Labeling in the Dark: Exploring Content Creators' and Consumers' Experiences with Content Classification for Child Safety on YouTube

Renkai Ma
College of Information Sciences and Technology, Pennsylvania State University
USA
renkai@psu.edu

Zinan Zhang
College of Information Sciences and Technology, Pennsylvania State University
USA
zzinan@psu.edu

Xinning Gui
College of Information Sciences and Technology, Pennsylvania State University
USA
xinninggui@psu.edu

Yubo Kou
College of Information Sciences and Technology, Pennsylvania State University
USA
yubokou@psu.edu

ABSTRACT

Protecting children's online privacy is paramount. Online platforms seek to enhance child privacy protection by implementing new classification systems into their content moderation practices. One prominent example is YouTube's "made for kids" (MFK) classification. However, traditional content moderation focuses on managing content rather than users' privacy; little is known about how users experience these classification systems. Thematically analyzing online discussions about YouTube's MFK classification system, we present a case study on content creators' and consumers' experiences. We found that creators and consumers perceived MFK classification as misaligned with their actual practices, creators encountered unexpected consequences of practicing labeling, and creators and consumers identified MFK classification's intersections with other platform designs. Our findings shed light on an interwoven network of multiple classification systems that extends the original focus on child privacy to encompass broader child safety issues; these insights contribute to the design principles of child-centered safety within this intricate network.

CCS CONCEPTS

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

KEYWORDS

COPPA, content creation, child privacy protection, content creator, child safety

DIS '24, July 01–05, 2024, IT University of Copenhagen, Denmark

ACM ISBN 979-8-4007-0583-0/24/07
https://doi.org/10.1145/3643834.3661565

ACM Reference Format:
Renkai Ma, Zinan Zhang, Xinning Gui, and Yubo Kou. 2024. Labeling in the Dark: Exploring Content Creators' and Consumers' Experiences with Content Classification for Child Safety on YouTube. In *Designing Interactive Systems Conference (DIS '24), July 01–05, 2024, IT University of Copenhagen, Denmark.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3643834.3661565

1 INTRODUCTION

Protecting children's online privacy is paramount. Notable laws, such as the Children's Online Privacy Protection Act (COPPA) in the US [24] and the General Data Protection Regulation (GDPR) in Europe [39], underscore this commitment. The emphasis on safeguarding children's online privacy stems from their inherent vulnerabilities and limited capacity to understand the risks of sharing personal information (e.g., name, browsing habits), such as threats like online harassment and cyberbullying [2, 89], as well as the inappropriate commercial use of children's information for targeted advertising [41], which COPPA and GDPR aim to mitigate. As these risks evolve due to technical advancements, such as recommendation algorithms [59] and advertising technologies [34], particularly with the growing popularity of social media, legal scholars have called for a more nuanced, modernized update of COPPA (e.g., [37, 67, 74]). Meanwhile, HCI researchers (e.g., [2, 41, 82, 98]) have striven to design a secure online environment for youth and minors.

To enhance child privacy protection, several online platforms in the US have implemented new classification systems in their existing content moderation practices. Content moderation refers to the organized practice of screening user-generated content to determine its appropriateness for a certain platform [76]. The changes in moderation practices, including the incorporation of new classification systems, emerged in response to fines for violating COPPA. In 2019, TikTok was fined $5.7 million by the Federal Trade Commission (FTC) for violating COPPA [25]. In response, TikTok launched "TikTok for Younger Users," a restricted version that shows algorithmically curated content and restricts minors from generating public videos or comments [83]. Similarly, after a $170 million FTC

fine [33], YouTube launched the "made for kids" (MFK) classification system [91], which requires content creators to classify videos, disabling consumers' data tracking [92]. Specifically, MFK videos restrict consumers from making comments and impact creator content's monetization performance, a type of moderation termed as "demonetization" [21, 53, 60].

However, traditional content moderation practices focus on managing content (Roberts, 2019) rather than user privacy. While moderation practices contribute to general child safety by classifying content as inappropriate for children (e.g., [3, 5, 70]), relatively little research examines how these moderation practices directly contribute to child privacy protection, especially preventing online platforms from collecting children's personal information. Also, social media platforms implement child privacy laws like COPPA [24] by restricting content creation and consumption (e.g., [83, 92]). However, prior work (e.g., [2, 8, 55, 82, 89]) primarily focuses on the roles of parents, adolescents, and children in child privacy on social media, leaving a gap in understanding the roles of creators and consumers. Given these research gaps, we chose YouTube's MFK classification as a case study to explore creators' and consumers' experiences with the implementation of child privacy protection. YouTube has been one of the most popular platforms among children [38] and, after receiving one of the heaviest FTC fines for violating COPPA, has made some of the most notable initiatives in child privacy protection design, such as MFK classification [91, 92] and YouTube Kids app, compared with other platforms like TikTok [25, 83] and Facebook Messenger [26]. So, we ask: **How do content creators and consumers experience the MFK classification system on YouTube?**

Guided by our research question, we used reflexive thematic analysis to qualitatively analyze relevant online discussions posted in YouTube-related communities on Reddit. Creators and consumers observed misalignment between MFK classification and their actual practices, resulting in false positives and false negatives of MFK classification. Creators experienced unintended consequences when manually practicing MFK classification. Both creators and consumers further observed that MFK classification intersected with other platform designs, especially classification systems, through coordination, inconsistency, and conflict. Drawing from and extending Bowker and Star's classification theories [16], we discuss how our findings indicate an interwoven network of classification systems that extends the MFK classification's original focus of child privacy to a broader issue of child safety (e.g., inappropriate content and advertisements). We thus put forward the design principles of child-centered safety in such an interwoven network of classification systems.

Our study utilizes social media discussions to delve into creators' and consumers' experiences associated with YouTube's MFK classification system, reflecting on how it shapes child safety. Rather than cataloging the exhaustive account of creators' motivations or how they use the MFK system to protect child privacy, our study prioritizes a practical lens to reveal a wide range of user reactions, strategies, and challenges underneath their experiences with MFK classification as YouTube and FTC enforce creators to directly implement child privacy protection through the MFK classification [33, 92]. This is vital for understanding the intricate ways creators contribute to child safety on the platform. Our findings also highlight the practical application of the MFK system and its influence on content consumption, thereby underpinning the necessity for policy and platform design to authentically reflect the dynamics among creators, consumers, and the platform. Our exploratory study can thus contribute to the HCI and designing interactive systems (DIS) communities with four insights:

- We offer an empirical account of content creators' and consumers' experiences with the designs of child privacy protection on commercial and social media platforms like YouTube, enriching existing HCI literature on children's online privacy (e.g., [55, 82, 89, 98]).
- We show how social media and commercial platforms like YouTube leverage content moderation and creators' classification practices for child privacy protection, deepening HCI literature concerning moderation practices (e.g., [48, 50, 79]) and content creators' interactions with moderation designs (e.g., [49, 53, 62]).
- Drawing from and expanding on Bowker and Star's classification theories [16], we show an interwoven network of classification systems on YouTube and its broad effects on child safety beyond the original focus of child privacy.
- In such an interwoven classification network, we lay out actionable design principles of child-centered safety on platforms like YouTube.

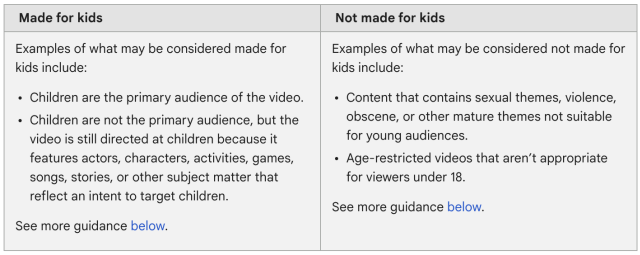
2 BACKGROUND: COPPA AND "MADE FOR KIDS" (MFK) CLASSIFICATION SYSTEM ON YOUTUBE

COPPA [24] is a US federal law established in 1998 to protect the privacy of children under 13, requiring US-based websites or services to obtain parental consent before collecting children's personal information, such as full names, home addresses, email addresses, IP addresses, behavioral data, photos, and recordings [9, 24].

To comply with COPPA, YouTube has carried out several measures over time. In 2008, it implemented an age-restriction classification to limit specific videos to viewers over 18. It introduced the YouTube Kids app/platform in 2015, providing a curated environment for kids [4]. In 2019, after a $170 million FTC fine for violating COPPA [33], YouTube launched "made for kids" (MFK) [91] on January 6, 2020. This requires all creators on the platform to manually classify their videos as MFK or non-MFK [92] , with guidelines provided in Table 1 and the classification interface shown in Figure 1. YouTube uses machine learning algorithms to label all unclassified videos or override creators' classifications and claims it disables consumers' data tracking on MFK videos to enhance child privacy [91, 92, 94].

3 RELATED WORK

We discuss two literature groups: children's online privacy and classification in content moderation. We discuss how the former distinguishes between interpersonal and commercial dimensions of children's online privacy. We further introduce Bower and Star's classification theories and discuss how they help understand the classification in content moderation. This sets a solid ground for us to explore users' experiences with YouTube's MFK classification.



Made for kids	Not made for kids
Examples of what may be considered made for kids include: • Children are the primary audience of the video. • Children are not the primary audience, but the video is still directed at children because it features actors, characters, activities, games, songs, stories, or other subject matter that reflect an intent to target children. See more guidance below.	Examples of what may be considered not made for kids include: • Content that contains sexual themes, violence, obscene, or other mature themes not suitable for young audiences. • Age-restricted videos that aren't appropriate for viewers under 18. See more guidance below.

Figure 1: MFK policy of how to decide a video between MFK and non-MFK [91].

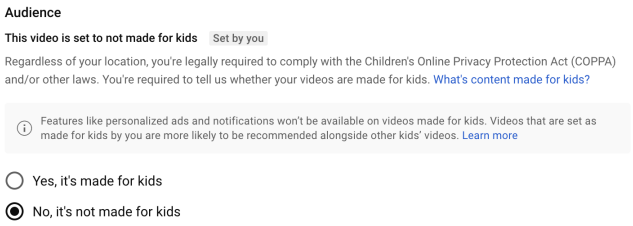


Figure 2: Screenshot of how creators manually classify a video as MFK or non-MFK on the YouTube Studio dashboard [92].

