Contestability For Content Moderation

KRISTEN VACCARO, University of California San Diego, USA
ZIANG XIAO, University of Illinois Urbana-Champaign, USA
KEVIN HAMILTON, University of Illinois Urbana-Champaign, USA
KARRIE KARAHALIOS, University of Illinois Urbana-Champaign, USA

Content moderation systems for social media have had numerous issues of bias, in terms of race, gender, and ability among many others. One proposal for addressing such issues in automated decision making is by designing for contestability, whereby users can shape and influence how decisions are made. In this study, we conduct a series of participatory design workshops with participants from communities that have experienced problems with social media content moderation in the past. Together with participants, we explore the idea of designing for contestability in content moderation and find that users' designs suggest three fruitful, practical avenues: adding representation, improving communication, and designing with compassion. We conclude with design recommendations drawn from participants' proposals, and reflect on the challenges that remain.

CCS Concepts: • **Human-centered computing** → *Human computer interaction (HCI)*; Social media.

Additional Key Words and Phrases: algorithmic experience; contestability; content moderation; participatory design

ACM Reference Format:
Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability For Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 318 (October 2021), 28 pages. https://doi.org/10.1145/3476059

1 INTRODUCTION

When Lance Brown Eyes' Facebook account was suspended, he had a good idea of what had gone wrong [90]. At the time, Facebook had a "real name" policy that required that users to use their "*real name as it would be listed on your credit card, driver's license or student ID*" [75]. Some users had issues with this policy in theory (arguing that they should be allowed to use pseudonyms, stage names, and so on), but there were even larger problems with the implementation. The algorithmic systems that identified accounts for violating the policy systematically suspended Native Americans accounts, for users with last names like Lone Hill or Brown Eyes [56]. Many users appealed individual decisions, but the automated systems weren't changed. Lance himself issued a call to action in an interview with the Washington Post: "*They let me change my name back, but what about you and all the others they discriminated against? Our people need to know they can fight back. The more of us stand up, they will change*" [90].

The community was eventually able to persuade Facebook to change the decision making process, but did so by turning to the press. But recently, an alternative mechanism — *contestability* — has been proposed to allow users to shape the decision making from *within* the system. Contestability

Authors' addresses: Kristen Vaccaro, kv@ucsd.edu, University of California San Diego, USA; Ziang Xiao, zxiao5@illinois.edu, University of Illinois Urbana-Champaign, USA; Kevin Hamilton, kham@illinois.edu, University of Illinois Urbana-Champaign, USA; Karrie Karahalios, kkarahal@illinois.edu, University of Illinois Urbana-Champaign, USA.

2573-0142/2021/10-ART318
https://doi.org/10.1145/3476059

has been defined as a set of "*mechanisms for users to understand, construct, shape and challenge model predictions*" [65]. Importantly, it has been defined as being "in band" for the system; unlike appeals which may be asynchronous, pursued through outside channels, or otherwise externalized, contestability is built into the system to support iteration on the decision making process.

Providing this is important in principle because it supports the co-construction of the decision making process. It is important in practice because these kinds of content moderation systems have been plagued by failures: in addition to the "whitewashing" experienced by Native Americans, social media platforms have been found to have other racial biases [5, 111], to censor LGBTQ+ users [20, 74] and Aboriginal women [2], and to restrict the exposure of users with disabilities [100]. In addition to limiting social connection and self-expression, these effects can also have significant harms on users' professional lives, as artists, for example, have argued [89].

By allowing users to co-construct decision making processes, social media systems may be able to avoid some of these troubling issues. In this study, we explore what users value in content moderation systems and how they envision groups like themselves shaping decision making processes. We conduct a series of participatory design workshops to allow everyday users to design solutions, and recruit from — and situate workshops in the experiences of — members of communities that have suffered from problems with content moderation. We conduct nine workshops, three from each of three groups previously harmed by content moderation: Black, Indigenous, and people of color (henceforth BIPOC), those in the LGBTQ+ community, and artists.

Using a series of participatory design activities, we capture the values participants think are important and explore participants' ideas for how to design for contestability in these systems. We find that participants have a diverse of values they consider important for the design and evaluation of content moderation systems – ranging from diversity and inclusion to safety and security. However, these values drive a small set of fruitful, practical avenues for designing for contestability: adding representation, improving communication, and designing with compassion.

2 RELATED WORK

In this section, we summarize the problems that content moderation creates for social media users and drive its choice as a system to improve. We also motivate the idea of designing for contestability and our use of participatory design.

2.1 Content Moderation

As social media platforms grew, they experienced an onslaught of problematic content, including self-harm, violence, sexual content, and hate speech. Content moderation systems attempt to filter out such content, ideally before it reaches any users. In order to remove problematic content, most platforms use a hybrid approach, where some content is automatically flagged by algorithms and other content is analyzed only after a user manually flags it [98]. In most cases, the content is reviewed by a human moderator, though recently (in response to the COVID-19 pandemic), some platforms have shifted to almost entirely automated review [83]. But while content moderation is well-intended, the process can harm moderators and decisions can harm users.

In recent years, the public [5, 8, 34, 57, 80, 110] and research community [23, 33, 41, 49, 64, 82, 98] have noted the problems content moderation can cause. Issues include the inconsistency and unfairness of decisions [41, 68, 81, 99, 103, 116], the harms moderators experience as part of their work [41, 98, 107, 110], and whether content moderation is (or can be) used to educate users [59, 86].

The potential for harm is particularly true for those in marginalized communities [8, 16, 20, 57, 79]. For example, content moderation systems have suppressed content shared by disabled, queer, and fat creators [7], which can isolate them and limit work opportunities [10]. Similarly, they have cut people with eating disorders off from community support and reproduced conformity to

certain body images [33]. And shadow-banning and deplatforming of sex workers can exacerbate inequalities, chill speech, and disrupt movements for change [8, 9]. These workshops focus on the experience three communities, all of which have experienced harms of content moderation in the past: BIPOC users, LGBTQ+ users, and artists. While many other communities have also faced frequent issues, the harms experienced by these three groups are common and well-researched.

BIPOC: When users call attention to racist content or hate speech, many have their content removed or accounts suspended. This has been particularly true for racial minorities [5, 57]. For example, when the writer Ijeoma Oluo suffered racist attacks, she posted screenshots of the messages; but it was her account that was suspended [46]. Issues also arise when communities seek to reclaim formerly pejorative terms, but face suspensions or deletions. And researchers at Instagram found that BIPOC users' accounts were 50% more likely to be automatically disabled [111].

LGBTQ+: Users from the LGBTQ+ community, particularly rural youth, have a long history of using social computing as a way to build community, understand their bodies and develop their gender expression [44]. However, their content is often removed for containing nudity [20, 79]. For example, prior to its 2018 policy changes, Tumblr "*allowed erotic content needed for intersectional trans community building*" [47], but began aggressively removing content and suspending users after that policy change [77]. As with BIPOC users, those who attempt to engage in counter speech or reclaim perjorative terms can be silenced [78, 112]. Finally, platforms have restricted the visibility of LGBTQ+ users and hashtags [37, 100].

Artists: While not traditionally marginalized in society in the same ways as BIPOC and LGBTQ+ communities, artists have encountered repeated issues with content moderation. Content moderation systems frequently take down artists' accounts for containing nudity, even when nudity in artwork is permitted by platform policy [29]. Artists have argued that since social media is so important for their profession, this moderation has a "chilling effect," where they have learned to self-censor their work to ensure their accounts remain active [89].

With participants from these three previously harmed groups (BIPOC, LGBTQ+, and artists), we explore how to design for contestability in content moderation systems.

2.2 Designing for Contestability

Systems designed for contestability allow users to shape and influence decision-making processes. The idea of contestability in technology systems can be traced back to early expert and mixed-initiative systems where experts negotiate with or correct the system to optimize its output [42, 87, 106]. More recently, researchers have returned to this idea, arguing that the algorithmic experience can be improved by allowing users more of a voice in how decision are made [115, 119].

Researchers have argued – in the context of new GDPR requirements – that proper protection data subjects' rights is feasible only if there are means for contesting decisions based on automated systems [3]. In addition to explainability and transparency, some have argued that expert decision-support systems can foster critical, generative, and responsible engagement among users, algorithms, system designers, those subject to decisions, and general public by desinging for contestability [85].

This has been borne out in practice. In high stakes domains, allowing practitioners to contest decisions can make systems more understandable, useful, and accountable [54]. Similarly, systems used by lawyers to predict which documents should be shared during the discovery process require transparent and configurable processes to ensure proper decision making [65]. Anonymous networks users have also called for a contestable mechanism to address abuses [1] And contestable design universally improved perceived fairness in an algorithmic system for distributing goods, by allowing people to realize the inherent limitations of decisions [71].

However, while recent work has proposed designing for contestability as an approach for social computing systems [116], most work on contestability has focused on expert users. In this project,

we instead explore how everyday users could take part in contestable systems, using participatory design methods to generate visions for how these users could shape content moderation.

2.3 Participatory Design in HCI and Machine Learning

Human-computer interaction research has drawn on participatory design methods for decades [6, 27, 38, 62]. Participatory design is the process of involving groups of users in the development of systems to better support diverse user interests and goals [72, 84, 102, 108]. By involving users in the design process, the participatory design method aims to avoid many of the unwanted effects of technology as seen from the users' point of view [108].

Unlike other forms of user experience research, participatory design is often leveraged as a generative design approach, addressing people's need and desire to create solutions rather than evaluate existing approaches [102]. It is intended to empower participants as it learns from them; when engaging with those marginalized in the past, this kind of empowerment is particularly important [51]. Participatory design is also a value-centered design approach that embeds democratic values into its practice [102]. Participatory design practices encourage participants to share and incorporate their values at the early stages of the design process [117, 118]. And research on ethical AI has highlighted the important role of values in understanding users' needs [15]. We draw on this rich literature to design our methods — particularly the "Convivial Toolbox" developed by Sanders and Stappers [102] — to address the unwanted effects of content moderation experienced by users.

Early participatory design in Scandinavia supported workers' self-determination of local working conditions and policy [36, 84]. The approach has now been used in many fields of HCI research (e.g., software development [43], educational technology [63, 67], internet of things [91] and health [35]).

Recently, participatory approaches have gained popularity for the design of the algorithmic experience. Researchers argue that participatory approaches enable users and designers to work together while negotiating the challenges algorithmic systems pose to society [13, 52]. To address concerns of surveillance and automation, researchers collaborated with community groups to produce materials promoting awareness and offering advocacy strategies [60]. For child welfare services, participatory design workshops surfaced families' perceptions of the current algorithmic system and developed strategies to decrease discomfort with it [14]. Through participatory design workshops, people from marginalized communities designed new social technologies together with researchers [40, 48, 50, 114] and identified future technologies to support their needs [48]. Our study uses a similar approach, to design the algorithmic experience of content moderation, together with communities who have been harmed by it in the past. In doing so, we investigate:
RQ1: What values do participants think are most important for the design and evaluation of content moderation systems?
RQ2: What design ideas do participants generate when designing for contestability in content moderation systems?

