“Defaulting to boilerplate answers, they didn’t engage in a genuine conversation”: Dimensions of Transparency Design in Creator Moderation

RENKAI MA, Pennsylvania State University, USA
YUBO KOU, Pennsylvania State University, USA

Transparency matters a lot to people who experience moderation on online platforms; much CSCW research has viewed offering explanations as one of the primary solutions to enhance moderation transparency. However, relatively little attention has been paid to unpacking what transparency entails in moderation design, especially for content creators. We interviewed 28 YouTubers to understand their moderation experiences and analyze the dimensions of moderation transparency. We identified four primary dimensions: participants desired the moderation system to present moderation decisions saliently, explain the decisions profoundly, afford communication with the users effectively, and offer repairment and learning opportunities. We discuss how these four dimensions are mutually constitutive and conditioned in the context of creator moderation, where the target of governance mechanisms extends beyond the content to creator careers. We then elaborate on how a dynamic, transparency perspective could value content creators’ digital labor, how transparency design could support creators’ learning, as well as implications for transparency design of other creator platforms.

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

KEYWORDS: Creator moderation; content moderation; YouTuber; moderation design

ACM Reference format:

Renkai Ma and Yubo Kou. 2022. “Defaulting to Boilerplate Answers, they Didn’t Engage in a Genuine Conversation”: Dimensions of Transparency Design in Creator Moderation. In *PACM on Human Computer Interaction,* Vol. 7, CSCW, Article 44, April 2023. ACM, New York, NY, USA. 26 pages. https://doi.org/10.1145/3579477

Author’s addresses: Renkai Ma (renkai@psu.edu) and Yubo Kou (yubokou@psu.edu), College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA, 16802, USA

2573-0142/2023/4 – Article 44 $15.00

https://doi.org/10.1145/3579477

1 INTRODUCTION

To construct safe online environments, online platforms moderate inappropriate content [30], such as harassment [77], terrorism speech [55], or misinformation [38]. In recent years, transparency in moderation system has become a concerning issue as scholars and society at large raise questions about its decision-making and broader ramifications. Transparency refers to the requirement that a sociotechnical system, including humans or non-human agents in it, should be accountable [23] for offering visibility of its complexity [4,22]. Regarding moderation transparency, legal scholars are concerned about how content rules are defined and enforced ambiguously [73,76]; social scientists point out the inherent opacity of moderation empowered by machine learning algorithms [27,29]; and HCI researchers further explore how design could help enhance moderation transparency [37,48,74].

Particularly, prior work has tended to hold that explanation and moderator-user communication as two primary design solutions. Given users' moderation experiences, researchers focused on offering detailed explanations of moderation decision-making [52,56,69] and appeal procedures [20,56,74] to users. Other work focusing on community moderation practices in communities (e.g., subreddits or Twitch [36,41,66]) stressed that human moderators offer communications to users. However, relatively few studies have dug deeper into how different procedures/stages of moderation (e.g., rule articulation, rule enforcement, moderation decision-making, or appeal) may entail different transparency design requirements.

Besides, prior discussion of transparency design is often situated in the context of *content moderation*, where moderation actions such as content removal or account suspension [35,37,56,69,74] are means to manage the appropriateness of speech content such as posts and comments on Reddit or tweets on Twitter. However, due to the platformization and monetization of creative labor [18,45,59], creator platforms such as YouTube and TikTok tend to practice *creator moderation*, where interleaved moderation mechanisms seek to manage not just content but also creators' careers such as visibility [53], identity expressions [6], and revenue [52]. As a result, content creators often encounter a plethora of moderation decisions (e.g., demonetization [9,52], video content shadowban [5,62]). So, investigating moderation transparency dimensions from creators' perspectives could not only describe dimensions of transparency design in moderation procedures but also shed light on unique transparency concerns in creator moderation. In this study, we use transparency design of moderation and dimensions of moderation transparency interchangeably.

We chose YouTube as our study site, the largest creator platform at the time of writing [24,28], and solicited YouTubers' (i.e., video content creators) moderation experiences to uncover dimensions of moderation transparency in creator moderation. YouTube provides a unique scenario of understanding experiences of moderation: it is among the first to allow users to earn advertising (ad) income from their content creation [34]. When YouTube deems a YouTuber violates specific content policies, the moderation system would decrease or remove their future ad income. These economic impacts might affect YouTubers heavily, especially those whose livelihoods rely on content creation on YouTube [10,39,52]. Given the context of creator moderation on YouTube, we aim to understand how YouTubers' moderation experiences reveal transparency designs of the moderation system. We ask:

What dimensions of moderation transparency do YouTubers desire based on their moderation experiences?

To answer this question, we interviewed 28 YouTubers who had experienced creator moderation. Through an inductive qualitative analysis [50,72], we found four critical dimensions of moderation transparency. Participants hoped the moderation system to present moderation saliently, explain moderation profoundly, afford communication effectively, and offer learning opportunities. Based on these findings, we discuss how creator moderation needs to maintain dynamic moderation transparency that both balances transparency efforts and moderation impacts to value YouTubers' labor. We further discuss how transparency design in moderation systems could support YouTubers' learning. Ultimately, we derive implications for transparency design on YouTube and other platforms affording creator monetization.

Our contributions to the moderation literature are multi-fold: (1) we offer fine-grained dimensions of transparency design requirements for creator and content moderation systems; (2) we contributed a conceptual, creator-centered understanding of how different non-human (e.g., chat function) and human agents (e.g., creator support teams, human reviewers) acted as stakeholders in maintaining and improving moderation transparency in different moderation phases; (3) we put forward translatable design considerations of moderation transparency on YouTube and relevant creator platforms.

2 RELATED WORK

In this section, we situate our study in prior work that defines transparency and discusses how the lack of transparency appears in moderation systems and procedures. Then, we introduce approaches that prior work explored or proposed to enhance moderation transparency.

2.1 Transparency Design in Moderation

"Transparency," a term implying openness and communication [70], allows information to be discernible and legible for individuals to "uncover the true essence of a system" [15]. However, transparency is beyond revealing information. Many scholars have argued that transparency requires a sociotechnical system, both humans and non-human agents in it, to be accountable [23] for offering visibility and possibilities of observing its complexity [4,22]. Transparency has been considerably studied in how explanations could support it [14,40,60] and how to make users aware of [19], trust [44], and perceive the fairness [51] of algorithmic systems. Researchers have also started to unpack the transparency of moderation [29,36,48,49].

On social media, moderation systems are designed for governance purposes and oftentimes exert punitive actions on users [7,30]. However, researchers have reported various transparency issues in procedures when platforms exert such punitive actions. First, the content that is deemed inappropriate might not be clearly defined before moderation decisions are issued [64]. Legal scholars have criticized that although social media platforms make their content policies publicly, platforms do not fully consider the context of content (e.g., localized meaning, identities of speakers and audiences) when assessing whether the content is appropriate [73,76]. HCI researchers have specifically examined 15 platforms' content policies and found there was no consensus in defining what counts as online harassment and the extent of force to moderate content deemed as harassment [58]. When these vague content policies are translated into content rule enforcement, there might be a lack of clarity in adjudicating moderation cases [76].

Second, human moderators, as one group of actors enforcing content rules, might issue moderation decisions in murky ways. Researchers found that moderators issued inconsistent moderation decisions, which consequentially hindered users from understanding the boundaries in content rules [17]. Users also complained moderators did not consider or understand the complex cultural contexts of their content [69]. To mitigate such human errors and the cost of operating human moderation, many social media platforms have turned to algorithms to support or automate moderation decision-making [27,29].

However, algorithm-driven moderation also has its own transparency challenges. For example, moderation decisions might not be issued with sufficient rationale disclosure. Suzor et al. uncovered that some users on Facebook were not informed of punishments, e.g., content removal or account suspension, on their issuance [69]. Users experienced automatic account bans, while Facebook did not disclose which policy they violated or how they violated a specific policy [56]. Vaccaro et al. further discovered that users requested explanations on why Facebook's automated moderation system issued inconsistent content removal between them and other users [74].

Furthermore, such opacity might become more complicated when human moderators participate in the sociotechnical procedures of moderation involving algorithms. When moderators use automated moderation tools, some moderators cannot fully understand the reasons behind actions made by the tools [36]. Even when moderators became proficient in using automated tools, they did not use the tools to operate content rules clearly. They might further conceal the critical information from users, such as who, either moderators or automated moderation, made moderation decisions [41].

Last, transparency issues may persist even after the moderation decisions have been issued to users. Once users believe moderation decisions are mistakenly applied, they could rely on "appeal" to request human reviewers to re-examine and provide explanations on prior decisions [80]. However, this does not mean appeal is always an effective way to improve transparency. Juneja et al. found that many subreddits do not have established appeal procedures, so users had to contact human moderators through other communication methods to request detailed reasons for moderation decisions [41]. When platforms provide an appeal option, it's not necessarily effective for users either. West, for example, found that Facebook users held negative attitudes towards appeal explanations because those explanations frequently reiterated earlier moderation decisions without new information [56]. Feuston et al. uncovered that users on Instagram drafted collective petitions against opaque appeal explanations [20].

This variety of work, as discussed, has shown that transparency issues exist in different moderation procedures, such as content rule articulation, rule enforcement or moderation decision-making, and moderation re-examination (e.g., appeal). However, much work tends to focus on a single or a few parts of moderation procedure, and relatively little attention has been paid to holistically examining how different parts of moderation procedures may entail different transparency design requirements in moderation. Our study aims to fill this research gap.