3.1 The Protection of Children's Online Privacy

Protecting children's online privacy is challenging. Nissenbaum defines privacy as "a right to appropriate flow of personal information" [66]. While there is no simple definition of children's online privacy, several HCI studies (e.g., [55, 63, 97]) commonly mention that children interpret online privacy as recognizing the sensitivity of personal information and the need to share it selectively online. Children's increasing use of social media poses risks to their privacy. Internet safety researchers warn the public that children lack an understanding of privacy and tend to disclose too much information on social media [68], while parents might hold different perspectives on it. Researchers found that parents generally adopt a passive approach in mediating their children's device use, information sharing, and privacy education [55], but their involvement and concern increase when their children's data is collected by social media platforms [35]. For children, Ghosh et al. found that they disliked apps that were overly restrictive of their privacy, negatively impacting their relationships with parents, while children liked apps that supported self-regulation [40]. As stressed by Badillo-Urquiola et al. [10], children value personal agency and privacy rather than constant parental consent.

Under such diverse evidence concerning children's online privacy protection, two distinguishable perspectives have emerged in prior HCI research: *interpersonal privacy* and *commercial privacy*. First, interpersonal privacy describes how children's personal information is created, accessed, and multiplied through their social connections [59]. HCI researchers have focused on "sharenting," parents' practices of sharing personal information of their children online (e.g., [6–8, 15, 56, 65]). Amon et al. found that the size of parents' social networks positively affects their parental sharing frequency [8]. Ammari et al. found that parents negotiate with each other and extended family members to establish sharing boundaries on social media (e.g., whether people can see or share their children's information) [6]. This line of work shows that different social connections of children shape the flow of their personal information online.

Second, commercial privacy concerns how online platforms gather and analyze children's personal information for business purposes (e.g., [1, 64, 69]), such as targeted advertisements [11]. Recently, HCI researchers have started to explore how children understand their commercial privacy. Goray and Schoenebeck found that children have limited awareness of whether online advertisers collect their data and what types of personal information are retained by social media platforms [41]. Zhao et al. also found that while children can identify privacy risks of oversharing with their social connections online, they remain incapable of identifying online tracking and targeted advertisements [98]. Similarly, Sun et al. uncovered that children tend to characterize data tracking as operated by humans rather than analytic tools on social media platforms and less consider the platforms that collect and process their data as privacy threats [82].

Our study aims to understand the impacts of content creation and consumption on children's commercial privacy. In the US, social media platforms like YouTube [91] and TikTok [83] have implemented child privacy laws like COPPA [24] to limit content creation and consumption, mitigating the risks of collecting child creators' and consumers' personal information. However, relatively little is known about how content creators and consumers, as important stakeholders in child privacy protection, perceive or respond to these measures. This study aims to fill this research gap.

3.2 Classification Theory and Classification Systems in Content Moderation

At the heart of content moderation is a classification task, where platforms organize a vast array of user-generated content (UGC) — from hate speech to misinformation — into categories defined in platforms' moderation policies (e.g., YouTube [96], Facebook [32]) to determine its appropriateness for the given platforms (Roberts, 2019). This classification is underscored by the moderation processes discussed in prior literature (e.g., [13, 36, 42]), where moderation policies establish the criteria for content classification to identify and flag policy violations [57]. Platforms typically employ human moderators for this classification task (Roberts, 2019) or develop complex algorithms to detect and categorize policy violations [45]. These algorithmic systems, refined through learning from prior moderated content, are then applied to detect new, non-compliant content [29, 42]. When moderation algorithms encounter potential classification issues (e.g., inaccurate classification), human moderators are called upon for final adjudication, typically resulting in moderation decisions like content removal [47, 81] or account suspension [44, 85].

This systematic classification in moderation profoundly resonates with Bowker and Star's classification theories [16]. Classification describes organizing things into categories, the metaphorical or literal "boxes," which significantly shape our experiences, knowledge production, and social interactions. For example, Bowker and

Star discussed an example of Dewey's library scheme, which assigns classification numbers (e.g., index) to new books in a library to allocate them to the appropriate location based on subjects. They highlighted how it facilitates knowledge access and shapes our cognition and understanding. While Bowker and Star mention that such classification is ubiquitous, it is typically invisible and only becomes noticeable when it breaks down.

Bowker and Star also stress that no classification system can fully capture the world's complexity [16]. So does moderation when it confronts the variety of content. HCI and legal scholars criticize that moderation policies across platforms lack granularity in defining the appropriateness of content [71, 88]. Similarly, Jhaver et al. found that moderators reassessed old moderation policies and articulated new ones to refine classification practices [48]. Existing moderation policies might fail to capture too nuanced human behaviors, as stressed by Jiang et al. [50].

Bowker and Star's classification theories thus offer a lens to better understand moderation experience – the experiences of those subjected to moderation decisions. They conceptualized the notion of torque to describe how classification systems influence individuals' lived experiences – "where the 'time' of the body and of the multiple identities cannot be aligned with the 'time' of the classification system" [16]. In moderation, HCI research reflects this classification misalignment. Vaccaro et al. highlighted that Facebook users felt moderation was inconsistently applied, leading to undue account suspensions [85]. Similarly, Haimson et al. reported that sexually minority individuals felt marginalized by the misclassifications of their surgery content [44]. Ma and Kou found that YouTube creators faced financial frustration when algorithms misclassified their gaming content as violent [62].

In this study, we apply Bowker and Star's classification theories to unpack the complexities of multiple classification systems that coexist with YouTube's child privacy protection implementation (e.g., MFK classification). While some prior work delves into moderation practices of classifying content as inappropriate for general child safety (e.g., [3, 5, 70]), relatively little research examines how these practices directly contribute to child privacy protection – preventing online platforms from collecting children's personal information. Additionally, prior HCI and CSCW literature (e.g., [36, 44, 62, 86]) has touched upon users' experiences with moderation, with a particular focus on content creators on YouTube [53, 61, 62]. However, the connection between these experiences and child privacy remains unexplored. Through the case of YouTube's MFK classification system, designed to align with child privacy laws [9, 24], we aim to reveal both creators' and consumers' experiences with this system. Recently, legal scholars [12, 23, 87] have suggested that YouTube creators might inadvertently misclassify their content, and business researchers [51] have found that child-directed content creators have reduced content quality due to the MFK classification. Given these concerns, it's necessary to empirically study the experiences of both creators and consumers with the MFK classification.

4 METHODS

4.1 Data Collection

In this study, we chose online discussion data from YouTube-specific subreddits as the data source for two reasons. First, analyzing online discussions from subreddits is a common approach for data collection in HCI research (e.g., [30, 58, 60]), especially when accessing targeted participants like content creators is challenging. Second, unlike interviews that rely on human memories, online discussions offer real-time insights into users' experiences as they share online, aiding our understanding of their interactions with MFK classification. This data source enabled us to capture a broad view of experiences interacting with MFK, encompassing user reactions, perceptions, and challenges.

Thus, we choose relevant subreddits for data collection, including r/youtube, r/youtubers, r/newtubers, and r/partneredyoutube. That was because, after searching "YouTube" on Reddit, these subreddits had noticeably larger community sizes: 1.1 million, 235 thousand, 328 thousand, and 57.6 thousand members, respectively, than others, except biased communities such as r/fuckYTCOPPA and r/BannedYouTube. Also, the four subreddits' self-descriptions are highly relevant in understanding both creators' and consumers' experiences. For example, r/youtube's "About Community" states it is for general discussions about YouTube, and the other three focus on content creation. Please note that we received approval from our institution's Institutional Review Board (IRB) before data collection.

Our data collection involved four steps. First, in February 2023, based on YouTube's moderation policies [91, 92] and the first author's domain knowledge about YouTube, we gathered a set of keywords, including "COPPA" and "made for kids," the two most saliently relevant terms, as well as "kids friendly," "child friendly," "kid oriented," "child oriented," "kid," and "parent," where the latter two were not strictly relevant but can potentially help fetch more data for analysis. Second, using these keywords, we fetched 1,819 threads and 54,020 associated comments from the four subreddits through the package 'RedditExtractoR' on R [75]. This R package helped search the keywords in titles, posted texts, or comments and returned the search results. We then skimmed all threads and selectively read the comments to examine the richness of the dataset and collect more keywords for further data collection. Third, we identified more keywords, including "Age 13", "FTC," "children," "children primary audience," and "MFK," leading to 745 more threads and 9,814 associated comments. The final dataset included 2,564 threads and 63,834 comments, stored and analyzed in Google Sheets, focusing on submission texts and comments. Fourth, since FTC fined YouTube after September 2019 [33], we removed threads and comments posted before January 1, 2019. This resulted in 528 threads and 8,295 comments for data analysis (see column "Data posted after January 2019" in Table 1).

Before data analysis, there was one step of data preprocessing. We recognize that a platform like YouTube involves both content creation and consumption by users. However, for the purpose of this study, we need to differentiate between creators and consumers to understand their respective behaviors and perspectives. Thus, we started by assuming all users on YouTube are content consumers.

This was based on the understanding that using YouTube inherently requires users to consume content, even at the most basic level, such as reading video titles or understanding platform functionalities, before engaging with more substantial content like the videos themselves. Then, for content creators, we set a "creator self-disclosure" criterion: Posters, either thread or comment posters in the four subreddits, were categorized as creators if they clearly disclosed their creator identity by mentioning keywords like "my audience," "my channel," "my video," or other keywords that are easy for us to identify their creator role.

Otherwise, we considered the posters to be consumers. We acknowledge that this method may not perfectly capture all creators, especially those not explicitly mentioning their creator role. However, given the larger proportion of consumers in online communities and the need for a practical method of categorization, this approach provides a functional way to differentiate between the two roles for our analysis.

4.2 Data Analysis

The research team applied reflexive thematic analysis (RTA) [18, 19] inductively to analyze the whole dataset. RTA is a "theoretically flexible method" for developing, analyzing, and interpreting patterns in qualitative data [18], incorporating researchers' experiences, pre-existing knowledge, and social positions to analyze the data critically.

