"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

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

How social media platforms could fairly conduct content moderation is gaining attention from society at large. Researchers from HCI and CSCW have investigated whether certain factors could affect how users perceive moderation decisions as fair or unfair. However, little attention has been paid to unpacking or elaborating on the formation processes of users' perceived (un)fairness from their moderation experiences, especially users who monetize their content. By interviewing 21 for-profit YouTubers (i.e., video content creators), we found three primary ways through which participants assess moderation fairness, including equality across their peers, consistency across moderation decisions and policies, and their voice in algorithmic visibility decision-making processes. Building upon the findings, we discuss how our participants' fairness perceptions demonstrate a multi-dimensional notion of moderation fairness and how YouTube implements an algorithmic assemblage to moderate YouTubers. We derive translatable design considerations for a fairer moderation system on platforms affording creator monetization.

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

KEYWORDS: content moderation; creator moderation; fairness perception; algorithmic moderation; moderation experience; moderation fairness; YouTuber

ACM Reference format:

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. In *PACM on Human Computer Interaction,* Vol. 6, CSCW, Article 425, November 2022. ACM, New York, NY, USA. 28 pages. https://doi.org/10.1145/3555150

1 INTRODUCTION

In a creator economy where numerous content creators make a living on social media platforms such as YouTube and Twitch [13,47,80], the fairness of content moderation decisions has significant and growing impacts on their livelihoods. Video content creators claimed that Facebook cut their thousands of dollars compared with the typical advertising income they earned from videos [22]; creators from marginalized groups complained their ad income was unfairly reduced on YouTube compared with others [3,5,70], and YouTube

This work is partially supported by the National Science Foundation, under grant no. 2006854.

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

overturned over 2.2 million erroneous copyright claims that had already shared creators' ad income with copyright owners [77]. This thread of news also shows how content moderation could be fairly conducted on social media platforms is gaining attention from society at large.

Recent research from HCI and CSCW has investigated end-users' fairness perceptions after they experience content moderation, especially on whether and how certain factors could affect such perceptions. For example, Jhaver et al. discovered that users who received explanations of content removal tended to perceive moderation as fair on Reddit [43]. Vaccaro et al. tested how users experienced different appeal explanations written by either algorithms or humans on Facebook and found that their perceived fairness increased after receiving all types of explanations [92]. However, what is relatively less discussed in the literature is the formation processes of fairness perception: how users develop their perceived (un)fairness from experiences with content moderation. In this study, we argue fairness perception means that users invoke the notion of fairness to describe their experiences with not only the decisions they receive but also decision-making processes they experience [29,58,91].

Thus, we draw from the interpretive lens of procedural justice/fairness to investigate how users develop their fairness perceptions. Procedural justice refers to a decision-making process considering voice comprehensively and allocating resources [37]. Individuals' fairness perceptions are impacted by the quality of their experience with decision-making processes and not only with outcomes from such processes [91,103]. Specifically, individuals' social contexts (e.g., time, people) [60] help them compare the *consistency* of decision-making processes. And the extent of their representation (i.e., *voice*) in decision-making [61] could affect their fairness perceptions. This paper aims to situate this interpretive lens in content moderation on YouTube to analyze and unpack YouTubers' fairness perceptions on moderation. We ask:

How do YouTubers' fairness perceptions generate from moderation experiences?

To answer this question, we interviewed 21 YouTubers who monetized their video content[2] and had experiences with YouTube moderation. Through an inductive qualitative analysis [55,90], we found three primary ways through which participants assessed moderation fairness, including their voice in algorithmic visibility decision-making processes, equality across their peers, and consistency across moderation decisions and policies. Thus, moderation fairness is best viewed as a multi-dimensional notion. That is, our participants developed fairness perceptions from temporal, social, and technical dimensions of moderation.

The focus on fairness perception and the findings about how participants experienced the complex, bureaucratic procedures behind an unequivocal moderation decision allow us to observe how content moderation on YouTube manifests not as a single class of algorithms, but an algorithmic assemblage. Algorithmic assemblage refers to a mixed "infrastructure that supports implementation, maintenance, use, and evolution of algorithms, data, and platforms" [74]. On YouTube, algorithms of different purposes such as copyright, visibility, and

[2] Those YouTubers signed YouTube Partner program (YPP, https://support.google.com/youtube/answer/72851?hl=en) to be eligible to acquire ad income. So, they accordingly might encounter a series of content moderation that amateur YouTubers or viewers would not meet, such as demonetization [3,6]. In this paper, we refer to YouTubers as those creators who joined the YPP.

monetization work together to moderate our YouTuber participants. Thus, participants who experienced moderation could feel impacts from multiple sources, which tend to have ripple effects that are multiple and overlapping on their future content creation and monetization endeavors. The algorithmic assemblage, in turn, renders source of accountability less traceable should an unfairness issue arise, as well as the issue of moderation fairness a multi-dimensional notion.

This study contributes to the HCI and CSCW literature with multi-fold insights: (1) to our knowledge, this study is among the first to contribute a nuanced account of for-profit content creators' interactions with content moderation; (2) we reveal how the assemblage of various classes of algorithms serve platform governance purposes; (3) we offer a conceptual understanding of how a notion of multi-dimensional moderation fairness develops from moderation experiences; (4) we articulate transferable design considerations for fair moderation systems on platforms affording creator monetization; (5) we highlight for future study: beyond understanding user perspectives on moderation outcomes, we should not ignore whether and how users have different qualities of experiences in moderation processes.

2 BACKGROUND: Monetization and Algorithmic Content Moderation on YouTube

Our study site is YouTube, the largest video sharing platform and a place for video content creators to perform profitable content creation. This process of converting video creation to profits refers to "monetization." YouTube offers YouTube Studio dashboard [104], as shown in Figure 1 (left), for YouTubers to understand their channel/content analytics (e.g., monetization statues, viewer engagement rates). As a primary source for monetization, ad income is calculated from cost per 1,000 impressions (CPM) (i.e., the unit income per 1000 views), multiplying by the proportion of a video's total views to 1,000 [105]. So, viewership matters to creators' ad income. Along with viewership, according to YouTube's policies [106], viewers' engagement is also critical to YouTubers' ad income. That is because more viewer engagement might allow a video to become "viral" through YouTube's recommendation algorithms [107]. Thus, more viewer engagement could generate more profits.

YouTube Studio dashboard presents whether certain moderation happens. For example, when a YouTuber clicks the Content tab in Figure 1 (left), they can observe all the videos' analytics in a catalog style. Then, YouTubers might further find a specific video is imposed with 'limited ads' (i.e., yellow dollar symbol), as shown in Figure 1 (right). Thus, their future ad income of such videos will decrease, namely rendering demonetization punishments [2,21,62].

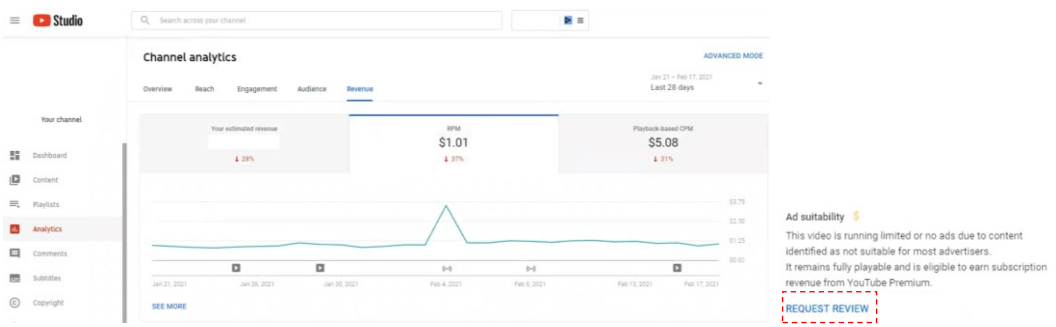

Fig. 1. YouTube Studio dashboard (left) and notification of 'limited ads' moderation decision (right).

YouTube largely utilizes algorithms to conduct video removal or account suspension [35] and moderation decisions that are less severe than these are also majorly adjudicated by moderation algorithms. For example, journalists and researchers reported that YouTube applies machine learning (ML) algorithms to detect copyright infringement and issue ad suitability (e.g., 'limited ads') decisions [108] based on the metadata of videos (e.g., titles, thumbnails, descriptions, captions, etc.) [3,36,73]. YouTube also uses automatic tools to hide videos that it deems as potentially mature under 'restricted mode' [109]. However, only when YouTube provides opportunities for "request review," as shown in Figure 1 (right), or appeal [110] can YouTubers initiate appeals to reverse potential false-positive moderation decisions. Hence, human reviewers on YouTube take important roles in auditing and reconciling platform-wide decisions made by moderation algorithms. YouTube's algorithmic moderation here provides a nuanced scenario for understanding how users interact with an algorithmic moderation system.

3 RELATED WORK

We introduce notions of justice/fairness and how researchers contextualize the notions in different settings. As fairness perception plays a part in social media users' interactions with moderation systems, we situate our work in past literature discussing algorithmic content moderation and fairness in moderation.

3.1 Procedural Justice and Fairness Perception

Given algorithms' growing role in decision-making processes in many aspects of our society, such as hiring and insurance, researchers have expressed their reflections on algorithmic fairness [15,49,56,96]. The key value underlying this line of work aligns with the notion of fairness.

Fairness is defined on different ontological bases. Some define algorithmic fairness as biases that coexist within the mathematical models [23,32]. To narrow the gap between true values and algorithms' expected values [65], researchers expand the model's training process to involve as many outliers as possible. Researchers have strived to reach proportionate classification performance by involving various groups' factors such as genders, race, education level, or computational literacy [15,38,96,98] in facial recognition algorithms [1,67], job hiring procedures [25,63], and criminal justice systems (e.g., risk score [14,51]). Growing efforts have also been made to improve algorithms' mathematic models by involving the notion of fairness into classifiers' formulation [27,101]. Outside of academia, ML practitioners further focus on collecting comprehensive datasets for model training and avoiding biased manual labeling[3] to ensure algorithmic fairness [41]. However, focusing on purely mathematical formulation is not the sole way to define algorithmic fairness [95]. A "fair" algorithm developed in a fair mathematical and economic setting could be perceived as unfair

[3] In the algorithmic model's training process, there is no ground truth to identify if specific historical decisions or manual labels offered by crowdsourcing is right and unbiased.

by end-users from different social contexts [57]. Furthermore, people's perceptions of fairness and trustworthiness on algorithms could be affected by algorithmic accuracy [14], and people might judge human decision-making as fairer than algorithmic one [56]. Thus, there is no absolute definition of algorithmic fairness to align with existing mathematical definitions [15,94].