3 METHODS

Our study uses a generative approach — participatory design — to create solutions for how users can shape and influence algorithmic content moderation systems. We recruit participants from groups that have encountered problems with social media content moderation in the past: BIPOC users, LGBTQ+ users, and artists. Each workshop focused on one community member's experience; we describe each case study before describing the rest of the workshop protocol.

3.1 Case Studies

Case studies were selected to share an individual's experience of a more widespread problem within each community, drawn from news reporting on the topic. Two cases addressed populations that

have been marginalized and discriminated against more broadly (BIPOC and LGBTQ+ users), the final case addressed artists, whose professional life can be harmed by content moderation.

BIPOC: Many users have content taken down or accounts suspended when they call out or call attention to racist content or hate speech. This case study focused on the experience of Francie Latour, as shared in the Washington Post [57]. Latour was grocery shopping when a man directed "*a profanity-laced racist epithet*" at her two young sons. Latour turned to Facebook to vent about the experience and shared the hateful words the man had said, saying: "*I couldn't tolerate just sitting with it and being silent. I felt like I was going to jump out of my skin, like my kids' innocence was stolen in the blink of an eye.*" However, within 20 minutes, Facebook deleted her post, with a brief message that her content violated Facebook's standards. Only two friends had gotten the chance to voice their anger and support for her. This case study focused on the harms to BIPOC users by not being able to share their experiences of racism and receive peer support.

LGBTQ+: Many users have content taken down or accounts suspended when they share content containing nudity. This has been particularly true among LGBTQ+ communities that seek to build body positive and sex positive communities online. This case study focused on the experience of Nyx Serafino, as reported by The Guardian [55]. As a gender-fluid sex worker, she struggled with her identity and childhood abuse. However, she found community on Tumblr, saying, "*It was a great place to mix art and adult content. I could put out my perspective on things, post a song, and feel comfortable in my own skin. It took a long time for that to happen for me.*" But after Tumblr's change in policy, its automated systems began aggressively flagging content [101] and over 20% of its user base left the platform [113]. This case study focused on the harms to LGBTQ+ users by not being able to build this kind of body positive, sex positive community.

Artists Users also have also have content taken down or accounts suspended when they share their artwork that contains nudity. The problem has become so recognized that professional artists were solicited for help by Instagram. This case study focused on the experience of Betty Thompkins, an American painter renowned for her Feminist art, who shared her experience in an interview with ART News [104]. Her account was suspended after sharing a catalog page featuring one of her explicit paintings. She said "*This is our job as artists: to break the rules. That's what makes it art—it doesn't conform.*" This case study focused on both the harms to artists' livelihood and the potential for chilling effects on what type of art is created.

3.2 Recruitment

Participants were recruited primarily online[1], through online newsletters, online contacts for community organizations (e.g., Women of Color, U of I Pride, 8 to CREATE) and through direct contacts. Rather than aiming for a sample representative of the United States, we instead recruited from three communities previously harmed by content moderation: BIPOC, LBTQ+, and artists.

Participants completed a brief initial survey including demographic questions. Three questions focused on these communities; one asked if the participant worked as an artist or shared their artwork online, one whether they identified as LGBTQ+ or as part of the LQBTQ+ community, and the final asked if they identified as a racial minority. Using participant responses, workshop groups were developed around these three questions. Of 96 responses to the initial survey, 33 people answered no (or no response) to all three questions and were excluded, and 30 others either could not be scheduled or did not attend their scheduled group. Each participant took part in a single workshop. We recruited participants from populations that have been previously harmed, but included participants who had not personally experienced content moderation. Nevertheless, 67% of participants had personally experienced content moderation or heard a friend's experience.

[1]Due to the COVID-19 pandemic

As Finch and Lewis write in their chapter on focus groups: "*In studies researching sensitive subjects, the shared experience of 'everyone in the same boat' is particularly important to facilitate disclosure and discussion*" [97]. Since the topics covered in our study were somewhat sensitive (having to do with racial slurs, nudity, etc.), we ensured that all participants within a group responded the same way to at least one of the community questions. Then, because "*some diversity in the composition of the group aids discussion*" [97], we used the remaining demographic questions (age, gender, household income, social media accounts, etc.) to ensure some diversity within the groups. This process facilitated disclosure by ensuring a degree of commonality, but also enough diversity that differences can be drawn out in the discussion.

We ran nine workshops, three for each of the three case studies. The workshops had an average of 3.7 participants (mode = 3, with a range of 2–6). The demographic composition of each workshop and overall is included in Table 1. Workshops lasted two hours each and were conducted over two weeks in mid-June 2020. Participants were paid $20/hour for their participation, including an estimated half hour for the pre-workshop preparations and activities.

3.3 Workshop Protocol

Each workshop consisted of three activities: slides, design activities and group discussions. Participants prepared for the workshop by completing a "home workbook". The home workbook was designed as a sensitizing activity, where "*in the period preceding the group session [...] the participant gets a feeling for the goals and topic of the study, collects personal experiences and increases his or her understanding*" [102]. Because we recruited participants for their community affiliation, rather than prior experience of content moderation, this pre-workshop work helped participants prepare and develop their thinking about content moderation. All three case studies were shared with participants in their home workbook. The one case study aligned with the group's population was reviewed at the beginning of the workshop. The home workbook also encouraged users to explore their relationship to social media, think through their opinions on different kids of content, and

Group	Case	P (#)	Age		Gender[a] (%)		Race (%[b])				
			M	Range	M	F	A[c]	B	H	N	W
1	LGBTQ+	4	41	29–58	25	75	–	–	–	–	100
2	LGBTQ+	3	36	19–64	33	66	–	–	–	–	100
8	LGBTQ+	3	21	20–21	33	66	66	–	33	–	33
3	BIPOC	3	42	36-48	33	66	–	100	–	33	33
6	BIPOC	6	21	19–22	16	83	50	50	–	–	–
7	BIPOC	6	21	19–24	33	66	33	–	66	–	–
4	Artists	2	35	24–45	50	50	–	–	–	–	100
5	Artists	3	21	20–22	33	66	33	–	–	–	66
9	Artists	3	36	24–61	33	66	66	–	–	–	33
Overall		*3.7*	*28.7*	*19–64*	*30*	*70*	*33*	*18*	*15*	*3*	*42*

[a]No participant selected 'Non-binary/third gender' or 'Other'

[b]Percentage of respondents who reported that race or ethnicity; values may add up to more than 100%

[c]A - Asian, B - Black or African American, H - Hispanic, Latino or Spanish, N - American Indian or Alaska Native, W - White. No participant selected 'Other'

Table 1. Workshop Participant Demographics

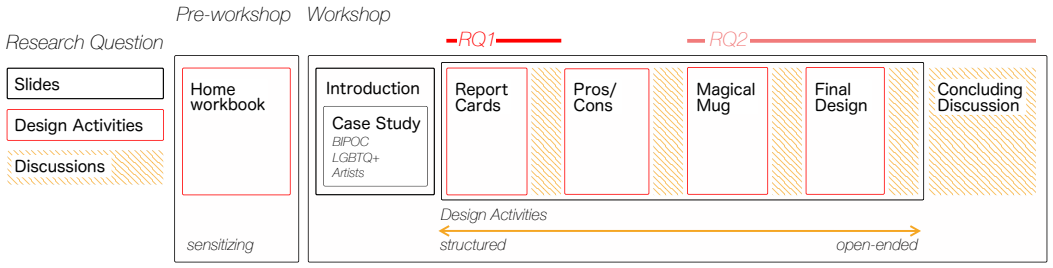


Fig. 1. Study Design. The study included both sensitizing activities prior to the workshop and a series of design activities (interleaved with group discussions) that were increasingly open-ended.

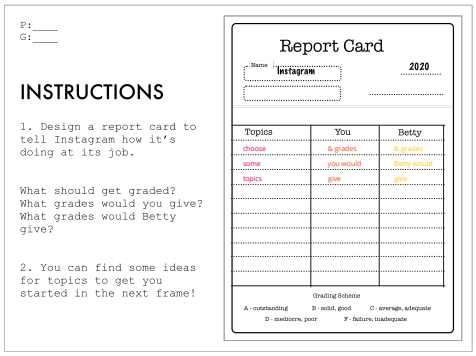
reflect on their own and others' experiences with content moderation, through writing, drawing, and playful activities (like making memes).

After introductions and a brief review of the workshop's case study, the workshop itself featured four design activities, interleaved with group discussions (as shown in Figure 1). The design activities were developed with two goals in mind: 1) establishing participants comfort and rapport with the group and 2) building up their ability to design a solution to the problems of content moderation. The first activities were highly structured to provide more support to participants during their initial engagement; activities became increasingly open ended as the workshop progressed. After each activity, participants shared their work and thoughts in a discussion with the group. A closing discussion after the final design activity asked participants to identify any common themes that had emerged and whether there were any aspects or concerns that we had not talked about during the workshop. The full protocol as well as the materials used for design activities are included in the supplementary materials. Since each case study featured a different protagonist and social network, all design prompts were tailored for the case study.

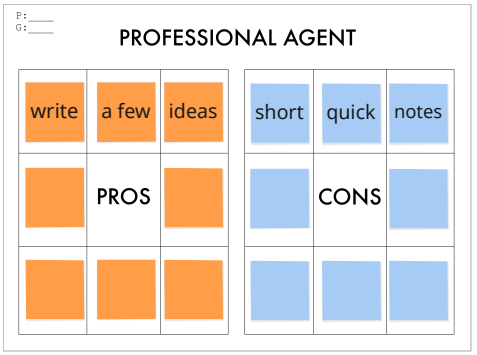
Report Cards: During the first activity, participants filled out a report card for the social network featured in the case study. This activity was designed to elicit participants' values and to allow participants to share those values with others in the workshop. Eliciting values is an important aspect of design research and has been successfully approached with a wide variety of design activities, including cultural probes, scenarios and card activities [118]. This activity was used to elicit the values of each participant, which are used to answer RQ1. But it was included as the first activity of the workshop (rather than in the home workbook), so that participants could share and establish a common set of values as a group. As the first design activity in the workshop, it was also designed to be highly familiar and highly structured.

Participants were provided a template report card (Figure 2a), which was personalized to fit the case study for their group (e.g., evaluating Instagram). Participants were asked fill in what they would grade the social network on: "*Imagine yourself in the role of a teacher, and [Instagram] is the student. But unlike students who get graded on things like Math and Science and Geography, you're going to decide what [Instagram] should get graded on.*" To have them think critically about the case study and their relation to it, participants provided the grade *they* would give the social network for that topic and the grade they thought the protagonist of the case study would give. Since some participants initially struggled to generate terms on their own, an "idea list" of over 100 values (e.g., adaptability, adventurousness, assertiveness) was also included nearby to help them get started. Participants also generated many of their own terms.