Data analysis involved four steps. First, two researchers familiarized themselves with the threads dataset and resolved confusion about the contexts that the dataset mentioned (e.g., what "age-restriction" is, how MFK classification will turn off video features such as commenting and playlist). Second, they individually screened the data and assigned initial codes in Google Sheets to represent ideas expressed in the dataset that can answer the research question. They held weekly meetings with two other senior researchers to discuss each code's relevance and correspondence with original quotations. The dataset included much data that were not directly related to MFK classification due to a wide range of keyword searching, so a critical criterion of deciding data point's relevance for coding was whether it discussed about MFK classification, such as how creators talked about their perceptions and reactions to COPPA or FTC and how they share perspectives about these with others. For example, some creators shared knowledge of creator growth, such as "the advice about views/subs vs. watch time is solid," which is unrelated to our research question.

Third, the two researchers examined the relevance of the data for coding and, meanwhile, continually assigned initial codes to the dataset and conducted rounds of coding to identify what patterns (i.e., subset themes) are reflected by initial codes. This process identified subset themes from the initial codes. For example, the researchers assigned the initial code, "MFK misaligned with parents' expected involvement, and they did not acknowledge MFK's labeling," to the quote, "I am a parent myself, and it is my job.…It is up to me to know or accept that a company like Google might scrape data targeted at me and my kids." This code is conceptually related to other codes about how other content consumers consider MFK classification and its outcomes misaligning with their collective understanding, so we grouped them together under one theme, "MFK misaligned with consumers' collective understanding and recognition," and further reported these similar codes together in Section 5.1, Misalignment between MFK Classification and Practice. The criterion for grouping initial codes into subset themes was if multiple codes consistently appeared and shared underlying concepts that can answer our research questions. For example, codes capturing user efforts to circumvent and not consider content as MFK labeled, such as "avoidance strategies for using restricted features" and "parental adjustments to content access," were consolidated under the theme of Section 5.1.2 User-Driven False Negatives of MFK. This theme reflects how users actively navigate around the MFK system's limitations to maintain their engagement with content in ways that defy its restrictions. For another example, we grouped codes describing creators' struggles with MFK classification, such as "uncertainty in content classification" and "challenges with vague guidelines," under the theme Section 5.2 Unexpected Consequences of Practicing MFK Classification as Creators. This theme reflects how broad MFK policies complicate content management for creators.

In the last data analysis step, the research team continued assigning codes to the data and grouping codes into subset themes until theoretical saturation [43]. This indicated a high percentage of initial codes within more than half the volume of the dataset and a minimal increment of initial codes after screening and analyzing around 60% of the dataset (see column "Screened and analyzed data for theoretical saturation" in Table 2). In other words, the team reached the point where no particularly new codes or themes emerged from the dataset [43]. Last, in the weekly meetings, all four researchers consolidated similar subset themes into overarching themes and discarded subset themes deemed thin without enough initial codes. Eventually, data analysis led to a thematic map to answer the research question sufficiently through three Findings from Sections 5.1 to 5.3. In reporting findings, we ensured our data's anonymity by removing YouTube channel names and paraphrasing the original quotes. Please also note that while Bowker and Star's classification theories provide a valuable conceptual framework for understanding classification systems on YouTube, they did not directly drive our data analysis.

4.3 Researcher Positionality

Our interpretation of YouTube creators' and consumers' experiences is based on our positionality [73], including social roles, intellectual history, and lived experience. The first two authors are amateur video content creators on YouTube, with the first author closely in touch with and frequently engaging with more than ten creators across content categories, such as gaming, animation, and beauty, for over three years. The other two authors, while not content creators themselves, are seasoned consumers of creators across education, gaming, and entertainment. Regarding intellectual history, the first and last authors have been researching content creators, creator-audience relationships, and YouTube since 2020, equipping the research team with rich domain knowledge. This combination of hands-on experience and academic insight positions us well to perform reflexive thematic analysis for this study.

Table 1: Data Preprocessing (Section 4.1) and Analysis for Theoretical Saturation (Section 4.1)

Data source	Data posted after January 2019		Screened and analyzed data for theoretical saturation	
	Quantity of threads	Quantity of comments	Quantity of threads	Quantity of comments
r/newtubers	168	1,915	14	347
r/partneredyoutube	25	210	6	85
r/youtube	245	5,396	39	4,423
r/youtubers	90	774	4	130
Total	528	8,295	63	4,985

5 FINDINGS

We found that the YouTube creators and consumers perceived MFK policy enforcement as misaligned with their actual community practices (Section 5.1), creators shouldered the unexpected burden of MFK classification (Section 5.2), and both creators and audiences observed three types of intersection of MFK classification with other platform designs, including coordination, inconsistency, and conflict (Section 5.3).

5.1 Misalignment between MFK Classification and Practice

YouTube creators and consumers have observed a misalignment between MFK classification and actual practices, resulting in false positives and false negatives: (1) the MFK classification system incorrectly flags videos as suitable for children, and (2) MFK videos do not receive the intended level of recognition by creators and consumers.

5.1.1 Perceived False Positives of MFK Classification. False positives of MFK classification occur when videos classified as MFK do not match the two criteria set out by the MFK policies—namely, that the primary or intended audience is children [91, 92, 94]. One such discrepancy occurs when an MFK video should have been classified as age-restricted (i.e., for people over 18) [90] instead of MFK. A consumer posted:

> *I just found a video that is labeled as [made] for kids, and yet it is age-restricted based on community guidelines. It's a Fritz the Cat episode where there's a Nazi bunny, and swastikas are shown a lot in it. There's clearly a bug in the AI that makes the AI fail to consider age restriction status. (consumer; r/youtube)*

This content consumer perceived an MFK video should have been classified as age-restricted because the video contained "Jojo Rabbit," a comedy about a German boy who imagines his friend is Hitler. This video thus had intense violence, death, and anti-Semitism and was not appropriate for children to watch, as the above consumer perceived.

Besides, creators themselves might produce false positives, as evidenced by one who admitted:

> *This whole "made for kids" thing is so dumb. I've uploaded Overwatch futa porn and marked it as "made for kids," but nothing has happened to me. (creator; r/youtube)*

The video posted by the creator above contained adult-oriented material from the video game "Overwatch," which should not have been marked as MFK. They further stressed the gap in the oversight mechanisms of YouTube's MFK classification in correcting creators' mislabeling.

When inappropriate content is classified as MFK, some consumers intuitively blame creators. For example, a consumer commented when a news video of a massive shooting is classified as MFK:

> *100 % on the uploader. They either checked the wrong box, or they blanketed their entire channel as for kids (which would be really stupid for news channels to do). (consumer; r/youtube)*

This consumer stressed two ways of avoiding false positives, including (1) videos within one YouTube channel needed attentive classification from creators, and (2) the MFK classification should not automatically apply MFK tags to all new videos, even though a creator set their whole channel as MFK.

However, creators observe algorithms cannot make nuanced classifications on their videos. A creator posted:

> *I keep getting YouTube setting my ESL language videos specifically targeted at teens and young adults to MADE FOR KIDS… my channel is targeted at kids, yes, (….) but basically, my channel is 80 percent made for kids, and 20 percent not made for kids. (creator; r/youtube)*

This case underscored the limitations of YouTube's algorithms in discerning nuanced consumer targets, leading to false positives. This case also shows how algorithms undermined creators' original discretion in classification, which is different from what MFK policies expect [92].

Some viewers thus believed responsibility lies with both the consumers and creators to avoid misclassifications, as a viewer commented: *"Message the creator. They need to uncheck the 'made for kids' box in their video settings"*.

5.1.2 User-Driven False Negatives of MFK. False negatives refer to instances where content is classified as MFK and thus should be restricted under the MFK guidelines [92, 94] but is not recognized or treated as such by creators and consumers. Specifically, creators and consumers have felt that the MFK classification limits their ability to engage with content as they please. Thus, they devise methods to maintain their autonomy, creating false negative actions that treat videos as if they are not MFK, even though they are, as labeled by the MFK classification system.

One such method involves the use of playlists, a feature that allows users to organize, curate, and share videos in a specific order. Despite MFK restrictions, a viewer shared:

> *I just tested this and found that I can still add Made for Kids videos to playlists from the search results, except for the actual video page! Hover over the video's title, click on the three dots that appear on the right side, and click on Save to Playlist. (…) It's 2022, and I've been using this method since COPPA started, and it has consistently worked for me. (consumer; r/youtube)*

This consumer above not only utilized a system flaw to create a false negative but also shared the knowledge with others, thus spreading the practice. Especially as this consumer validated its effectiveness in 2020 and 2022, the MFK classification did not enforce moderation policies and implement function changes on YouTube over time.

Children, recognized as a unique group of content consumers, often venture into areas not covered by MFK classification, potentially leading to unintended data collection. For example, a creator wrote:

> *This proposed rule (MFK) won't change anything. Kids can just use a parent's account via iPad/phone/TV/computer or laptop, so all that's really happening is the creators are being punished.(creator; r/youtube)*

In this case, as YouTube applied MFK policies [91] at a video level rather than considering the broader context of children's media interaction habits, the creator above claimed that it would be easy to create false negatives of MFK when children access their non-MFK content inadvertently.

Another creator questioned, *"How are YouTube and the FTC gonna deal with kids commenting on Not Made for Kids videos and adding them to playlists?"*This query pointed out that children have been engaging with content outside the MFK classification, hinting at the widespread user-driven creation of false negatives.

Parents deem that MFK classification undermines their autonomy in managing their children's content consumption experience. For example, a parent expressed their frustration:

> *I am a parent myself, and it is my job to parent my kid how I see fit. It is up to me to know or accept that a company like Google might scrape data targeted at my kids and me. Just like most laws like this, they always start out with good intentions, but you know how the saying goes. (consumer; r/youtube)*

This parent highlights two issues that might create false positives of MFK. First, the MFK classification above did not offer a parent consent option, which was misaligned with COPPA requirements [24, 33]. Second, different from how parents can make autonomous, informed decisions on their children's well-being [54], this parent above complained about the lack of control over their kids' data privacy, showing a disconnect between policy and practical parental needs.