Moving beyond the technical and mathematical definitions of fairness, HCI and CSCW researchers' investigations of fairness have engaged with concepts and theories from philosophy and social sciences. One of the most famous notions is John Rawls' *Justice as Fairness*, where he maintains that all individuals ought to be guaranteed equal rights of liberty and thus achieve different positions [31,71]. At the same time, social organizations need to contextualize individual differences to allocate resources to secure the value of equity [64].

This justice notion stresses outcome equality and resonates well with *procedural justice*. Procedural justice focuses on a fair decision-making process considering voice comprehensively and allocating resources [37]. People's fairness perception is impacted by the quality of their experience in decision-making processes and not only the results of the processes [91,103]. In detail, when people commit wrong acts by violating policies or laws, the authority in the juridical system should allow these violators to *voice* their side of stories and then makes a neutral judgment [60]. If people could express their voice before decisions are made, it could enhance the possibility of reaching the deemed equitable decisions [89]. Also, when people believe that their voice is included in decision-making processes, they tend to believe that they have a say in the outcomes they would experience later, thus increasing perceived fairness [61]. Without voice or control in decision-making processes, people may develop different fairness perceptions when comparing their expectations of outcomes, others' experiences, and what they think others experience with their own [60]. This reflects a key criterion in procedural justice: *consistency*, meaning "similarity of treatment and outcomes across people or time or both" [91]. Thus, voice and consistency constitute the major procedural justice notions.

Fairness perception is conditioned upon various social contexts [84]. Education, for example, could be a significant predictor to understand fairness perceptions of algorithms [76]. Wang et al. [96] found that users with lower education levels tended to perceive favored outcomes as fairer than users with higher education levels. Van Berkel et al. [15] added that users having higher algorithmic literacy, especially the females with higher education levels, tended to perceive algorithms as less fair. Beyond education, Lee et al. [59] suggested attention to resource allocation because when people receive more efficient resource distribution, they might perceive algorithms as fairer. Especially when factors such as race and gender [49] or loan repayment rates in a loan approval system, have been involved in algorithms [78], users might perceive acquiring resources proportionally as fairer. Besides, researchers have also pointed out that algorithms' clarity or transparency (e.g., outcome explanations) is a crucial condition in helping users recognize the fairness of decision-making processes [58]. This line of diverse work presents the importance of contextualizing fairness in specific social contexts and recognizing its plurality [57].

3.2 Algorithmic Moderation and Fairness in Content Moderation

Early day content moderation relied upon manual intervention to deal with identified deviant behaviors [19,52,53]. Manual or human moderation plays a role in "structuring participation in a community to facilitate cooperation and prevent abuse" [39]. However, manual moderation is time-consuming and costly [39,82]. Reviewing a sheering volume of potentially harmful content would also bring psychological harm to human workers [72,85], engender emotional labor [26,97], and cause epistemic ramifications [66]. Thus, social media platforms have increasingly implemented algorithms in their moderation systems [7,24,44].

However, moderation algorithms have intrigued various concerns from the public. For example, the international society is concerned that automatic tools need to be designed as accountable to protect human rights (e.g., free speech) [83]. Journalists stressed that algorithms could not perfectly enforce content policies given users' complex language use, cultural background, and intentions of generating content [82]. Even social media platforms admitted that over-reliance on algorithms hurts more innocent users, and they thus reverted to deploy additional human moderators [12].

In this background of concerning about algorithmic moderation's issues, prior researchers have started to investigate users' perspectives on moderation after they experience it. One of academic attention is focused on the fairness aspect of moderation. Prior researchers have uncovered the unfairness of moderation appeared in users' marginalization. For instance, Haimson et al. uncovered that sexual and racial minority groups who expressed their personal identities tended to experience content removal higher than others [40]. Sybert found that sexual minority groups contested NSFW (not safe for work) content ban by posting content that condemned Tumblr's new platform policies and legitimacy of moderation [88]. Furthermore, Feuston et al. found pro-eating (Pro-ED) order users considered account suspension as unfair because they lost opportunities of reflecting on their personal content and receiving community support [30]. Even though Pro-ED users were found to circumvent moderation, Gerrard has stressed that moderation should impartially protect such an already marginalized group [33].

The lack of transparency in moderation decision-making might incur perceived unfairness. Suzor et al. [86] uncovered users felt largely unfair about their content removal or account suspension because they did not receive detailed explanations on what rules they violated. Researchers also found that YouTubers with small fanbase considered moderation punishments were unfairly imposed on them compared with large channels [21], and they requested explanations for moderation decision-making processes [62].

Research from HCI and CSCW has also investigated users' fairness perceptions on moderation cases. Lampe et al. [53] have found that when working as voluntary meta-moderators, users can rate moderation decisions as either "fair" or "unfair" on Slashdot to reverse unfair moderation decisions or remove low-quality moderators. More recent research has tended to focus on how certain factors could influence users' fairness perception given their moderation experiences. For instance, Jhaver et al. [43] discovered that users on Reddit who experienced content removal with receiving explanations perceived the moderation decision as fair. In the phase of contesting moderation, Vaccaro et al. [92] simulated Facebook's appeal process and uncovered that after receiving explanations of appeals, users' fairness perceptions improved.

The variety of work above shows that (1) prior researchers have considered unfairness of moderation as biased decisions for marginalized people and (2) for users in general, prior work has been focused on whether and how users generate a binary conclusion between fairness or unfairness. However, relatively little attention has been paid to unpacking or elaborating on the formation processes of fairness perception: how users develop their perceived (un)fairness after experiencing content moderation. Thus, we aim to fill this research gap.

4 METHODS

We conduct a qualitative study by interviewing 21 YouTubers with a semi-structured interview protocol and use inductive qualitative analysis to analyze the whole interview dataset.

4.1 Data Collection

We conducted 21 interviews with YouTubers who had experienced YouTube moderation, as shown in Table 1. After obtaining the Institutional Review Board (IRB) board's approval in our institution, we used both purposeful sampling and snowball sampling [90] to recruit participants to join this study. For the purposeful sampling, we created a recruitment website describing the criteria of this study: recruiting a YouTuber who is over 18 years old and has experienced YouTube moderation (e.g., 'limited ads,' 'age restriction,' etc.). We disseminated this website on social media platforms such as Reddit (e.g., r/youtube, r/PartneredYoutube) and Twitter. We searched keywords related to YouTube moderation, such as 'demonetization' on Twitter, to directly message YouTubers who have shared their moderation experiences in their tweets. For the snowball sampling, we let interviewed YouTubers introduce YouTubers in their community who also experienced moderation to join our study. As shown in Table 1, we recruited YouTubers from various content categories (e.g., games, technology, entertainment, education) with a wide scope of subscription numbers (i.e., fanbase ranging from ∼2k to ∼497k on interview dates). In this study, nearly every participant was compensated with a 20 dollars gift card, except three proactively and firmly expressed their will not to be recompensated.

Table 1. YouTubers' demographic information. Subscription number (fanbase) was accordingly collected on the date of the interviews. Work status is self-identified by YouTubers depending on their time spent on video creation. Career refers to the consistent period of video creation for their primary channel on YouTube by the date of interviews. Category refers to the channel's content category, which is defined by YouTube (note: animation is under the entertainment category). Recruit indicates a participant is either recruited by purposeful sampling (coded as 0) or snowball sampling (1). "N/A" means that our participants did not disclose the information.

#	Subscription #	Age	Work Status	Nationality	Race	Gender	Career (yrs.)	Category	Recruit
P1	~ 25.8k	18	part-time	US	White	Male	0.5	Games	0
P2	~ 21.3k	23	full-time	US	White	Male	5	Games	0
P3	~ 6.6k	40	part-time	England	White	Male	3	Travel	0
P4	~ 52k	28	part-time	US	Black	Female	6	People	0

P5	~ 4.33k	19	part-time	England	White	Male	5	Technology	0
P6	~ 268k	29	full-time	US	White	Male	9	Animation	0
P7	~ 84.7K	29	full-time	US	White	Male	3	Games	0
P8	~ 177k	32	part-time	US	White	Male	4	History	0
P9	~ 365k	28	full-time	Germany	White	Male	2	Entertainment	0
P10	~ 23.1k	38	part-time	Mexico	Hispanic	Female	3	Education	0
P11	~ 292k	29	part-time	Brazil	White	Female	12	Entertainment	0
P12	~ 2.02k	21	part-time	England	White	Male	3	Education	0
P13	~ 124k	19	full-time	US	Hispanic	Male	4	Entertainment	0
P14	~ 88.6k	28	part-time	Colombia	Hispanic	Male	2	Education	1
P15	~ 12.6k	29	part-time	Mexico	Hispanic	Male	6	Education	1
P16	~ 35.5k	29	part-time	Mexico	Hispanic	Female	4	Technology	1
P17	~ 5.7k	21	part-time	US	N/A	Male	8	Entertainment	0
P18	~ 26.8k	29	part-time	Mexico	Hispanic	Female	3	Education	1
P19	~ 53.9k	32	part-time	Mexico	Hispanic	Female	3	Technology	1
P20	~ 497k	25	full-time	US	N/A	Male	2	Entertainment	0
P21	~ 230k	22	part-time	US	White	Male	7	Animation	0

We held interviews as well as recorded and transcribed them through Zoom. The duration of each interview ranged from 28 minutes to 94 minutes (Average = 66.5), with the median equaling 63.5 minutes. The duration of conducting interview procedures, i.e., data collection, was from Jan to Mar 2021. Before starting every interview, we requested and acquired verbal consent from participants, confirming their willingness to join this study. Also, we informed them that their personal information would be anonymous and protected, and they reserved the right to withdraw from the interviews whenever they wanted. We followed a semi-structured interview protocol (see Appendix A) to conduct each interview with YouTubers. Based on prior work (e.g., [43,62,92]) that discussed users' moderation experiences, we designed a part of questions to understand how YouTubers experience and handle moderation. We further designed another part given the functional dimensions on YouTube [105,111] to investigate YouTubers' fairness perceptions. In the process of conducting interviews, once we found intriguing points related to our research questions or unique experiences that need to be elaborate, we put forward probes, i.e., asking follow-up questions. Additionally, many of our participants shared their screens during Zoom interviews or sent screenshots through emails to permit us to use them. This significantly supplemented our data collection.

