmad, Meriem Beloucif, Saif M Mohammad, Oumaima Hourrane, et al. 2023. AfriSenti: A Benchmark Twitter Sentiment Analysis Dataset for African Languages. In *4th Workshop on African Natural Language Processing*.

[59] Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, Abinew Ayele, Saif M Mohammad, and Meriem Beloucif. 2023. SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval). *arXiv preprint arXiv:2304.06845* (2023).

[60] Evelyne Musambi and Cara Anna. 2023. Facebook content moderators in Kenya call the work 'torture.' Their lawsuit may ripple worldwide. *AP News* (June 2023). https://apnews.com/article/kenya-facebook-content-moderation-lawsuit-8215445b191fce9df4ebe35183d8b322

[61] Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan Van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 2144–2160. https://doi.org/10.18653/v1/2020.findings-emnlp.195

[62] Shuo Niu, Keegan Veazey, Phoenix Pagan, and Abhisan Ghimire. 2022. Understanding Hate Group Videos on YouTube. In *CSCW'22 Companion: Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*. Association for Computing Machinery, New York, NY, USA, 32–36. https://doi.org/10.1145/3500868.3559465

[63] Safiya Umoja Noble. 2018. Algorithms of oppression. In *Algorithms of oppression*. New York university press.

[64] Safiya Umoja Noble. 2018. Algorithms of Oppression: How Search Engines Reinforce Racism. https://nyupress.org/9781479837243/algorithms-of-oppression

[65] Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 907–914.

[66] Oluyemi Ogunseyin. 2023. FG seeks top Nigerian researchers to co-create National AI strategy. *Guardian Nigeria News - Nigeria and World News* (Aug. 2023). https://guardian.ng/news/fg-seeks-top-nigerian-researchers-to-co-create-national-ai-strategy

[67] Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David I Adelani. 2023. How good are Large Language Models on African Languages? *arXiv preprint arXiv:2311.07978* (2023).

[68] Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the Everything in the Whole Wide World Benchmark. http://arxiv.org/abs/2111.15366 arXiv:2111.15366 [cs].

[69] Google Transparency Report. 2023. *Views - Violative View Rate*. Technical Report. Google. https://transparencyreport.google.com/youtube-policy/views?hl=en

[70] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 131–141. https://doi.org/10.1145/3351095.3372879

[71] Sarah T. Roberts. 2021. Behind the Screen. https://yalebooks.yale.edu/9780300261479/behind-the-screen

[72] Niamh Rowe. 2023. 'It's destroyed me completely': Kenyan moderators decry toll of training of AI models. *The Guardian* (Aug. 2023). https://www.theguardian.com/technology/2023/aug/02/ai-chatbot-training-human-toll-content-moderator-meta-openai

[73] Christian Sandvig, Kevin Hamilton, K. Karahalios, and Cédric Langbort. 2014. Auditing Algorithms : Research Methods for Detecting Discrimination on Internet Platforms. https://www.semanticscholar.org/paper/Auditing-Algorithms-%3A-Research-Methods-for-on-Sandvig-Hamilton/b7227cbd34766655dea10d0437ab10df3a127396

[74] Sarita Schoenebeck, Amna Batool, Giang Do, Sylvia Darling, Gabriel Grill, Daricia Wilkinson, Mehtab Khan, Kentaro Toyama, and Louise Ashwell. 2023. Online Harassment in Majority Contexts: Examining Harms and Remedies across Countries. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–16. https://doi.org/10.1145/3544548.3581020

[75] Farhana Shahid and Aditya Vashistha. 2023. Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony?. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–18. https://doi.org/10.1145/3544548.3581538

[76] Internet Society. 2020. Global Internet Shutdowns. (2020). https://pulse.internetsociety.org/shutdowns?search=ethiopia#events

[77] Ivan Srba, Robert Moro, Matus Tomlein, Branislav Pecher, Jakub Simko, Elena Stefancova, Michal Kompan, Andrea Hrckova, Juraj Podrouzek, Adrian Gavornik, and Maria Bielikova. 2022. Auditing YouTube's Recommendation Algorithm for Misinformation Filter Bubbles. *ACM Transactions on Recommender Systems* (Oct. 2022), 3568392. https://doi.org/10.1145/3568392 arXiv:2210.10085 [cs].