2.2 Design Approaches to Moderation Transparency

HCI and CSCW researchers have explored approaches to enhance moderation transparency. First, existing studies have focused on designing explanations to enhance moderation

transparency in community moderation. In certain online communities, such as subreddits and Discord servers, where users themselves enforce localized content rules [13,21,68], human moderators have used bots to improve moderation transparency [41]. For instance, human moderators in Discord servers built bots to warn or punish users by sending selected reasons to assure transparency of moderation decision-making [43]. On Reddit [12,36] and Twitch channels [65], human moderators also relied on bots to automatically adjudicate moderation cases and offer explanations, such as direct messages, to users. These bots also informed users that they were the decision-makers for punishments such as comment or post removal.

Second, based on users' experiences of moderation, researchers have suggested specific types of explanations. For example, Jhaver et al. have suggested offering explanations in comment forms rather than flairs (i.e., short tags attached to user posts) can effectively help users learn more about content rules on Reddit [37]. Also, a recent study has argued that the game League of Legends needs to offer players statistics such as ban rate to better understand why they experience moderation [46]. In explanation provision, researchers have called for platforms to explore what a high-quality moderation explanation is for end-users, especially those who have been punished [48].

Last, CSCW researchers have conducted design activities to embed transparency in moderation system design. For instance, Vaccaro et al. conducted participatory design workshops to allow marginalized people to design values, including transparency, of existing moderation systems to offer communication with human moderators [75]. Wright et al. designed an interactive tool, Recast, to visualize algorithmic moderation's decision-making on textual content's toxicity. This design could help end-users learn about moderation's criteria for adjudicating user content [78].

In sum, much research in HCI and CSCW has considered offering explanations as a primary approach to enhance moderation transparency. Based on experiences of moderation, researchers have focused on offering explanations of moderation decision-making [31,48,52] and appeal procedures [20,74] to users. However, relatively little attention has been paid to how different moderation procedures may entail different dimensions of moderation transparency. Also, along with the trend of understanding content creators' moderation experiences [5,62,79], there is a lack of work, in the HCI and CSCW fields, on how creators perceive moderation transparency in its different procedures. Our study thus fills the research gap by unfolding the dimensions of transparency in a holistic cycle based on users' moderation experiences. In the next section, we will introduce our study context, creator moderation on YouTube, and how it provides a nuanced scenario to distill the dimensions of moderation transparency.

3 BACKGROUND: Creator Moderation on YouTube as Study Context

YouTube is among the first online platforms to allow users to earn advertising income from their content creation [34] (i.e., creator monetization). YouTubers, namely users who create video content, can join the YouTube Partner Program (YPP) [81] to monetize (i.e., receive advertising income) their videos. Nowadays, at least six million YouTubers globally are eligible to earn money from the platform [24], and some of their livelihoods lean heavily on content creation [10,39,52]. These YouTubers essentially provide digital labor [59], contributing to the platform economy, where their key performance metrics (i.e., income, viewership, audience

engagement) and practices such as content creation and engagement with audiences become the YouTube platform's commodity to be bid and traded with intermediates like advertisers. While YouTubers receive partial income generated from this commercial process, many have unionized and requested transparency of platform-wide income distribution and better labor conditions on YouTube [71].

Moreover, YouTubers in YPP might encounter moderation that amateur YouTubers or general users who consume content would not meet. For instance, once YouTube deems a YouTuber's video violates advertiser-friendly content guidelines [82], it will issue ad-suitability moderation decisions, "limited ads" or "ineligible," claiming the video or YouTube channel is not suitable for advertisers. YouTubers who receive these moderation decisions will have their future ad income decreased or deprived. YouTubers have coined the term "demonetization" to describe such moderation decisions [9].

Algorithms play a central role in detecting and sanctioning video content on a massive scale on YouTube. As YouTube's content policies state, many moderation decisions, such as "limited ads" [83] or hiding videos under the "restricted mode" filter [84], are made by machine learning (ML) algorithms. News reports and academic work have reported how YouTube uses ML algorithms to moderate YouTubers: YouTube's Content ID system allows copyright holders to upload their audio and video that further become digital fingerprints in the system. The system then would classify whether newly uploaded content from other YouTubers matches the existing digital fingerprints to decide copyright infringement [29]. Also, YouTube uses ML classifiers to identify whether specific segments in the metadata of a video (e.g., titles, thumbnails, descriptions, captions, etc.) should receive demonetization punishments [1,61]. Thus, moderation on YouTube is not only about how the platform evaluates the appropriateness of content but also how multiple governance mechanisms exert control over YouTubers.

YouTube's moderation system thus presents a sociotechnical system for us to understand moderation transparency. YouTube broadly uses machine learning algorithms to issue specific moderation decisions such as copyright violation and ad suitability decisions (e.g., 'limited ads') based on the metadata of content (e.g., title, thumbnail, etc.) [1,29,61]. According to YouTube's content policies [83], when YouTubers believe that they have been mistakenly moderated [85], they can request a human review or initiate an appeal to have a human reviewer re-check their content. Alternatively, YouTubers in YYP could have access to contact the creator support team (e.g., through emails) [86] and directly discuss or resolve moderation issues, as YouTube's policy states that they help "get answers on account and channel management questions." YouTube could also simply use the real-time "Chat" function [86], a chatbot-like communication method, on YouTube Creator Studio dashboard to contact the creator support team. Hence, human reviewers play important roles in auditing algorithmic decisions after they are issued by moderation algorithms. However, it remains relatively unknown how transparency has been exhibited in such sociotechnical system of creator moderation on YouTube.

4 METHODS

4.1 Data Collection

We recruited and interviewed 28 YouTubers who had already experienced moderation. Appendix A details their demographic information. We obtained our institution's Institutional Review Board (IRB) board's approval prior to the recruitment. We recruited YouTubers through purposeful and snowball sampling [72]. Using purposeful sampling, we set the criteria of recruiting participants as YouTubers who are over 18 years old and have experienced moderation in the past. Based on the criteria, we created a recruitment website and posted it in relevant YouTuber communities, such as subreddits, r/youtube, and r/youtubers, as well as relevant Discord servers. We also proactively reached out to YouTubers if they had posted their moderation experience on social media such as Twitter, Facebook, or Reddit. Specifically, we used several keywords (e.g., demonetization, limited ads [9,42]) related to moderation on YouTube to randomly search for YouTubers meeting our criterion to ask if they would like to participate in our study. Then, through snowball sampling, we encouraged the already interviewed participants to introduce other YouTubers who meet our criteria to join this study.

This study was geared towards determining transparency issues in moderation before data collection was initiated. That meant we created a semi-structured interview protocol to do so and conducted 28 interviews from January to October 2021. Two sections constituted our interview protocol. One section asked warm-up questions such as YouTubers' demographic information, content category, frequency of creating content, and the range of advertising income they usually received. The other section focused on their moderation experiences, diving deep to investigate potential moderation transparency issues. For example, we asked YouTubers questions such as (1) what moderation did you experience? (2) Can you explain how did that happen? (3) did YouTube provide explanations or reasons for that? If so, what is that? (4) how do you think of these explanations? (5) did this moderation affect your video's performance? If so, how? (6) how did this affect your community that created similar content with you? (7) have you requested human review or appeal? If so, what results did you receive? (8) except for appeal, how did you handle this moderation punishment? (9) how effective do you handle this? After each question, we asked probes (i.e., instant follow-up questions) to dive deeper.

All interviews were conducted, recorded, and transcribed through Zoom. The average length of each interview was 61.5 minutes. After each interview, nearly every participant was reimbursed with a $ 20 gift card, except three explicitly refused compensation. Participants also sent us screenshots after interviews, largely supplementing our interview dataset.

4.2 Data Analysis

We performed an inductive thematic analysis on the dataset [8]. Two researchers independently read through and familiarized themselves with the interview dataset. Then they assigned initial codes to ideas that reflected in certain data unit such as a paragrah or sentence in the interview transcripts. The researchers also held regular meetings to discuss about each initial codes' definitions and address disagreement amomg them. After coding all data, the first author conducted rounds of coding by identifying patterns (i.e., a theme that can cover several codes) among the codes and further grouped these themes into overaching,

higher-level themes that can cover a gounp of themes. This process ended up with generating four overaching themes, as shown from Sections 5.1 to 5.4.

5 FINDINGS

We found four dimensions of moderation transparency reflected in our participants' moderation experiences. They desired YouTube's moderation system to (1) present moderation decisions saliently, (2) offer moderation explanations profoundly, (3) afford communitive support from human agents, and (4) help them practically learn about moderation.

5.1 Salience in Presenting Moderation Decisions

Salience means that the moderation system should deliver decsion in a noticeable way to raise users' awareness of it. Our participants desired to receive notifications regarding all kinds of moderation decisions. Otherwise, participants would be largely doubtful and frustrated if they only found out moderation decisions later or through other means. For example, "limited ads" (i.e., demonetization) reduces the future ad income of a YouTuber's video, and prior work has discussed this moderation punishment's opacity [9,52]. However, in participants' experiences, the "limited ads" decision could be coupled with other unknown penalties. P23 described:

> It seems my channel is on a kind of list in which, as soon as I go live (on YouTube Live), it will always be demonetized after seconds [on Studio dashboard]. I can prove this because I use a software called StreamYard; I can send the same [recorded live-streaming] content to my friends' channels. They don't get limited ads, but I do. (…) It would be ideal they let me know that I am on some sort of blacklist. [P23]

StreamYard is a live streaming platform that provides streaming services and can be connected with other social media (e.g., YouTube, Facebook, Twitch) to live stream. YouTube Live is a live streaming option internally on YouTube. Studio dashboard is a panel provided for YouTubers to understand their video's statuses of visibility, monetization, restriction, user engagement, and so on. In P23's case, he was confused about his channel's monetization status and suspected that he was punished for a certain reason. He confirmed this suspicion by comparing how monetization statuses of the same videos, either live streaming or recorded ones, appeared on his friends' channel. The transparency issue here lies in P23's inability to know how moderation systems classified his channel, meaning if he was on an underlying blacklist to receive "limited ads." He desired to have a notification of any form of punishment against his channel. P23's case highlighted how, in creators' experiences, several penalty mechanisms, including "limited ads" and "blacklist," might work together to moderate creators, but the penalties resulting from these interleaved mechanisms were not made salient to the creators.