4.2 Data Analysis

We conducted an inductive qualitative analysis to analyze all interview data [55] by NVivo 12. NVivo 12 stored all interview transcript data, and two researchers read through it all. They first obtained an initial impression of the size of the dataset and YouTubers' moderation experiences. In a weekly meeting, they discussed the initial understanding and agreed to start the 'open coding' process, given the richness of the dataset. Then, two researchers separately returned to the dataset and assigned discrete codes, either to sentences or paragraphs. The purpose of this step is to convert textual data into condensed codes. During three weeks of open coding and weekly meetings, two researchers discussed and resolved their disagreements on their assigned codes, which then were altered to describe the data appropriately. After completing opening coding, two researchers started the iterative 'axial

coding' process. They re-read the quotes and associated open coding codes and classified these codes together into higher-level concepts (i.e., axial coding codes).

After two researchers moved back and forth between codes and associated data, they identified the connections between axial coding codes (i.e., categories), combined them into themes, and dissolved disagreements weekly. Concepts and ideas from the procedural justice scholarship informed the axial coding process. We looked at whether some axial coding codes could align with notions in procedural justice (e.g., voice, consistency). If so, we paid attention to how they could be gathered into higher-level themes to end the selective coding process. In the final preparation of presenting findings, the researchers removed themes that did not have enough data to support them. This data analysis process ended by generating overarching themes from axial coding codes to answer the research question.

5 FINDINGS

We found how our participants developed fairness perceptions from their moderation experiences: (1) they encountered unequal moderation treatments through cross-comparisons; (2) they observed that the moderation system made inconsistent decisions, processes, system actions or those inconsistent with content policies; (3) they did not have voice or control in multiple algorithmic visibility decision-making processes. Both inconsistency and the lack of voice violate the core principles in procedural justice.

5.1 Equality in Comparing Moderation Treatment

Equal treatment matters to fairness perception [102]. Equality perception refers to how YouTubers actively compared their moderation treatments with others to assess moderation fairness. YouTube's public statement [112] endorses the value of equality by claiming that nearly all YouTube's policies are equally applied to YouTubers. However, our participants thought differently and claimed that they could observe unequal moderation actions through cross-comparison. For example, several small YouTubers[4] stated that large YouTubers received preferential treatments. P4, a small YouTuber creating fitness videos, described her experience and reasoning to us:

> There are female YouTubers creating fitness videos with million-plus subscribers, so I'm not sure what difference is between their content and mine, other than the person itself. (…) I'm a black woman, so I don't know if race has a part to play in it, but I just want to point it out that if YouTube is flagging their content, at least, the women that are in control of those channels are gonna speak on it, the way that many YouTubers do. So, it leads me to believe that they are not experiencing the demonetization issues, but I am [experiencing] in such a large volume and a high frequency. [P4]

P4 performed observations and cross-comparisons on large YouTubers' experiences in the same content category with her. She assumed that those large YouTubers would complain in their videos and let viewers know if they experienced YouTube moderation. Since those she

[4] In this paper, we define 'small YouTuber' as those in the YouTube Partner program having more than 1,000 but less than 100K subscribers. Large YouTuber correspondingly refers to those with more than 100K subscribers.

observed did not complain, she inferred that they did not experience the highly frequent moderation as her. Thus, she felt it unfair that she received disproportional moderation, different from large YouTubers.

Prior work has uncovered that sexual minority and small YouTubers felt demonetization punishment was unfair compared with large YouTubers [21,62]. Beyond that, our participants further reported fairness issues related to support resources they had access to upon receiving penalties. Small YouTubers felt unfair when receiving less support to remedy moderated content. For example, P17 described:

> If you contact a YouTuber with 100K or more, they are usually assigned a partner manager. With that power, they can communicate with someone who works at YouTube directly and get better specifications to fix their videos. With someone like myself who's under that threshold, usually I might have to talk to a content creator [large YouTuber], whom I know and say: "Hey, can you ask your partner manager or whatever?" and sometimes they say yes and sometimes they say no, [they are not] willing to help me. [P17]

Partner managers refer to YouTube-hired experts helping YouTubers grow their channels. As claimed by YouTube, "we determine partner eligibility according to channel size, channel activity, and adherence to YouTube" for inclusion [113]. P17 observed that this service also helped large YouTubers repair moderation issues. So, he sensed YouTube moderation's unfairness from acquiring unequal distribution of support resources (e.g., communication, human reviewers' support) to solve moderation issues given different channel sizes.

Echoing the sentiment, some participants with a large fanbase also acknowledged that YouTube might issue unfair moderation on others, especially the small ones. P6, a large YouTuber creating anime videos, told us:

> One thing that's helped a bit is the form that I can fill out. That gives me a breakdown of the guidelines, and I'm able to figure out what is overall acceptable on YouTube. And then I apply that to other videos that I see and be like, hey, this is being unfair to this, this should have been fine like what the example of [YouTube A (a small anime YouTuber)], his video obviously should not have been marked as 'made for kids' because there's literally some blood. There's some humor that kids would be completely confused. It was very obviously targeted toward young adults. [P6]

In this example, the form that P6 mentioned refers to the self-certification function for YouTubers to self-report whether their new content complies with content policies before uploading it [111]. 'Made for kids' is a moderation tag (in audience settings) directing specific videos primarily for viewers under 13 years old and disabling a series of features at both video and channel levels [114] (e.g., comment close and channel membership close [4]). P6, a relatively large anime YouTuber, observed uneven moderation treatments when applying the knowledge that he acquired from using self-certification to other small YouTubers who created similar videos. He thus considered moderation as unfair for small YouTubers because he thought they should not have been tagged 'made for kids.'

In sum, our participants attributed unequal moderation treatments to their different fanbases or identities. This perceived inequality existed in moderation decisions, the

frequency or severity of moderation practices they encountered, and disproportionate resources acquired to repair moderation issues. Thus, our participants felt their moderation experiences as unfair after their cross-comparisons.

5.2 Consistency within Algorithmic Moderation Decisions

YouTubers' fairness perception hinges on consistency across moderation decisions and policies. However, our participants perceived the unfairness when they observed that the moderation system made inconsistent decisions or ones inconsistent with content policies.

5.2.1 Consistency between Moderation Algorithms and Content Policy

YouTube relies on machine learning algorithms to enforce content policies to moderate videos through metadata (e.g., titles, thumbnails, descriptions, captions, etc.) [3,36,73]. And YouTubers have developed folk theories regarding what texts would cause moderation to happen [3,68]. Our participants described how such knowledge informed them of misalignment between algorithmic decisions and content policies, resulting in the feeling of unfairness. P7, a YouTuber creating games-related videos, said:

> The demonetized one [video] was a video game related to World War Two, and it included Hitler [in the captions]. I was talking about World War Two or Germany in this context. YouTube is very sensitive [to violent content], even though it is a video game. Of course, they don't really care. (…) Basically, they didn't explain anything. It was very opaque. [P7]

In the above example, P7 referred to the 'demonetized one' as a video with 'limited ads' [108], denoting that few or no advertisers would like to place ads on the video. YouTube's advertiser-friendly content guidelines [115] state that "violence in the normal course of video gameplay is generally acceptable for advertising, but montages, where gratuitous violence is the focal point, is not." P7 deduced the keyword "Hitler" in his video might be associated with violence, but he considered his video unrelated to violence at all. Thus, he felt it unfair that YouTube's moderation failed to be in line with content policies.

Prior work uncovered that Facebook inconsistently applied content rules among users [54,92]. In a similar vein, our YouTubers participants were particularly concerned about the consistency between the actions of moderation algorithms and content policies. They considered YouTube moderation as unfair when moderation algorithms seemingly took more time to implement updates in content policy. For instance, P10, a YouTuber creating science and education videos, described:

> I felt unfair because it (YouTube) only stated that they didn't find it suitable for their advertisers. (…) During the same week or two, all of our channels were demonetized on COVID-19 videos, so I was constantly in touch with other YouTubers, and we concluded that the problem was the COVID-19 situation (…). I think it may have been from May to July [, 2020]. We would all still make videos telling people to take care, but we would not say 'COVID-19'; (…) our community just used euphemisms that kept us on the safe side. [P10]

Limited explanations of 'limited ads' motivated P10 to collaborate with her community to observe, exchange information, and discuss their past experiences. P10 and her community showed their benevolent intentions of complying with advertiser-friendly guidelines, and the content policy explicitly claimed support for science and education videos. YouTube updated its policy, allowing YouTubers to discuss COVID-19 without 'limited ads' [116]. However, before that time, the community collaboratively questioned moderation's fairness due to conflicts between the intentions they assumed their videos had, namely disseminating scientific knowledge of COVID-19 and algorithm's classification based on the keywords of their videos.

Prior work suggested that inconsistent enforcement of content policies might lead to chaotic communities [79]. However, our participants such as P10 showed they did not insist on the negative unfairness perception; rather, they proactively managed to solve their moderation issues through community-wide support.

Moreover, participants experienced unfairness when anticipated moderation actions did not take place. For example, P7 and P11 shared similar experiences:

> There's nudity in the game [video] I forgot to censor, several parts of nudity, and it's completely monetized. It has almost one million views, and there's literally nudity. It's not accurate at all. The [moderation] algorithm is absolutely terrible on YouTube. [P7]

> Recently I did a video where I cursed (…). The whole video was me cursing at them. And it didn't get demonetized. I was a bit shocked how this didn't get demonetized. (…) [YouTube is in the] really bad accuracy. When I do high-value projects, I get demonetization. When I do this trolling, I don't get demonetized. [P11]

In the above two examples, while creating different categories of content, P7 and P11 both felt it unfair that YouTube did not issue 'limited ads' to videos that they perceived as not advertiser-friendly. Also, YouTube demonetized the videos which P11 presumed as advertiser-friendly. Both participants showed they understood what content would violate advertiser-friendly guidelines [115]. They then attributed the moderation conduct that was not in line with content policies to the issues of algorithmic accuracy. The above cases presented that the perceived fairness emerged when moderation systems failed to consistently implement content policies.