Pros/Cons: The second activity had participants evaluate pros and cons of four possible designs for allowing users to shape and influence content moderation. This activity was designed to



(a) Report Card Activity Template



(b) Pros/Cons Activity Template

Fig. 2. Design Activity Templates. The earlier, more structured design activities used templates: 2a, a report card that participants filled out with the values they would use to evaluate a content moderation system, and 2b, "sticky notes" to brainstorm pros and cons of different alternatives to content moderation systems.

encourage participants to broaden their thinking beyond the existing approach of social networks. Participants were briefly introduced to four approaches (current system, structured application, community forum, and professional agent), described in Table 2. The reason for including each approach is noted in Table 2, but in general alternatives were chosen to include a breadth of possible approaches, particularly those suggested in prior work [70, 86, 116].

Participants were given a template for each approach, as shown in Figure 2b. The template included eight sticky notes for pros and eight for cons of each approach. This activity was inspired by the "crazy eights" brainstorming activity that encourages designers to generate many ideas [69], to encourage our participants to think as broadly as possible about potential strengths and weaknesses of each possibility and to begin thinking about alternatives not currently in use.

Approach	Summary	Reason for inclusion
Current system	Individual user clicks "request review" and are emailed when a final decision is made.	Ensures everyone is aware of and thinks critically about how current systems work
Structured form	Individual user fills out a form that indicates the kinds of information taken into account	Introduces context information to decisions, eases difficulty of reading standards documents
Community forum	Groups of users discuss content moderation approaches, similar to the Yelp community forum	Introduces community and consensus-focused approach, rather than individual action
Professional agent	A agent makes an argument on the user's behalf, as an actor's agent negotiates contracts	Introduces the idea of professionalism and expertise, could improve success rate

Table 2. Pros/Cons Systems. Participants evaluated four alternatives in the pros and cons activity.

The later activities began generating potential designs; these (with the closing discussion) are used to answer RQ2. As participants had a better sense of how to approach design activities, no template was provided for the final two tasks.

Magical Mug: The third activity had participants design a "magical mug" that could help the protagonist of their group's case study as they went through that experience. The goal was to begin having participants generate their own solutions, while still providing some structure. The design task was developed iteratively in discussion with design faculty, after considering several alternatives. The prompt evolved from one generated by the "*What Should I Design*" generator [58], with the "magical" element included to ensure participants did not fixate on how platforms currently approach these problems, but instead tried to develop an ideal solution.

The researcher leading the workshop briefly reiterated the situation of the case study, and then asked participants to design the mug, saying, "*the magical mug can help her – in whatever way you want. And since it's magical, you can have it do whatever you want!*" In the task instructions, participants were encouraged to use the lens of one of the pros: that a structured application can help someone know what to say. While many participants explored ideas beyond this, the instruction and the specific task of designing a mug were included to provide some structure.

Final Design: The final activity gave participants the most open-ended prompt:

> *Design any kind of system or interaction or app or physical thing you want – with the idea that it can help communities or groups of users, people like [Nyx], to shape, influence, or improve how content moderation is done at [Tumblr].*

Participants were encouraged to design with the previous activities in mind: to include as many pros and as few cons from the Pros/Cons activity, or to draw on the strengths they saw in each others' Magical Mugs. As we note in our results, participants were highly attuned to the incentives of the platform, so groups were also told to imagine the social network supported the design.

3.4 Moving Online

Initial pilots of this study were conducted in person. After the growth of the global COVID-19 pandemic, further pilots and the workshops we report on were conducted online. This led to major changes. All workshops were conducted via Zoom[2], an enterprise video communication system. All the design artifacts were generated in Miro[3], which is an online collaborative whiteboarding platform. While Zoom includes built-in whiteboarding functions, Miro allows us to provide templates, archive participants' designs, and provides a better multi-user experience (allowing participants to zoom in on their work area, for example).

Participants typically used two windows, one for the design artifacts in Miro and one for the audio and video in Zoom. However, several participants had constraints that meant participating via mobile devices or using different devices for the meeting and for the design activities. As a result, the researcher coordinating the workshop also shared the design artifacts through Zoom when they were being discussed. While most participants were familiar with videoconferencing in general, the online whiteboarding was a new experience for many. A few participants were unable to learn the new technology of Miro in time for their scheduled workshop. While Miro itself offers little in the way of onboarding videos, we did share an overview video developed by a third party. We also offered one-on-one tutorial sessions with any interested participants to go over Miro.

The shift online also led to changes in participant behavior. For example, initial pilot studies were full of rich turn taking. In an online videoconferencing system, signaling interest and coordinating turn taking is much more difficult. In the eventual workshops, we found that many participants

[2]https://zoom.us/
[3]https://miro.com/

communicated all of their ideas at once, and were less likely to break in to share thoughts as later participants spoke. Smaller workshops tended to have more back-and-forth interactions and might be a better choice for online environments.

Challenges around connectivity and the technology were significant. Many participants had issues with low-bandwidth connections. As a result, several could participate only with audio or occasionally dropped out of the workshop. In one notable case, the participants actually began referring to each other by participant ID rather than name, because that was shown on screen and was most salient. But while it is clearly preferable to share video for participants to get to know each other, that places greater demands on participants who lack access.

As with other teleconferencing settings, we found that interruptions increased after moving online. Participants occasionally stood up or stopped to communicate with roommates or family. In general, these interruptions were brief and did not interrupt the flow of the workshop. There were also conversations and even fights in the background, so their focus may be lower than in other settings. But in some cases, where participants were willing, participating from their homes could add to the workshop. For example, one participant walked the group around their home, to share the artwork they had tried to post on social media before their account was suspended. Thus, future online participatory design workshops might consider the potential provided by the unusual access to personal spaces as a way to balance the added challenges of conducting workshops online.

3.5 Limitations

There are some limitations to this approach. While 67% of participants had prior personal experience with content moderation, the study design is likely to influence the values and suggestions they share. If a participant had previously experienced content moderation related to self-promotion, for example, a value like "diversity" might be less salient. Similarly, because not all participants had previously experienced these harmful forms of content moderation personally, many participants distinguished between what they and the case study protagonist would evaluate in the activity. We suggest that these values and designs are specific to these harmful forms of content moderation, and less harmful forms should be studied as well. Further, while all participants used social media (on average reporting use of 5.8 platforms), not every participant used the platform featured in their case study. Finally, our study is small and involves self-selected participants, so we may miss important subgroups within these communities. For example, no participants selected 'Other' or 'Non-binary' in the LGBTQ+ groups. Future studies could cover these communities more comprehensively.

3.6 Ethical Considerations

Researchers have recently begun attending more carefully to ethical considerations in the practice of participatory design. One recent effort outlined four principles of central importance: free and informed participation, minimizing risk of harm, maximizing outcomes and benefits, and supporting appropriate empowerment [61]. These principles can be challenging to employ in practice.

For example, part of acheiving informed consent requires recognizing that "*design is not a universally understood practice*" and that organizers need to take care when sharing the project details, activities, and goals [61]. In some cases, researchers have identified issues with particular activities, for example, brainstorming with markers and colored pencils, which participants can perceive as "*infantilizing and belittling*" [51]. To minimize these concerns, when participants chose to "write" a design instead of sketching it out, the research team supported that decision. Similarly, when participants were asked to engage in 'blue sky' brainstorming like the magical mug (which some can perceive as a "*luxury practice*" [51]), the research team shared why: to avoid focusing on what platforms currently provide. However, even with careful consideration, challenges remain. For example, one principle highlights supporting fair and appropriate empowerment, which can

"*involve trying to find ways to distribute power more equally amongst stakeholders*" [61]. In our case, the powerful stakeholders are the social media platforms themselves. While we can add our voice to calls to attend to these users, we cannot enforce or demand those changes.

4 ANALYSIS

All nine workshops were recorded and transcribed. The results take the form of: the transcript for the workshop and the paired design artifacts. Results were analyzed using qualitative coding.

To address RQ1 (understanding the values that participants considered most important), the report card design artifacts were analyzed using the "spreadsheet method," as described in [102]. This process first counts the number of occurrences of each value (e.g., 'honesty') on the report cards. The values are then clustered to group related terms. Two researchers independently performed the clustering and iterated on the groupings until they reached agreement. The results present only the major clusters, though the full set is included in the supplementary materials. We analyze the artifact rather than the transcript for this research question because some values may not have been covered in the group discussion due to time constraints rather than importance.

To answer RQ2 (understanding how users approach designing systems to allow user influence and control), we analyze the magical mug and final design activities as well as the closing discussion. The designs participants created were analyzed together, as we found many of the same themes emerged. A randomly selected subset of designs (five per activity) is included in the supplementary materials. Iterative open coding was used to identify themes in the designs participants created and how they discussed them [12]. A large initial set of codes were discussed by two authors; 18 of the most common were categorized into three broader themes. Both codes and themes were discussed by two authors until agreement was reached [31].

5 RESULTS

5.1 RQ1 Values For Content Moderation Systems

To address RQ1, we analyze the report cards that participants generated in the first activity. Participants had a very diverse set of values that they consider important for the design and evaluation of content moderation systems. The most common clusters are shown in Table 3, though only those where five or more participants contributed.[4]

Likely influenced by the case studies that participants were reflecting on, the most commonly shared value was inclusivity. Twenty distinct participants – almost two thirds of all participants – mentioned concepts around diversity, tolerance, and inclusiveness. As one participant described this in the discussion:

> *I put tolerance and inclusiveness. Because it is... I think these questions have to do with large communities of people with different ideas about what is acceptable. So, that's the question of, to what degree do we tolerate differences, even if we don't agree with them? And inclusiveness is similar, but it has to do with how much people feel included in this community that they voluntarily joined.* [P12]

The second most common cluster considered a very different direction: the competence of the platform itself. Unlike the inclusiveness cluster which featured only a few, very common terms, participants used a wide variety of terms to describe the aspects of competence that they valued: being reliable, accurate, rigorous, efficient, useful, and so on. For example, one person who thought quality was important defined that as, "*In a sense of what it does to show you in your newsfeed, of how it uses AI to really customize and show you stuff that keeps you active on the site, that keeps you wanting to scroll*" [P16]. For these users platforms need to be well designed and executed: "*there*

[4]Other clusters did have more total words, but were contributed to disproportionately by a single person.

Inclusivity	**20**[a]
Inclusiveness	10[b]
Diversity	6
Tolerance	4
Openness	3
Belongness	1
Discrimination	1

Competence	**19**
Usefuleness	5
Accountability	4
Consistency	3
Professionalism	2
Quality	2
Accuracy	2
Competent	1
Reliable	1
Carefulness	1
Rigor	1
Discipline	1
Efficiency	1
Responsibility	1

Communication	**14**
Transparency	10
Communicable	2
Honest	2
Clear policy	1
Shadowbanning	1
Trigger warnings	1
Quick response	1

Equality	**10**
Fairness	7
Equality	2
Democraticness	1

Security	**9**
Security	5
Safety	3
Privacy	1
Privacy policies	1

Compassion	**7**
Empathy	2
Support	2
Compassion	1
Sympathy	1
Sensitivity	1
User Support	1

Fun	**7**
Fun	3
Enjoyable / fun	1
Enjoyment	1
Enthusiam	1
Interesting or not	1
Interests/Hobbies	1

Freedom	**5**
Freedom	5

Pro Social	**5**
Making a difference	3
Community	3
Justice	1
Socially responsible	1
Community building	1

[a]Number of unique participants contributing to that cluster. One participant may contribute multiple terms to a cluster.
[b]Number of unique participants who include that value

Table 3. The values users consider most important in the design and evaluation of content moderation systems, where at least five different users included the term in their "report card" for the social network.

are so many platforms out there there must be something about this one that made them want to use it" [P1]. And as has been shown in prior work [25], participants have strong opinions about how the news feed curation operates: "*I think the algorithm accuracy is pretty bad also. I had Instagram back when it was still chronological timeline, and I always wished that that would come back*" [P11].