[78] Barney Glaser Strauss, Anselm. 2017. *Discovery of Grounded Theory: Strategies for Qualitative Research*. Routledge, New York. https://doi.org/10.4324/9780203793206

[79] Angelika Strohmayer, Jenn Clamen, and Mary Laing. 2019. Technologies for Social Justice: Lessons from Sex Workers on the Front Lines. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–14. https://doi.org/10.1145/3290605.3300882

[80] Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, et al. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*. 26–41.

[81] Tsegamlak Terefe and Dereje Hailemariam. 2017. Entropy estimation and entropy-based encoding of written Amharic language for efficient transmission in telecom networks. In *2017 IEEE AFRICON*. 238–244. https://doi.org/10.1109/AFRCON.2017.8095488 ISSN: 2153-0033.

[82] Kurt Thomas, Patrick Gage Kelley, Sunny Consolvo, Patrawat Samermit, and Elie Bursztein. 2022. "It's common and a part of being a content creator": Understanding How Creators Experience and Cope with Hate and Harassment Online. In *CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3491102.3501879

[83] Zoe Thomas. 2023. Big Tech Relies on Outsourcing. Lawsuits in Africa Could Upend That. - Tech News Briefing - WSJ Podcasts. https://www.wsj.com/podcasts/tech-news-briefing/big-tech-relies-on-outsourcing-lawsuits-in-africa-could-upend-that/aea41e18-85a3-4f67-9a51-7d8b378f4fb8

[84] Daricia Wilkinson and Bart Knijnenburg. 2022. Many Islands, Many Problems: An Empirical Examination of Online Safety Behaviors in the Caribbean. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–25.

[85] Daricia Wilkinson and Bart Knijnenburg. 2022. Many Islands, Many Problems: An Empirical Examination of Online Safety Behaviors in the Caribbean. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–25. https://doi.org/10.1145/3491102.3517643

[86] Erik Wingrove-Haugland and Jillian McLeod. 2022. Not "Minority" but "Minoritized". *Teaching Ethics* (Jan. 2022). https://doi.org/10.5840/tej20221799

[87] John Woodhouse. 2022. Regulating online harms. (2022). https://researchbriefings.files.parliament.uk/documents/CBP-8743/CBP-8743.pdf

[88] Eva Yiwei Wu, Emily Pedersen, and Niloufar Salehi. 2019. Agent, Gatekeeper, Drug Dealer: How Content Creators Craft Algorithmic Personas. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–27. https://doi.org/10.1145/3359321

[89] Seid Muhie Yimam, Abinew Ali Ayele, and Chris Biemann. 2019. Analysis of the Ethiopic Twitter dataset for abusive speech in Amharic. *arXiv preprint arXiv:1912.04419* (2019).

[90] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian Davison, April Edwards, and Lynne Edwards. 2009. Detection of harassment on Web 2.0. (Jan. 2009).

[91] Leon Yin and Aaron Sankin. 2020. Google Ad Portal Equated "Black Girls" with Porn – The Markup. https://themarkup.org/google-the-giant/2020/07/23/google-advertising-keywords-black-girls Section: Google the Giant.

[92] Zheng-Xin Yong, Cristina Menghini, and Stephen H Bach. 2023. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446* (2023).

[93] Youtube. 2023. *How does YouTube manage harmful content?* Technical Report. Youtube. https://www.youtube.com/intl/ALL_ca/howyoutubeworks/our-commitments/managing-harmful-content/

[94] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 1568–1575. https://doi.org/10.18653/v1/D16-1163

A GUIDING INTERVIEW QUESTIONS

In this section, we present the guiding questions we used during our interviews for Study 1. Since it was a semi-structured interview, we used those questions as a guide and used participants' responses to probe further into a particular theme. We arrived at this set of questions after continuous feedback from colleagues, specifically on making the questions open-ended to avoid leading interviewees to an answer.

Demographic Questions

- What is your age range?
- What languages do you speak?
- Are you a student? Professional? What field?