Lastly, our participants noticed that YouTube might run different moderation algorithms on several services. For instance, P9, a YouTuber who also did live-streaming, said to us:

> I had the live stream [on YouTube Live] where I took videos that were already uploaded on YouTube; (…) they were monetized, and I played them in a live stream 24/7 (…). It did so for two months, but then suddenly, I received a warning from YouTube saying that my live stream violated the community guidelines, which doesn't make any sense to me even though I can appeal. [P9]

YouTube Live is a live streaming service provided by YouTube; P9 displayed the already published videos on the YouTube live. Those videos were in all normal statuses on YouTube, which thus gave him a policy-level signal that his videos did not violate any rules. However, he found that YouTube moderated his live streaming after two months of displaying the already

published pre-recorded videos. He felt it unfair why YouTube ran an inconsistent level of content moderation between YouTube and YouTube Live while both should moderate content equally given policies [117].

5.2.2 Consistency in Moderation Explanations

Participants' feelings of unfairness arose when they received inconsistent moderation explanations from different entities in the moderation system. For example, P2, a YouTuber making reaction videos, said to us:

> My channel was deleted for spam scams or commercially deceptive content, which is incredibly unfair because there's nothing in that category that I did. It happened literally after a third copyright strike (warning). (…) "If you think this was done incorrectly, email this back, and we'll review it." [P2]

Reaction video is a type of video content where a YouTuber video record their real-time reactions (facial expressions, physical postures, language, etc.) when watching other people's video. YouTube's content policy [118] claims that once YouTubers receive a third copyright strike (i.e., warning), their channels will be suspended. P2 experienced inconsistency that he expected a copyright strike, but YouTube's explanation cited the reason for generating scams to suspend his channel. Hence, he felt the reason for suspending his channel was unreasonable and unfair. YouTube is clearly aware of the potential limits in their algorithms and thus suggests that YouTubers like P2 could contact human moderators to correct potentially disagreed moderation decisions.

Participants also received explanations that were inconsistent after they contacted human reviewers (e.g., creator support [119]) behind the moderation system. For example, P13, a YouTuber making comedy videos, said to us:

> My bigger video got taken down by scams, that was ridiculous, so I kept @ing them on Twitter. Four days later, I received an email from creator support. (…) "from our internal team, your video was intended to incite violence or dangerous illegal activities." So that's completely different from 'scams or deceptive practices.' I was like, please explain to me how my video was intended to incite illegal activities that have an inherent risk of serious physical harm or death now. [P13]

The example above showed the inconsistency between the algorithmic explanation and the human explanation that P13 received. The first explanation already invoked P13's sense of unfairness; the inconsistency between the first and second exacerbated this feeling. P13 claimed that his video did not intend to promote violence and thus urged a convincing and detailed moderation explanation to alleviate his sense of unfairness.

5.2.3 Consistency between Different Algorithms

YouTubers noticed an inconsistency between different algorithms in YouTube moderation. Some participants reported how YouTube's moderation algorithm and monetization algorithm were inconsistent. For instance, P3 described:

> I know that (YouTube) they're playing advertisements because people will leave a comment saying there are three ads in this video. (…) This is a video that had the yellow dollar sign. So, it's amazing that they claimed it'd been known to [be] demonetized, yet they're still putting ads on these videos. [P3]

The yellow dollar symbol refers to 'limited ads.' A video with 'limited ads' should have few or no ads. However, P3 found that a high quantity of ads was placed on his video with 'limited ads,' but P3 could not earn much from it. P3 thus developed perceptions of unfairness.

Similarly, YouTube's moderation algorithms and visibility/recommendation algorithms were not always consistent. Without any notifications or explanations, YouTube could incoherently limit the direct visibility of videos, leaving YouTubers to test whether they experience content moderation on their own. For example, P21 said to us by showing one piece of evidence from Figure 2.1 to 2.3:

> [When I] turned on the restricted mode filter, half of my videos just disappeared. (…) There's no label from the creator (Studio) dashboard. You have to @ YouTube on Twitter to access the appeal [form] because some YouTubers don't even know it exists. That's how I got access to this form where I can send them the video links to appeal to this restricted mode. Still, again, since they (YouTube) don't tell you that your videos [are] unrestricted to begin with, they will not tell you if your video is restricted correctly. [P21]

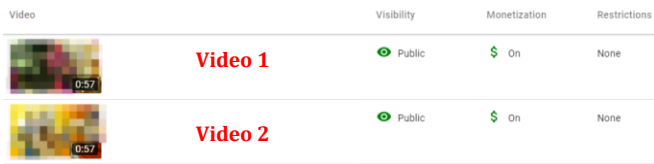

Fig. 2.1. Two videos are in all normal statuses on the YouTube Studio dashboard.

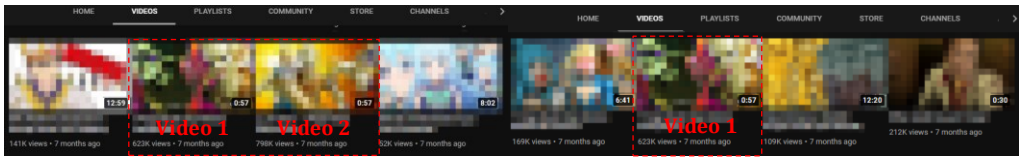

Fig. 2.2. When the restricted mode is off.

Fig. 2.3. When the restricted mode is on.

In the above example, 'restricted mode' is a content filter that prevents viewers from videos containing potentially mature content. At a policy level, P21's video had all normal statuses (i.e., visibility, monetization, and restriction), indicating his video was neither 'age-restricted' nor 'made for kids' nor unfriendly for advertisers. In other words, he deemed that his video did not violate any content policies. However, YouTube blocked the direct visibility of videos for audiences who opened 'restricted mode' or without accounts signed in. P21 was only able to find this inconsistent moderation by switching to another YouTube account and testing with the 'restricted mode,' as shown from Figure 2.1 to 2.3. Thus, he perceived it unfair that YouTube conducted inconsistent moderation on his videos.

5.3 Voice in Algorithmic Visibility Decisions

Procedural justice assumes that voice in decision-making processes enhances people's perceived fairness and is more likely to produce an equitable outcome [89]. Voice means allowing "people an opportunity to provide inputs to decision maker" [16]. Prior work has stressed that platforms need to have users' voice/representation in moderation decision-making for designing contestability for moderation [93] or embedding procedural justice in it [29]. Users' feelings of fairness could improve if users feel their voice has been heard [92]. Resonating with this line of work, we found our participants wanted their voice to be heard either in moderation-decision making or after the issuance of moderation decisions. But beyond that, we uncovered that when YouTubers realized their voice or input was not heard, the effects of moderation-related algorithms had been already taking actions, triggering ripple negative effects on their YouTube channels. So, their perceived unfairness arose. For instance, P21 shared with us his comparative evidence (see Figure 3) between a properly monetized video and two videos with different types of moderation decisions:

> When [my] videos are in the limited ads, they just take your video completely out of the algorithm. (…) they will not push it in browse features; they will not push it in your suggested videos; they will not push it through notifications. People can still search it up. They can find it on your channel, but it's incredibly unlikely to find it that way, which is terribly unfair. [P21]

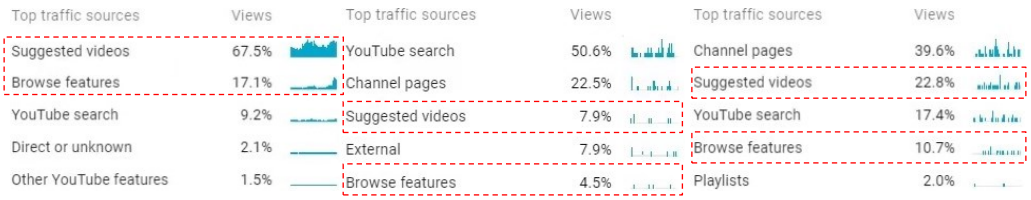

Fig. 3. Different traffic sources of videos with full monetization (left), 'limited ads' (middle), and tags of 'made for kids' (right).

In this example, the moderation tag, 'made for kids,' directs videos' visibility to audiences under 13 years old and disables specific videos features (e.g., commenting). Such moderation tag, according to YouTube, should not have affected his visibility in search and recommendation algorithms [120]. However, by observing and analyzing the videos' different online traffic sources, P21 found that once he received either 'limited ads' or 'made for kids,' YouTube algorithmically constrained the visibility of his videos (i.e., diminished 'suggested video' rate and 'browse feature'[42] in Figure 3). This ultimately led to less watch retention time and subscription increment. P21 showed us that he could tell and describe moderation's various negative impacts on him. However, none of the algorithmic visibility decisions had considered his voice, so his perceived unfairness arose.

Besides the algorithmic impacts caused by one moderation decision each time (e.g., either 'made for kids' or 'limited ads'), YouTube moderation overlapped between different types of it to harm participants' performance metrics. For example, P8 said:

> This was a (…) limited ads video got age-restriction it but beyond that. They also delisted it from search, so the best way to tell that is to type in my channel's name and

> the name of the video. And if it doesn't show up in that search thing, there's something wrong here, especially when it's a video that got about 20,000 views in the first day. It's obvious shadowban because you cannot find it. [P8]

In this example, 'age-restriction' refers to the moderation decision where YouTube sets videos unviewable to "users under 18 years of age or signed out;" at a policy level, an age-restricted video might also be considered not advertiser-friendly (i.e., 'limited ads'), which is what P8 experienced. However, YouTube further disabled the public searchability of P8's video, which exceeded his common understanding. P8 used 'shadowban' to describe the phenomenon that YouTube seemed to block the direct visibility of his video from non-subscribers. This case presented that P8 was unable to express his voice before the multiple algorithms made decisions and took effect, arising his perceived unfairness.