The lack of nuanced control is further emphasized by another parent's request for more selective content filtering:

> *I want an app that will give me the ability to select the shows/channels *I* want my kids to be able to see. Whether that app is YouTube, YouTube Kids, or whatever, I don't care. For example, my son is 13 and big into Fortnite. I want him to be able to watch specific YouTubers that do Fortnite while excluding others. (consumer; r/youtube)*

This case underscored the deficiencies in the current MFK classification design, which did not afford parents the active role they seek in the content classification process and inadvertently prompted them to create workarounds that could lead to the creation of false negatives — videos that the MFK classification system would categorize as not suitable for children being treated by parents as acceptable for their children's content consumption.

5.2 Unexpected Consequences of Practicing MFK Classification as Creators

The unexpected consequences of MFK classification refer to the predicament creators need to overcome in their content creation, compared to the time before YouTube enforced MFK policies [91, 94]. YouTube expects creators to practice MFK classification: *"We rely on you to tell us if your content is intended for kids because you know your content best. We trust you to set your audience accurately"* [94]. However, creators feel uncertain about what accurate labeling is, which prompts them to exert additional effort to standardize content creation or reduce their passion for content creation.

5.2.1 Uncertainty Regarding Proper Classification. While YouTube explicates which content categories are MFK [91], creators still struggle to understand how to apply these policies to classification practices, especially when their videos are filled with different, nuanced content elements. For example, video game is a subject matter generally directed at kids, as stated by MFK policy [91]. However, this generalization does not account for the diverse genres and themes within gaming, many of which may not be suitable for children. This one-size-fits-all approach to classification presents challenges for creators. A creator posted:

> *I don't know if my video game videos are "made for kids" or not due to the lack of clarification. Even if games like Call of Duty are violent and look realistic, they can say it is kid-directed as the games are animated. And what about Fortnite, Minecraft, etc.? Kids can watch anything, and anything can be made for kids. Some are just clearer than others. (creator; r/youtube)*

This creator struggled to make a classification decision given the abundance of content elements such as game types, some extent of violence, and visuals between reality and animation. This struggle showed that the simple term "game" in MFK policy cannot sufficiently cover the actual complexity of videos and creators' practices in measuring their videos. Besides, content elements that are not directly measurable also pose obstacles. For example, a game creator posted:

> *If I play Minecraft, which is a game *directed to kids*, but I myself am aiming for a young adult+ audience because my humor is a bit unsuitable for kids, I'm in a very grey area if you consult those guidelines. (creator; r/youtubers)*

Minecraft is a popular sandbox game. This creator thought their humorousness as a creative part of their Minecraft gaming videos made the videos inappropriate for children. Meanwhile, they complained that MFK policies did not explain the extent or kind of creativity that can make a gaming video MFK [91].

As a viewer observed, this vagueness might lead to improper, careless classification practices: *"I think a lot of people marked their own channels as made for kids, fearing that if they didn't, they'd be chased up/sued."* This viewer shared that many creators' confusion or struggle was exacerbated by fears of legal repercussions.

5.2.2 Extra Labor for Disagreed Classifications. Creators face additional work when disagreeing with MFK classifications made by YouTube. A creator posted:

> *Or YouTube set it themselves [through machine learning algorithms]. I had a video I had to manually set back to Not Made for Kids 4 times. Every once in a while, I go through my videos and double-check to make sure YouTube didn't make the decision on its own again. (creator; r/youtube)*

This creator manually labeled back and forth and frequently checked out other videos' statuses on the YouTube Studio dashboard to make sure the automatic MFK classification made sense to them.

Given that one of the two primary criteria for classifying a video as MFK is "children are the primary audience of the video" [91], many creators assume that when a video is labeled as MFK, the primary consumers are already kids. When facing the other criterion, "children are not the primary audience, but the video is still directed at children," creators will likely assume the criterion YouTube chooses for its classification decision randomly. For example, a creator posted:

> *Literally, almost every video I've seen marked [by YouTube] for kids is not intended for kids, likely leading to alienating their main audience by disabling comments and probably bringing in a child audience that the video, in some ways more than others, is clearly *not* intended for. I had to bring it to one user's attention through another video that was not marked as such to remove the marker, as it was clearly not intended as such. (creator; r/youtube)*

Without classification explanations, the creator disagreed that the videos labeled by the platform should be classified as MFK. Then, they felt compelled to draw online traffic of older consumers to MFK videos to remove the labels because they did not want the negative impacts of the labels. Some creators even mentioned their channel-level efforts:

> *My original channel was going to be a kid channel. I did more research into monetizing made-for-kids videos. If I were monetized, I'd make little to no money at all, and it absolutely wouldn't be worth it. I'd keep up with educating parents but make it geared toward the parent. (creator; r/NewTubers)*

This creator above performed the extra labor by changing their whole channel's content category to avoid MFK classification because of the low and unpromising profitability of content creation associated with it.

5.2.3 Reduced Motivation for Creation. MFK classification often demotivates creators from creating content. A creator highlighted the challenge of establishing a kid's channel in the unpromising profitability and fanbase: *"Unless it's a hobby or an insanely huge channel, I don't see the point of having a kid's channel. No way to get one off the ground in this environment."*

The ambiguity of MFK policies further discourages creators. A creator posted:

> *This foggy space between what we are and are not allowed to do disheartened and uninspired me to start uploading anything at all. Some of the things I would like to make would be safe for all. Some would only be safe for a more mature crowd. Do I walk the tightrope and throw some truly unique and fun ideas out the window in case I tread in the wrong territory or make a claim that a bot deems incorrect? "For kids" does not mean the same thing as "safe for kids." (creator; r/youtube)*

This case showed that creators clearly understood the difference between MFK and "safe for kids." They were afraid that the YouTube platform's algorithms would mix up these concepts and misclassify their "safe for kids" videos as MFK. Due to this fear, the creator hesitated and considered not investing more creativity in their content creation.

5.3 Intersection of MFK Classification with Other Platform Designs

MFK classification is not mutually exclusive with other platform designs. Instead, creators and consumers find it intersected with other designs, especially other classification systems, in three ways: coordination, inconsistency, and conflict. While these intersections sometimes align with protecting children's privacy, creators and consumers often observe them as failing to do so or even negatively impacting user interests.

5.3.1 Coordination. Coordination refers to how MFK classification does not work alone but works with other classification systems for child privacy protection. A viewer posted: *"How do I get my features back as a watcher, not an uploader (creator)? I am not a kid and would like to use all the features of YouTube."* This viewer disagreed that such coordination between video function disablement and MFK classification should be applied to their content consumption experience as an adult. Besides consumer experience, creators also found that MFK classification coordinates with other content moderation classifications to influence their MFK classification decisions. For example, a creator discussed with a viewer:

> ***Creator**: My Hot Wheels review channel is for adult collections but is completely family-friendly. But because they are a "kid toy" that appeals to kids (according to MFK policies), I'm probably going to have to label them "made for kids," and therefore, what is the point of trying to grow a channel which is never going to amount to anything.*
> ***Consumer**: Start swearing?*
> ***Creator**: I would do that, but it's in a kid-friendly game because if I curse, they would flag my video or terminate my channel. (r/youtube)*

When following MFK policies to classify a video as MFK, the creator in the above case was more likely to receive less income and audience engagement from the video. But if the creator cursed in a video and thus labeled it as non-MFK, they would violate other content policies (e.g., community guideline [96]) beyond the MFK one, to lose videos or even the whole channel. Thus, the creator weighed the risk of content removal higher than the limited growth due to MFK classification, highlighting the difficult position creators are in.

Creators further voiced dissatisfaction about how data tracking coordinates and is tied to content moderation:

5.3.2 Inconsistency. Inconsistency means how the MFK classification works with other designs in a way inconsistent with what is stated in the MFK moderation policies [91, 94]. For example, a newbie creator posted: "What's even weirder is that when I set a video as made for kids on the first video that I made, it still allows comments." As comment disablement is a designed change given the MFK policy [94], the creator here felt surprised that it did not work in the designed way. This posed the risk of collecting behavioral data of potential kid consumers.

Such inconsistency also appears in non-MFK videos. For example, a creator shared:

> *My friend was trying to turn notifications on for my channel and got a warning saying, "This action is turned off for content made for kids." But the thing is, I never selected a video or my channel for kids. Why does it happen, and how to fix it? (consumer; r/youtube)*

"Notification" in this example refers to the bell icon beside the YouTube channel, which can notify consumers of new videos, and YouTube turns it off on MFK channels according to the policies [94]. The inconsistency existed in two phases of MFK classification. First, in its decision-making phase, if the creator assumed their channel was non-MFK, then YouTube's notification bell worked inconsistently on their channel. Second, in the sense-making phase of MFK classification results, the notification bell as a notification system only notified the consumers of MFK videos they watched but did not notify the creators who created these videos. This was inconsistent with MFK policy, where YouTube notifies creators of MFK videos that are classified by the platform [94].

Suppose the inconsistency in the last case is potentially attributed to creators' lack of awareness that their videos are classified as MFK by the platform algorithms. In that case, other creators flag the explicit inconsistency when they already classify their channels as non-MFK. For example, a creator mentioned: "I changed the setting on my channel to a hard not made for kids and set all my videos to that as well, but the problem persisted."

A creator further highlighted MFK classification's inconsistency with monetization algorithms/classifications:

> *I have several videos that were manually changed to "for kids" after I had published them. Interestingly, their CPM only dropped by 1/3. On the other hand, I have published videos as "for kids," and their CPM is 1/4 of what a normal video would be. (creator; r/PartneredYoutube)*

As MFK policies [94] explain, there will be no targeted ads on MFK videos, meaning only contextual ads will be placed, generating lower ad income. However, the creator above recognized inconsistent ad income performances between MFK and non-MFK videos, indicated by CPM (the net amount of ad income for every 1,000 ad impressions). Such inconsistency between MFK classification and monetization algorithms further implied uncertainty about whether YouTube's ad placement system (i.e., how YouTube places ads) consistently worked with MFK policy enforcement.