Even when participants did receive notifications, they hoped that the notifications could be delivered in cautionary and noticeable ways. When experiencing demonetization in individual videos, our participants desired explicit notifications. For example, P3 said to us:

> I seemed to recall a long time ago; there may have been at least one email [that my video received "limited ads"]. But most of the time and they never tell you you've

> been demonetized on a particular video. You have to look in the Studio [dashboard], and you see that the dollar sign is yellow. [P3]

By receiving notifications inconsistently, P3 desired the moderation system to send emails to notify him whenever "limited ads" punishments happened to him. Without email notifications, he felt it hard to notice whether his individual video's ad income decreased or deprived unless he frequently checked his YouTube Creator Studio dashboard.

Prior literature has reported that some users did not receive notifications of content removal or account suspension on Facebook, Twitter [56,69], and Reddit [41]. Adding to such findings, we found our participants further complained that they did not realize that they were punished until informed by other people. For instance, P13 received "limited ads" and said:

> I noticed my demonetized videos didn't get picked up, and I asked my fans a lot. I gave them a poll and said, "hey did you watch my new video? I hope you got notifications." (…) 33% yes 67% no. I go into its comments, "oh, I didn't get notifications." (…) And they said, "no notification didn't come up in my recommendation [feeds]." (…) I did not know that YouTube didn't actually show my [demonetized] video to people who subscribed to me. [P13]

YouTube Creator Studio dashboard shows whether the viewership of a video, an indicator of visibility, is higher or lower than a typical record. However, P13 did not notice his video was less notified or recommended to his viewers until they told him so. Thus, he hoped that the moderation system should help him recognize whether he experienced the punishment of decreased visibility.

In sum, participants desired the moderation system to present them with punishments in noticeable ways, raising their awareness of the issuance of punishments. When participants found out about their punishments from sources other than YouTube, such as themselves or their viewers, they experienced confusions and frustrations.

5.2 Explainability for Moderation

Explainability refers to both willingness and clarity that YouTube's moderation system explains why creators experience moderation. Prior work has stressed the importance of offering explanations to punished users [41,48] and called for moderation explanations to be attached to content rules for moderated users' better understanding [35]. Extending this thread of research, we identified three primary dimensions that YouTubers frequently invoked to talk about their expectations for an explanation from the moderation system. In other words, participants paid attention to not only whether but also how the moderation system provided explanations. Specifically, our participants desired explanations that (1) are provided in a proactive and timely fashion, (2) specify detailed reasons behind moderation decisions, and (3) offer actionable information for repairing problematic content.

5.2.1 Proactivity

Proactivity refers to the proactiveness and timeliness of the moderation system to offer an explanation. For example, P8, who posted a collaborative video, said to us:

> Across the collab [video with other YouTubers], everyone was commenting like, "Why are you deleting my comment? I said nothing wrong." And then, like all of us were

> experiencing this, even the actual channels in the collab were having their comments deleted [by YouTube] for no apparent reasons. We still haven't gotten an answer. [P8]

YouTube did not proactively offer explanations for deleting P8's and his peers' comments. This caused misunderstandings between P8 and his viewers. P8, his viewers, and YouTuber peers all desired the moderation system to offer timely explanations.

Furthermore, participants desired timely explanations after they initiated appeals. For example, P19 said:

> The majority of times, it ("limited ads") was not correct, and we asked for a review, but you know that the first [several] hours when you open a video are very important for the [recommendation] algorithm. That is bad for us because we're in the time that they give us a response or we lost money; then I feel powerless. [P19]

P19 waited for a long time for an explanation of her appeal. Because she assumed "limited ads" would affect whether recommendation algorithms promote her content in a normal way, she felt frustrated with losing income without knowing how and why she violated content rules. She thus desired human reviewers to timely offer such explanations.

Moreover, some participants requested timely moderation explanations, given the coordination between YouTube and copyright owners. For example, P24, a YouTuber who received a copyright claim, said:

> The difficulty is that due to the way the YouTube system is set up, you can challenge a company's copyright claim (Content ID claim), but they have like 30 days to respond. So, they can just wait out the clock on; they can basically force you to wait a whole month to see [if] it (video) fits [copyright laws], and it's just a matter of waiting for it got to monetize. [P24]

When copyright-related algorithms (i.e., ContentID system [87]) detect potential copyright infringement on YouTubers' videos, they automatically issue Content ID claims to those YouTubers. When this happens, YouTubers' ad income in these videos will be shared with copyright owners. P24 frequently experienced such situation, so he desired the copyright owners could timely respond to him. The sooner the owner manually identifies false-positive copyright infringement and informs YouTube, the less ad income P24 will lose.

5.2.2 Depth

Depth refers to how participants disliked "superficial" explanations that only describe what rules they have violated and expected an in-depth, logical argument that led to the conclusion of rule violation. Lack of depth in explanations would fail to support participants' reasoning and sense-making about their punishment, as well as behavioral improvement in the future. For example, P22, a YouTuber creating anime content, said:

> I think [YouTube's] reasoning is very arbitrary. For example, the animation itself (Video 1) is unavailable in restricted mode. But the side-by-side comparison of the exact same video (Video 2) next to the original video is perfectly available with identical tips and nearly identical titles. The only changing the title is [the word,] "comparison," at the end of this video. So, whatever label they're using for restricted mode is entirely arbitrary. [P22]

'Restricted mode' is a content filter to make potentially mature content invisible to viewers who are under 18 years old or without accounts signed in [84]. P22 initially understood that YouTube identified his Video 1 as mature content (e.g., adult content), which implicitly explained why such video was hidden by the 'restricted mode' filter. However, the early hidden explanations he received were too generic to help him fully understand why Video 2, with similar content, received different treatments. Thus, he desired more in-depth explanations about why such inconsistency of moderation happened to resolve his frustration and confusion.

When participants came to appeal or request human review on moderation decisions, a lack of depth also lay in appeal explanations. For example, P8, a YouTuber who made history education content, encountered demonetization, as shown in Figure 1:

Fig. 1. Requesting "human review" for algorithmic decision (left) and human reviewer selected a moderation explanation (right).

> I had a video that was about the sectional crisis, the 1850s, and how the slavery debate developed through that. (…) They (YouTube) said disagree. It has a focus on dramatic acts or eating things that shouldn't be eaten. (…) It is ridiculous. There's no food in it. I mean, it's literal stock footage of history. [P8]

As shown in Figure 1 (left), P8 requested human review on his Creator Studio dashboard for a "limited ads" decision that was made originally by moderation algorithms. And a human reviewer selected "harmful or dangerous acts" from the advertiser-friendly content guidelines [82] as a moderation explanation. This interaction is static, meaning that a YouTuber did not directly communicate with a human reviewer regarding prior moderation decisions. And the explanation provided, as P8 thought, failed to fit his niche content type in history and to explain how specific historical incidents were related to dangerous acts. With only help articles or videos (see Figure 1 right) provided, P8 implied the need for contextualized explanations on which historical footage of his video violated policies to better support his reasoning.

5.2.3 Actionability

Actionability refers to how participants desire explanations that are detailed enough for them to repair and render the content acceptable. For example, P3, a YouTuber who created travel content, experienced video takedowns and said:

> When I get a manual privacy thing (privacy compliant) [from viewers], the [explanation's] wording is terrible because they (YouTube) don't tell you where, [in] the video, [the complaint] is happening and who complained. (…) Can you tell me where I can remove that segment? By then, it's too late, so it's almost like they don't want that video to be there. [P3]

According to YouTube's policies [88], audiences can manually initiate privacy complaints to YouTube if they find any videos that expose their personal information (e.g., image, voice). In the above example, P3 encountered content removal by viewers' privacy complaints. He desired the moderation system could offer opportunities for him to repair his content by pre-emptively pointing out where the clips violated audiences' privacy in his videos. So he could edit the content to have videos not be taken down.

Prior work has found that punished/moderated users needed explanations to reform or repair their behaviors [48]. Adding to that, we found lack of such actionable explanations could further become participants' labor of guesswork and repairment. For example, when the moderation system issued a "limited ads" decision, P25, a YouTuber creating film compilation content, said:

> You know they (YouTube) have all these standards, but they don't tell you specific reasons and where it is. They just slap you with either the yellow limited ads or no monetization. And the problem is that I'm left to guess what the problem is, and then I have to keep re-editing it like 17 versions. Every version takes like an hour or more to export and upload, as I'm trying to guess what the problem is. [P25]

The moderation system did not specify why P25's video violated what clause of advertiser-friendly content guidelines [82]. Such lack of details led to his repetitive labor of reproducing and testing the video because he could only guess what clips in his video were deemed problematic. He desired more details from the moderation system to flag these clips so that he could repair his videos effectively.