Another large cluster dealt with the communication that the platform provides with its users, focusing on issues of transparency, clear policy, and quick responses. Participants also included some aspects (rather than values) that they did not agree with, like shadowbanning. Many participants thought platforms performed poorly on that front:

> *Whilst Instagram does a good job at explaining like what their content is, they do not do a good job of telling the users why their thing is being specifically taken down. They'll just be like, 'Oh, you broke one of the codes of conduct,' or, 'Your post has been flagged for not meeting the community guidelines,' but then won't go into specifics or communicate with them how to avoid this in the future* [P15],

But others disagreed and mentioned prior experiences with platforms communicating their policies and standards: "*I feel like I get sometimes posts from Tumblr, like their actual blog about their policies, so I feel sometimes they do an okay job of at least telling people what the policies are*" [P29].

Finally, another large cluster addressed ideas of equality, fairness, and democratic values. Often in the discussions this emerged as an idea concerned with resisting censorship, though in some cases participants mentioned that once a user has one piece of content flagged the problem can become recurring, "*I personally think that I've never had an issue with it, but I don't really post anything that controversial. But I have noticed a lot of people that don't get treated very fairly, they get things removed a lot*" [P13] and another participant replied "*I definitely agree with [P13], I've seen it a lot with other people. And what I've also been thinking about is when somebody has their account flagged, or something gets taken down, I feel they're more prone to have it happen again to them*" [P15]. Compared to the value of inclusivity, participants were concerned more about process when they discussed the value of equality, "*I also put fairness. And it's more talking about the content review and how fair the process is to artists and people of different professions*" [P14].

Many of these values: communicating to and from the platform, supporting diverse participation, taking democratic approaches, as well as others from the smaller clusters like compassion and pro sociality emerge in many of the designs that participants develop in the later portions of the workshop, which we report results from next.

5.2 RQ2 How Users Design for Contestability

We highlight three major themes we identified from participant designs and the group discussions: representation, communication and compassion (Figure 3).

5.2.1 Representation. One of the most common themes that participants identified was the idea of representation: "*something I'm concerned with is also that people have representation on whatever the spaces or platform they're using*" [P6]. Participants shared *many* ways of achieving this (Figure 3), which can be grouped into *direct* or *indirect* forms of representation.

Forms of Representation: Some suggested very direct forms of representation as part of building community consensus into content moderation, for example, "*a randomly selected jury duty*" [P25] or "*vote*" [P5, P10] to establish the norms and policies of content moderation: "*Randomly select 10 people and honestly, there's a lot of people on Facebook, so you could pick 1000 people and then you say this, 'Should this be allowed on Facebook?'*" [P25]. Participants who agreed with this approach appreciated that many users could be involved and noted that it is a "*good idea to have*

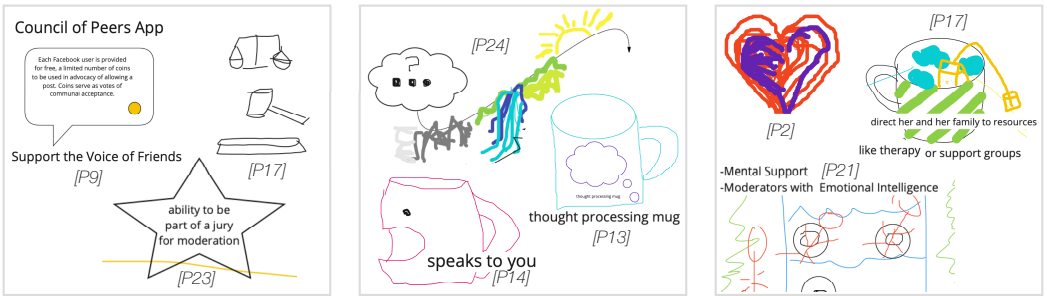


Fig. 3. Themes of Designing for Contestability in Content Moderation Systems. Participants generated a diverse set of ideas, with many focused on adding representation, improving communication, and designing with compassion. While many participants suggested practical changes (e.g., P9 designing a "Council of Peers" app to signal communal acceptance), others generated metaphorical or conceptual approaches (e.g., P24's idea of helping users emerge from a fog of confusion to clarity).

a mix of different people" [P23]. These suggestions typically involved having all users engaged directly in decision making.

Others suggested less direct forms of representation, where instead community members volunteer or are elected to represent their peers, like "*some sort of a focus group*" [P7] or "*a board*" [P6]. One participant suggested: "*Maybe have like community-based arbitration, where they're vetted, super users who can actually review it and say, 'Is this actually a problem or is this just a mistake in the algorithm?'*" [P4], while another suggested "*People can run for being on this board, there is some way that people can either be elected to it or group kind of process, and to help set the standards, but also a place to bring disputes about the content*" [P7]. In these suggestions, fewer users would direct weigh in on decisions about content moderation, but would be represented by fellow community members (rather than the platform itself). The forms of representation that participants suggested could have major impacts on the challenges they identified with the approach, as we discuss later. First, however, we address the motivations participants identified for this approach.

Motivations for Representation: Participants justified their calls for representation from both theoretical and practical perspectives. On the theoretical side, two participants referred to the democratic principles that underlie this kind of representation: "*I like the idea that feeling democratic*" [P12]. Others argued that users have a right to representation: "*Representation is necessary and that users have a say 'cause users of platforms are the reasons they flourish, the reason anyone's able to make any kinds of money*" [P6].

However, many participants approached the idea pragmatically. While they mentioned many potential benefits, they were often tied to a concern with *cultural competence*, by which we mean the ability to be aware of and sensitive to cross-cultural differences. Several groups argued that cultural knowledge was essential for evaluating content.

A number of participants discussed the importance of cultural competence in relation to important content that violated platform policy. These workshops were conducted during the Black Lives Matter protests of early 2020, and one participant shared how videos containing violence could be important:"*Most of us would have known nothing about George Floyd [...] had those social media platforms been censoring and not showing*" [P9]. This participant tied the importance of these videos to the scarcity of other forms of access:

> *For me, I think that social media platforms have to understand the difference that they are making in terms of the oppressed groups in society today [...] so that's why I said, don't silence the voices of the oppressed. One of the beautiful things about oppressed groups is you use what's available to try to make the difference in your situation. And so someone in China or India muting that, it also serves as a double injury* [P9].

Precisely because marginalized communities may lack other forms of power and access, the ability to make full use of social media becomes incredibly important. And while many platforms have exceptions for "newsworthy" content, everyday users do not have the privilege to decide what is "news" or have their content exempted.

Even beyond high impact content, participants agreed that this ability to deal with local community norms could be productive. For example, in considering language that might be considered offensive, "*I have family members, women family members, that refer to each other using B. I don't particularly like it, [...] but again, I'm not the one to tell you not to say it, because it doesn't impact me in the same way that it impacts them.*" [P9]. Even within a family, some can have very different expectations of behavior, and forms of moderation that can incorporate these local differences would be useful. Nevertheless, some participants also noted challenges around who should weigh in, which we will discuss shortly.

Similarly, representation can help manage rapid changes in community norms. One participant from one of the LGBTQ+ groups described himself as "*in that generation that I'm still struggling with 'they', so I try to use 'they' but I don't know*" and noted that a representation-based approach could deal with change better than others: "*to determine the approach that it's always changing that represents the diverse aspects of the queer community*" [P7].

Finally, participants argued that this approach could increase trust in the platform and the efficacy of the moderation. Participants saw trust arising naturally, because even indirect representatives would be close to the community: "*these are good people who we trust to make decisions [...] who would be more community-based than company-based*" [P4]. In addition, some suggested representative moderation could be more effective at changing behavior, because users would listen to their community in a way they would not listen to a social media platform:

> *[Your friends] might say, "Hey, you shouldn't be talking like that on the platform." It's nice when your friends check you in that way. But if Facebook is just deleting things before they can even see it, you might not think that what you did was bad. You might just blame the algorithm or you might blame Facebook.* [P8]

If users knew instead that their community was represented in the decision – even directly weighed in on that decision – "*you might actually be able to reflect on your actions*" [P8]. Participants saw this self-reflection as crucial for real change; some users might be "*easily swayed by whatever the majority opinion is,*" but communities actually acheive more "*if the person actually makes the choice themself wanting to change, it has more of a lasting impact that defines who they are as a person. And they become more of an ally and advocating for what's right*" [P16]. Thus, participants saw representation and a voice for their community in moderation decisions as a way to reflect but also shape their communities.

Challenges of Representation: While groups saw clear benefits for designing for representation, participants identified a number of challenges in the discussions as well. Several groups highlighted one well known risk of direct representation: the tyranny of the majority. In one group that discussed direct representation with users voting, one participant argued that "*just because the majority agrees with something doesn't mean it's right. I know there's some things in the past, that a majority have agreed and it's now in the present day morally un-opposable [sic]*" [P27]. Similarly, another group noted that "*having a homogeneous group of people making a decision, a collective decision for an identity that is usually the minority, doesn't really benefit the minority at all and continues to make minority voices not heard*" [P19].

This also connected to a broader challenge many groups identified of who would weigh in. Participants in one group argued that to achieve the benefits of cultural competence, representatives would need to knowledge of specific, local communities: "*American culture, which is completely different, and then Black-American culture, which is completely different, or Spanish-American culture, which is completely different*" [P10]. Some went even further, arguing that within these communities important intersectional identities and experiences emerge:

> *Me, as a man of color, when women of color are talking about issues that are pertinent to their sex and their gender, who am I to say anything other than 'How can I support?' I can't speak on it, I shouldn't try to correct that. You may even have a similar identity. I'm in my late 40s, so even if I see another black man, but he's 22, it's different. And so I can't say, 'Oh, you don't know...' No, he's a young man. So yeah, it's very complex, and that needs to be taken into consideration* [P9].

This question of whether someone without shared identity can weigh in became an important point of disagreement. Some argued that content moderation should be self-determined by only those with personal experience: "*You can read about all sorts of different things and still not know anything*

about it, really. If you're not part of that group and experience it for yourself personally, then this difference is just not for you to even speak about" [P10]. But others suggested that it was possible to learn and know without personal experience: "*We all come from different backgrounds and different identities, so I think that there should be multiple people on the table who are responsible for explaining these different aspects and different perspectives*" [P19]. This question of whether representation should be centralized or distributed out to individual communities was central and unresolved.