5.3.3 Conflict. Conflict arises when the MFK classification operates simultaneously with other platform designs, especially other classifications, which compromises child privacy and safety protection – what the MFK classification intends to implement. For example, a creator mentioned:

> *Because people under 13 can still go on the main site. If they (YouTube) made a 13+ requirement, and not logging in with a mature account could get you on the kid's website, that would be perfectly fine. It kind of sucks right now (with MFK classification on the main site). (creator; r/youtube)*

This creator discussed two designs about the consumers on YouTube sequentially. First is an open-access consumer model where most YouTube videos are publicly accessible to users without registering or logging in with an account. The second is the age verification classification, where the consumers need to be over 13 years old to register an account on YouTube. So, the conflict in the above case was that the open-access consumer model allowed potential kids to watch videos on YouTube without an account, while the MFK classification intended to prevent potential kids from accessing non-MFK videos and getting their data tracked by the platform.

Such conflict is not rare. A creator discussed the paradox of a video being simultaneously MFK and age-restricted:

> *Age restriction classification has not changed; it still means and does the same thing. Made For Kids classification now only means what roughly COPPA intended was to stop data collection and surgical ad targeting aimed at kids. Now, since You are aware that video should be restricted (due to language), You might worry that kids will watch it (since it's known as a cartoon), and You will get in trouble with COPPA. (creator; r/youtube)*

Age restriction refers to a binary classification that creators need to practice, and that can indicate whether their videos are only for people over 18; otherwise, the YouTube platform will label the videos on behalf of creators [90]. Here, the creator grappled with the conflict between age restrictions meant for adult content and MFK regulations designed to protect children, potentially endangering the intended consumers' safety.

MFK classification intends to prevent consumers' data on MFK videos from being collected by the platform, and no targeted ads will be placed [92]. However, this intent is conflicted with how YouTube places the ads. A viewer posted:

> *They probably still collect data from clicking ads [on MFK videos], though, like those ones in the up-next feed that are designed to look like kids' videos but, when*

clicked, will take your child off to a third-party website. (consumer; r/youtube)

This viewer mentioned a possibility where kids' data can be collected by the ads they watched and clicked on. This conflicted with children's privacy protection. Beyond content itself, MFK policies did not consider how ad placement impacts potential kids, not to mention that sometimes the ads are harmful, as a viewer said: *"'try out a Slavic wife, see what happens' might be a little inappropriate for children."*

"YouTube Kids," a separate platform from the regular, main YouTube for consumers under the age of 13, is also perceived to be repetitive with MFK classification. For example, a creator elaborated, *"There is an app called YouTube Kids. If they can't handle two apps at once, then shut down YouTube Kids because (with made for kids) they clearly want kids and everyone else to use the normal YouTube."* This creator implied that MFK classification and the YouTube Kids platform should be integrated so they would not be confused about which one was for child privacy protection.

Consumers also notice such conflict. For example, a consumer noticed kid creators who seemed to be under 13 were active on YouTube: *"Technically, it's not allowed, but YouTube doesn't ban those under 13 creators for some reason. (You can find so many under 13 creators on YouTube)."* That meant MFK classification was potentially practiced by those under 13, and thus, their data might be collected from the platform as both a creator and viewer.

6 DISCUSSION

The YouTube platform initially proposed MFK classification and its associated moderation policies [91, 92, 94] to respond to the allegations from and legal tension with FTC about violating children's online privacy [33]. Although both FTC and YouTube believe MFK classification can alleviate this legal tension, it unexpectedly creates new tension among the platform, creators, and consumers, complicating the efforts for child safety. That means, while the MFK classification has helped prevent the collection of children's personal information—enhancing child privacy—it has also inadvertently exposed children to broader safety risks, including inappropriate video content and advertising. This section thus will discuss how the YouTube platform positions MFK classification in an interwoven network of classification systems. Using and expanding on Bowker and Star's classification theories [16], we will unpack this network's structure and impacts on child safety. This section will close with design principles for child-centered safety.

6.1 An Interwoven Network of Multiple Classification Systems and Its Broad Impacts on Child Safety

Bowker and Star have performed an extensive analysis of single and static classification systems through examples like apartheid's racial classification in South Africa [16]. However, our analysis diverges in a significant aspect: Digital and social media platforms like YouTube design a complex interplay of multiple, dynamic classifications. On YouTube, MFK classification intersects with other classifications like advertising restrictions and monetization algorithms, influencing content creation, monetization, and audience consumption patterns. These patterns, in turn, influence platform algorithms and trends, showing that YouTube's classification systems are interconnected.

Against this background, our findings shed light on a designed, interwoven network of classification systems that operate interdependently, in tension, and dynamically. First, the classifications on YouTube – age-restriction, MFK, and age verification are not standalone but directly impact child safety (see blue arrows in Figure 3). As Section 5.3 shows, age-restriction (i.e., content classification for consumers only over 18) was labeled concurrently with MFK on videos, leaving creators uncertain if the goal is to limit viewership to adults over 18 or to protect children by disabling data tracking. This is compounded by YouTube's open-access consumer model, which allows content viewing without an account, thus obscuring the presence of viewers under 13. This critical child safety concern undermines age verification on YouTube, which approves users to be on YouTube if they are over 13. The interdependence of these classifications often goes unnoticed by creators until it negatively affects creators' content, monetization, and audience engagement, resonating with Bowker and Star's concept of infrastructural inversion [16], where the underlying classifications only become visible during conflicts or breakdowns. Similarly, extending prior HCI research on such classification breakdowns [14, 44, 78], our study brings to light not only the classification at work but also the route from its structure to impacts, compounding the experiences of users, including creators, consumers, and children.

Second, classification systems on YouTube pull different entities' contention (see red arrows in Figure 3), including YouTube, creators, and consumers. Bowker and Star highlight the concept of boundary objects, referring to "objects that both inhabit several communities of practice and satisfy the informational requirements of each of them," which manages the tensions among diverse perspectives [16]. MFK classification is a boundary object: It is meant to protect children's privacy but is interpreted diversely. Our findings (e.g., Section 5.2) show that creators interpreted it as a negotiation between their vested interests and external requirements from YouTube or FTC. Parents, however, felt it limited their agency in content selection for their kids (e.g., Section 5.1.2), while YouTube's algorithms used this label to curate content, including ads, for viewers (e.g., Section 5.1.1). This multiplicity reflects the engagement of diverse users with children's online privacy measures, as seen in prior HCI research (e.g., children [55, 82], developers [31]. Our study further underscores a scale challenge: Engagement with YouTube's network of classifications extends beyond mere videos and MFK labels to a broader array of policies (e.g., community guidelines [96], algorithms, and interfaces (e.g., YouTube Studio [93]), forming what Bowker and Star conceptualized as a boundary infrastructure, "objects that cross larger levels of scale than boundary objects" [16]. As creators assign the MFK labels, they do not merely classify a video but interact with YouTube's entire classification ecosystem, affecting everything from video uploads to content recommendations and ad placements. This again shows how a singular classification, when entwined with others, can have profound implications for users, particularly children.

Third, classification systems operate dynamically on YouTube (see green circle arrows in Figure 3). Drawing on Bowker and Star's concept, infrastructural inversion [16], where classification systems are reformed in response to breakdowns, we observe, on YouTube,

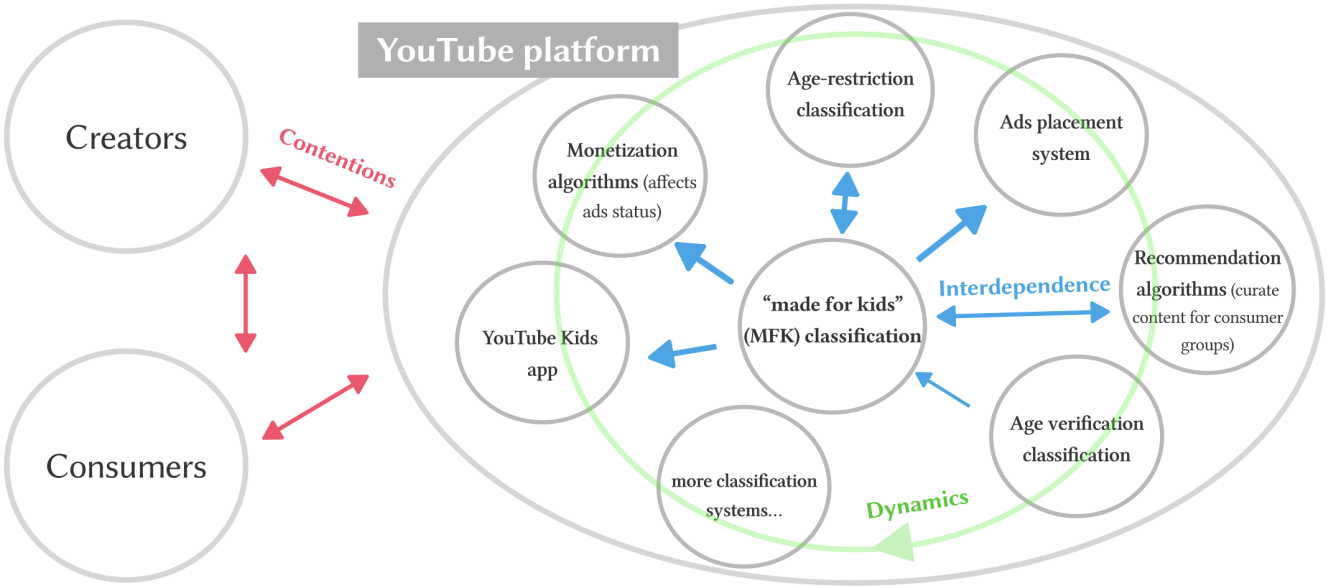


Figure 3: An interwoven network of classification systems impacting child safety on YouTube is indicated by our findings. Blue arrows show interdependence between classifications, red arrows highlight contention among creators, the platform, and consumers, while green arrows show how classifications transition within the network. Double-headed arrows denote bidirectional relationships. For instance, red arrows indicate that consumers note malicious ads placed by the platform, but meanwhile, they cannot comment on MFK videos. Bolder blue arrows signify stronger relationships (e.g., MFK classification mutually affects age-restriction criteria).

that the breakdowns don't necessarily bring classification modifications. Instead, different entities navigate breakdowns through the existing classification network: Platform algorithms correct or override classifications like MFK and age restrictions (e.g., Section 5.3.2), creators make or reverse their mislabeling (e.g., Section 5.2.2), and consumers point out different types of misclassifications (e.g., Section 5.1.1). This also differs from prior work, where classifications adapted to nuanced user behaviors (e.g., updated moderation policies [27, 88]), classifications on YouTube dynamically shift, often transitioning from one label to another over time, while the underlying network of classifications may remain unchanged.