Furthermore, once experiencing moderation, participants found that moderation algorithms even affected their future channel performance, so they doubted YouTube moderation's fairness. For example, P13 shared his observation with us from Figure 4.1 to 4.2:

> I gotta get off this call (moderation), and now I'm working on my next video, which I'm trying to make extra great. Because I want to hopefully get back in the [recommendation] algorithm again, but I can see this video that's like eight minutes that I put several hours of editing into, as I spent almost three days straight editing it, literally not being pushed out. It's unfair because I can compare it to my last omega (popular) video's views, and I can look at an omega video from a few months ago. [P13]

Visibility	Monetization	Restrictions	Date ↓	Views	Comments	Likes (vs. dislikes)
Public	On	None	Feb 15, 2021 Published	5,858	211	99.4% 1,457 likes
Public	On	None	Feb 8, 2021 Premiered	12,931	256	99.0% 1,615 likes
Public	On	None	Jan 29, 2021 Published	13,863	253	98.5% 1,558 likes
Public	On	None	Jan 25, 2021 Published	34,287	540	99.3% 4,268 likes
18+ only	Limited	Age restriction +1 more	Jan 15, 2021 Published	244,750	1,312	99.0% 21,079 likes
Public	On	None	Jan 4, 2021 Published	563,114	2,905	97.7% 50,179 likes

Fig. 4.1. The underperformance of future videos.

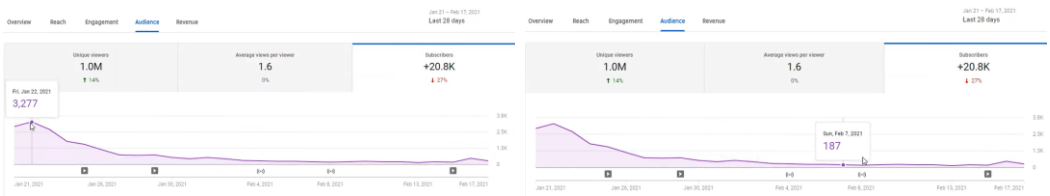

Fig. 4.2. Channel's longitudinal changes of subscription before (left) and after (right) moderation.

In the above example, once receiving a combined moderation decision of 'limited ads' and 'age-restriction,' P13 found that all his new videos, even without these penalties, underperformed in viewership and viewer engagement metrics, as shown in Figure 4.1. Thus, by comparing with his past similar videos, he felt it unfair that YouTube did not promote his videos even though they had all normal statuses to monetize. P13 further observed that his subscription quantity increments largely dwindled at a channel level once experiencing moderation, as shown in Figure 4.2. However, YouTube neither showed how the moderation system made decisions to render P13's channel underperform in different visibility metrics nor had his voice in making these algorithmic decisions, implying his perceived unfairness.

P20's thought after he experienced moderation summarized the perceived unfairness where our participants felt their voice was not heard by moderation algorithms:

> "When your videos get limited ads, your other videos may underperform [in metrics]. There's so much that can be hard to wrap your head around because you have to [assure] at least two algorithms work correctly, the monetization and the recommendation algorithm." [P20]

As P20 noted, YouTubers needed to navigate between different 'black-box' algorithms. From a procedural justice perspective, they were not given the power to express their voice before algorithms made moderation decisions, and thus, they failed to comprehend why the ripple effects happened.

So, in sum, our participants encountered a complex set of moderation decisions resulting from multiple classes of algorithms, as well as a series of algorithmic impacts beyond what they understood as fair. The lack of voice in these decision-making processes aroused their feeling of unfairness.

6 DISCUSSION

Through an interview study with 21 YouTubers, we unpack and elaborate on the formation processes of how our participants develop their fairness perceptions from experiences with content moderation on YouTube: they reported (1) unequal moderation treatments, (2) inconsistent moderation procedures and outcomes or inconsistency between these and content policies; (3) their voice was scarcely involved before or after algorithmic visibility decision-making processes.

Building upon these findings, we will discuss how a multi-dimensional notion of moderation fairness generates from our participants' fairness perceptions of moderation processes and outcomes. We will further discuss how YouTube moderation demonstrates an algorithmic assemblage: not a single algorithm but different classes of algorithms moderate content on YouTube. Ultimately, we put forward design considerations for the moderation systems on a platform like YouTube that affords monetization for content creators.

6.1 Moderation Fairness: A Multi-Dimensional Notion

Many prior studies have explored the result of perceived fairness in moderation with a binary question, whether users felt moderation as fair or unfair upon receiving moderation explanations [43,92] or upon having their participation in the adjudication of moderation

cases [29]. Probing deeper into YouTuber's moderation experience, we unpacked and elaborated on the processes of how YouTubers developed such fairness perceptions from the consistency, equality of, and their voice in moderation procedures and outcomes.

Fairness perception is not an isolated experience but involves comparison. We discovered that our participants compared their moderation outcomes and the associated time of happening, severity, resource allocations for repairing content, and subject characteristics (e.g., fanbase) with those of other YouTubers. In the view of procedural justice, people perceive fairness from consistency based on comparisons across time, people, or both [91]. Our participants individually and collectively compared moderation actions, outcomes, and impacts with the claims of content policies. This thus extended a binary question of whether users feel moderation as fair or unfair to how their perceived (un)fairness was generated from moderation experiences. In other words, fairness perception concerned not only a conclusion about whether the system was fair or not but also the cognitive activities such as comparison and evaluation through which they reached the conclusion.

Voice inclusion, both prior to and after authorities make decisions, matters to fairness perception [60,91]. Our empirical findings pointed to participants' representation, or lack thereof, in moderation decision-making processes. They could observe how various moderation-related algorithms made decisions and affected their metrics beyond the initial scope. However, they were unable to have any voice in decision-making processes to obtain a sense of control over moderation decisions. So, their perceived unfairness arose. This lack of understanding, communication, and control for moderation outcomes indicated the importance of recognizing YouTubers' voice in moderation decision-making processes.

Moderation fairness on YouTube thus presents as a multi-dimensional notion as summarized in Figure 5:

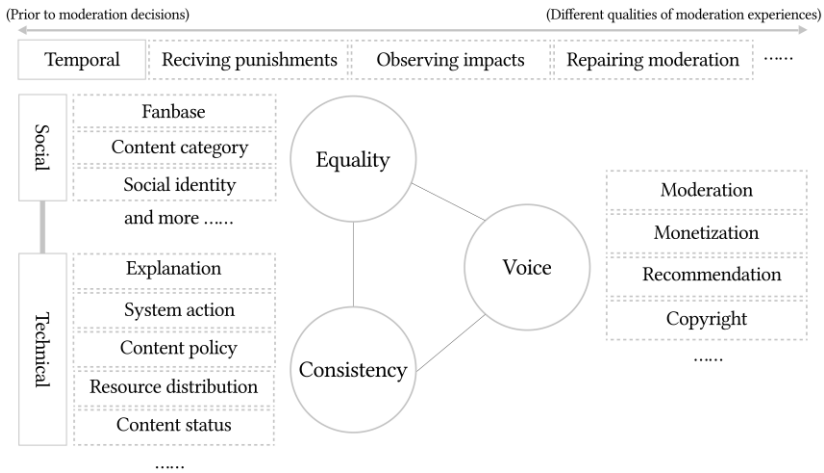

Fig. 5: Multi-dimensional notion of fairness in YouTube moderation.

- First, it has a temporal dimension in terms of at what stage YouTubers invoked the fairness of moderation. Participants could invoke the concept of fairness when they just received moderation decisions or penalties, when they sought resources to repair them,

when they tracked their channels' later performance, or when they observed how future new videos with all normal statuses underperformed in metrics.

- Second, it concerns the social context where YouTubers situate their fairness perception. A social context here denotes a group of people who shares similar traits and thus allow cross-comparisons. Our participants shared several distinct social contexts, including groups of participants who they were familiar with, who created content in the same category (e.g., fitness, education), and a network of YouTubers with various fanbases who experienced punishments. Different social contexts allowed them to assess the fairness of interactions they had with the moderation system on YouTube.
- Third, a technical dimension of moderation fairness surfaced. Our participants experienced (1) inconsistency either in moderation explanations, decisions, system-actions, or those inconsistent with content policies, (2) unequal moderation treatments with others because they attributed inconsistent technical actions to their subjective characteristics, and (3) unbalanced support from different algorithms for unmoderated videos and moderated ones as well as between themselves and others.
- Collectively, when experiencing moderation processes over time, our participants had different qualities of moderation experiences. They developed various fairness perceptions because they experienced different impacts of moderation and made sense of it differently; such cognitive activities from different participants showed social, temporal, and technical dimensions of moderation.

YouTubers' moderation experiences contour the multi-dimensions of moderation fairness, complicating the understanding of moderation. Prior studies found that YouTubers felt it unfair that they experienced disproportionate demonetization penalties [21,62] or recommendation rates compared with large YouTubers [100]. Beyond YouTubers' comparisons on outcomes or decisions they encounter, this study further painted a more comprehensive picture of how their fairness perceptions were generated from technical, social, and economic dimensions of moderation. When examining the consistency of moderation, prior research has discussed inconsistent moderation decisions from human moderators [11,86], and journalists have uncovered the unfair moderation on users' visibility [5]. Beyond inconsistent moderation decisions, we further highlighted how moderation fairness on YouTube demonstrated temporal dimensions of moderation, arousing our participants' perceived unfairness. Such perceptions were generated when participants had different qualities of moderation experiences in processes (e.g., observing ripple impacts, handling moderation), as shown in Figure 5.

For future work in HCI and CSCW, this multi-dimensional moderation fairness indicates the importance of studying moderation experiences: beyond understanding moderation outcomes, researchers should not ignore how end-users have different qualities of experiences in social, temporal, and technical dimensions of content moderation. For users who experienced moderation outcomes (e.g., content removal, demonetization), as our participants' moderation experiences shown, their fairness perceptions originated not only from the punishments/decisions but also from measuring the equality, consistency of, and their representation in moderation processes.

6.2 Content Moderation as an Algorithmic Assemblage

Commonly discussed moderation algorithms are those that directly impact content, such as removing a video that contains hateful speech. However, moderators might use a bricolage of tools to do the moderation work in online communities like subreddits [26]. Meanwhile, platform users could also experience multiple moderation mechanisms at work. Our findings showed that what moderates our participants on YouTube is not just a sole moderation algorithm. For instance, when the ad or monetization algorithms issued a 'limited ads' (i.e., demonetization) decision, the video would remain intact but would no longer be featured in search or recommendation. The visibility algorithm could further reduce the visibility of a whole channel for the YouTuber in question. On YouTube, it is thus an algorithmic assemblage composed of different classes of algorithms that work together to moderate content creation. When YouTubers interact with the algorithmic assemblage, they essentially interact with the organization that implements it [8,74,75]. So, when interacting with the algorithmic assemblage of YouTube moderation, YouTubers participants generated their fairness perceptions in negotiations for content creation favors from YouTube. That was because they cared about the consequences of not involving their voice in moderation decision-making, which harm their content visibility and monetization.