However, copyright moderation (i.e., Content ID system [87]) on YouTube is a distinctive moderation category where the moderation system does offer actionable explanations. The explanations will detail videos' timestamps where moderation algorithms or copyright owners claim copyright infringements. This means that YouTubers could trim out, replace, or mute any segment that is deemed as a copyright violation. Even so, several of our participants complained that copyright-related explanations oftentimes only presented an incomplete list of problematic segments. For example, P2 said:

> I have to go into my editing software and blur out that segment [with a Content ID claim], and then I have to re-render that video which takes time re-upload it. And then they'll tell me about a different segment. [For] a 20-minute-long video, that's going to take 15 to 20 minutes to re-render it like 20 plus times, so you can figure out how many hours go into a video that I will not get a single penny of ad revenue from. [P2]

Once a YouTuber's video receives a Content ID claim (i.e., copyright infringement claim), their ad income from this video would be shared or fully transferred to copyright owners. In the above case, because moderation algorithms or copyright owners did not point out complete potential infringements, P2 had to repeatedly re-render the video whenever a new

Content ID claim was issued. P2 desired the moderation system to point out segments, including audio or video clips, that had been identified as copyright violations all at once.

In sum, we found our participants desired profound and timely moderation explanations. It meant the moderation system needed to indicate reasons for its actions, offer in-depth explanations for issuing punishments, and point out actionable solutions for repairing content.

5.3 Communicativeness of Human Agent

Communicativeness refers to the effective communication provision from human agents behind the moderation system. Previous work has reported that users had difficulties directly contacting human moderators [41,75]. Adding to that, we elaborated on dimensions of communicativeness: participants desired direct and easier access to human agents (e.g., human reviewers) and high-quality responses from them. These desires showed moderation transparency was further requested to support YouTubers' content creation and monetization growth, which general users might not always request for the same purposes.

5.3.1 Access

Access refers to the easiness of contacting human agents such as human reviewers or creator support [86] on the platform. For example, P25 said:

> I am not a huge channel but not a small channel; you know I have over 50,000 subs at this point. And I have never been able to speak to a human reviewer at YouTube, even when I appealed my last ["limited ads"] video. They're like unicorns. (…) It would be nice if they were a little more communicative. [P25]

In the above quote, P25 explained how he desired better access to human reviewers. However, he suspected that such access hinged on YouTubers' channel sizes, and he did not pass the threshold. Consequently, he felt that human reviewers were rare, like "unicorns." P25 hoped the moderation support could be more communicative so that he could more easily reach human reviewers.

Given such hardship of accessing human reviewers on YouTube, many participants tried to reach out to reviewers from other platforms. For instance, P13, who experienced "limited ads," said:

> I've seen through this that creator support is useless in order to get a response out of YouTube. [So,] you have to make as much noise with my fans on Twitter as possible, like as much noise as you can make YouTube Team notice [your problems]. [P13]

P13 and his subscribers collectively reached out to human agents on Twitter for attention because they assumed that YouTube's Twitter account was an easier source they could reach out to. P13's case implied his desire to obtain contact with YouTube through internal communication methods on YouTube instead of Twitter.

5.3.2 Responsiveness

Responsiveness means human agents' effectiveness in replying to participants regarding their moderation decisions. For example, P23, who experienced "limited ads," used the "chat" function:

> So, I'm using the chat section that the content creators have. I talked to the people, and you get the feeling that they're not listening to you. There's just a robot, or

> somebody would just be getting paid to just type answers. They kept asking me, again and again, to go and check the guidelines. So, I've done that. Please don't tell me to do that again because I know that. [P23]

"Chat" is a function embedded in creator support [86] to have YouTubers directly real-time chat with human reviewers. P23 communicated back and forth with creator support/human reviewers regarding how to resolve moderation issues. P23 assumed that he could contact them to re-check his demonetized video. However, the responses from "Chat" appeared to be robotic without giving any concrete help on his video, such as re-reviewing it or providing explanations. Thus, he desired more interactive human communication to effectively solve his moderation issues.

Some participants desired human reviewers' responses that could clarify the reasons why the moderation system decided to reverse moderation decisions. For example, P6 said:

> I had one video that got demonetized. It was just an animation test of my character, and the YouTube algorithm automatically marked that as demonetized ("limited ads"). When I contacted [creator] support because I wasn't able to actually appeal, they were able to go through and re-enabled monetization, but they never really told me specifically what they found. [P6]

YouTubers are unqualified to appeal for a video when algorithmic decisions have been confirmed by human reviewers or YouTubers' last appeal fails. In P6's case, creator support successfully reversed the moderation decision that happened in this way. However, reviewers did not disclose why they reversed it and how P6 could prevent such moderation from happening again in the future. So, P6 implied he needed effective responses to learn from his past moderation punishment by communicating with human reviewers.

Similarly, when our participants received YouTube's responses on third-party platforms, e.g., Twitter, they still remained ineffective. For instance, P26, who created ASMR videos, experienced "limited ads" as YouTube deemed them as sexually suggestive and said:

> On Twitter, they always say sorry, I'm sorry you feel that way or saying like hey, we'll look into those things. (…) It's pretty frustrating for the [ASMR] community. It would be nice if they didn't have generic responses to our inquiries, it would be nice for them to be like, hey [this is] for you to improve your channel is. [P26]

ASMR (autonomous sensory meridian response) is a video content type for reaching relaxing and sedative sensation effects by producing placid sound. In the above case, P26 and her community assumed YouTube's platform support on Twitter would help them solve "limited ads" issues. However, they frequently received polite remarks but no responses regarding what ways to improve her content and channel in the niche category to be acceptable on YouTube. Her experience showed a desire to receive such effective responses.

5.4 Learning for Creator Moderation

Learning refers to opportunities that the moderation system provides for participants to grow and learn to become sound community members. After YouTubers experience moderation, YouTube's moderation system might occasionally provide educational resources (e.g., help articles of content rules, videos), as examples shown in Section 5.2.2. Prior work has also called for instructional content rules instead of tedious "Terms of Service" [56]. However,

beyond these resources, our participants desired to learn about moderation at practical levels, from guidance on content rules to instructions for newcomers to community learning. Thus, after such learning, they could create the "right" (e.g., monetizable) content without moderation.

5.4.1 Rule Guidance

Rule guidance means clear guidance on content policies that the moderation system applies. For example, P24 spoke for his community in educational content:

> A lot of us don't tend to talk too much to each other about it (moderation) because we're all just aware that YouTube's guidelines are very vague, and they (YouTube) can intentionally enforce them as they (YouTube) see fit. So, for most of us, it's just a matter of it go to demonetize what you are going to do. None of us have much control over it. YouTube gives you a bare minimum amount of guidance. It tends to be very vague general things that could be interpreted in a million different ways. [P24]

P24 stressed a collective perception that YouTube's content policies were vague because the system did not disclose how moderation algorithms or human agents translate policies to moderation practices. So, P24 argued policies were flexible to be interpreted in various ways from their moderation experiences. However, he and his community still showed the desire to receive guidance in learning what content is unacceptable under content policies.

Content policies also remained relatively abstract, meaning that participants failed to connect moderation decisions they experienced with a specific clause in content policies. Some participants desired the moderation system to offer rule guidance on niche content categories because the moderation practices they experienced appeared to be ambiguous. For example, P25, who creates film compilations, said:

> One [moderation of "limited ads"] that always kills me is violence. I'm doing horror movies; it's not graphic law enforcement footage, but there's violence in it. They (YouTube) don't address this; there's no context. And I'm always trying to understand what the rules are. I don't want to break YouTube's rules. As YouTube tells me like, hey, you absolutely cannot show this thing. I'm happy to take it out. [P25]

YouTube's advertiser-friendly content guidelines [82] allow videos to contain "fighting violence excerpts from an action movie (e.g., scripted content)," but there is no specific guideline about to what extent YouTubers can use violent content from horror movies. Therefore, in P25's case, he felt confused and frustrated that he encountered demonetization because of using violent clips from horror movies. P25 thus desired contextualized examples for his content category, horror movies, in content policies and to guide him to learn from moderation decisions.

5.4.2 Newcomer Education

Newcomer education means the instruction provision from the moderation system when participants were still newcomers on YouTube. Many participants' ad income got largely decreased or deprived because they knew little about moderation on YouTube when they were newcomers. They wished the moderation system could have provided instructions when they started to monetize content, so they did not have to experience moderation. For example, P4 experienced Content ID frequently claims when she was a newcomer:

> In the beginning, I got a lot of strikes (Content ID claims) because I knew nothing about copywriting from YouTube. It's a very gray area to understand, especially if you're new to YouTube. So, I made a lot of issues with using content that wasn't originally mine. And it would detect, say music, for example, and then they would tell me that this video is no longer monetizable to you. [P4]

In the beginning, P4 knew little about copyright and reused copyrighted content such as music. As a result, she was punished multiple times with Content ID claims. She felt there were lots of challenges in learning about copyright policy on YouTube. Thus, she desired newcomer education from YouTube that could help her understand copyright issues better.

Some participants further desired functional designs that could offer moderation knowledge. For instance, P28, who experienced channel demonetization, said:

> The reason was because of the reused content, and besides that, I'm also very stupid, so I used someone's consent and made a compilation. Primarily I think it's my own fault, (…) It would be great if the self-certification function was there [to read the rules] from the beginning. [P28]

Self-certification is a tool for YouTubers to self-check whether their video complies with YouTube's advertiser-friendly guidelines [82] before it is published. In P28's case, he realized his mistakes and desired YouTube could provide the self-certification tool when he was a newcomer. Because when he was punished, there was no self-certification tool for him to learn more about content policies. He thus realized his mistakes until he actually experienced channel demonetization. P28's experience showed the necessity of offering more such educational designs to newcomers, especially to help them grow to be sound community members.