About use of YouTube

- What do you use YouTube for?
- What kind of content do you consume on YouTube?
- What features of YouTube do you use?
- Do you have your own channel?
- Do you post on your own channel?
- Do you comment on videos?
- Do you like videos?
- Do you subscribe to channels? What factors do you consider when you decide to subscribe?
- What kinds of channels do you subscribe to?

YouTube as a platform

- In your opinion, what is a good substitute for YouTube?
- What kinds of trends do you observe over the years as you use YouTube?
- How would you describe your YouTube homepage? Do you think everyone's home page is the same?
- How do you think YouTube works?

Reporting

- What type of content have you reported?
- What do you think is YouTube's policy for reporting?
- How do you go about reporting?
- Why have you not reported yet?

Language related questions

- In what language do you use YouTube? Why?
- In what language do you comment on YouTube? Why?
- In what language do you consume YouTube content? Why?

About Harmful content

- What kinds of harmful content have you experienced on YouTube?
- How would you describe the way you reach these types of content? Through search, recommendation?
- What does recommendation mean?
- How do you think YouTube recommendations work?
- Have you ever reported videos on YouTube? How?
- Have you seen a difference in the type of harmful content you get depending on the language you are using?

B MINIMIZING HARM

We iteratively shared drafts of the paper with participants who were concerned with de-identification and refrained from sharing details about participants per their comfort level. While this limited the demographic information we can share, our participants' safety and concern are our top priority. Additionally, we minimized potential biases in our study by clarifying to participants they can refuse to answer any question and only delve into details depending on their comfort level. One participant, after stating she had been exposed to sexual videos on YouTube and giving a general description, declined to go into specific detail about the types of sexual videos she encountered. In this case, we instead focused on how she was exposed to such videos and what strategies she used to mitigate harm. We make it clear that declining to answer any question will not affect their compensation in our screening form, consent form, and during the interview sessions. We also acknowledge the privilege we had in stepping away from the data when the mental load was too much.

B.1 Documentation of harmful text presentation the HarmCheck framework.

In this section, we use HarmCheck[33] as a framework to scrutinize how we presented the harmful content in our research.

Risk of harm protocol There is a risk of harm to data subjects whose images and names have been used in the distribution of sexual content and to subjects whose identities have been described using slurs in the data. There is a risk of harm to the researchers who have to label and analyze the data. The study included frequent check-ins, 10-20 minute conversations over the impact of the data labeling process on the researcher's mental health, and support and understanding for taking breaks from the project. There is a risk of exposing harm to those sharing work and home environments. The research was conducted either in isolation (outside of the lab) or with a content warning sent to lab mates. Additionally, we restricted ourselves to only processing text data in Amharic in the lab (with content warnings still going out) since no other lab mate could read or understand the language. There is also a risk of harm to the reader from exposure to explicit, distressing, and/or offensive images. We try to minimize this risk by providing content warnings and blurring images.

Preview A content warning is included directly after the abstract and is highlighted in red. It is placed before any in-depth discussion of harmful content and describes the nature of the content. Captions in tables and figures also include content warnings to warn against the type of harmful content included.

Distance We avoid propagating harm by minimizing the examples we provide and replacing them with identifiers. In cases where interview subjects describe the content they were exposed to in detail, we avoid further propagation of harm by reducing direct quotes in those scenarios and instead giving general descriptions.

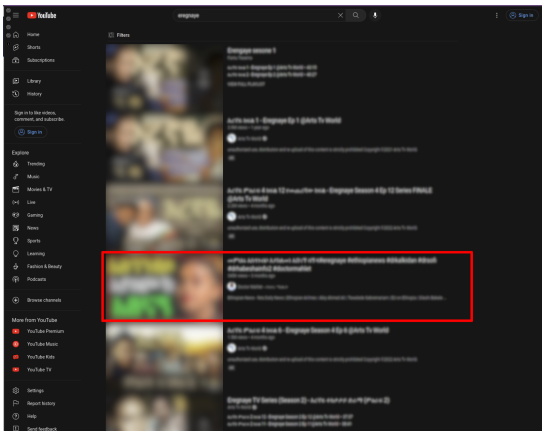
Disclaim In order to clearly indicate that the images and text come from the platform, we used screenshots (which we then blurred) that show the platform's interface.