Prior moderation studies have explored modes and algorithms built for the purpose of moderation [34,36]. Moderation algorithms issue punishments that could be directly felt, such as content removal or account termination. Visibility algorithms do not issue direct punishments. Instead, they govern YouTubers through multi-layer visibility limits ranging from content-hiding for certain groups of viewers (e.g., 'restricted mode') to public searchability ban to overall recommendation rate deduction. When these outcomes appeared inconsistently, our participants realized that their voice was barely included in the visibility algorithm's decision-making processes. YouTubers cannot observe the impacts on their performance of channels and videos until their reduced visibility has taken effect for a while.

Amateur or small YouTubers were found to perceive recommendation algorithms as unfair [100], and large YouTubers who obtained more knowledge had greater visibility than YouTubers who did not [17,18]. Viewing from the angle of content moderation, our study distinctively showed how one or more moderation decisions intertwined with algorithmic visibility. That is, underneath moderation decisions, our participants experienced unfair visibility deduction at transverse (i.e., videos, channels) and longitudinal levels, both and respectively. Meanwhile, human moderators and moderation algorithms released inconsistent moderation decisions, adding instability to their future content visibility and the labor of solving moderation issues.

Users wished to obtain control strategies to make their presence more visible by algorithms (e.g., Twitter [20]). However, growing work has pointed out that users disproportionately lack control of the harm caused by visibility algorithms (e.g., filtering, searching, ranking) [9,28,46], which might occasionally hamper marginalized people's recovery online [30]. Similarly, as our participants' experiences showed, observing and analyzing algorithmic visibility outcomes did not indicate that YouTubers have their voice heard by YouTube's human moderators or moderation algorithms either before or after receiving moderation decisions. Valuing individual voice in content moderation procedures, at the same time, has been a rising consideration for researchers, especially in designing contestability for moderation with various values such as fairness and transparency [93]. To design such contestability on

YouTube, moderated YouTubers should be considered as relevant stakeholders in moderation decision-making. That is because YouTubers joining the YouTube Partner program essentially have a contract relationship with YouTube. In this contact relationship, their activities of creating and monetizing video content become YouTube's commodity to earn ad income from advertisers. So, valuing YouTubers' voice could manifests the value of interactional justice [50], where the social media company, Google/YouTube, values their digital or creative labor, YouTubers [69].

A lack of voice in algorithmic decision-making might further deepen the existing inequality between users. Prior studies or news reports have initially discovered how YouTubers perceived demonetization penalties as unfairly imposed on them [3,21,62]. We, through this study, moved a step forward to show YouTubers' lack of their voice in multiple algorithmic decisions, especially those moderation penalties they discovered on their own. Various visibility deductions and inconsistent moderation decisions further constituted an unequal ad income distribution.

6.3 Design Considerations

While YouTube is an authority to implement platform governance practices, it should embrace diverse voices in decision-making processes to ensure procedural justice. YouTubers' content creation and livelihoods rely on the platform; at the same time, their value and voice for the platform should be fairly recognized by the decision-maker, YouTube (e.g., the moderation system). Otherwise, they might develop different perceptions of fairness from their moderation experience, as our participants did.

Offering moderation explanations could sometimes relieve users' perceived unfairness [43], but it is conditioned by the platform's will to first disclose whether and how moderation happens. Prior studies have uncovered that letting users know the benevolent intentions behind algorithmic decision-making could increase their perceived fairness [58]. However, we discovered that certain YouTube's moderation practices were presented inapparent where YouTubers can only know they experienced such moderation through their own tests. As P21 and other minority people experienced [81], the silent moderation happened on videos with all normal statuses, which further showed inconsistency with content policies. To resolve perceived unfairness, we thus suggest that YouTube should disclose whether YouTuber's specific videos are invisible under 'restricted mode,' which ought to be listed in the Studio dashboard. Also, if invisible, YouTube should clarify the reasons and how to repair such moderation issues in steps instead of publicly claiming "will not respond" to YouTubers even though they appeal [109]. This suggestion resonated with many researchers' call on making moderation more transparent [45,48].

For specific functionality designs, end-users' voice and input should be valued in an algorithmic assemblage of moderation decision-making processes. For example, as shown in our findings, YouTube created a 'self-certification' (i.e., ad suitability) function in 2019. It aimed to improve the accuracy of ML algorithms by acquiring YouTubers' input to issue fairer ad-suitability decisions in line with advertiser-friendly content guidelines [115]. However, our findings showed that YouTube moderation was inconsistent in issuing and enforcing ad-suitability decisions. Also, YouTubers further felt unfair about inconsistent decisions such as 'made for kids' and being invisible under 'restricted mode.' Thus, we suggest self-certification

functionality should further consider more classes of user input for moderation decision-making processes because YouTube moderation cannot be simply referred to as 'limited ads' moderation. Hence, asking YouTubers whether and why their video is made for kids before listing it, instead of choosing options in audience settings, can give YouTubers chances to predict and acquire a sense of control over how acceptable their video is. This change could inform platform developers or engineers about the way how users want their voice involved in moderation decision-making. That is, resonating with the research trend of embedding procedural justice in moderation [29,92], to allow moderated users' inputs to impact moderation decision-making (e.g., having their inputs as training data for moderation algorithms before decisions are really made).

Also, this study provides transferable design considerations for content moderation on other platforms. Many platforms nowadays afford monetization for content creators to benefit from creating content on platforms such as Facebook, Twitter, TikTok, Instagram, and YouTube. Based on our findings, we argued that creators are entitled to know whenever moderation decisions affect their visibility and monetization because moderation is conducted through algorithmic assemblages. For example, YouTube educates YouTubers on how to improve their visibility through recommendation algorithms, giving them knowledge and control of delivering content to audiences [87]. However, YouTube did not teach them how to repair moderation issues brought by visibility algorithms. When moderation decisions were made, we found that both visibility of the moderated video and future new videos with all normal statuses became less favored by recommendation algorithms, which presented a conflict with content policies. So, YouTubers might feel it unfair that they could not have their voice in visibility algorithmic decisions.

These findings resonated with "shadowban," the phenomenon that researchers have largely discussed. It was reported that minority people (e.g., woman, sexual minority)'s visibility was largely decreased by moderation algorithms on Reddit [99] and Instagram [10]. However, our findings (e.g., Section 5.3) have been extending this line of work regarding what impacts of visibility decrease (or shadowban) meant to YouTubers. That is, like users on Instagram or Reddit, YouTubers might lose chances of free expression or self-disclosure. But moreover, they needed to undergo potential loss in monetization due to visibility decrease. So, not only moderation algorithms were found to be an assemblage structure but also creators' negative experiences where the impacts generated from this structure were also correlated.

Thus, we suggest specific procedures that platforms affording monetization for creators could apply to disclose how moderation decisions would affect creators. First, using YouTube as an example, when a YouTuber receives a 'limited ads' decision, YouTube should inform them of how such a decision affects their Cost per 1,000 impressions (CPM) [105] instead of only claiming a statement that most advertisers would not place ads. Second, YouTube should let creators know whether such a decision affects their visibility (e.g., reach to audiences, audience engagement). Last, YouTube should inform creators of how this impact will be calculated to the monetization mechanism, especially for the YouTubers who have revered false-positive moderation decisions. These strategies could be transferred to other platforms affording monetization for content creators.

7 LIMITATION AND FUTURE WORK

The 21 YouTube partners we interviewed might not represent the experience of all YouTubers who perform either profitable or amateur video creation. Algorithmic moderation on YouTube is a complex process involving a massive-scale user base, so our findings cannot represent the moderation operations or fairness perceptions of all. Especially those YouTubers or a team working for one YouTube channel who have a much larger fanbase (e.g., more than 1 or 10 million subscribers) might have different experiences and perceptions of YouTube moderation. However, with the nature of interview studies for qualitative research, we do not intend to produce generalizability but yield a novel and insightful understanding of how YouTubers perceive the fairness of YouTube moderation. So, future studies could investigate how larger or amateur creators (i.e., ones who cannot monetize) experience moderation from severer punishments such as account suspension or content removal.

Although our participants are from relatively different backgrounds (e.g., countries, ages groups), they did not express much about how their cultural background might affect their moderation experiences. They tended to describe personal interactions with YouTube's moderation systems. This was because, we assumed, the technologies on YouTube are relatively universal. While the cultural difference in interactions with moderation was not our focus in this study, we did recognize this could be a future study.

Also, in our study, we found several YouTubers experienced moderation majorly due to copyright infringements. Copyright cases involve copyright laws regarding 'fair use' and one more stakeholder, copyright owners, in the moderation procedures. The owners can take part or all YouTubers' ad income earned from a video that violates copyright [36]. Thus, it would be more complex than other types of moderation. Future research could explore the relationship between YouTuber's monetization and YouTube copyright moderation involving the platform, YouTubers, and copyright owners.

8 CONCLUSION

Research in HCI and CSCW has started to investigate users' content moderation experiences. Beyond focusing on specific moderation decisions (e.g., content removal), we argue that we should not ignore how users experience moderation processes and how they generate fairness perceptions from different dimensions of moderation, as the multi-dimensions of moderation fairness showed on YouTube. This study with 21 YouTubers uncovered that participants' perceived fairness arose when they measured equality, consistency, and their representation in moderation decision-making processes and outcomes. Their experiences show that on YouTube, not a sole class of moderation algorithm but an algorithmic assemblage moderates YouTubers' profitable content creation, implementing platform governance. To design a fairer moderation system, users' voice needs to be involved in moderation decision-making process and allowed to impact this process. That is because, though moderation systems are relatively universal, users' moderation experiences could be different due to their temporal, social, or technical contexts. We thus call for action: future research should not ignore whether and how moderated users have different qualities of experiences in moderation processes.

ACKNOWLEDGMENTS

We thank the associate chairs and anonymous reviewers for their constructive feedback and suggestions. We thank Xinning Gui for her feedback to the iterations of this paper. We also appreciate 21 YouTubers' participation and support for this study. Lastly, this work is partially supported by the NSF, under grant no. 2006854.