The complexities within this classification network could first risk child privacy. Previous work has highlighted the design inadequacies of social media in adhering to child privacy laws like COPPA (e.g., [37, 74]), noting parents' roles in circumventing age checks [17]. YouTube's MFK classification, which involves creators directly in classifying content for child consumers, contrasts traditional age verification or filtered browsing [95]. The open-access consumer model of YouTube further complicates child privacy efforts, allowing unaccounted viewership, including by children. This makes it nearly impossible for creators to identify if a viewer is a child. Although creators lack data on consumers under 13 on the YouTube Studio dashboard [93], MFK policies [91, 94] still require them to label content based on its primary intention for this age group. This inconsistency, coupled with both the platform's and creators' drive to maximize growth (e.g., visibility and income), makes child privacy protection more challenging.

The MFK classification network can further expose children to safety issues. Prior research has assessed moderation effectiveness – how it restricts the proliferation of inappropriate content (e.g., [22, 79, 84]). Our study highlights a more critical concern: When YouTube relies on content creators to classify their content for moderation [94], it is evident that they aren't professional moderators or labelers, often leading to errors and repeated relabeling, increasing the exposure of problematic videos and advertisements. Our findings thus reveal a fundamental flaw in merging child privacy protection with moderation – making content an indicator for consumer demographics. As the MFK misclassifications intersect with other labels, coupled with children's unpredictable content consumption, this flaw extends the original focus of MFK classification on children's data privacy into child safety on YouTube.

6.2 Designing for Child-Centered Online Safety in the Network of Classification Systems

On social media platforms like YouTube, children's online safety involves different stakeholders, such as content creators, consumers, platform designers, and policymakers. However, our findings reveal a divergence in understanding and approaching child safety across these groups. Content creators and consumers, particularly, face discrepancies in how they experience content creation and consumption: They wanted to weigh the moderation challenges and impacts posed by MFK classification with the necessity for them to protect children's privacy. Besides, as prior work in legal and HCI fields has criticized platforms for inconsistent enforcement of

moderation policies (e.g., [44, 77, 85, 88]), our study supplements how such inconsistency originates: Platform designs work differently from what moderation policies state, unpredictably intersect with other designs, and thus undermine the force of moderation in regulating inappropriate content, including ads. Especially when the MFK classification enforces policies and operates, we found it ignores the design of parental consent and participation, showing a discrepancy that extends beyond platform governance issues to policy gaps between platform policymaking and COPPA [24].

Thus, we highlight two design principles that are key to enhancing child-centered online safety:

The first design principle is the multi-stakeholder principle in child safety. This entails giving visibility to both stakeholders who directly get involved in child safety and those who play an indirect role in it. On the one hand, in platforms' content moderation, the classification work is often invisible [20]. Our findings show that this invisibility can obscure the efforts of content creators for child safety. Bowker and Star highlight the critical need for visibility in classification systems—not only to understand and recognize the work that goes into them but also to critically examine their impacts [16]. When it comes to MFK policies [91], it thus means bringing to light the invisible labor that underpins child safety within the classification network where MFK classification is part.

On the other hand, a singular safety design needs to acknowledge the influence of multiple stakeholders. Prior DIS researchers have examined parental use of AI-assisted or technology-based decision-making [52, 54] or how children themselves use chatbots [72] to enhance safety. Our findings show that while the YouTube platform offers MFK classification as an important child safety function, there's a noticeable gap in the willingness and ability of consumers and creators to participate in MFK classification. **Design implication**: Thus, platforms like YouTube should support creator-consumer collaboration to positively influence child safety and avoid biases when implementing protection measures for children.

The second design principle is the systems thinking principle in child safety. Systems thinking refers to "seeing interrelationships rather than things, for seeing patterns rather than static snapshots" [80]. It emphasizes making sense of the interconnectedness and patterns of change within a system rather than viewing parts in isolation [80]. In design, our study informs two aspects of this notion: The internal aspect, which speaks to the interaction among system parts/components, and the contextual aspect, which is how the system or its parts interact with the world, such as people.

On the one hand, our study shows the necessity of acknowledging the interconnectedness among safety designs like MFK classification and other classification systems. For example, our findings showed that MFK classification, as one type of safety design, was interwoven with other platform classifications, such as monetization algorithms, advertising settings [21, 61, 62], and age verification [37]. These connections can increase the risk of exposing child consumers to safety issues. **Design implication**: To solve these issues, we do not suggest that MFK classification should operate in isolation. Rather, in policy enforcement, the interwoven network of classification systems should enhance transparency to users regarding the distinctions among various classifications. This approach enables different stakeholders to examine if child safety designs align with COPPA, preventing child data collection without parental consent. This, thus, positions users in a fair environment for content creation and consumption without unexpected impacts from other classifications.

On the other hand, our study highlights the potential for innovating safety designs to enhance child protection effectively. **Design implication**: While age verification systems typically confirm a consumer's age during the account registration stage, we suggest they should also periodically verify ages in the content consumption stage, especially when there's a significant shift in consumption patterns or an increase in child-oriented content consumption. Implementing such a safety design could hold platforms more accountable for child safety, given our findings, where creators on YouTube would not be aware if children under 13 consume their content due to the open-access content consumption model.

Expanding on the interaction among multiple stakeholders with safety designs, such designs should facilitate collaborative practices of child safety. Our study found that although engaged viewers identified mislabeling in MFK classifications, they lacked a mechanism for reporting these issues and stopping the spread of harmful content. Additionally, our findings reveal that many parents are excluded from participating in MFK classification or selecting content for their children. We thus propose the below three **design changes**:

- Social media platforms like YouTube should enhance their flagging options to enable the reporting of perceived MFK misclassifications.
- Furthermore, platforms should provide creators with educational resources, including contact points and workshops, to ensure their content aligns with MFK policies and avoid mislabeling.
- Informed by prior literature that advocates for co-using digital devices [28, 46] between parents and children, we propose introducing a co-watching feature on platforms like YouTube, encouraging both parties to decide if they wish to consume videos together in real time.

7 LIMITATION AND FUTURE WORK

This study has a few inevitable limitations. First, a few prior HCI studies have investigated children's privacy-related experiences on several social media platforms (e.g., [82, 98]). That means children and parents could be a future focus in understanding how they experience MFK classification on YouTube or similar child privacy-preventing technologies across platforms. Similarly, future work can further focus on child or teenager creators and how they experience privacy-preventing designs on platforms like YouTube. Second, the method we used, analyzing online discussions, might be subjected to the misreported experiences with MFK classification on Reddit. For example, there might be creators' mislabeling of content when they did not share such experiences as mislabeling in our data. Recognizing this potential limitation, we will dive deeper into future work with parents, kids, creators, and consumers through methods like participatory design workshops. Also, as YouTube's MFK is heavily influenced by COPPA regardless of region [92], we recognize future work that can delve into a localized understanding

of child privacy in places complying with GDPR or age-appropriate design.

8 CONCLUSION

This study delves into creators' and consumers' experiences with YouTube's MFK classification system, focusing on the broad implications of its implementation. We uncover a spectrum of user reactions, strategies, and challenges in navigating the MFK system. Our findings contribute to a deeper understanding of MFK as more than a technical measure: We identify an interwoven network of classification systems centered on MFK classification. Using and extending Bowker and Star's classification theories, we unpack how such a network challenges child safety (e.g., content moderation effectiveness, data tracking, malicious ad placement). We conclude by laying out the design principles of child-crenated safety on commercial and social media platforms like YouTube.

ACKNOWLEDGMENTS

We thank the associate chairs and anonymous reviewers for their insightful feedback and suggestions. This work is partially supported by the NSF, under grant no. 2326505.

REFERENCES

[1] Amelia Acker and Leanne Bowler. 2018. Youth Data Literacy: Teen Perspectives on Data Created with Social Media and Mobile Devices. In *Proceedings of the Annual Hawaii International Conference on System Sciences*. 1923–1932. https://doi.org/10.24251/HICSS.2018.243

[2] Zainab Agha, Karla Badillo-Urquiola, and Pamela J. Wisniewski. 2023. "Strike at the Root": Co-designing Real-Time Social Media Interventions for Adolescent Online Risk Prevention. *Proc ACM Hum Comput Interact* 7, CSCW1 (April 2023), 149. https://doi.org/10.1145/3579625

[3] Syed Hammad Ahmed, Muhammad Junaid Khan, H. M. Umer Qaisar, and Gita Sukthankar. 2023. Malicious or Benign? Towards Effective Content Moderation for Children's Videos. In *Proceedings of the International Florida Artificial Intelligence Research Society Conference, FLAIRS 36*. https://doi.org/10.32473/flairs.36.133315

[4] Davey Alba. 2015. Google Launches "YouTube Kids," a New Family-Friendly App. https://www.wired.com/2015/02/youtube-kids/

[5] Sultan Alshamrani, Ahmed Abusnaina, Mohammed Abuhamad, Daehun Nyang, and David Mohaisen. 2021. Hate, Obscenity, and Insults: Measuring the Exposure of Children to Inappropriate Comments in YouTube. In *The Web Conference 2021 - Companion of the World Wide Web Conference, WWW 2021*. 508–515. https://doi.org/10.1145/3442442.3452314

[6] Tawfiq Ammari, Priya Kumar, Cliff Lampe, and Sarita Schoenebeck. 2015. Managing children's online identities: How parents decide what to disclose about their children online. In *Conference on Human Factors in Computing Systems - Proceedings*. 1895–1904. https://doi.org/10.1145/2702123.2702325

[7] Tawfiq Ammari and Sarita Schoenebeck. 2015. Understanding and supporting fathers and fatherhood on social media sites. In *Conference on Human Factors in Computing Systems - Proceedings*. 1905–1914. https://doi.org/10.1145/2702123.2702205