Respect We blurred images to protect subject identity. We also removed all personally identifiable information (including location) from interviews and from data we collected. Additionally, for interview subjects concerned with their data protection, we shared the anonymized table (Table 1) and made sure they were okay with the presentation. We also do not release any data publicly. We avoid presenting data that shows groups being described with pejorative terms and instead describe the general phenomenon or use identifiers. We also use placeholders for celebrity names and TV shows we used in our queries out of respect for individuals and to protect individuals and groups from further harm.

C SCREENSHOTS

In this section, we provide some screenshots from Study 1. In Fig. 1, we show how sexual videos occur in search results for benign queries. We found that sexual videos are not limited to just recommendation and search results; in Fig. 2, when the query was the name of an Ethiopian celebrity, the "People also watched" section showed a video with a sexual thumbnail and title. As seen in Fig. 3, we also found that for a verified channel that posts sexual content in Amharic, the recommended videos were all sexual videos from the same channel, a clear violation of the policy to suppress content which does not abide by the community guidelines. Additionally, we noted that when we set our VPN location to the UK, we got a privacy notice about data collection and were offered options to limit the extent of data collection by YouTube (see Fig. 6). We did not receive this notice in any other location where we collected data from. During the labeling process, searching for video titles found in our collected data returned actual, non-Ethiopian pornographic videos (see Fig.4). In the recommendations, we observed a range of different types of videos. As seen in Fig. 9, recommendations include, but are not limited to, sexual videos in other languages and sexual scenes cut from movies.

(a) Screenshot showing the results for searching for Doctor in Ge'ez script.



(b) Screenshot showing the results for searching for a famous TV show in Latin script.

Figure 1: CW: Discussion of sexual content. In Fig. we present a screenshot for the search results for the query "Doctor" written in Ge'ez script. The first four results are from a medical YouTube channel by Ethiopian doctors, an entertainment talk show, a news channel, and a talk show featuring a psychiatric doctor. Then, the fifth result (highlighted in red box) is a sexual video with a picture of an Ethiopian girl and explicit sexual writing on the thumbnail. The title says 'He made me [EXPLICIT WORD] 7 times'. The video is from a channel that has a name that starts with "Dr." Similarly, in Fig. 1b, the first three results for a famous TV show are episodes of the TV show, and the fourth result (highlighted in red box), is a sexual video with similar characteristics as the one in Fig. 1.

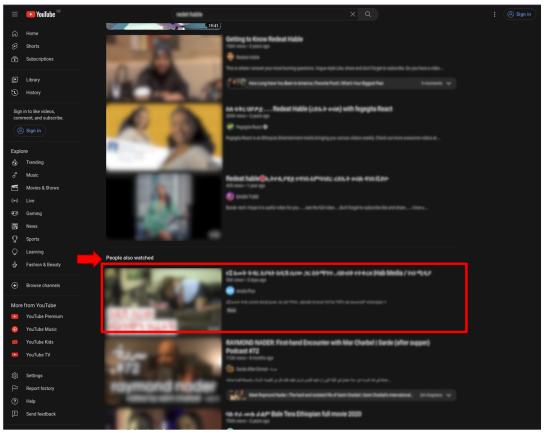


Figure 2: CW: Discussion of sexual content. Scrolling down the search results for a famous Ethiopian celebrity. Sexual videos are not limited to direct responses to search queries. Here, we found a sexual video that has two people engaged in a sexual act on a sofa with the title indicating a person cheating on her husband with a satellite dish maintenance person in the "People also watched" section.