5.4.3 Community

Our participants stressed the need to learn from a community composed of peer YouTubers about how the moderation system works and how to make acceptable content. For example, P8 described:

> There's also the history YouTubers Slack now. It has something like 40 or 50 channels on there. And there are those personal connections that you make over the years, and all of that, you know, we talk to each other; we look over each other's stats and discuss what it could possibly mean. So, it's a very community-driven effort to understand this kind of stuff (moderation) because YouTube just doesn't explain. [P8]

Because of the few effective explanations offered by the moderation system, P8 collaborated with other YouTubers in the history content community to understand punishments and their impacts. This showed P8's desire for collective learning about moderation and applying their collective sense-making into practice. Thus, they could grow to be experienced in handling moderation issues.

When the moderation system enforced new content policies, participants also learned from public conversations between YouTubers and human agents behind the moderation system. For example, P5 said:

> They (YouTube) just won't tell you a lot about the policies, or it could just be too late on updating a policy on a date, or they may update policies for a certain percentage of

> users. (…) So, you can often see it on Twitter; the team YouTube Twitter account is very active in addressing a lot of complaints, and they also tend to offer some explanations as to why certain users may have been demonetized, so people usually like go to Twitter to study all those different case studies. [P5]

Because, as P5 claimed, YouTube disproportionally updated and implemented content policies, he desired to learn from moderation experiences shared by other YouTubers' tweets and their conversations with YouTube's human reviewers on Twitter. How P5 learned from collective sense-making on Twitter showed there was in lack of internal spaces on YouTube (e.g., comment section under content policies) for him to discuss new content policies with his peers.

6 DISCUSSION

Prior researchers have discussed specific design solutions to improve moderation transparency, such as explanations [20,52,56,69,74] and communications between punished users and human moderators on platforms such as subreddit [36,41] and Twitch [66]. Building upon this thread of research, our study presents a holistic picture of transparency design, situated in the context of creator moderation on YouTube. We described four primary dimensions for the transparency design, including salience in the presentation of punishment, explainability for moderation punishment, communicativeness of human agents, and learning for content moderation, as summarized in Figure 2.

By presenting a taxonomy of fine-grained dimensions for moderation transparency, we not only unpack the notion of transparency in the unique context of creator moderation, but also provide translatable insights for transparency design of other moderation systems. If content moderation is about catching and punishing "bad actors" who could have malicious intents [7], creator moderation on YouTube builds upon creators' socioeconomic stakes in the platform economy in which creators share the goals of making "advertiser-friendly" videos. Situated in this context, our study is among the early attempts to document what content creators experienced and needed with regard to moderation transparency. For example, while prior work has discussed the importance of salient notifications of moderation decisions [35,56,69], we further broke down what salience entails in creators' interaction with moderation and pointed out that such salience was contingent on the complexity of known and unknown moderation decisions (i.e., a combination of limited ads and perceived blacklist).

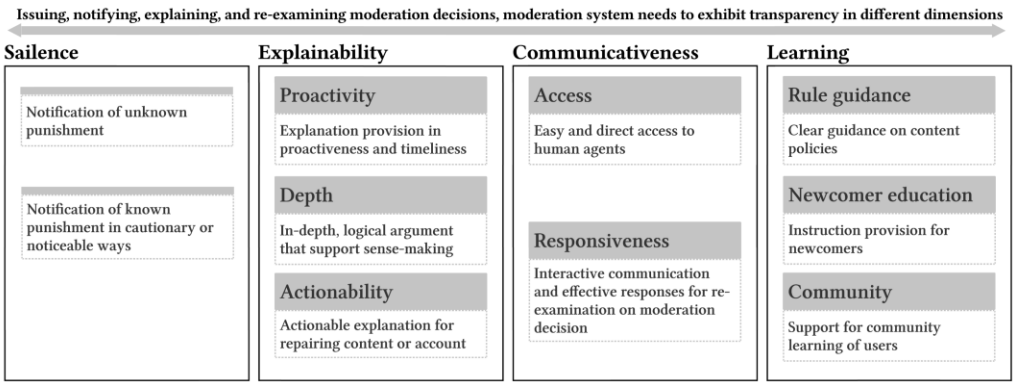

Fig. 2. Dimensions of moderation transparency of creator moderation on YouTube.

Participants' moderation experiences demonstrate that transparency should be a critical ethical value that a moderation system holds. Their experiences help us reflect on the extent of transparency that moderation procedures should maintain. In the following subsection, we will discuss how creator moderation on YouTube needs to maintain dynamic transparency.

6.1 Dynamic Transparency in Creator Moderation

Our study of transparency design is situated in the context of commercial, algorithm-driven creator moderation. Content creators' stake in moderation is categorically different from that of Reddit or Twitter users, as the former derive a livelihood from the platformization and monetization of their creative labor [39,42,52]. Thus, content creators like YouTubers in our study are economically incentivized to appease the creator moderation system [54], regardless of its complexity and opacity. This, in turn, renders the transparency issue of creator moderation even more pressing. Specifically, in this study, we frame participants' experiences with creator moderation in light of dynamic transparency, a dynamic process between transparency efforts such as aggregation, circulation, and interpretation of information, and governance impacts [15,32]. In other words, transparency design involves more than one-off information provisions.

Our findings about participants' interactions with YouTube's creator moderation stress dynamic transparency as a multi-stage process. Issues of transparency unfold at the multiple stages of creator-moderation interaction, such as the reception of punishments, the comparison and consulting with community members, and the appeal process. Thus, the transparency experience, or how participants encounter, perceive, interpret, and deal with transparency issues in their interactions with moderation systems, is not limited to just one stage but persists through the entire moderation cycle.

Such dynamic transparency invites critical reflections on the provision of moderation explanations to enhance moderation transparency for end-users who experienced moderation [36,48,74]. Moderation explanation as a direct way that platforms communicate with end-users about moderation decisions has largely been reported to be deficient or opaque [37,46,52,56,69] or needed in a particular phase of moderation [20,74]. Adding to this line of work, we point to the significance of improved procedures as to how explanations are offered: (1) in punishment notification, punishments can be presented saliently; (2) in punishment explanation, participants hoped to receive structural explanations ranging from rule violation reasons to in-depth explanations to ones can help improve content; (3) in requesting re-examinations (e.g., appeal), participants asked for explanations to fit their contexts of content. The Santa Clara Principles (SCP) also articulate such procedures [89] by calling for platforms' transparency and accountability. What is beyond the SCP and other relevant public attention is how moderation transparency is cited by our participants as a primary reason to request the protection of creators' labor and livelihoods. As content creators, our participants hoped to smoothly build up and engage with their communities of audiences, continue creating monetizable content, avoid unexpected income loss, and obtain the platform's support for these goals and activities.

However, a complex procedure requirement for moderation transparency surfaced from our participants' moderation experiences. That requirement is coordination between human agents (e.g., reviewers) and moderation algorithms through bi-directional communication. On

the one hand, our participants expected less lengthy documents from the platform that explains algorithmic decision-making, but more accessible, effective interactions with human agents. Prior work has pointed out human moderators avoided offering reasons for content removal (e.g., [41,69]). On YouTube, our participants felt frustrated and confused about how to reach out to human reviewers successfully. They suspected having a small fanbase as a factor in preventing direct dialogues with human reviewers. Even when they finally managed to connect to the reviewers, participants desired to have more interactive and effective communications (e.g., effective feedback) regarding how to solve moderation issues for videos, especially those from niche content types.

On the other hand, dynamic transparency not only means disclosing information but also supporting users' sense-making [15]. When a platform like YouTube affords human reviews on algorithmic decisions [83], it remains questionable whether they could successfully identify algorithmic errors that frequently happen to marginalized groups (e.g., [1,2]) and explain algorithmic decision-making effectively. Human moderators on Reddit could be flexible in adapting moderation decision-making to nuanced cases when moderators notice cases of algorithmic decision-making overlooking key contextual information [66]. Resonating with that, our participants desired human reviewers to offer context-specific explanations of how they confirmed the true positiveness of algorithmic decisions and how the decisions corresponded to content rules. Instead of asking for types of explanations [35,48], what our participants requested was more about a configuration of enhancing transparency: they desired humans to coordinate moderation processes (e.g., moderation decision-making, appeal, communication support) and collaborate with algorithms to support their personal or collective sense-making. For example, participants such as P8 and P23 desired human reviewers to disclose details of how they confirmed algorithmic decisions as true-positive accurate to adjudicate moderation cases and offer further explanations.

However, such sociotechnical coordination does not mean unlimited, full disclosure from the platform. Instead, the goal of enhancing moderation transparency is to conduct effective moderation that would be beneficial and instructional for users and communities. It is known that when aspects of algorithms go public, users might game algorithmic systems [16]. Prior researchers have also pointed out a grey area of moderation: pro-eating disorder users circumvented moderation by various actions that made their content appear to be unproblematic [11,20,25]. It thus becomes hard to decide whether to disclose moderation decision-making details to users or moderate problematic content that might suppress marginalized communities. However, this concern becomes a nonissue largely in the context of creator moderation. Our participants from niche content types (e.g., anime, education) also encountered such suppression, but their moderation experiences offered a direction for the YouTube platform. That is, how to support users' learning could be a key determinant of deciding the proper extent of transparency. Participants showed the extent of transparency they desired was based on the aim of growing to be a sound community member and work with the creator moderation system to fit in the platform. In detail, they desired to understand how content rules were applied by human reviewers and moderation algorithms and how to relieve and handle moderation impacts on their content and channels. After this transparency was collaboratively offered by humans and algorithms, they desired to avoid violating content rules in the future for better content creation experiences. Then, they hoped to grow and fit in

the platform to smoothly learn about and conduct content creation, sharing, and monetization and engage with viewers.