[8] Mary Jean Amon, Nika Kartvelishvili, Bennett I. Bertenthal, Kurt Hugenberg, and Apu Kapadia. 2022. Sharenting and Children's Privacy in the United States: Parenting Style, Practices, and Perspectives on Sharing Young Children's Photos on Social Media. *Proc ACM Hum Comput Interact* 6, CSCW1 (April 2022). https://doi.org/10.1145/3512963

[9] National Archives. 2013. PART 312—Children's Online Privacy Protection Rule. https://www.ecfr.gov/current/title-16/chapter-I/subchapter-C/part-312

[10] Karla Badillo-Urquiola, Diva Smriti, Brenna Mcnally, Evan Golub, Elizabeth Bonsignore, and Pamela J Wisniewski. 2019. "Stranger Danger!" Social Media App Features Co-designed with Children to Keep Them Safe Online. In *Proc 18th ACM Int Conf Interact Des Child*. https://doi.org/10.1145/3311927

[11] Emmanuelle Bartoli. 2010. Children's data protection vs marketing companies. *International Review of Law, Computers & Technology* 23, 1-2 (January 2010), 35–45. https://doi.org/10.1080/13600860902742612

[12] Stephen Beemsterboer. 2020. COPPA killed the video star: How the YouTube settlement shows that COPPA does more harm than good. https://publish.illinois.edu/illinoisblj/files/2020/06/12-Stephen-COPPA.pdf

[13] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 10540. 405–415. https://doi.org/10.1007/978-3-319-67256-4_32

[14] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and its consequences for online harassment: Design insights from HeartMob. *Proc ACM Hum Comput Interact* 1, CSCW (November 2017), 1–19. https://doi.org/10.1145/3134659

[15] Lindsay Blackwell, Jean Hardy, Tawfiq Ammari, Tiffany Veinot, Cliff Lampe, and Sarita Schoenebeck. 2016. LGBT parents and social media: Advocacy, privacy, and disclosure during shifting social movements. In *Conference on Human Factors in Computing Systems - Proceedings*. 610–622. https://doi.org/10.1145/2858036.2858342

[16] Geoffrey C. Bowker and Susan Leigh Star. 2000. *Sorting things out: Classification and its consequences*. MIT press.

[17] Danah Boyd, Eszter Hargittai, Jason Schultz, and John Palfrey. 2011. Why parents help their children lie to Facebook about age: Unintended consequences of the "Children's Online Privacy Protection Act". https://firstmonday.org/ojs/index.php/fm/article/download/3850/3075

[18] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qual Res Psychol* 3, 2 (January 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[19] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qual Res Sport Exerc Health* 11, 4 (August 2019), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

[20] Jie Cai, Donghee Yvette Wohn, and Mashael Almoqbel. 2021. Moderation visibility: Mapping the strategies of volunteer moderators in live streaming micro communities. In *IMX 2021 - Proceedings of the 2021 ACM International Conference on Interactive Media Experiences*. 61–72. https://doi.org/10.1145/3452918.3458796

[21] Robyn Caplan and Tarleton Gillespie. 2020. Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy. *Soc Media Soc* 6, 2 (2020). https://doi.org/10.1177/2056305120936636

[22] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29, 4 (March 2022). https://doi.org/10.1145/3490499

[23] Stuart Cobb. 2020. It's COPPA-Cated: Protecting Children's Privacy in the Age of YouTube. https://heinonline.org/HOL/Page?handle=hein.journals/hulr58&id=997&div=&collection=

[24] Federal Trade Commission. 1998. Children's Online Privacy Protection Rule ("COPPA"). https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa

[25] Federal Trade Commission. 2019. Musical.ly, Inc. https://www.ftc.gov/news-events/news/press-releases/2019/02/video-social-networking-app-musically-agrees-settle-ftc-allegations-it-violated-childrens-privacy

[26] Federal Trade Commission. 2023. FTC Proposes Blanket Prohibition Preventing Facebook from Monetizing Youth Data. https://www.ftc.gov/news-events/news/press-releases/2023/05/ftc-proposes-blanket-prohibition-preventing-facebook-monetizing-youth-data

[27] MacKenzie F. Common. 2020. Fear the Reaper: how content moderation rules are enforced on social media. *International Review of Law, Computers & Technology* 34, 2 (May 2020), 126–152. https://doi.org/10.1080/13600869.2020.1733762

[28] Sabrina L. Connell, Alexis R. Lauricella, and Ellen Wartella. 2015. Parental Co-Use of Media Technology with their Young Children in the USA. *J Child Media* 9, 1 (2015), 5–21. https://doi.org/10.1080/17482798.2015.997440

[29] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media Soc* 18, 3 (March 2016), 410–428. https://doi.org/10.1177/1461444814543163

[30] Dipto Das, Carsten Østerlund, and Bryan Semaan. 2021. "Jol" or "Pani"?: How Does Governance Shape a Platform's Identity?. In *Proc ACM Hum Comput Interact*, Vol. 5. https://doi.org/10.1145/3479860

[31] Anirudh Ekambaranathan and Jun Zhao. 2021. Money makes the world go around: Identifying barriers to beter privacy in children's apps from developers' perspectives. In *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3411764.3445599

[32] Facebook. 2019. Facebook Community Standards. https://transparency.fb.com/policies/community-standards/

[33] Federal Trade Commission. 2019. Google and YouTube Will Pay Record $170 Million for Alleged Violations of Children's Privacy Law. https://www.ftc.gov/news-events/news/press-releases/2019/09/google-youtube-will-pay-record-170-million-alleged-violations-childrens-privacy-law

[34] Gavin Feller and Benjamin Burroughs. 2021. Branding Kidfluencers: Regulating Content and Advertising on YouTube. *Television & New Media* 23, 6 (October 2021), 575–592. https://doi.org/10.1177/15274764211052882

[35] Yang Feng and Wenjing Xie. 2014. Teens' concern for privacy when using social networking sites: An analysis of socialization agents and relationships with privacy-protecting behaviors. *Comput Human Behav* 33 (April 2014), 153–162. https://doi.org/10.1016/J.CHB.2014.01.009

[36] Jessica L. Feuston, Alex S. Taylor, and Anne Marie Piper. 2020. Conformity of Eating Disorders through Content Moderation. *Proc ACM Hum Comput Interact* 4, CSCW1 (May 2020). https://doi.org/10.1145/3392845

[37] Shannon Finnegan. 2019. How Facebook Beat the Children's Online Privacy Protection Act: A Look into the Continued Ineffectiveness of COPPA and How to Hold Social Media Sites Accountable in the Future. https://heinonline.org/HOL/Page?handle=hein.journals/shlr50&id=838&div=&collection=

[38] Jeremy Gan. 2023. YouTube reportedly dominates competition as top social media platform for children. https://www.dexerto.com/youtube/youtube-reportedly-dominates-competition-as-top-social-media-platform-for-children-2264857/

[39] GDPR. 2018. General Data Protection Regulation (GDPR). https://gdpr-info.eu/

[40] Arup Kumar Ghosh, Karla Badillo-Urquiola, Shion Guha, Joseph J. Laviola, and Pamela J. Wisniewski. 2018. Safety vs. surveillance: What children have to say about mobile apps for parental control. In *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3173574.3173698

[41] Cami Goray and Sarita Schoenebeck. 2022. Youths' Perceptions of Data Collection in Online Advertising and Social Media. *Proc ACM Hum Comput Interact* 6, CSCW2 (November 2022). https://doi.org/10.1145/3555576

[42] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data Soc* 7, 1 (January 2020). https://doi.org/10.1177/2053951719897945

[43] Greg Guest, Emily Namey, and Mario Chen. 2020. A simple method to assess and report thematic saturation in qualitative research. *PLoS One* 15, 5 (May 2020). https://doi.org/10.1371/JOURNAL.PONE.0232076

[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. In *Proc ACM Hum Comput Interact*, Vol. 5. https://doi.org/10.1145/3479610

[45] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving Google's Perspective API Built for Detecting Toxic Comments. https://arxiv.org/abs/1702.08138v1

[46] Isil Oygur Ilhan, Yunan Chen, and Daniel A. Epstein. 2023. Co-designing for the Co-Use of Child-Owned Wearables. In *Proceedings of IDC 2023 - 22nd Annual ACM Interaction Design and Children Conference: Rediscovering Childhood*. 603–607. https://doi.org/10.1145/3585088.3593868

[47] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did you suspect the post would be removed?": Understanding user reactions to content removals on reddit. *Proc ACM Hum Comput Interact* 3, CSCW (November 2019), 1–33. https://doi.org/10.1145/3359294

[48] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction* 26, 5 (July 2019), 1–35. https://doi.org/10.1145/3338243

[49] Shagun Jhaver, Quan Ze Chen, Detlef Knauss, and Amy Zhang. 2022. Designing Word Filter Tools for Creator-led Comment Moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3491102.3517505

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

[51] Garrett Johnson, Tesary Lin, James C. Cooper, and Liang Zhong. 2024. COPPAcalypse? The Youtube Settlement's Impact on Kids Content. *SSRN Electronic Journal* (March 2024). https://doi.org/10.2139/SSRN.4430334

[52] Anna Kawakami, Venkatesh Sivaraman, Logan Stapleton, Hao Fei Cheng, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. "Why Do I Care What's Similar?" Probing Challenges in AI-Assisted Child Welfare Decision-Making through Worker-AI Interface Design Concepts. In *DIS 2022 - Proceedings of the 2022 ACM Designing Interactive Systems Conference: Digital Wellbeing*. 454–470. https://doi.org/10.1145/3532106.3533556

[53] Sara Kingsley, Proteeti Sinha, Clara Wang, Motahhare Eslami, and Jason I. Hong. 2022. "Give Everybody a Little Bit More Equity": Content Creator Perspectives and Responses to the Algorithmic Demonetization of Content Associated with Disadvantaged Groups. *Proc ACM Hum Comput Interact* 6, CSCW2 (November 2022). https://doi.org/10.1145/3555149

[54] Susanne Kirchner, Dawn K. Sakaguchi-Tang, Rebecca Michelson, Sean A. Munson, and Julie A. Kientz. 2020. This just felt to me like the right thing to do": Decision-Making Experiences of Parents of Young Children. In *DIS 2020 - Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 489–503. https://doi.org/10.1145/3357236.3395466

[55] Priya Kumar, Shalmali Milind Naik, Utkarsha Ramesh Devkar, Marshini Chetty, Tamara L. Clegg, and Jessica Vitak. 2017. "No Telling Passcodes Out Because They're Private.". *Proc ACM Hum Comput Interact* 1, CSCW (December 2017). https://doi.org/10.1145/3134699

[56] Priya Kumar and Sarita Schoenebeck. 2015. The modern day baby book: Enacting good mothering and stewarding privacy on facebook. In *CSCW 2015 - Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing*. 1302–1312. https://doi.org/10.1145/2675133.2675149

[57] Cliff Lampe and Paul Resnick. 2004. Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04*. ACM Press, New York, New York, USA.