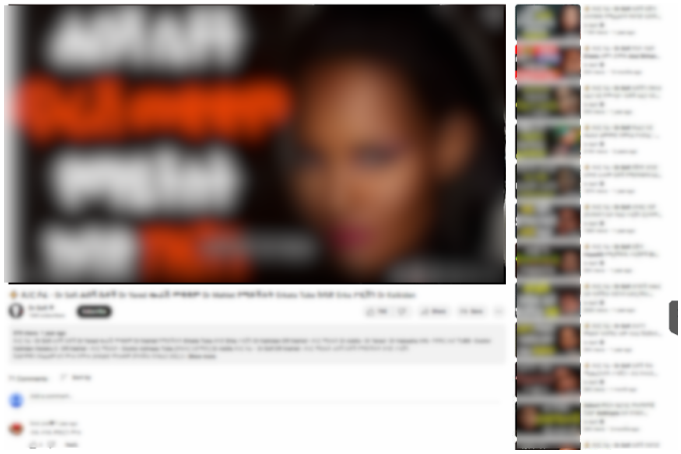


Figure 3: CW: Discussion of sexual content. Screenshot showing recommendations for one of the sexual videos opened to collect recommendation data. This video is from a verified channel that exclusively posts sexual videos in Amharic. All the recommendations for the one video opened from this channel are other sexual videos all from the same channel. The videos would have an image of a woman, often Ethiopian, and explicit sexual writing in Amharic. The channel also uses a 'Dr.' title in their channel name.

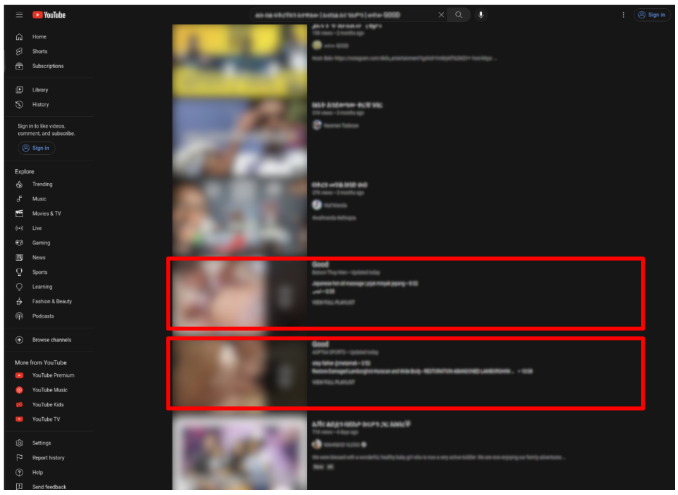
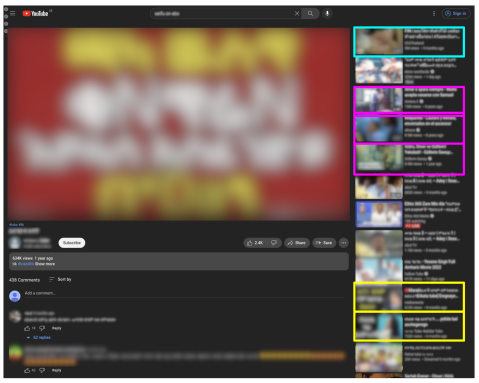
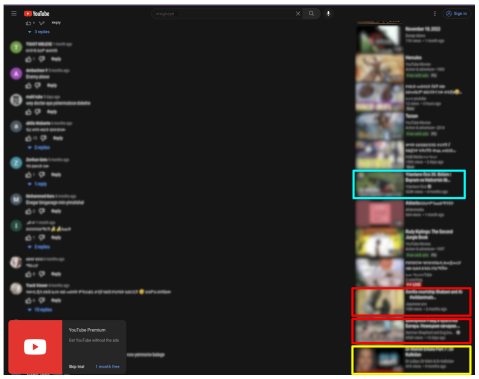


Figure 4: CW: Discussion of sexual content. In trying to label some videos from the collected data, we searched the titles of some videos on the YouTube interface. The results for the Ge'ez-based titles would sometimes return graphically explicit videos. In this case, the search query included the word "GOOD" which phonetically is similar to a word in Amharic used to describe astonishment. The fourth and fifth search results (highlighted in red boxes) were playlists with the title "Good" but had video thumbnails of actual pornographic videos.



(a) Screenshot showing recommended videos once a sexual video is opened.



(b) Screenshot showing recommended videos as one scroll through the recommendations.

Figure 5: CW: Discussion of sexual content. Fig. 5a shows a screenshot of the recommendation list for an Amharic sexual video we opened for data collection. The video has explicit Amharic writing and uses lexical variation by mixing Ge'ez and Latin characters. Fig. 5b shows screenshots of recommended videos as we scroll down the recommendations for an Amharic sexual video. Recommendations include sexual videos in other languages (highlighted in teal), sexual scenes cut from movies (highlighted in pink), Amharic sexual videos from other channels (highlighted in yellow), and videos of animals mating (highlighted in red).