We observed a mismatch between content creators' needs for dynamic transparency versus the rather static transparency design in YouTube creator moderation. As a result, the platformization of creative labor on YouTube inevitably solicits extra, unpaid labor from creators to deal with transparency issues. Thus, we see an urgent need for creator moderation to adopt a dynamic view of transparency to support creative labor, so that creators like YouTubers and platforms could efficiently reach common ground based on higher-quality communication and educational designs of moderation. YouTubers thus are more likely to effectively understand where the boundaries in content rules are. Then, their key performance metrics, such as income, audience engagement, visibility, and more are less likely to be under unexpected range. YouTubers as creative workers [9,59] fuel the platform economy and growth through different activities such as socioeconomic content creation [52] and engagement with audiences. As their creative labor was better supported, this, in return, would benefit all, including the platform, intermediates (e.g., advertisers), audiences, and creators from monetizable content that follows content rules.

6.2 Support for Learning as Transparency Design

Besides governance and punitive purposes, researchers have called for a moderation system with learning opportunities, e.g., moderation explanations, to help users learn more about their moderation decisions [35,37,52,56,74]. Our participants' moderation experiences revealed learning as an important dimension in moderation transparency design. They desired a transparency design that could instruct them to create the "right" content. Ultimately, one of the goals of transparency design is to help YouTubers to learn or to understand and internalize what is made visible to them. Thus, they could at least obtain chances to grow to be productive creators and sound community members.

Content rule requires end-users to understand how to generate or create the "right" content, but platforms rarely instruct them on how to do so. Prior moderation research has largely focused on how content rules or norms are articulated [13,21] or applied [41] on social media. However, our participants' moderation experiences on YouTube demonstrated their learning needs and the educational potential of creator moderation. In responding to critiques that content rules lack consensus in defining unacceptable content [76], we argue the "right" content needs to be defined in a context-specific rather than abstract fashion. Our participants showed their desire to interpret content rules in their niche content types to help them understand why they experienced moderation and create the "right" content in the future.

Such guidance of content rules further points to designs that can help users learn about moderation. Prior studies have discussed the educational roles of moderation explanations in instructing users' future behaviors [37,74] and helping them learn about content rules [35]. Adding to that, we found participants on YouTube needed rule learning in more design components. They requested (1) instructions before moderation algorithms refer to policies to make decisions, (2) clarifications of what rules algorithms used to decide content as unacceptable and how reviewers confirmed these decisions, and (3) learning resources on content rule and its update. End-users like creators who request moderation transparency might not equate with trolls [16,41] described in prior work aiming to game the system. Instead, they, at least our participants, wished to learn more about creator moderation to have

their key performance metrics within the expected range and have their creative labor rewarded in an expected way.

Participants also exerted agency to seek or create learning resources on their own. They engaged in collective activities to support each other's learning on content moderation. Much research in HCI and CSCW has described how people create online communities of practices (CoP) to share different levels of expertise to become experts in the data science field [67], micro-entrepreneurs on Airbnb [33], or user experience professionals in organizations [47]. Similarly, CoP also was crafted when our participants encountered opaque moderation. Participants such as P8 and P5 engaged with other YouTubers in public and exclusive communities to gain knowledge about moderation decisions and how content policy updates impacted them. Participants' collective learning showed YouTubers' agency of learning from their peers and further implied the necessity for YouTube to support their learning.

6.3 Design Considerations for Transparency Design of Moderation

To construct transparency designs of moderation, a platform like YouTube, where YouTubers create different types of content, should consolidate content rule articulation first. Rules need to be unambiguously defined before moderation decisions or actions, especially elaborated on niche content types for YouTubers such as P8, P25, and P26. To conduct it, the platform could arrange, for example, participatory workshops for creators from different content types to brainstorm on what content rules fit their communities' contexts. After content rules are defined clearly and comprehensively, moderation algorithms or human agents (e.g., reviewers) need to refer to specific clauses in content rules to make decisions or offer explanations. Then, users could successfully connect their moderation decisions with the clauses in content rules.

Second, moderation system designs could consider enforcing content rules with precautionary mechanisms. Some HCI studies have demonstrated an analogy between algorithmic systems and "street-level" bureaucracy that algorithms fail to adapt decision-making to novel or unpredictable cases [3,57]. Our participants' moderation experiences demonstrated a direction: moderation algorithms could be precautionary in informing decisions before errors really happened, as what newcomer participants experienced. Algorithms could detail what rules were used to make decisions for different content types. Then, there would be relieved labor of both human agents and YouTubers to explain algorithmic decision-making, preventing more conflicts of understanding on moderation decisions from happening.

Third, the sociotechnical system of moderation, including both human agents and algorithms, should bridge communication with end-users. Responding to the call about formalizing communication between users and human moderators (e.g., [41,75]), we argue that communication should be bi-directional, as we discussed in section 6.1. On the one hand, moderation should be conducted openly, meaning that both known and unknown punishments could be clearly notified, timely recognized, and profoundly explained by the moderation system. Then, users could easily contact human moderators and engage in interpersonal communications effectively to request re-examinations on moderation cases. On the other hand, moderation transparency involves not only information disclosure but also configurations of supporting users to do sense-making. Thus, similar to general solutions mentioned by Flyverbom about enhancing transparency [22], YouTube could arrange public

meetings or Q&A sessions with creators from different content types or geographic areas. This could help them better understand how at the organizational level, what content YouTube, e.g., human reviewers, or engineers who developed monetization or recommendation algorithms, consider unacceptable to YouTube communities. At the same time, YouTubers could suggest ways of how human viewers could collaborate with algorithms to offer communication with them. In this sense, the support for YouTubers' learning could extend what content rules articulate and allow YouTubers to understand how they could grow together with the platform.

Content creators like YouTubers could be encouraged to learn to be productive community members through different designs of moderation transparency. Prior work has stated platforms' content rules, like the term of services or community guidelines, are not practically educational [56] and suggested moderation explanations grounded with them for better moderation transparency [35]. Besides precautionary mechanisms of algorithmic moderation and bi-directional communication, our study with 28 YouTubers further put forward some new possibilities. On the one hand, as we found that newcomers were not sufficiently instructed to get familiar with content rules, a learning opportunity implicitly appears to all YouTubers who have not experienced moderation yet. YouTube could proactively disseminate educational resources (e.g., instructional videos, pop-ups, or interactive cues/markers when YouTubers upload videos) instead of waiting for moderated YouTubers to report or appeal potential false-positive moderation decisions. On the other hand, YouTube could consider supporting online communities of practices (CoP) features on the YouTube Creator Studio dashboard. As we uncovered that YouTubers frequently communicated with each other about how to resolve moderation issues that negatively impact their income and livelihoods, their community-wide connections and engagement should be supported. YouTubers who are in the same content categories or geographics could share their sense-making or concerns about moderation together on the platform, along with creator support team's observation.

These design considerations above could ultimately be translatable for other creator platforms. Nowadays, YouTube is not the only one that affords monetization for creators. Twitter, Facebook, TikTok, and more are eager to grow their creator monetization initiatives or new creator economy products. Although moderating the sheering scale of content appears to be challenging [26,63], platforms should never neglect how their profits could align with supporting creators to learn about creator moderation. Enhancing transparency with communication and learning in different moderation phases or procedures would not only help creators better fit in communities but also share a growing platform economy with them.

7 LIMITATIONS AND FUTURE WORK

The YouTubers we interviewed did not represent all YouTubers' moderation experiences. Six million YouTubers (who have over 1,000 subscribers) nowadays [24] might have various moderation experiences because they create different types of content or locate in specific geographic areas. This study thus tends to generate nuanced qualitative insights into how moderation experiences could better inform transparency designs for moderation systems. This also provides future research directions to understand how specific groups (i.e., content type, areas) of YouTubers experience content moderation. In this sense, future work could consider using quantitative methods (e.g., surveys) to analyze how, for example, ASMR YouTubers interact with moderation systems and their challenges in solving moderation

issues. Also, as we did not aim to put forward generalizable factors or frameworks from our qualitative findings, there can be future quantitative work cross-validating our findings through surveys with creators and exploratory factor analysis.

8 CONCLUSION

Nowadays, in the presence of complex sociotechnical structures to implement content moderation, more attention has been accumulated on enhancing moderation transparency. Along with endeavors in HCI and CSCW uncovering users' experiences of moderation, we unpack and elaborate on the dimensions of moderation transparency from 28 YouTubers' experiences with creator moderation. As creator moderation is not simply about how platforms evaluate the appropriateness of content but also how multiple governance mechanisms exert control over creators, our participants desired transparency designs of moderation in plurality. They desired the platform could present moderation saliently, explain moderation profoundly, afford communication responsibly, and offer learning opportunities practically. Reflecting on their moderation experiences, we argue that creator moderation needs to maintain dynamic transparency to value YouTubers' labor. The key determinant of such balance lies in how to support YouTubers' learning about content creation, monetization, moderation, and more. We also discuss design claims of transparency design in creator moderation.

ACKNOWLEDGMENTS

This work is partially supported by the National Science Foundation, under grant no. 2006854. First, we appreciate 28 creators' participation in this study and shared their stories with us. We thank Dr. Xinning Gui for her feedback on this study. Last, we appreciate 1AC's guidance and all anonymous reviewers' valuable feedback.