[58] Tianshi Li, Elizabeth Louie, Laura Dabbish, and Jason I. Hong. 2021. How Developers Talk About Personal Data and What It Means for User Privacy. *Proc ACM Hum Comput Interact* 4, CSCW3 (January 2021), 1–28. https://doi.org/10.1145/3432919

[59] Sonia Livingstone, Mariya Stoilova, and Rishita Nandagiri. 2019. Children's data and privacy online: growing up in a digital age: an evidence review. http://www.lse.ac.uk/my-privacy-uk

[60] Renkai Ma and Yubo Kou. 2021. "How advertiser-friendly is my video?": YouTuber's Socioeconomic Interactions with Algorithmic Content Moderation. *PACM on Human Computer Interaction* 5, CSCW2 (2021), 1–26. https://doi.org/10.1145/3479573

[61] Renkai Ma and Yubo Kou. 2022. "I am not a YouTuber who can make whatever video I want. I have to keep appeasing algorithms": Bureaucracy of Creator Moderation on YouTube. https://doi.org/10.1145/3500868.3559445

[62] Renkai Ma and Yubo Kou. 2022. "I'm not sure what difference is between their content and mine, other than the person itself": A Study of Fairness Perception of Content Moderation on YouTube. *Proc ACM Hum Comput Interact* 6, CSCW2 (2022), 28. https://doi.org/10.1145/3555150

[63] Emily McReynolds, Sarah Hubbard, Timothy Lau, Aditya Saraf, Maya Cakmak, and Franziska Roesner. 2017. Toys that listen: A study of parents, children, and internet-connected toys. In *Conference on Human Factors in Computing Systems - Proceedings*. 5197–5207. https://doi.org/10.1145/3025453.3025735

[64] Kathryn C. Montgomery, Jeff Chester, and Tijana Milosevic. 2017. Children's Privacy in the Big Data Era: Research Opportunities. *Pediatrics* 140 (November 2017), S117–S121. https://doi.org/10.1542/PEDS.2016-1758O

[65] Carol Moser, Tianying Chen, and Sarita Y. Schoenebeck. 2017. Parents' and children's preferences about parents sharing about children on social media. In *Conference on Human Factors in Computing Systems - Proceedings*. 5221–5225. https://doi.org/10.1145/3025453.3025587

[66] Helen Nissenbaum. 2004. Privacy as Contextual Integrity. *Washington Law Review* 79 (2004). https://heinonline.org/HOL/Page?handle=hein.journals/washlr79&id=129&div=16&collection=journals

[67] Anna O'Donnell. 2020. Why the VPPA and COPPA Are Outdated: How Netflix, YouTube, and Disney Can Monitor Your Family at No Real Cost. https://heinonline.org/HOL/Page?handle=hein.journals/geolr55&id=471&div=&collection=

[68] Gwenn Schurgin O'Keeffe, Kathleen Clarke-Pearson, Deborah Ann Mulligan, Tanya Remer Altmann, Ari Brown, Dimitri A. Christakis, Holly Lee Falik, David L. Hill, Marjorie J. Hogan, Alanna Estin Levine, and Kathleen G. Nelson. 2011. The Impact of Social Media on Children, Adolescents, and Families. *Pediatrics* 127, 4 (April 2011), 800–804. https://doi.org/10.1542/PEDS.2011-0054

[69] Luci Pangrazio and Neil Selwyn. 2018. "It's Not Like It's Life or Death or Whatever": Young People's Understandings of Social Media Data. *Social Media and Society* 4, 3 (July 2018). https://doi.org/10.1177/2056305118787808/ASSET/IMAGES/LARGE/10.1177_2056305118787808-FIG1.JPEG

[70] Kostantinos Papadamou, Antonis Papasavva, Savvas Zannettou, Jeremy Blackburn, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Michael Sirivianos. 2020. Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 522–533. https://doi.org/10.1609/ICWSM.V14I1.7320

[71] Jessica A. Pater, Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. 2016. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work*. 369–374. https://doi.org/10.1145/2957276.2957297

[72] Lara Schibelsky Godoy Piccolo, Pinelopi Troullinou, and Harith Alani. 2021. Chatbots to Support Children in Coping with Online Threats: Socio-technical Requirements. In *DIS 2021 - Proceedings of the 2021 ACM Designing Interactive Systems Conference: Nowhere and Everywhere*. 1504–1517. https://doi.org/10.1145/3461778.3462114

[73] Dongxiao Qin. 2016. Positionality. *The Wiley Blackwell Encyclopedia of Gender and Sexuality Studies* (April 2016), 1–2. https://doi.org/10.1002/9781118663219.WBEGSS619

[74] Irwin Reyes, Primal Wijesekera, Joel Reardon, Amit Elazari Bar On, Abbas Razaghpanah, Narseo Vallina-Rodriguez, and Serge Egelman. 2018. "Won't Somebody Think of the Children?" Examining COPPA Compliance at Scale. In *The 18th Privacy Enhancing Technologies Symposium (PETS 2018)*. 63–83. https://doi.org/10.1515/popets-2018-0021

[75] Ivan Rivera. 2019. CRAN - Package RedditExtractoR. https://cran.r-project.org/web/packages/RedditExtractoR/index.html

[76] Sarah T. Roberts. 2019. *Behind the Screen: content moderation in the shadows of social media.*

[77] Barrie Sander. 2019. Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation. *Fordham Int Law J* (2019).

[78] Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis and image labeling services. *Proc ACM Hum Comput Interact* 3, CSCW (November 2019), 33. https://doi.org/10.1145/3359246

[79] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 111–125. https://doi.org/10.1145/2998181.2998277

[80] Peter M. Senge. 1990. *The Fifth Discipline: The art and practice of the learning organization*. Broadway Business. https://books.google.com/books/about/The_Fifth_Discipline.html?id=wg9DG42quXEC

[81] Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community. *Proc ACM Hum Comput Interact* 3, CSCW (November 2019), 163. https://doi.org/10.1145/3359265

[82] Kaiwen Sun and Carlo Sugatan. 2021. They see you're a girl if you pick a pink robot with a skirt: A qualitative study of how children conceptualize data processing and digital privacy risks. In *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3411764.3445333

[83] TikTok. 2019. TikTok for Younger Users. https://newsroom.tiktok.com/en-us/tiktok-for-younger-users

[84] Milo Z. Trujillo, Samuel F. Rosenblatt, Anda Jáuregui Guillermo De, Emily Moog, Briane Paul, V. Samson, Laurent Hébert-Dufresne, and Allison M. Roth. 2021. When the Echo Chamber Shatters: Examining the Use of Community-Specific Language Post-Subreddit Ban. https://doi.org/10.48550/arxiv.2106.16207

[85] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "At the End of the Day Facebook Does What It Wants": How Users Experience Contesting Algorithmic Content Moderation. In *Proceedings of the ACM on Human-Computer Interaction*. 1–22. https://doi.org/10.1145/3415238

[86] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability For Content Moderation. *Proc ACM Hum Comput Interact* 5, CSCW2 (October 2021), 28. https://doi.org/10.1145/3476059

[87] Heather Wilson. 2020. YouTube Is Unsafe for Children: YouTube's Safeguards and the Current Legal Framework Are Inadequate to Protect Children from Disturbing Content. https://heinonline.org/HOL/Page?handle=hein.journals/sjel10&id=237&div=&collection=

[88] Richard Ashby Wilson and Molly K. Land. 2020. Hate Speech on Social Media: Content Moderation in Context. *Conn Law Rev* 52 (2020). https://heinonline.org/HOL/Page?handle=hein.journals/conlr52&id=1056&div=28&collection=journals

[89] Pamela Wisniewski, Heng Xu, Mary Beth Rosson, and John M. Carroll. 2017. Parents just don't understand: Why teens don't talk to parents about their online risk experiences. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. 523–540. https://doi.org/10.1145/2998181.2998236

[90] YouTube. 2023. Age-restricted content. https://support.google.com/youtube/answer/2802167?hl=en

[91] YouTube. 2023. Determining if your content is "made for kids.". https://support.google.com/youtube/answer/9528076?hl=en

[92] YouTube. 2023. Frequently asked questions about "made for kids.". https://support.google.com/youtube/answer/9684541?hl=en#zippy=%2Chow-do-i-know-if-my-content-is-not-made-for-kids

[93] YouTube. 2023. Navigate YouTube Studio. https://support.google.com/youtube/answer/7548152?hl=en

[94] YouTube. 2023. Set your channel or video's audience. https://support.google.com/youtube/answer/9527654?hl=en&ref_topic=9689353&sjid=16427619472020172874-NA#

[95] YouTube. 2023. Your YouTube content and Restricted Mode. https://support.google.com/youtube/answer/7354993?hl=en

[96] YouTube. 2023. YouTube Community Guidelines & Policies. https://www.youtube.com/howyoutubeworks/policies/community-guidelines/

[97] Leah Zhang-Kennedy, Christine Mekhail, Sonia Chiasson, and Yomna Abdelaziz. 2016. From nosy little brothers to stranger-danger: Children and parents' perception of mobile threats. In *Proceedings of IDC 2016 - The 15th International Conference on Interaction Design and Children*. 388–399. https://doi.org/10.1145/2930674.2930716

[98] Jun Zhao, Ge Wang, Carys Dally, Petr Slovak, Julian Edbrooke-Childs, Max Van Kleek, and Nigel Shadbolt. 2019. 'I make up a silly name': Understanding children's perception of privacy risks online. In *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290605.3300336