There were other challenges participants identified in considering who would weigh in, particularly around bias in who would be involved. Participants identified challenges both of those who would *not* participate ("*if it was random, you always have that person that might skip out on it*" [P23]) and those who would ("*more extremists on the issue might be more inclined to self-select and actually volunteer their opinion*" [P27]). Nevertheless, participants argued that well-meaning people who would be involved exist and could be recruited: "*there are certainly people who believe in the power of social media ...[who are] more interested in creating meaningful change that stems from a platform like this. So I would say that both people who, it's their job but also people who find meaning and purpose in kind of advising others and helping on their own time*" [P26]. But they noted that some guidelines were needed, both before they participate, "*I do think that the people need to be vetted*" [P26], and after "*[you need] ramifications, repercussions rather, if you aren't diligent with your reasoning*" [P27] to address the challenges of laziness or extremism.

Finally, echoing the issue that some users "*might skip out*," many users thought representation in content moderation could be a challenge due to the amount of work (and kind of work) users would need to do. One participant pointed this out as he was describing his own idea: "*I feel like it would take a lot more work [chuckle] then, now that I just thought about it*" [P29]. Similarly, the kind of content this would expose everday users to could be a problem: "*Originally, when I was doing the workbook thing, I did read previously about the human moderators and how much personal distress and damage they go through to have to moderate these kinds of things [...] it's tough to have to ask people to view those kinds of materials*" [P29]. Some even argued that using this approach could destroy the platform entirely: "*if you have something to say about that content that's being reviewed or it's personally affecting you, but if everybody had to participate [...] it would probably crash the social media and nobody would use it*" [P28]. So while many groups saw great potential in the idea of designing content moderation for representation, they also identified major (and potentially fatal) challenges at the same time.

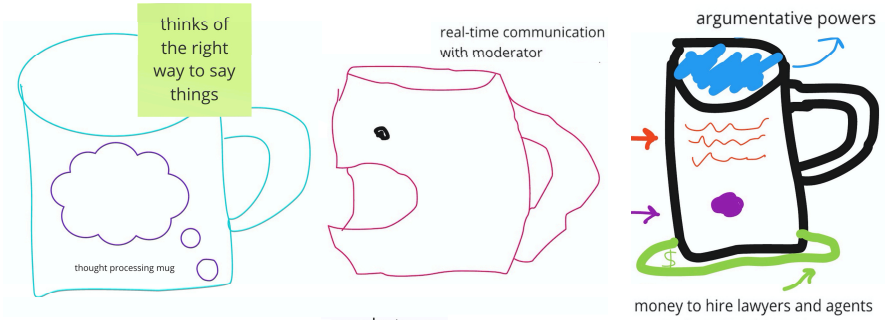


Fig. 4. Designing for Communication. Some communication focused on argumentation (P13), while others focused on access (P14). P29 noted that the argumentation is also tightly connected to resources.

5.2.2 Communication. A second common theme that appeared in many designs for shaping and influencing content moderation was that of supporting users' communication with the platform.

Forms of Communication: In thinking through communication with the platform, groups discussed both questions of *whether* users can communicate with the platform – which we refer to as access – and *how* users communicate with the platform – which we refer to as argumentation.

Many groups focused on the goal of helping users formulate an argument: "*It could also help you find the right way to say things*" [P13]. These suggestions focused on two primary approaches: improving users' persuasion and improving platforms' transparency. One participant from an artist group shared her ideas about how both sides of this communication could work – communication to the platform by users, "*she would be able to provide evidence of others who find her work inspirational, or educational, like there's a hashtag called 'Draw this in your style', [...] provide evidence of a community that she's helping to cultivate*" and communication from the platform, "*and then alternatively, [...] the ability for the platform to provide evidence of the people she's offending with her work, or the effect that this work is having on this person's life who got offended, maybe did their child see something*" [P11]. In many cases, participants saw this as scaffolding users' communication skills: helping them understand what the platform would consider and listen to, while providing the information they would need from the platform to argue for changes successfully.

The idea of access, or whether users *could* communicate with the platform, also emerged frequently in the designs participants generated. Typically these ideas were seen in contrast to the current system, where users are stuck "*waiting multiple days trying to communicate, while not getting anything back*" [P14]. Many participants introduced designs that would allow users to communicate directly to human content moderators: by allowing the user to "*have a chance to even talk to somebody on the phone, or do a Zoom call*" [P1], or "*call or text them personally to see what the issue was and talk to them face-to-face*" [P28]. In some cases this was tied to ideas for transparency and persuasion, for example, connecting to moderators who could "*[give] you the context and you get feedback immediately and they can just assist you on your way*" [P14]. Interestingly, some went beyond this direct access to the moderators and instead argued for direct access to the power brokers at the platform. For example, one participant designed a magical mug where "*you'd pour all your problems and worries into the mug,*" and the worries would evaporate and "*when evaporated would go to Instagram and they would see them in their mugs at their next board meeting and have to talk about them because it has filled up their entire mug*" [P15]. Others appreciated this approach, because "*it's nice that it directly influences the higher-ups instead of the employees, which was usually never the ones to blame, they're just following the guidelines*" [P14]. So while some participants believed that improved access to the moderators could improve communication, others argued access to decision-makers is actually what is needed.

Motivations for Communication: When groups discussed the value of communication, they focused primarily on the practical benefits it would offer. One participant did introduce more theoretical reasons for focusing on communication. He explicitly contrasted the approach of focusing on argumentation with the idea of representation: "*I do find a lot of value in referring to what the community wants, the democracy aspect, [...but sometimes it's] incorrect or not reasonable*" [P12]. Instead, he argued, "*if the artist says, 'No, look, you're thinking about it the wrong way and I have an argument behind that.' That is what I kinda revert to. It seems kind of dry and rational, but...*" [P12] can offer a more justified and consistent approach. Despite this strong theoretical motivation, most groups focused on practical benefits.

In terms of access, they saw it as a way to fix the problem of "*speaking into a void*" that users have expressed in prior content moderation work [86, 116]. As one participant described his idea, "*I'm creating more out of a personal experience of I'm not... I'm usually pretty lazy about filing complaints or applications. I just never think that I would be heard*" [P14]. In terms of argumentation, many

participants saw this as a way to address the cold start problem, when users don't know what or how to say to persuade a social media platform: "*if you know exactly what to say, then it would give you a jump start [...] as opposed to having something where you just have a blank slate and you don't even know what to say or include, you don't know where to start*" [P30]. And in connection to the idea of compassion, which we discuss next, some noted that "*any time you're heated*" [P13] it can be hard to make a clear case or think creatively.

One participant argued that communication is key because "*These aren't laws. Most of the time, they're not federal laws that they're violating, these are just policies put forth by companies that can change*" [P29]. However, as in the case for representation, she highlighted the connection between users' communication with the platform and other forms of power and access. Since these policies are decided by the platform alone, users' ability to drive change depends on their resources. As she described it, "argumentative powers" incorporate resources and money as well as rhetoric: "*There's always a possibility of changing them if you have enough resources, if you have enough money, if you have the... I guess, the argumentative powers to make it happen*" [P29] (Figure 4). So designing for better communication offers the ability to address the "*inequities*" [P27] that exist in terms of how persuasive users can be to the platform.

Communication from the platform was also seen as a way to address inequities that can arise in content moderation itself. One participant made the case that technology can embed bias:

> *If you are using some kind of technology to do your content moderation: What is it? Who made it? Who else uses it? Because I know there's a lot of issues with the way technology is developed [...] Racism, sexism, these kind of things are embedded in technologies and coding, if the individual that created the code has that. So you need a lot more transparency* [P6].

But another noted that even human content moderators can have "*some implicit biases*" but with "*accountability and transparency,*" users "*should know exactly where the decision came from*" [P27]. It was striking that everyday technology users were so aware of potential challenges of automation and called for transparency, accountability, and fairness in the same language as researchers.

Challenges of Communication: Participants highlighted a number of challenges for designing for communication around content moderation systems. Many participants articulated major challenges around the idea of communication from the platform that a number of groups suggested. For example, one participant liked an idea but added you need to share information "*in a way that makes sense. It's not just a bunch of all of the work that goes into making an algorithm that you're like, 'I don't know what this means.' But just simple and concise to the extent it can be*" [P26]. Some went further, closely mirroring work by researchers, who have argued that "*transparency can intentionally occlude*" [4]. One participant shared that, "*I think one thing that they really need to get right is the transparency*," continuing:

> *If Instagram was like, 'Okay well we're unveiling a new content moderation thing.' It is very, very, very easy for someone who is very smart in computer science or tech to completely muddle up the entire thing with these terms* [P15].

Whether participants foresaw this as intentional or not, they anticipated the risks of asking for transparency and made it clear that only some forms of communication from the platform (clear, simple, concise) would actually be useful.

A second, even more pernicious challenge that participants identified was how this approach might run counter to the platform's incentives. Adding enough human moderators so that users could access them promptly would be expensive: "*if that was available to every single person on social media, that obviously would not work out.*" [P28]. While some groups brainstormed intermediaries, chatbots, and other measures that could substitute, they saw the cost to platforms as a major issue. This was an issue users often found easiest to solve in the magic mug activity, for example,

envisioning "*every time maybe she used the mug, she would get money. And she could use that money to hire lawyers and agents to force Tumblr to change*" [P29]. Another noted the fundamental problem of scale, where the platform could easily ignore the problems of any small minority of users. To solve these problems, "*it's gonna take a lot of manpower*" [P15], so "*it's very easy for Instagram as a big corporation to brush that off and be like, 'Okay, we're not gonna deal with this. We said what we said, if you don't like it, then just don't be on the app. We have 100 million other people who are on this app.'*" [P15]. This participant argued that you need a "*solution forces their hand, and the only way to do that is to overwhelm them*" [P15]. So while participants envisioned money as an easy way to drive communication and changes that differ from platform incentives, they also saw solidarity and communal efforts as an alternative approach.

5.2.3 Compassion. The most prevalent – and perhaps most surprising – theme was the very simple idea of increasing the love and compassion expressed by social media platforms (Figure 5).

Forms of Compassion: Groups' suggestions to increase love and compassion took a number of different forms: suggesting that platforms 1) emphasize sympathy when making decisions, 2) provide emotional and particularly mental health support, and 3) connect users impacted by content moderation to their family and/or community.

Perhaps the simplest change would be emphasizing empathy when making decisions. For example, one design called for "*more moderators with more emotional intelligence*" [P21] (Figure 3), which could help them listen and understand. One participant shared her own experience from college, "*I was a resident advisor and we always are trained on saying, 'I see you, I hear you.' And I think that that's something that needs to be taken into account more*" [P19], where moderators actually acknowledge that they see and hear the harms users experience.

A second, frequent suggestion recommended providing emotional and mental health support. This kind of emotional support was particularly common in the magical mugs, where the emotional fallout of a negative experience could be fixed by magic: "*Drink the contents of this mug and find peace*" [P22]. But some recommended more specific support for mental health: "*it would be really interesting to see Facebook or other social media advocating for more mental health resources, whether that be through virtual therapy sessions, or even just like more videos or content that's tailored with the like, 'How do you process what you're feeling?'*" [P16]. Others suggested that the platform itself would not need to provide this support but could "*[direct] her and her family to resources so that they could get the counseling or therapy*" [P17], connecting users with external mental health support.