REFERENCES

[1] Salem Hamed Abdurrahim, Salina Abdul Samad, and Aqilah Baseri Huddin. 2018. Review on the effects of age, gender, and race demographics on automatic face recognition. Visual Computer 34, 1617–1630. DOI:https://doi.org/10.1007/s00371-017-1428-z

[2] Julia Alexander. 2018. What is YouTube demonetization? An ongoing, comprehensive history. Polygon. Retrieved from https://www.polygon.com/2018/5/10/17268102/youtube-demonetization-pewdiepie-logan-paul-casey-neistat-philip-defranco

[3] 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

[4] Julia Alexander. 2019. YouTube is disabling comments on almost all videos featuring children. The Verge. Retrieved from https://www.theverge.com/2019/2/28/18244954/youtube-comments-minor-children-exploitation-monetization-creators

[5] 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

[6] Julia Alexander. 2020. YouTube is demonetizing videos about coronavirus, and creators are mad. Retrieved from https://www.theverge.com/2020/3/4/21164553/youtube-coronavirus-demonetization-sensitive-subjects-advertising-guidelines-revenue

[7] 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

[8] Mike Ananny. 2016. Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness. Sci. Technol. Hum. Values 41, 1 (September 2016), 93–117. DOI:https://doi.org/10.1177/0162243915606523

[9] 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

[10] 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

[11] Anna Veronica Banchik. 2020. Disappearing acts: Content moderation and emergent practices to preserve at-risk human rights–related content. New Media Soc. (March 2020), 146144482091272. DOI:https://doi.org/10.1177/1461444820912724

[12] Alex Barker and Hannah Murphy. 2020. YouTube reverts to human moderators in fight against misinformation. Financial Times. Retrieved August 4, 2021 from https://www.ft.com/content/e54737c5-8488-4e66-b087-d1ad426ac9fa

[13] Karissa Bell. 2021. How the pandemic supercharged the creator economy in 2021. Engadget. Retrieved from https://www.engadget.com/how-the-pandemic-supercharged-the-creator-economy-153050958.html

[14] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in Criminal Justice Risk Assessments: The State of the Art. Sociol. Methods Res. 50, 1 (2018), 3–44. DOI:https://doi.org/10.1177/0049124118782533

[15] Niels Van Berkel, Jorge Goncalves, and Daniel Russo. 2021. Efect of information presentation on fairness perceptions of machine learning predictors. CHI Conf. Hum. Factors Comput. Syst. Proc. (CHI 2021) (May 2021). DOI:https://doi.org/10.1145/3411764.3445365

[16] Robert J. Bies and Debra L. Shapiro. 2017. Voice and Justification: Their Influence on Procedural Fairness Judgments. Acad. Manag. J. 31, 3 (November 2017), 676–685. DOI:https://doi.org/10.5465/256465

[17] Sophie Bishop. 2020. Algorithmic Experts: Selling Algorithmic Lore on YouTube. Soc. Media + Soc. 6, 1 (2020), 205630511989732. DOI:https://doi.org/10.1177/2056305119897323

[18] Sophie Bishop. 2021. Influencer Management Tools: Algorithmic Cultures, Brand Safety, and Bias. Soc. Media +

Soc. 7, 1 (March 2021). DOI:https://doi.org/10.1177/20563051211003066

[19] Amy Bruckman, Pavel Curtis, Cliff Figallo, and Brenda Laurel. 1994. Approaches to managing deviant behavior in virtual communities. Association for Computing Machinery, New York, New York, USA. DOI:https://doi.org/10.1145/259963.260231

[20] Jenna Burrell, Zoe Kahn, Anne Jonas, and Daniel Griffin. 2019. When Users Control the Algorithms: Values Expressed in Practices on Twitter. Proc. ACM Human-Computer Interact. 3, CSCW (November 2019), 1–20. DOI:https://doi.org/10.1145/3359240

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

[22] Ashley Carman. 2021. Facebook shorted video creators thousands of dollars in ad revenue. The Verge. Retrieved from https://www.theverge.com/2021/3/31/22358723/facebook-creators-video-revenue-estimate-tool-pages

[23] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Big Data 5, 2 (June 2017), 153–163. DOI:https://doi.org/10.1089/big.2016.0047

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

[25] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated Experiments on Ad Privacy Settings A Tale of Opacity, Choice, and Discrimination. In Proceedings on Privacy Enhancing Technologies, 92–112. Retrieved from http://www.google.com/settings/ads

[26] Bryan Dosono and Bryan Semaan. 2019. Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on reddit. CHI Conf. Hum. Factors Comput. Syst. Proc. (CHI 2019) (May 2019), 1–13. DOI:https://doi.org/10.1145/3290605.3300372

[27] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In ITCS 2012 - Innovations in Theoretical Computer Science Conference, ACM Press, New York, New York, USA, 214–226. DOI:https://doi.org/10.1145/2090236.2090255

[28] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User attitudes towards algorithmic opacity and transparency in online reviewing platforms. In CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), Association for Computing Machinery, New York, NY, USA, 1–14. DOI:https://doi.org/10.1145/3290605.3300724

[29] Jenny Fan and Amy X. Zhang. 2020. Digital Juries: A Civics-Oriented Approach to Platform Governance. In CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2020), Association for Computing Machinery, New York, NY, USA, 1–14. DOI:https://doi.org/10.1145/3313831.3376293

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

[31] Andreas Follesdal. 2015. John rawls' theory of justice as fairness. In Philosophy of Justice. Springer Netherlands, 311–328. DOI:https://doi.org/10.1007/978-94-017-9175-5_18

[32] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. 2017. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the United States. Proc. Natl. Acad. Sci. U. S. A. 114, 50 (December 2017), 13108–13113. DOI:https://doi.org/10.1073/pnas.1700035114

[33] 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

[34] 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

[35] Google. YouTube Community Guidelines enforcement – Google Transparency Report. Retrieved from https://transparencyreport.google.com/youtube-policy/removals?hl=en&total_removed_videos=period:Y2019Q1;exclude_automated:all&lu=total_removed_videos

[36] 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

[37] Jerald Greenberg and Robert Folger. 1983. Procedural Justice, Participation, and the Fair Process Effect in

Groups and Organizations. Basic Gr. Process. (1983), 235–256. DOI:https://doi.org/10.1007/978-1-4612-5578-9_10

[38] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. Web Conf. 2018 - Proc. World Wide Web Conf. WWW 2018 (April 2018), 903–912. DOI:https://doi.org/10.1145/3178876.3186138

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

[40] 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 Human-Computer Interact. 5, CSCW2 (October 2021). DOI:https://doi.org/10.1145/3479610

[41] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miroslav Dudík, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need? In CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), Association for Computing Machinery, New York, NY, USA, 1–16. DOI:https://doi.org/10.1145/3290605.3300830

[42] Daniel James. 2020. What Are YouTube Browse Features? Tubefluence. Retrieved from https://tubefluence.com/what-are-youtube-browse-features/

[43] 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 Human-Computer Interact. 3, CSCW (November 2019), 1–33. DOI:https://doi.org/10.1145/3359294

[44] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. ACM Trans. Comput. Interact. 26, 5 (July 2019), 1–35. DOI:https://doi.org/10.1145/3338243

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

[46] Shagun Jhaver, Yoni Karpfen, and Judd Antin. 2018. Algorithmic Anxiety and Coping Strategies of Airbnb Hosts. Proc. 2018 CHI Conf. Hum. Factors Comput. Syst. (2018). DOI:https://doi.org/10.1145/3173574

[47] 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

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

[49] Maria Kasinidou, Styliani Kleanthous, Plnar Barlas, and Jahna Otterbacher. 2021. "I agree with the decision, but they didn't deserve this": Future Developers' Perception of Fairness in Algorithmic Decisions. FAccT 2021 - Proc. 2021 ACM Conf. Fairness, Accountability, Transpar. (March 2021), 690–700. DOI:https://doi.org/10.1145/3442188.3445931

[50] Tae Yeol Kim and Kwok Leung. 2007. Forming and reacting to overall fairness: A cross-cultural comparison. Organ. Behav. Hum. Decis. Process. 104, 1 (September 2007), 83–95. DOI:https://doi.org/10.1016/J.OBHDP.2007.01.004

[51] Keith Kirkpatrick. 2016. Battling algorithmic bias. Communications of the ACM 59, 16–17. DOI:https://doi.org/10.1145/2983270

[52] Cliff Lampe and Erik Johnston. 2005. Follow the (Slash) dot: Effects of Feedback on New Members in an Online Community. Proc. 2005 Int. ACM Siggr. Conf. Support. Gr. Work - Gr. '05 (2005). DOI:https://doi.org/10.1145/1099203

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

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

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

[56] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management: https://doi.org/10.1177/2053951718756684 5, 1 (March 2018). DOI:https://doi.org/10.1177/2053951718756684

[57] Min Kyung Lee and Su Baykal. 2017. Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, ACM, New York, NY, USA. Retrieved from http://dx.doi.org/10.1145/2998181.2998230

[58] 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 Human-Computer Interact. 3, CSCW (November 2019), 26. DOI:https://doi.org/10.1145/3359284

[59] Min Kyung Lee, Ji Tae Kim, and Leah Lizarondo. 2017. A Human-Centered Approach to Algorithmic Services: Considerations for Fair and Motivating Smart Community Service Management that Allocates Donations to Non-Profit Organizations. Proc. 2017 CHI Conf. Hum. Factors Comput. Syst. (2017). DOI:https://doi.org/10.1145/3025453

[60] Gerald S. Leventhal. 1980. What Should Be Done with Equity Theory? Soc. Exch. (1980), 27–55. DOI:https://doi.org/10.1007/978-1-4613-3087-5_2

[61] E. Allan Lind, Ruth Kanfer, and P. Christopher Earley. 1990. Voice, Control, and Procedural Justice: Instrumental and Noninstrumental Concerns in Fairness Judgments. J. Pers. Soc. Psychol. 59, 5 (1990), 952–959. DOI:https://doi.org/10.1037/0022-3514.59.5.952