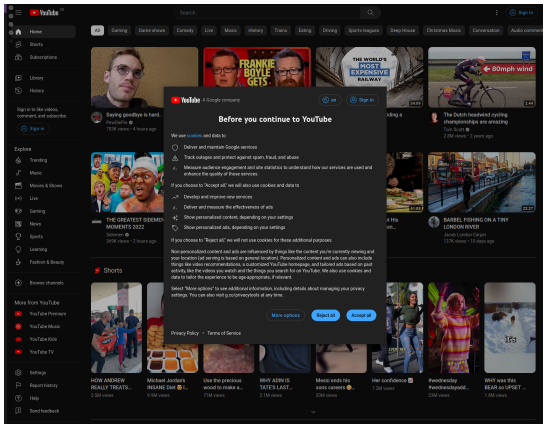


Figure 6: Screenshot showing privacy notice for opening YouTube when the VPN location was changed to the UK. The first thing on the YouTube interface was a prompt about data collection through cookies and options to accept all, reject all or ask for more options. YouTube did not display this notice for any of the other four locations.

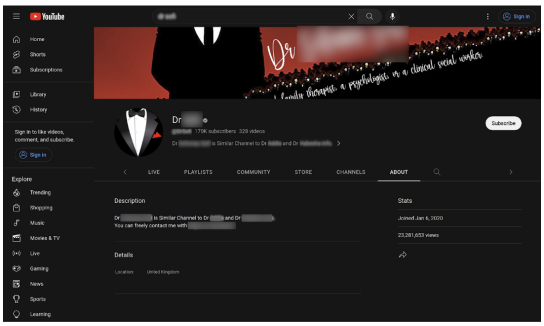


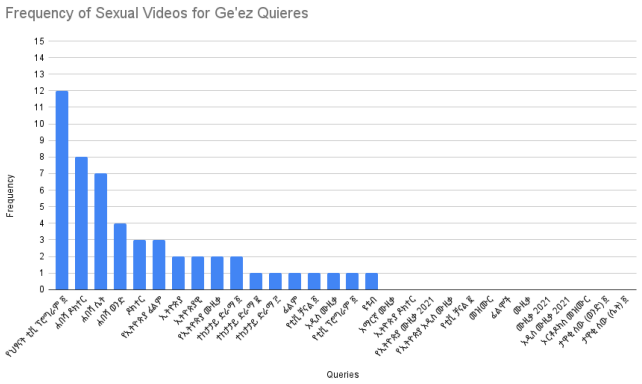
Figure 7: Screenshot showing a YouTube-verified channel that exclusively posts sexual videos in Amharic. The channel name (blurred to protect the identity of the creator) starts with "Dr." followed by a popular Ethiopian name. Further in the description, the channel states "Dr. [NAME_1] is similar channel to Dr. [NAME_2] and Dr. [NAME_3]. You can freely contact me with [TELEGRAM LINK]". The channel location is set in the United Kingdom and has over 23 million views.

D ADDITIONAL EVIDENCE FOR STUDY 2

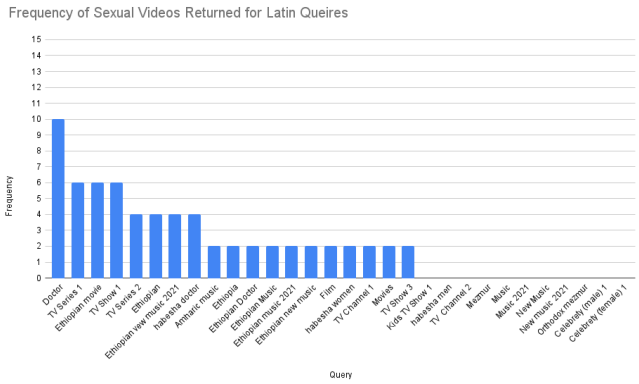
In this section, we provide further evidence to support our findings in Study 2. In Fig. 8, we present the full list of queries we assembled as described in Sec. 3.2. In Fig. 9a and Fig. 9b, we show the distribution of policy-violating sexual content returned for the queries we assembled for Study 2. In Table 5, we show the percentage of each of the categories of policy-violating videos in recommendations.