REFERENCES

[1] Julia Alexander. 2019. YouTube moderation bots punish videos tagged as 'gay' or 'lesbian,' study finds. The Verge. Retrieved from https://www.theverge.com/2019/9/30/20887614/youtube-moderation-lgbtq-demonetization-terms-words-nerd-city-investigation

[2] Julia Alexander. 2019. LGBTQ YouTubers are suing YouTube over alleged discrimination. The Verge. Retrieved from https://www.theverge.com/2019/8/14/20805283/lgbtq-youtuber-lawsuit-discrimination-alleged-video-recommendations-demonetization

[3] Ali Alkhatib and Michael Bernstein. 2019. Street–level algorithms: A theory at the gaps between policy and decisions. In CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), Association for Computing Machinery, New York, NY, USA, 1–13. DOI: https://doi.org/10.1145/3290605.3300760

[4] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media Soc 20, 3 (March 2018), 973–989. DOI: https://doi.org/10.1177/1461444816676645

[5] Carolina Are. 2021. The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram. Fem Media Stud (2021). DOI: https://doi.org/10.1080/14680777.2021.1928259

[6] Sophie Bishop. 2018. Anxiety, panic and self-optimization: Inequalities and the YouTube algorithm. Convergence: The International Journal of Research into New Media Technologies 24, 1 (2018), 69–84. DOI: https://doi.org/10.1177/1354856517736978

[7] Lindsay Blackwell, Mark Handel, Sarah T. Roberts, Amy Bruckman, and Kimberly Voll. 2018. Understanding "bad actors" online. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 1–7. DOI: https://doi.org/10.1145/3170427.3170610

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

[9] 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). DOI: https://doi.org/10.1177/2056305120936636

[10] Bobo Chan. 2020. Is Being A YouTuber Still Lucrative? Jumpstart Magazine. Retrieved from https://www.jumpstartmag.com/is-being-a-youtuber-still-lucrative/

[11] Stevie Chancellor, Jessica Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. #thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16, ACM Press, New York, New York, USA. DOI: https://doi.org/http://dx.doi.org/10.1145/2818048.2819963

[12] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A Cross-Community Learning-Based System to Assist Reddit Moderators. Proc. ACM Hum.-Comput. Interact. 3, CSCW (November 2019). DOI: https://doi.org/10.1145/3359276

[13] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. Proc ACM Hum Comput Interact 2, CSCW (November 2018), 1–25. DOI: https://doi.org/10.1145/3274301

[14] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. DOI: https://doi.org/10.1145/3290605

[15] Lars Thøger Christensen and George Cheney. 2015. Peering into Transparency: Challenging Ideals, Proxies, and Organizational Practices. Communication Theory 25, 1 (February 2015), 70–90. DOI: https://doi.org/10.1111/COMT.12052

[16] Nicholas Diakopoulos and Michael Koliska. 2017. Algorithmic Transparency in the News Media. Digital Journalism 5, 7 (August 2017), 809–828. DOI: https://doi.org/10.1080/21670811.2016.1208053

[17] Bryan Dosono and Bryan Semaan. 2019. Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on reddit. In Conference on Human Factors in Computing Systems - Proceedings, Association for Computing Machinery, New York, New York, USA, 1–13. DOI: https://doi.org/10.1145/3290605.3300372

[18] Brooke Erin Duffy. 2020. Algorithmic precarity in cultural work. Communication and the Public 5, 3–4 (September 2020), 103–107. DOI: https://doi.org/10.1177/2057047320959855

[19] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "I always assumed that I wasn't really that close to [her]": Reasoning about Invisible Algorithms in News Feeds. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (April 2015), 153–162. DOI: https://doi.org/10.1145/2702123

[20] 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). DOI: https://doi.org/10.1145/3392845

[21] Casey Fiesler, Jialun Aaron Jiang, Joshua McCann, Kyle Frye, and Jed R. Brubaker. 2018. Reddit rules! Characterizing an ecosystem of governance. 12th International AAAI Conference on Web and Social Media, ICWSM 2018 (2018), 72–81.

[22] Mikkel Flyverbom. 2016. Digital Age| Transparency: Mediation and the Management of Visibilities. Int J Commun 10, 0 (January 2016), 13.

[23] Archon Fung, Mary Graham, and David Weil. 2007. Full Disclosure: The Perils and Promise of Transparency.

[24] Matthias Funk. 2020. How Many YouTube Channels Are There? Retrieved from https://www.tubics.com/blog/number-of-youtube-channels/

[25] Ysabel Gerrard. 2018. Beyond the hashtag: Circumventing content moderation on social media. New Media Soc 20, 12 (December 2018), 4492–4511. DOI: https://doi.org/10.1177/1461444818776611

[26] Tarleton Gillespie. 2018. Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press. Retrieved from https://www.degruyter.com/document/doi/10.12987/9780300235029/html

[27] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale: https://doi.org/10.1177/2053951720943234 7, 2 (August 2020). DOI: https://doi.org/10.1177/2053951720943234

[28] GMI Blogger. 2022. YouTube Statistics 2022. Retrieved October 24, 2022 from https://www.globalmediainsight.com/blog/youtube-users-statistics/

[29] 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), 205395171989794. DOI: https://doi.org/10.1177/2053951719897945

[30] James Grimmelmann. 2015. The Virtues of Moderation. Yale Journal of Law and Technology 17, (2015). Retrieved from https://heinonline.org/HOL/Page?handle=hein.journals/yjolt17&id=42&div=&collection=

[31] 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. Proc ACM Hum Comput Interact 5, CSCW2 (October 2021). DOI: https://doi.org/10.1145/3479610

[32] Hans Krause Hansen, Lars Thøger Christensen, and Mikkel Flyverbom. 2015. Introduction: Logics of transparency in late modernity: Paradoxes, mediation and governance. https://doi.org/10.1177/1368431014555254 18, 2 (April 2015), 117–131. DOI: https://doi.org/10.1177/1368431014555254

[33] Maya Holikatti, Shagun Jhaver, and Neha Kumar. 2019. Learning to Airbnb by Engaging in Online Communities of Practice. Proc ACM Hum Comput Interact 3, CSCW (November 2019), 228. DOI: https://doi.org/10.1145/3359330

[34] Nicholas Jackson. 2011. Infographic: The History of Video Advertising on YouTube. The Atlantic. Retrieved from https://www.theatlantic.com/technology/archive/2011/08/infographic-the-history-of-video-advertising-on-youtube/242836/

[35] 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. DOI: https://doi.org/10.1145/3359294

[36] 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. DOI: https://doi.org/10.1145/3338243

[37] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter?: User behavior after content removal explanations on reddit. Proc ACM Hum Comput Interact 3, CSCW (2019). DOI: https://doi.org/10.1145/3359252

[38] Shan Jiang, Ronald E Robertson, and Christo Wilson. 2019. Bias Misperceived: The Role of Partisanship and Misinformation in YouTube Comment Moderation.

[39] Lin Jin. 2020. The Creator Economy Needs a Middle Class. Harvard Business Review. Retrieved from https://hbr.org/2020/12/the-creator-economy-needs-a-middle-class

[40] Hilary Johnson and Peter Johnson. 1993. Explanation Facilities and Interactive Systems. In Proceedings of the 1st international conference on Intelligent user interfaces.

[41] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. 2020. Through the looking glass: Study of transparency in Reddit's moderation practices. Proc ACM Hum Comput Interact 4, GROUP (January 2020), 1–35. DOI: https://doi.org/10.1145/3375197

[42] D. Bondy Valdovinos Kaye and Joanne E. Gray. 2021. Copyright Gossip: Exploring Copyright Opinions, Theories, and Strategies on YouTube: Soc Media Soc 7, 3 (August 2021). DOI: https://doi.org/10.1177/20563051211036940

[43] Charles Kiene, Jialun Aaron Jiang, and Benjamin Mako Hill. 2019. Technological Frames and User Innovation: Exploring Technological Change in Community Moderation Teams. Proc. ACM Hum.-Comput. Interact. 3, CSCW (November 2019). DOI: https://doi.org/10.1145/3359146

[44] Rene F. Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2390–2395. DOI: https://doi.org/10.1145/2858036.2858402

[45] Susanne Kopf. 2020. "Rewarding Good Creators": Corporate Social Media Discourse on Monetization Schemes for Content Creators. Soc Media Soc 6, 4 (October 2020), 205630512096987. DOI: https://doi.org/10.1177/2056305120969877

[46] Yubo Kou. 2021. Punishment and Its Discontents: An Analysis of Permanent Ban in an Online Game Community. Proc ACM Hum Comput Interact 5, CSCW2 (October 2021). DOI: https://doi.org/10.1145/3476075

[47] Yubo Kou and Colin M. Gray. 2018. "What do you recommend a complete beginner like me to practice?" Proc ACM Hum Comput Interact 2, CSCW (November 2018). DOI: https://doi.org/10.1145/3274363

[48] Yubo Kou and Xinning Gui. 2020. Mediating Community-AI Interaction through Situated Explanation: the Case of AI-Led Moderation. Proc ACM Hum Comput Interact 4, CSCW2 (October 2020), 1–27. DOI: https://doi.org/10.1145/3415173

[49] Kyle Langvardt. 2017. Regulating Online Content Moderation. Georgetown Law Journal 106, (2017). Retrieved from https://heinonline.org/HOL/Page?handle=hein.journals/glj106&id=1367&div=39&collection=journals

[50] Ralph LaRossa. 2005. Grounded Theory Methods and Qualitative Family Research. Journal of Marriage and Family 67, 4 (November 2005), 837–857. DOI: https://doi.org/10.1111/j.1741-3737.2005.00179.x

[51] Min Kyung Lee, Anuraag Jain, Hae J.I.N. Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. Proc ACM Hum Comput Interact 3, CSCW (November 2019), 26. DOI: https://doi.org/10.1145/3359284

[52] 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. DOI: https://doi.org/https://doi.org/10.1145/3479573

[53] 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. DOI: https://doi.org/10.1145/3555150