Fig. 5. Designing for Compassion. Many designed around ideas of compassion and love, saying, "*the goal is to to have love and peace*" [P2] and how much it can mean to say "*I'm sorry, [and] that you're being seen*" [P19].

Finally, some suggested connecting users to their families and communities. Participants expressed both the need from the user, "*sometimes when you go through those situations, you feel like you're alone, so just know that you're not alone, everyone has to deal with idiots like this too*" [P8]. But they also shared how when they are able to provide such emotional support, they benefit as well: "*So I like to get joy and support someone after they've gone through something shitty and just let them know you're there with them*" [P6].

Motivations for Compassion: Participants motivated this call for compassion in part because content moderation often occurs when users share an initial negative experience. These experiences "*can be very triggering [... and require] proper emotional support, because at the end of the day, we're all human and words could definitely hurt*" [P20]. But they also noted that moderation itself is a negative experience. One participant had been temporarily suspended after posting his artwork:

> *The experience was very, very bad, and I was literally shocked. It took me like two, three days to figure out what is happening because I was literally shocked. I mean, as an artist, I'm very emotional person so I was like, 'What the fuck happened? I can't believe it'*" [P32].

After experiencing moderation, participants argued that users can feel "*forgotten*" [P21] and if they are suspended, "*one of the worst things about that experience would be losing that sense of community you have*" [P6]. For these reasons, increasing love and compassion assume paramount importance.

Challenges for Compassion: Again, as in the discussions around communication, participants noted that love and compassion can conflict with current platform incentives. When one participant described her magic mug, she described it as "*filled with compassion and understanding [...] by wizards who are balanced and listen with their hearts instead of their profit motive*" [P4], and continued:

> *A lot of these social media companies, even if they try to come across as this fair, open space, their main motivation is profit in some fashion. Whether that is from users buying into their system or from advertisers buying in order to push their products, there's a profit motive basis for everything. And for people to genuinely be heard and to feel like they have a voice, I think that at least that profit motive needs to be reduced as the driving force, because to my mind, profit motive trumps compassion all the time* [P4].

Participants saw great value in the idea of compassion, but also saw how difficult it could be to achieve when pitted against concrete goals that social media platforms have as a business. If designing for compassion does not increase revenue, it is hard to justify when "*there's a profit motive basis*" for every decision. This recognizes the potential need to continue looking outside of social media systems when seeking to drive change.

5.3 Connecting Values to Designs

Our results found that participant's designs for contestability in content moderation are largely consistent with their values, though some interesting changes emerge. The most commonly shared values – diversity and inclusiveness – drive suggestions for representation: "*it would be a good idea to have a mix of different people. Just because the input would overall benefit society, since society as diverse as a whole*" [P23]. Similarly, the values of transparency and competence are reflected in designs that call for better communication. For example, participants suggest platforms inform users about the decision-making process and helping them construct strong arguments, "*Instagram should be able to provide evidence for their case*" [P11]. Lastly, users' values of compassion and pro sociality directly translate to demands in their designs for greater love and compassion from social media platforms: "*the goal is to to have love and peace*" [P2].

Despite the overall strong connection between participants' values and designs, a few notable exceptions arose. In one case, a new value became a focus of many designs: user control. As one participant described it in the closing discussion, "*I think control for the users [...] was the most*

common theme to me" [P7]. Participants often discussed this control as a user's ability to choose what they would see and who would view their content, a "*method where viewers can block this kind of content, but she can also access other viewers [...] it's a magical solution for both sides somehow*" [P29]. Several groups discussed it as a form of consent, where platforms could "*allow people to post the things that maybe wouldn't able to be fit into the guidelines right now, and those who have consented to see it would be able to see it*" [P13]. This new value may have emerged as a result of our study design; by explicitly asking participants to think about how they would choose to design content moderation systems, we may have made them more aware of ideas of choice, agency, and control.

Finally, while seven participants mentioning the value of fun initially, it was mentioned rarely and only in passing in the designs participants generated. Again, this may be a function of the study design. Our study focused on case studies that cause substantial harmful consequences for their protagonists, and may have encouraged participants to adopt more serious approaches. In different circumstances, however, shaping and influencing content moderation systems can be fun. Users have created memes about content moderation systems to protest questionable decisions [105], which can attract public attention and awareness [92]. So while this value was less commonly expressed in participants' designs, it may still offer important avenues for future approaches.

5.4 Differences Between Groups

Finally, we examined whether our results could offer any insight into differences in discussions for those previously harmed in our specific context of content moderation (e.g., the artists) and those harmed more pervasively (BIPOC, LGBTQ+ users). Most themes that emerged were shared across many groups. Indeed, many participants shared thoughts (e.g., discussing the Black Lives Matter movement in LGBTQ+ groups) and shared identities with other groups (e.g., in one BIPOC group nearly every participant shared artwork online). But a few interesting distinctions arose.

One distinction was that artists were already very familiar with moderation and workarounds. As one filmmaker noted, "*before we publish or release our film, we have already experienced moderation because the Netflix or PBS will have different requirement for us [...] we always experience that moderation*" [P31]. And while these ideas emerged in other groups as well, artists – particularly professional artists – were very familiar with strategies for sharing their work while evading content moderation. For example, by hosting their work other platforms, like Etsy or even pornography platforms, and altering images in ways that make it difficult for platforms to detect.

Groups of artists were also far more likely to consider definitions ("*what is art?*") than other groups. For example, stating that the protagonist of the case study "*had a definition of art, and I don't feel like I have one*" [P12] and "*art, it's so widespread and nobody knows where the boundary is, and what are we supposed to do and we're not supposed to do*" [P32]. This drove lengthy conversations, like thought experiments of when something transitioned from vulgarity to art. And while some suggested criteria, others wondered how social media could hope to "*distinguish art from the other stuff, 'cause I'm not good at that even as a human being*" [P33] and artist.

While artists engaged in these debates during the workshops, the BIPOC and LGBTQ+ groups were more likely to see the value of having "*tough conversations*" [P3] on social media. For example, one participant developed a metaphor where part of the platform would be "underground," where "*[you know] what you're in for, which is not always gonna be comfortable for you, necessarily* [P3]. Participants in another group discussed how important it is to engage in critical thinking and dicussion, "*my main message was just to get rid of the group think mindset* " [P20] and that it is important "*not to bully or anything like that, but just to be an educator*" [P20]. While this echoes the idea that representation can be used to shape communities through reflection, these groups also emphasized that difficult conversations on social media needed to be part of that change.

Finally, several participants in the BIPOC and LGBTQ+ groups connected their concerns to issues with larger social ideologies and beliefs like heteronormativity, patriarchy, and racism. For example, in considering the LGBTQ+ case study, one participant noted that "*it's a total patriarchal approach because specifically for the Tumblr post, it's female nipples*" [P5]. By connecting to larger challenges in society, these participants made it clear that solving issues around content moderation also involved making progress on these larger social issues.

The differences that emerged are likely to be at least to some extent due to the different case studies and the aspects they made most salient. Further, drawing on the concepts of inclusive design [22], we argue the ideas we generated can be helpful to social media users more broadly.

6 DISCUSSION

6.1 Representative Moderation

Social media platforms are not democratic institutions. But it is striking how little they solicit the public's opinions about what is acceptable. Instead, they decide for their users. But this level of supervision and control is unusual for adults to experience. Our work finds a growing demand for *representative moderation*, or a form of content moderation where users are empowered to have a say in how policies are made and how decisions are carried out[5]. In recent months, platforms have introduced approaches in this spirit [21, 28]. And recent work has found that civics-oriented approaches can improved the perceived fairness of content moderation [32], but our participants identified a number of additional benefits that representative moderation can offer, like cultural competence. These ideas can also serve as prompts to adapt automated approaches. For example, rather than have a single policy, platforms could infer "local" communities and their norms, similar to community-based approaches that have been successful in the past [18]. Still, even beyond those identified by participants, many challenges for representative moderation remain.

Many of the suggestions offered by participants could exacerbate problems identified in prior work, like the frustration users feel at perceived inconsistency [86, 116]. If platforms treat communities differently, this is likely to worsen. Adding representation could also raise questions of accountability. When existing models have problems — for example, targeting Native American users — there is a single organization responsible. Systems that embrace representational efforts may encounter fewer issues but more diffuse responsibility when those problems do arise [93].

Moderators encounter fatigue, trauma, and even PTSD when asked to evaluate potentially harmful content every day [30, 66, 98, 109, 110]. Any approach that asks users to weigh in on decisions is likely to expose some to similarly traumatizing content. In addition, these communities are underrepresented in society. If platforms ensure representatives of these communities are included, they will work (and potentially suffer) more than others. They may also be asked to speak on behalf of many – who might not share their identity, values, or goals. And as communities grow, efforts to use community-policing or other community-based moderation have run into challenges as the communities' norms weaken [17]. This again emphasizes the fact that collective approaches may be preferable to more individualized ones. Instead of asking a single user to represent their entire community, we need approaches where the community can weigh in as a group.

6.2 Communication

Participants suggested improvements for communication both *to users* by platforms and *to platforms* by users. Many participant suggestions were similar for both, particularly around providing transparency and context for decisions. This is particularly striking given the robust research agenda around efforts towards providing transparency and explanations [26, 88, 95, 96]. Recent

[5]While they did not define representative moderation, this term was first used by Boufoy-Bastick and Singh [11].

work has found these approaches can be effective for content moderation [59]. There seems to be ample room to adapt these approaches to address users concerns: providing context, demonstrating consistency in decision making, scaffolding users' arguments to the platform, and so on. One of the greatest opportunities is designing around *social explanations*, as called for in recent work [76].

In addition to platforms seeking to provide this transparency, however, platforms might benefit by providing opportunities for others to co-design content moderation. For example, there may be organizations or other third parties that understand specific domains and can design more effective moderation for them. For example, an arts organization might be better able to adjudicate and articulate reasoning for content moderation of art. Individuals could then opt-in to content moderation that addresses their particular set of needs. Rather than trying to resolve what counts as art, platforms could allow users to delegate to organizations with tailored goals and expertise.

6.3 Compassion

Compassion was perhaps the most surprising theme that participants returned to in their designs. Design considers empathy a central task [39, 53]. In his influential 'Designerly ways of knowing,' Cross distinguished design culture from existing cultures of sciences and the humanities, citing as core values: practicality, ingenuity and *empathy* [24]. Despite the uptake of "design thinking" in industry, some have questioned whether this has included empathy [39], while others have noted the challenges in achieving empathy [53]. This suggests that the task participants set for platforms will not be easy. However, platforms developing content moderation seem to focus primarily on malicious users and those gaming the system. Many users simply called for platforms to recognize the harms they do to users when they make mistakes – the disruption, dislocation, and isolation they can experience. And while most content moderation is not a crisis, on occasion it can be: one young man recently committed suicide after an account suspension prevented him from conducting business and repeated attempts to appeal the decision failed [94]. Platforms already provide mental health support in specific contexts like searches for self harm [73], eating disorders [45], and suicide [19]. Providing emotional support could one small way to begin to address users' calls for additional compassion in the design of content moderation. But more generally, HCI has a long history of developing methods, practices, and techniques for empathy, particularly empathy through dialogue [120]. So while some platforms have spoken with users harmed by past problems [29], both platforms and users could both benefit if they spoke with users proactively instead.