N	Amahric/Ge'ez Query	English/Latin Query	Type	Source
1	የህፃናት ቲቪ ፕሮግራም ፩	Kids TV Show 1	TV Show	Top Searches
2	ሐበሻ ዶክተር	habesha doctor	Benign term	From Tags
3	ሐበሻ ሴት	habesha women	Benign term	Previous work inspired
4	ሐበሻ ወንድ	habesha men	Benign term	Previous work inspired
5	ዶክተር	Doctor	Benign term	From Tags
6	የኢትዮጵያ ፊልም	Ethiopian movie	Benign term	Top Searches
7	ኢትዮጵያ	Ethiopia	Benign term	Top Searches
8	ኢትዮጵያዊ	Ethiopian	Benign term	Top Searches
9	የኢትዮጵያ ሙዚቃ	Ethiopian Music	Entertainment	Top Searches
10	ተከታታይ ድራማ ፩	TV Series 1	TV Drama	From Tags
11	ተከታታይ ድራማ ፪	TV Series 2	TV Drama	Top Searches
12	ተከታታይ ድራማ ፫	TV Series 3	TV Drama	Top Searches
13	ፊልም	Film	Entertainment	Top Searches
14	የቲቪ ቻናል ፩	TV Channel 1	Entertainment	Top Searches
15	አዲስ ሙዚቃ	New Music	Entertainment	Top Searches
16	የቲቪ ፕሮግራም ፩	TV Show 1	TV Show	Top Searches
17	ዩቱብ	YouTube	Benign term	Top Searches
18	አማርኛ ሙዚቃ	Amharic music	Entertainment	Top Searches
19	ኢትዮጵያ ዶክተር	Ethiopian Doctor	Benign term	From Tags
20	የኢትዮጵያ ሙዚቃ 2021	Ethiopian music 2021	Entertainment	Top Searches
21	የኢትዮጵያ አዲስ ሙዚቃ	Ethiopian new music	Entertainment	Top Searches
22	የቲቪ ቻናል ፪	TV Channel 2	Entertainment	Top Searches
23	መዝሙር	Mezmur	Religious	Top Searches
24	ፊልሞች	Movies	Entertainment	Top Searches
25	ሙዚቃ	Music	Entertainment	Top Searches
26	ሙዚቃ 2021	Music 2021	Entertainment	Top Searches
27	አዲስ ሙዚቃ 2021	New music 2021	Entertainment	Top Searches
28	ኦርቶዶክስ መዝሙር	Orthodox mezmur	Religious	Top Searches
29	ታዋቂ ሰው (ወንድ) ፩	Celebrety (male) 1	Famous Person	From Tags
30	ታዋቂ ሰው (ሴት) ፩	Celebrety (female) 1	Famous Person	From Tags
31		Ethiopian vew music 2021*	Entertainment	Top Searches

Figure 8: Search queries we curated to collect data for Study 1. In this table, we show the Ge'ez and the Latin versions of each query which were all used in our data collection. We curated the search queries based on (1) top YouTube search statistics for Ethiopia, (2) popular, benign terms from sexual videos we were getting during our initial data collection, and (3) benign phrases inspired by previous work that studies search engine bias. The last query on the Latin list does not have an Amharic equivalent since it includes the misspelled word 'vew' in place of 'new'; we still ran the query in English as it was in the Top YouTube searches. Following guidelines in HarmCheck, we replaced celebrity names with identifiers, to avoid exposing their identity and any potential association with harm.



(a) Distribution of Sexual videos per query for Ge'ez based search.



(b) Distribution of Sexual videos per query for Latin-based search.

Figure 9: Fig. 9a shows the distribution of sexual videos for each of the Ge'ez queries in our Study 2. We see that the highest number of sexual videos is returned for a query with a children's show name. Fig. 9b shows the distribution for the Latin queries in Study 2; we see that the highest number of sexual videos is returned for the query "Doctor".