[54] 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. In Companion Computer Supported Co-operative Work and Social Computing (CSCW'22 Companion). Retrieved from https://doi.org/10.1145/3500868.3559445

[55] Melissa J. Morgans. 2017. Freedom of Speech, the War on Terror, and What's YouTube Got to Do with It: American Censorship during Times of Military Conflict. Federal Communications Law Journal 69, (2017). Retrieved from https://heinonline.org/HOL/Page?handle=hein.journals/fedcom69&id=163&div=&collection=

[56] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. New Media Soc 20, 11 (2018), 4366–4383. DOI: https://doi.org/10.1177/1461444818773059

[57] Juho Pääkkönen, Matti Nelimarkka, Jesse Haapoja, and Airi Lampinen. 2020. Bureaucracy as a Lens for Analyzing and Designing Algorithmic Systems. Conference on Human Factors in Computing Systems - Proceedings (April 2020). DOI: https://doi.org/10.1145/3313831.3376780

[58] Jessica A. Pater, Oliver L. Haimson, Nazanin Andalibi, and Elizabeth D. Mynatt. 2016. "Hunger hurts but starving works": characterizing the presentation of eating disorders online. In Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW, Association for Computing Machinery, New York, NY, USA, 1185–1200. DOI: https://doi.org/10.1145/2818048.2820030

[59] Hector Postigo. 2016. The socio-technical architecture of digital labor: Converting play into YouTube money. New Media Soc 18, 2 (2016), 332–349. DOI: https://doi.org/10.1177/1461444814541527

[60] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 1–13. DOI: https://doi.org/10.1145/3173574.3173677

[61] Aja Roman. 2019. YouTubers claim the site systematically demonetizes LGBTQ content. Vox. Retrieved from https://www.vox.com/culture/2019/10/10/20893258/youtube-lgbtq-censorship-demonetization-nerd-city-algorithm-report

[62] Laura Savolainen. 2022. The shadow banning controversy: perceived governance and algorithmic folklore: Media Cult Soc (March 2022). DOI: https://doi.org/10.1177/01634437221077174

[63] Mark Scott and Mike Isaac. 2016. Facebook Restores Iconic Vietnam War Photo It Censored for Nudity. The New York Times. Retrieved from https://www.nytimes.com/2016/09/10/technology/facebook-vietnam-war-photo-nudity.html

[64] Joseph Seering. 2020. Reconsidering Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation. Proc ACM Hum Comput Interact 4, CSCW2 (October 2020), 1–28. DOI: https://doi.org/10.1145/3415178

[65] Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. 2018. The Social Roles of Bots: Evaluating Impact of Bots on Discussions in Online Communities. Proc ACM Hum Comput Interact 2, CSCW (November 2018). DOI: https://doi.org/10.1145/3274426

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

[67] Nischal Shrestha, Titus Barik, and Chris Parnin. 2021. Remote, but Connected: How #TidyTuesday Provides an Online Community of Practice for Data Scientists. Proc ACM Hum Comput Interact 5, CSCW1 (April 2021), 1–31. DOI: https://doi.org/10.1145/3449126

[68] Spandana Singh. 2019. Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content. Retrieved from https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/

[69] Nicolas P. Suzor, Sarah Myers West, Andrew Quodling, and Jillian York. 2019. What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. Int J Commun 13, (2019). Retrieved from https://ijoc.org/index.php/ijoc/article/view/9736

[70] Lars Thøger Christensen. 2002. Corporate communication: The challenge of transparency. Corporate Communications: An International Journal 7, 3 (September 2002), 162–168. DOI: https://doi.org/10.1108/13563280210436772/FULL/XML

[71] Kaitlyn Tiffany. 2019. The YouTuber union isn't really a union, but it could be a big deal. Vox. Retrieved from https://www.vox.com/the-goods/2019/7/30/20747122/youtube-union-fairtube-ig-metall-instagram

[72] Sarah J. Tracy. 2013. Qualitative Research Methods: Collecting Evidence, Crafting Analysis.

[73] Rebecca Tushnet. 2019. Content Moderation in an Age of Extremes. Case Western Reserve Journal of Law, Technology and the Internet 10, (2019).

[74] 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, Association for Computing Machinery, 1–22. DOI: https://doi.org/10.1145/3415238

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

[76] Richard Ashby Wilson and Molly K. Land. 2021. Hate Speech on Social Media: Content Moderation in Context. Conn Law Rev (2021). Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3690616

[77] Lindsey Wotanis and Laurie McMillan. 2014. Performing Gender on YouTube. Fem Media Stud 14, 6 (November 2014), 912–928. DOI: https://doi.org/10.1080/14680777.2014.882373

[78] Austin P. Wright, Omar Shaikh, Haekyu Park, Will Epperson, Muhammed Ahmed, Stephane Pinel, Duen Horng (Polo) Chau, and Diyi Yang. 2021. RECAST: Enabling User Recourse and Interpretability of Toxicity Detection Models with Interactive Visualization. Proc ACM Hum Comput Interact 5, CSCW1 (April 2021), 1–26. DOI: https://doi.org/10.1145/3449280

[79] Jing Zeng and D. Bondy Valdovinos Kaye. 2022. From content moderation to visibility moderation: A case study of platform governance on TikTok. Policy Internet 14, 1 (March 2022), 79–95. DOI: https://doi.org/10.1002/POI3.287

[80] How to Appeal. onlinecensorship.org. Retrieved from https://onlinecensorship.org/resources/how-to-appeal

[81] YouTube Partner Program overview & eligibility. YouTube Help. Retrieved from https://support.google.com/youtube/answer/72851?hl=en

[82] Advertiser-friendly content guidelines. YouTube Help. Retrieved from https://support.google.com/youtube/answer/6162278?hl=en#Adult&zippy=%2Cguide-to-self-certification

[83] Request human review of videos marked "Not suitable for most advertisers." YouTube Help. Retrieved from https://support.google.com/youtube/answer/7083671?hl=en#zippy=%2Chow-monetization-status-is-applied

[84] Your content and Restricted mode. YouTube Help. Retrieved from https://support.google.com/youtube/answer/7354993?hl=en-GB

[85] Appeal Community Guidelines actions. YouTube Help. Retrieved from https://support.google.com/youtube/answer/185111?hl=en&ref_topic=9387060

[86] Get in touch with the YouTube Creator Support team. YouTube Help. Retrieved from https://support.google.com/youtube/answer/3545535?hl=en&co=GENIE.Platform%3DDesktop&oco=0#zippy=%2Cemail

[87] What is a Content ID claim? YouTube Help. Retrieved from https://support.google.com/youtube/answer/6013276
[88] Protecting your identity. YouTube Help. Retrieved from https://support.google.com/youtube/answer/2801895?hl=en
[89] The Santa Clara Principles on Transparency and Accountability in Content Moderation. Retrieved from https://santaclaraprinciples.org/

A DEMOGRAPHIC INFORMATION OF PARTICIPANTS

Table 1. Participant information. Subscription # (fanbase) was collected on the date of the interviews. Status was identified by YouTubers themselves by the time spent on creating videos. Career refers to how long YouTubers consistently make videos for their primary channel. Category refers to the content category, which is defined by YouTube. "N/A" means the information that our participants chose not to disclose.

#	Sub #	Age	Status	Nationality	Race	Gender	Career	Category
P1	~ 25.8k	18	part-time	US	White	Male	5 months	Games
P2	~ 21.3k	23	full-time	US	White	Male	5 years	Games
P3	~ 6.6k	40	part-time	England	White	Male	3 years	Travel
P4	~ 52k	28	part-time	US	Black	Female	6 years	People
P5	~ 4.33k	19	part-time	England	White	Male	5 years	Technology
P6	~ 268k	29	full-time	US	White	Male	9 years	Animation
P7	~ 84.7K	29	full-time	US	White	Male	3 years	Games
P8	~ 177k	32	part-time	US	White	Male	3.5 years	History
P9	~ 365k	28	full-time	Germany	White	Male	2 years	Entertainment
P10	~ 23.1k	38	part-time	Mexico	Hispanic	Female	2.5 years	Education
P11	~ 292k	29	part-time	Brazil	White	Female	12 years	Entertainment
P12	~ 2.02k	21	part-time	England	White	Male	2.5 years	Education
P13	~ 124k	19	full-time	US	Hispanic	Male	4 years	Entertainment
P14	~ 88.6k	28	part-time	Colombia	Hispanic	Male	2 years	Education
P15	~ 12.6k	29	part-time	Mexico	Hispanic	Male	6 years	Education
P16	~ 35.5k	29	part-time	Mexico	Hispanic	Female	4 years	Technology
P17	~ 5.7k	21	part-time	US	N/A	Male	8 years	Entertainment
P18	~ 26.8k	29	part-time	Mexico	Hispanic	Female	3 years	Education
P19	~ 53.9k	32	part-time	Mexico	Hispanic	Female	3 years	Technology
P20	~ 8.8k	18	part-time	US	White	Male	2 years	Games
P21	~ 497k	25	full-time	US	N/A	Male	2 years	Entertainment
P22	~ 230k	22	part-time	US	White	Male	7 years	Animation
P23	~ 31.3k	48	part-time	Colombia	Latino	Male	5 years	Education
P24	~ 63.2k	31	part-time	US	White	Male	5 years	History
P25	~ 52k	48	part-time	US	White	Male	3 years	Film

P26	~ 5.51k	27	part-time	US	Asian	Female	1 year	Entertainment
P27	~ 60.6k	55	full-time	US	White	Male	2 years	Technology
P28	~ 21.4k	23	full-time	Demark	Mixed	Male	6 months	Entertainment

Received January 2022; revised July 2022; accepted November 2022.