7 CONCLUSIONS

Users' values are closely connected to the changes they would like to see to support users in shaping and influencing content moderation decision making. For example, the largest cluster of values, around diversity and inclusion, drive users to demand greater cultural competence and representation in content moderation. Similarly, users' interest in communication from the platform and equality help shape their demands to support argumentation and access – both by informing users about how decisions are made, but also by supporting requests in a way that addresses inequities connected to argumentative skill, like time, education and access. Finally some directly translate – where users value compassion and demand that platforms rebalance to prioritize compassion over their existing incentives. These results also suggest the importance of considering broader forms of justice when designing social computing systems. The importance that participants place on connecting to family and community seem natural given the original purpose of social computing. However, participants note that platforms often fail to build up these connections, particularly in times of harm due to content moderation. But is precisely at these times when relational justice would emphasize the importance of maintaining community and restorative justice the need to repair the harm.

8 ACKNOWLEDGEMENTS

We thank Sanmi Koyejo, Christian Sandvig, Hari Sundaram, and our anonymous reviewers for their thoughtful feedback on this work, and the NSF and Capital One for their funding support.

REFERENCES

[1] Wajeeha Ahmad and Ilaria Liccardi. 2020. Addressing Anonymous Abuses: Measuring the Effects of Technical Mechanisms on Reported User Behaviors. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[2] Leigh Alexander. 2016. Facebook's censorship of Aboriginal bodies raises troubling ideas of "decency". *The Guardian* (March 2016). https://www.theguardian.com/technology/2016/mar/23/facebook-censorship-topless-aboriginal-women

[3] Marco Almada. 2019. Human Intervention In Automated Decision-Making: Toward The Construction Of Contestable Systems. In *Proc. International Conference on Artificial Intelligence and Law*. 2–11.

[4] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society* (2018).

[5] Julia Angwin and Hannes Grassegger. 2017. Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children. https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms. ProPublica.

[6] Shaowen Bardzell. 2010. Feminist HCI: Taking Stock and Outlining an Agenda for Design. In *Proc. CHI*.

[7] Sam Biddle, Paulo Victor Ribeiro, and Tatiana Dias. 2020. TikTok Told Moderators: Suppress Posts by the "Ugly" and Poor. https://theintercept.com/2020/03/16/tiktok-app-moderators-users-discrimination

[8] Danielle Blunt, Ariel Wolf, Emily Coombes, and Shanelle Mullin. 2020. Posting Into the Void: Studying the Impact of Shadowbanning on Sex Workers and Activists. https://hackinghustling.org/posting-into-the-void-content-moderation/

[9] Ariel Bogle. 2019. Aussie sex workers were kicked off American websites so they built their own. https://www.abc.net.au/news/science/2019-06-22/fosta-sesta-laws-impact-australian-sex-workers-one-year-later/11229724

[10] Elena Botella. 2019. TikTok Admits It Suppressed Videos by Disabled, Queer, and Fat Creators. *Slate* (2019).

[11] Zach-Amaury Boufoy-Bastick and Lenandlar Singh. 2007. The Socio-Technical Indicator Model: Socially-Sensitive CMC Technology, with an Implementation of Representative Moderation. *International Journal of Computer, Information, and Systems Science, and Engineering* 1 (2007), 1.

[12] Richard E Boyatzis. 1998. *Transforming qualitative information: Thematic analysis and code development*. sage.

[13] Tone Bratteteig and Guri Verne. 2018. Does AI Make PD Obsolete? Exploring Challenges From Artificial Intelligence To Participatory Design. In *Proc. Participatory Design Conference*.

[14] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward Algorithmic Accountability In Public Services: A Qualitative Study Of Affected Community Perspectives On Algorithmic Decision-Making In Child Welfare Services. In *Proc. CHI*.

[15] Jenna Burrell, Zoe Kahn, Anne Jonas, and Daniel Griffin. 2019. When Users Control the Algorithms: Values Expressed in Practices on Twitter. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–20.

[16] Online Censorship. [n.d.]. OFFLINE-ONLINE: data visuals exploring how content moderation practices impact marginalized communities. https://onlinecensorship.org/content/infographics.

[17] @cfiesler (Casey Fiesler). 2020. "OK I'm going to weigh in on the tagging/content moderation conversation happening right now". https://twitter.com/cfiesler/status/1273300997801041920

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

[19] Qijin Cheng and Elad Yom-Tov. 2019. Do search engine helpline notices aid in preventing suicide? Analysis of archival data. *Journal of medical internet research* 21, 3 (2019).

[20] Alexander Cheves. 2018. The Dangerous Trend of LGBTQ Censorship on the Internet. *Out Magazine* (2018). http://www.out.com/out-exclusives/2018/12/06/dangerous-trend-lgbtq-censorship-internet

[21] Keith Coleman. 2021. Introducing Birdwatch, a community-based approach to misinformation. *Twitter Blog* (2021).

[22] Roger Coleman and Cherie Lebbon. 1999. Inclusive design. *Helen Hamlyn Research Centre, Royal College of Art* (1999).

[23] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (2016), 410–428.

[24] Nigel Cross. 1982. Designerly ways of knowing. *Design studies* 3, 4 (1982).

[25] Michael A DeVito, Darren Gergle, and Jeremy Birnholtz. 2017. Algorithms Ruin Everything: #RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In *Proc. CHI*. https://doi.org/10.1145/3025453.3025659

[26] Nicholas Diakopoulos and Michael Koliska. 2017. Algorithmic Transparency in the News Media. *Digital Journalism* (2017). https://doi.org/10.1080/21670811.2016.1208053
[27] Carl DiSalvo, Jonathan Lukens, Thomas Lodato, Tom Jenkins, and Tanyoung Kim. 2014. Making Public Things: How HCI Design Can Express Matters of Concern. In *Proc. CHI*.
[28] Evelyn Douek. 2021. The Facebook Oversight Board's First Decisions: Ambitious, and Perhaps Impractical. *Lawfare Blog* (2021).
[29] Maximiliano Duron. 2019. Instagram Holds Closed-Door Roundtable with Artists on Art and Nudity. (2019). https://www.artnews.com/art-news/news/instagram-censorship-roundtable-13431/
[30] Elizabeth Dwoskin, Jeanne Whalen, and Regine Cabato. 2019. Content moderators at YouTube, Facebook and Twitter see the worst of the web and suffer silently. https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price
[31] Victoria Elliott. 2018. Thinking About the Coding Process in Qualitative Data Analysis. *The Qualitative Report* (2018).
[32] Jenny Fan and Amy X Zhang. 2020. Digital Juries: A Civics-Oriented Approach to Platform Governance. In *Proc. CHI*.
[33] Jessica L Feuston, Alex S Taylor, and Anne Marie Piper. 2020. Conformity of Eating Disorders through Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–28.
[34] Juniper Fitzgerald and Jessie Sage. 2019. Shadowbans: Secret Policies Depriving Sex Workers of Income and Community. https://titsandsass.com/shadowbans-secret-policies-depriving-sex-workers-of-income-and-community/
[35] Eivind Flobak, Jo D Wake, Joakim Vindenes, Smiti Kahlon, Tine Nordgreen, and Frode Guribye. 2019. Participatory Design of VR Scenarios for Exposure Therapy. In *Proc. CHI*.
[36] Christine Floyd, Wolf-Michael Mehl, Fanny-Michaela Resin, Gerhard Schmidt, and Gregor Wolf. 1989. Out Of Scandinavia: Alternative Approaches To Software Design And System Development. *Human–Computer Interaction* 4, 4 (1989), 253–350.
[37] Chris Fox. 2020. TikTok admits restricting some LGBT hashtags. *BBC News* (September 2020). https://www.bbc.com/news/technology-54102575
[38] Sarah Fox, Mariam Asad, Katherine Lo, Jill P Dimond, Lynn S Dombrowski, and Shaowen Bardzell. 2016. Exploring Social Justice, Design, and HCI. In *Proc. CHI Extended Abstracts*.
[39] Andrea Gasparini. 2015. Perspective and use of empathy in design thinking. In *Proc. ACHI*.
[40] Cally Gatehouse, Matthew Wood, Jo Briggs, James Pickles, and Shaun Lawson. 2018. Troubling vulnerability: Designing with LGBT young people's ambivalence towards hate crime reporting. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.
[41] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
[42] G Anthony Gorry. 1973. Computer-Assisted Clinical Decision Making. *Methods of Information in Medicine* 12, 1 (1973), 45.
[43] Peter Leo Gorski, Yasemin Acar, Luigi Lo Iacono, and Sascha Fahl. 2020. Listen to Developers! A Participatory Design Study on Security Warnings for Cryptographic APIs. In *Proc. CHI*.
[44] Mary L Gray. 2009. *Out in the country: Youth, media, and queer visibility in rural America*. NYU Press.
[45] Rebecca Greenfield. 2012. Pinterest Has an Anorexia Problem Now. *The Atlantic* (March 2012). https://www.theatlantic.com/technology/archive/2012/03/pinterest-has-anorexia-problem-now-too/330315/
[46] Jessica Guynn. 2017. Facebook apologizes to black activist who was censored for calling out racism. *USA Today* (August 2017). https://www.usatoday.com/story/tech/2017/08/03/facebook-ijeoma-oluo-hate-speech/537682001/
[47] Oliver L Haimson, Avery Dame-Griff, Elias Capello, and Zahari Richter. 2019. Tumblr was a trans technology: the meaning, importance, history, and future of trans technologies. *Feminist Media Studies* (2019), 1–17.
[48] Oliver L Haimson, Dykee Gorrell, Denny L Starks, and Zu Weinger. 2020. Designing trans technology: Defining challenges and envisioning community-centered solutions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[49] Oliver L Haimson and Anna Lauren Hoffmann. 2016. Constructing and enforcing" authentic" identity online: Facebook, real names, and non-normative identities. *First Monday* (2016).
[50] Jean Hardy and Stefani Vargas. 2019. Participatory design and the future of rural LGBTQ communities. In *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion*. 195–199.
[51] Christina Harrington, Sheena Erete, and Anne Marie Piper. 2019. Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. *Proc. CSCW* (2019).
[52] Jaz Hee-jeong Choi, Laura Forlano, and Denisa Kera. 2020. Situated Automation: Algorithmic Creatures in Participatory Design. In *Proc. Participatory Design Conference*.
[53] Ann Heylighen and Andy Dong. 2019. To empathise or not to empathise? Empathy and its limits in design. *Design Studies* (2019).

[54] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E Imel, and David C Atkins. 2017. Designing Contestability: Interaction Design, Machine Learning, And Mental Health. In *Proc. DIS*. 95–99.