Video Label	Label Description	Ethiopia	Saudi Arabia	UAE	US	UK
Sexual	Narrations of sexual stories; audio recordings of sexual intercourse and phone sex; exposé videos with fully or partially naked people performing sexual acts; sexual videos in other, non-English languages; sex scenes cut from movies; videos of animals mating; rape advocating videos; documentation of rape acts; videos with sexually explicit thumbnails but educational content; non-consensual released sexual videos.	**45.15%**	15.78%	**35.52%**	18.67%	**38.25%**
Entertainment	Talk shows; games and competitions; and food channels	30.90%	**33.73%**	33.16	**34.09%**	30.96%
Politics	News; Commentary; Interviews	6.36%	13.59%	4.04%	3.43%	9.65%
Vlogs	Video blogs by college students; Ethiopian people living in the Middle East	0.30%	12.70%	8.08%	5.33%	11.84%
Educational	videos about taxes; high school textbook explanations; and a decent amount of language learning videos	8.18%	8.92%	5.38%	14.48%	0.91%
Religious	Songs and sermons from Ethiopian Orthodox Christian, Islam, and Evangelical Christian religions.	3.63%	7.34%	5.55%	11.81%	4.01%
Sports	News; Commentary; Recorded Streams; Reaction Videos.	0.45%	1.49%	1.34%	4.04%	0.18%
Motivational	Talks; Interviews; Narrations	2.12%	2.28%	2.69%	5.14%	1.82%
Relationship Advice	Podcasts; Interviews; Video essays	2.42%	1.69%	3.53%	2.29%	1.46%
Information	Where to buy things; Reviews	0.45%	2.48%	0.17%	0.57%	0.00%
Conspiracy Theories	Interviews; Video essays	0.00%	0.00%	0.50%	0.19%	0.91%

Table 5: CW: Discussion of sexual content and rape.Types of recommended videos for every policy-violating sexual video identified, by country. For each of the policy-violating sexual videos we identified, we categorize the videos recommended. We observed that for Ethiopia, UAE, and UK locations, clicking on policy-violating sexual content will likely lead to more sexual content being recommended by YouTube's algorithm. Of all locations, Ethiopia has the highest recommendation for sexual videos, with 45.15% of recommendations being sexual videos.

E THEMES FROM INTERVIEW STUDY

In this section, we provide themes and sub-themes that emerged from our interview study analysis and provide example open codes in Table 6.

Second-Level Themes	First-Level Themes	Example open codes
Users' search experience	experience based on search language	report better experience searching for English content
	experience based on writing script	indicate searching in Ge'ez is harder due to keyboard switching
	mixed, foreign languages in search results	got content in Hindi for Amharic query
Perceptions of how YouTube features and procedures work	perceptions about video recommendations	believes recommendations are usually based on watch history
	perceptions about Home Page	get political content recommendation on home page
	perceptions about trends	trend has more or less evolved with their taste
	perceptions of effects of VPN	search experience are the same with and without VPN
	perceptions about YouTube policy	believes explicit language is not allowed
	perceptions about reporting mechanisms	believes there is a queue for reports
Policy, culture, and fairness	questions on who makes decisions	questioning the cultural context 'sexual videos' are defined in
	role of culture in justification of content	notes videos use conservative culture to justify releasing degrading, non-consensual videos of women
	(anticipated) consequence of unaddressed policy-violating content	fears unintended exposure of younger generation to sexual content
YouTube platform usage and utility	default YouTube settings	default YouTube language is English
	time span of use	has been using YouTube for over 8 years
	comment feature	commented to defend a woman whose video was released without her consent
	like feature	likes to get the 'algorithm' to bring them similar content
	posting videos	has never posted their own video
	subscriptions	subscribes to channels after checking the full video list
	substitute platform	does not believe there is a substitute for YouTube
Harmful content	characters of sexual videos	has been exposed to non-consensual release of sexual video if a minor
	categories of other harmful content	has been exposed to graphic and violent content
	events that led to unintended exposure of sexual videos	got exposed while searching for a famous singer
	strategies against unintended exposure	uses multiple google accounts
Reporting practices	steps taken to report	chose sexually explicit from the provided categories
	outcomes of reporting	did not receive feedback on their email
	types of reported videos	ethnic hate speech video

Table 6: Second- and first-level themes, and example open codes from Study 2. Our analysis resulted in 936 unique open codes which were grouped to 26 first-level themes, which were further grouped into 6 second-level themes.