[62] Renkai Ma and Yubo Kou. 2021. "How advertiser-friendly is my video?": YouTuber's Socioeconomic Interactions with Algorithmic Content Moderation. PACM Hum. Comput. Interact. 5, CSCW2 (2021), 1–26. DOI:https://doi.org/https://doi.org/10.1145/3479573

[63] Claire Cain Miller. 2015. The Upshot: Can an Algorithm Hire Better Than a Human? The New York Times. Retrieved from https://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html

[64] Viginia Murphy-Berman, John J. Berman, Purnima Singh, Anju Pachauri, and Pramod Kumar. 1984. Factors affecting allocation to needy and meritorious recipients: A cross-cultural comparison. J. Pers. Soc. Psychol. 46, 6 (June 1984), 1267–1272. DOI:https://doi.org/10.1037/0022-3514.46.6.1267

[65] Arvind Narayanan. 2018. 21 fairness definition and their politics. ACM FAT* 2018 tutorial. Retrieved from https://shubhamjain0594.github.io/post/tlds-arvind-fairness-definitions/

[66] Casey Newton. 2019. The secret lives of Facebook moderators in America. The Verge. Retrieved from https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

[67] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science (80-.). 366, 6464 (October 2019), 447–453. DOI:https://doi.org/10.1126/science.aax2342

[68] Ocelot AI. 2019. Demonetization report. Retrieved from https://docs.google.com/document/d/18B-X77K72PUCNIV3tGonzeNKNkegFLWuLxQ_evhF3AY/edit

[69] 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

[70] Molly Priddy. 2017. Why Is YouTube Demonetizing LGBTQ Videos? Autostraddle. Retrieved from https://www.autostraddle.com/why-is-youtube-demonetizing-lgbtqia-videos-395058/

[71] John Rawls. 1971. A Theory of Justice.

[72] Sarah Roberts. 2016. Commercial Content Moderation: Digital Laborers' Dirty Work. Media Stud. Publ. (January 2016). Retrieved from https://ir.lib.uwo.ca/commpub/12

[73] 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

[74] Howard Rosenbaum. 2020. Algorithmic neutrality, algorithmic assemblages, and the lifeworld. AMCIS 2020 Proc. (August 2020). Retrieved from https://aisel.aisnet.org/amcis2020/philosophical_is/philosophical_is/6

[75] Howard Rosenbaum and Pnina Fichman. 2019. Algorithmic accountability and digital justice: A critical assessment of technical and sociotechnical approaches. Proc. Assoc. Inf. Sci. Technol. 56, 1 (January 2019), 237–244. DOI:https://doi.org/10.1002/PRA2.19

[76] Debjani Saha, Candice Schumann, Duncan Mcelfresh, John Dickerson, Michelle Mazurek, and Michael Tschantz. 2020. Measuring Non-Expert Comprehension of Machine Learning Fairness Metrics. In Proceedings of the 37th International Conference on Machine Learning, PMLR, 8377–8387. Retrieved from http://proceedings.mlr.press/v119/saha20c.html

[77] Mia Sato. 2021. YouTube reveals millions of incorrect copyright claims in six months. The Verge. Retrieved from

https://www.theverge.com/2021/12/6/22820318/youtube-copyright-claims-transparency-report

[78] Nripsuta Ani Saxena, Goran Radanovic, Karen Huang, David C. Parkes, Evan DeFilippis, and Yang Liu. 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. Association for Computing Machinery, Inc, New York, NY, USA. DOI:https://doi.org/10.1145/3306618.3314248

[79] 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

[80] Lucas Shaw. 2021. The Pandemic Has Been Very, Very Good for the Creator Economy. Bloomberg. Retrieved from https://www.bloomberg.com/news/newsletters/2021-08-29/the-pandemic-has-been-very-very-good-for-the-creator-economy

[81] Catherine Shu. 2017. YouTube responds to complaints that its Restricted Mode censors LGBT videos. TechCrunch. Retrieved from https://techcrunch.com/2017/03/19/youtube-lgbt-restricted-mode/

[82] 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/

[83] Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. 2018. Report of the Special Rapporteur to the General Assembly on AI and its impact on freedom of opinion and expression. United Nations Human Rights Office of The High Commissioner. Retrieved August 4, 2021 from https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ReportGA73.aspx

[84] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical notions vs. Human perception of fairness: A descriptive approach to fairness for machine learning. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, 2459–2468. DOI:https://doi.org/10.1145/3292500.3330664

[85] Miriah Steiger, Timir J. Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators the emotional labor of commercial moderation and avenues for improving support. CHI Conf. Hum. Factors Comput. Syst. Proc. (CHI 2021) (May 2021). DOI:https://doi.org/10.1145/3411764.3445092

[86] 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

[87] Lydia Sweatt. 2021. YouTube Algorithm Guide: How Your Videos Are Recommended to Viewers. vidIQ. Retrieved from https://vidiq.com/blog/post/how-youtube-algorithm-recommends-videos/

[88] Jeanna Sybert. 2021. The demise of #NSFW: Contested platform governance and Tumblr's 2018 adult content ban: New Media Soc. (February 2021). DOI:https://doi.org/10.1177/1461444821996715

[89] John Thibaut and Laurens Walker. 1978. A Theory of Procedure. Calif. Law Rev. 66, (1978). Retrieved from https://heinonline.org/HOL/Page?handle=hein.journals/calr66&id=555&div=36&collection=journals

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

[91] Tom R. Tyler. 1988. What Is Procedural Justice?: Criteria Used by Citizens to Assess the Fairness of Legal Procedures. Law Soc. Rev. (1988).

[92] 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

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

[94] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. Proc. 2018 CHI Conf. Hum. Factors Comput. Syst. (2018). DOI:https://doi.org/10.1145/3173574

[95] Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. IEEE/ACM Int. Work. Softw. Fairness 18, (2018). DOI:https://doi.org/10.1145/3194770.3194776

[96] Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. CHI Conf. Hum. Factors Comput. Syst. Proc. (CHI 2020) (2020). DOI:https://doi.org/10.1145/3313831

[97] Donghee Yvette Wohn. 2019. Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), Association for Computing Machinery, New York, New York, USA, 1–13. DOI:https://doi.org/10.1145/3290605.3300390

[98] Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeff Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. Proc. 2018 CHI Conf. Hum. Factors Comput. Syst. (2018). DOI:https://doi.org/10.1145/3173574

[99] Lucas Wright. 2022. Automated Platform Governance Through Visibility and Scale: On the Transformational Power of AutoModerator. Soc. Media + Soc. 8, 1 (February 2022). DOI:https://doi.org/10.1177/20563051221077020

[100] Eva Yiwei Wu, Emily Pedersen, and Niloufar Salehi. 2019. Agent, gatekeeper, drug dealer: How content creators craft algorithmic personas. Proc. ACM Human-Computer Interact. 3, CSCW (2019), 1–27. DOI:https://doi.org/10.1145/3359321

[101] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In 26th International World Wide Web Conference, WWW 2017, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1171–1180. DOI:https://doi.org/10.1145/3038912.3052660

[102] Conrad Ziller. 2017. Equal Treatment Regulations and Ethnic Minority Social Trust. Eur. Sociol. Rev. 33, 4 (August 2017), 563–575. DOI:https://doi.org/10.1093/ESR/JCX059

[103] Procedural Justice. Yale Law School. Retrieved from https://law.yale.edu/justice-collaboratory/procedural-justice

[104] YouTube analytics basics. YouTube Help. Retrieved from https://support.google.com/youtube/answer/9002587?hl=en

[105] Understand ad revenue analytics. YouTube Help. Retrieved from https://support.google.com/youtube/answer/9314357?hl=en

[106] How engagement metrics are counted. YouTube Help. Retrieved from https://support.google.com/youtube/answer/2991785?hl=en

[107] How to Develop a YouTube Monetization Strategy for Your Channel. Retrieved from https://www.tastyedits.com/youtube-monetization-strategy/

[108] "Limited or no ads" explained. YouTube Help. Retrieved from https://support.google.com/youtube/answer/9269824?hl=en

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

[110] 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

[111] YouTube Self-Certification overview. YouTube Help. Retrieved from https://support.google.com/youtube/answer/7687980?hl=en

[112] YouTube Community Guidelines & Policies. How YouTube Works. Retrieved from https://www.youtube.com/howyoutubeworks/policies/community-guidelines/

[113] YouTube Partner Manager overview. YouTube Help. Retrieved from https://support.google.com/youtube/answer/6361049?hl=en

[114] Watching "made for kids" content. YouTube Help. Retrieved from https://support.google.com/youtube/answer/9632097?hl=en

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

[116] Upcoming and recent ad guideline updates. YouTube Help. Retrieved from https://support.google.com/youtube/answer/9725604?hl=en#February2021

[117] Restrictions on live streaming. YouTube Help. Retrieved from https://support.google.com/youtube/answer/2853834?hl=en

[118] Copyright strike basics. YouTube Help. Retrieved from https://support.google.com/youtube/answer/2814000

[119] 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

[120] Discovery and performance FAQs. YouTube Help. Retrieved from

https://support.google.com/youtube/answer/141805?hl=en

A INTERVIEW PROTOCOL

Interview Protocol	
Guiding Research Question: How do YouTubers' fairness perceptions generate from moderation experiences?	
1. Warm-up questions	How old are you? What gender do you identify with? What ethnicity do you identify with? Which country do you locate in? How many years do you consistently create videos on YouTube? What content category/community of your channel is? Do you consider yourself as a part-time or full-time YouTuber, on your time spent on creating videos?
2. Investigating moderation experience on YouTube	What moderation or punishment did you experience? Can you explain how did that happen? What explanations did YouTube provide for it? How do you think of these explanations? Did this moderation affect your video's performance? If so, What metrics are affected? And how do you know that? How did this affect your channel's metrics? Have you noticed any impression rate difference after experiencing it? If so, could you elaborate the situation? How did this affect your community who create the same content with you? How do you handle moderation punishment? How effective is your coping behavior for it?
3. Understanding perceived fairness on YouTube moderation	Do you feel if it's fair for the moderation you experienced, and why? If exist, how do you feel about the impacts that your channel experience? Do you feel if it's fair for the procedure of fixing the moderation, and why? Do you feel if it's fair for the appeal process, and why? How do you consider YouTube Team's role in this process, and why? What did you learn from the moderation?

Received: January 2022, Revised: April 2022, Accepted: May 2022.