[55] Vivian Ho. 2018. Tumblr's adult content ban dismays some users: 'It was a safe space'. https://www.theguardian.com/technology/2018/dec/03/tumblr-adult-content-ban-lgbt-community-gender?__s=abutgqgfq3qtwvi5gq42

[56] Amanda Holpuch. 2015. Facebook still suspending Native Americans over 'real name' policy. *The Guardian* (February 2015). https://www.theguardian.com/technology/2015/feb/16/facebook-real-name-policy-suspends-native-americans

[57] Tracy Jan and Elizabeth Dwoskin. 2017. A white man called her kids the n-word. Facebook stopped her from sharing it. https://www.washingtonpost.com/business/economy/for-facebook-erasing-hate-speech-proves-a-daunting-challenge/2017/07/31/922d9bc6-6e3b-11e7-9c15-177740635e83{_}story.html The Washington Post.

[58] Chantal Jandard. 2015. The "What Should I Design" Generator. http://www.whatshouldidesign.com/

[59] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2018. Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 2 (2018).

[60] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and PM Krafft. 2020. Toward Situated Interventions For Algorithmic Equity: Lessons From The Field. In *Proc. FAT**.

[61] Janet Kelly. 2019. Towards ethical principles for participatory design practice. *CoDesign* 15, 4 (2019), 329–344.

[62] Os Keyes, Josephine Hoy, and Margaret Drouhard. 2019. Human-Computer Insurrection: Notes on an Anarchist HCI. In *Proc. CHI*.

[63] Sunyoung Kim, Muyang Li, Jennifer Senick, and Gediminas Mainelis. 2020. Designing To Engage Children In Monitoring Indoor Air Quality: A Participatory Approach. In *Proc. Interaction Design and Children*. 323–334.

[64] Kate Klonick. 2017. The new governors: The people, rules, and processes governing online speech. *Harvard Law Review* 131 (2017), 1598.

[65] Daniel N Kluttz and Deirdre K Mulligan. 2019. Automated Decision Support Technologies And The Legal Profession. *Berkeley Tech. Law Journal* 34 (2019), 853.

[66] Jason Koebler and Joseph Cox. 2018. The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People. (August 2018). https://www.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works

[67] Dev Raj Lamichhane and Janet C Read. 2020. Play It My Way: Participatory Mobile Game Design with Children in Rural Nepal. In *International Conference on Human-Computer Interaction*. Springer, 325–336.

[68] Kyle Langvardt. 2017. Regulating Online Content Moderation. *Georgetown Law Journal* 106 (2017), 1353.

[69] Iris Latour. 2020. The guide to mastering online brainstorming: 20 Brainstorming Techniques That Work. https://miro.com/guides/online-brainstorming/techniques-methods#20-crazy-eights

[70] Min Kyung Lee and Su Baykal. 2017. Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. In *Proc. CSCW*. ACM. https://doi.org/10.1145/2998181.2998230

[71] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[72] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. 2019. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–35.

[73] Kimberly Leonard. 2015. Is Social Media Making Self-Harm Worse for Teens? *US News* (May 2015). https://www.usnews.com/news/articles/2015/05/29/is-social-media-making-self-harm-worse-for-teens

[74] Sam Levin. 2017. As Facebook blocks the names of trans users and drag queens, this burlesque performer is fighting back. *The Guardian* (June 2017). https://www.theguardian.com/world/2017/jun/29/facebook-real-name-trans-drag-queen-dottie-lux

[75] Karyne Levy. 2014. Facebook Apologizes For 'Real Name' Policy That Forced Drag Queens To Change Their Profiles. *Business Insider* (2014). https://www.businessinsider.com/facebook-apologizes-for-real-name-policy-2014-10

[76] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proc. CHI*.

[77] Shannon Liao. 2019. After the porn ban, Tumblr users have ditched the platform as promised. The Verge. https://www.theverge.com/2019/3/14/18266013/tumblr-porn-ban-lost-users-down-traffic.

[78] Dottie Lux and Little Miss Hot Mess. 2017. Facebook?s Hate Speech Policies Censor Marginalized Users. *Wired* (August 2017). https://www.wired.com/story/facebooks-hate-speech-policies-censor-marginalized-users/

[79] Amber Madison. 2015. When Social-Media Companies Censor Sex Education. https://www.theatlantic.com/health/archive/2015/03/when-social-media-censors-sex-education/385576/. The Atlantic.

[80] Aaron Mak. 2018. Facebook content moderation rules: How company decides what to remove. https://slate.com/technology/2018/04/facebook-content-moderation-rules-how-company-decides-what-to-remove.html. Slate.
[81] Ariadna Matamoros-Fernandez. 2017. Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society* 20, 6 (2017), 930–946.
[82] J Matias, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, and Charlie DeTar. 2015. Reporting, reviewing, and responding to harassment on Twitter. *Available at SSRN 2602018* (2015).
[83] Louise Matsakis and Paris Martineau. 2020. Coronavirus Disrupts Social Media's First Line of Defense. Wired. https://www.wired.com/story/coronavirus-social-media-automated-content-moderation/.
[84] Michael J Muller and Allison Druin. 2012. Participatory Design: The Third Space In Human–Computer Interaction. In *Human Computer Interaction Handbook*. CRC Press, 1125–1153.
[85] Deirdre K Mulligan, Daniel Kluttz, and Nitin Kohli. 2019. Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions. *Available at SSRN 3311894* (2019).
[86] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383.
[87] David G Novick and Stephen Sutton. 1997. What is Mixed-Initiative Interaction. In *Proc. AAAI Spring Symposium*, Vol. 2. 12.
[88] Frank A Pasquale. 2011. Restoring transparency to automated authority. *Journal on Telecommunications and High Technology Law,* (2011).
[89] Delia Paunescu. 2019. Inside Instagram?s nudity ban. (2019). https://www.vox.com/recode/2019/10/27/20932915/instagram-free-the-nipple-photo-facebook-nudity-ban-art-reset-podcast
[90] Abby Phillip. 2015. Online 'authenticity' and how Facebook's 'real name' policy hurts Native Americans. *The Washington Post* (February 2015). https://www.washingtonpost.com/news/morning-mix/wp/2015/02/10/online-authenticity-and-how-facebooks-real-name-policy-hurts-native-americans/
[91] Alisha Pradhan, Ben Jelen, Katie A Siek, Joel Chan, and Amanda Lazar. 2020. Understanding Older Adults' Participation in Design Workshops. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.
[92] Ondřej Procházka. 2019. Making Sense of Facebook's Content Moderation: A Posthumanist Perspective on Communicative Competence and Internet Memes. *Signs and Society* 7, 3 (2019), 362–397.
[93] Adam Przeworski, Susan Carol Stokes Stokes, Susan C Stokes, and Bernard Manin. 1999. *Democracy, accountability, and representation*. Vol. 2. Cambridge University Press.
[94] Anniek Bao Qian Tong and Flynn Murphy. 2020. Young Shopkeeper?s Suicide After WeChat Ban Highlights App?s Influence Over 1.2 Billion Lives. *Caixin Global* (August 2020). https://www.caixinglobal.com/2020-08-28/young-shopkeepers-suicide-after-wechat-ban-highlights-apps-influence-over-12-billion-lives-101598510.html
[95] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proc. CHI*. ACM.
[96] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proc. KDD*. ACM.
[97] Jane Ritchie, Jane Lewis, Carol McNaughton Nicholls, Rachel Ormston, et al. 2013. *Qualitative research practice: A guide for social science students and researchers*. Sage.
[98] Sarah Roberts. 2016. Commercial Content Moderation: Digital Laborers' Dirty Work. In *The Intersectional Internet: Race, Sex, Class and Culture Online*, S. U. Noble and B. Tynes (Eds.). Peter Lang Publishing.
[99] Sarah T Roberts. 2018. Digital detritus: 'Error' and the logic of opacity in social media content moderation. *First Monday* 23, 3 (2018).
[100] Adi Robertson. 2019. TikTok prevented disabled users? videos from showing up in feeds. *The Verge* (2019). https://www.theverge.com/2019/12/2/20991843/tiktok-bytedance-platform-disabled-autism-lgbt-fat-user-algorithm-reach-limit
[101] Aja Romano. 2018. Tumblr is banning adult content. It's about so much more than porn. *Vox* (2018).
[102] Elizabeth B-N Sanders and Pieter Jan Stappers. 2012. *Convivial Toolbox: Generative Research for the Front End of Design*. BIS Publishers.
[103] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* (2019).
[104] Claire Selvin. 2019. Betty Tompkins Ousted from Instagram for Posting Classic Painting of Penetration. https://www.artnews.com/art-news/news/betty-tompkins-expelled-from-instagram-12463/
[105] Limor Shifman. 2014. *Memes in digital culture*. MIT press.
[106] Edward Hance Shortliffe and Bruce G Buchanan. 1985. *Rule-based Expert Systems: the MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley Publishing Company.
[107] Ellen Silver. 2018. Hard Questions: Who Reviews Objectionable Content on Facebook? And Is the Company Doing Enough to Support Them? https://newsroom.fb.com/news/2018/07/hard-questions-content-reviewers/

[108] Jesper Simonsen and Toni Robertson. 2012. *Routledge International Handbook Of Participatory Design.* Routledge.
[109] Olivia Solon. 2017. Facebook is hiring moderators. But is the job too gruesome to handle? https://www.theguardian.com/technology/2017/may/04/facebook-content-moderators-ptsd-psychological-dangers
[110] Olivia Solon. 2017. Underpaid and overburdened: the life of a Facebook moderator. *The Guardian* (May 2017). https://www.theguardian.com/news/2017/may/25/facebook-moderator-underpaid-overburdened-extreme-content
[111] Olivia Solon. 2020. Facebook ignored racial bias research, employees say. *NBC News* (2020). https://www.nbcnews.com/tech/tech-news/facebook-management-ignored-internal-research-showing-racial-bias-current-former-n1234746
[112] Kaitlyn Tiffany. 2017. Twitter criticized for suspending popular LGBTQ academic @meakoopa. *The Verge* (June 2017). https://www.theverge.com/2017/6/13/15794296/twitter-suspended-meakoopa-anthony-oliveira-controversy
[113] Kaitlyn Tiffany. 2019. Tumblr's First Year Without Porn. *The Atlantic* (2019).
[114] Alexandra To, Hillary Carey, Geoff Kaufman, and Jessica Hammer. 2021. Reducing Uncertainty and Offering Comfort: Designing Technology for Coping with Interpersonal Racism. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.*
[115] Kristen Vaccaro, Karrie Karahalios, Deirdre Mulligan, Daniel Kluttz, and Tad Hirsch. 2019. Contestability in Algorithmic Decision Making. In *Proc. CSCW Workshops.* ACM.
[116] 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 *Proc. CSCW.* ACM.
[117] Maja Van der Velden, C Mörtberg, J Van den Hoven, PE Vermaas, and I Van de Poel. 2014. Participatory design and design for values. *Development* 11, 3 (2014), 215–236.
[118] Richmond Y. Wong, Deirdre K. Mulligan, Ellen Van Wyk, James Pierce, and John Chuang. 2017. Eliciting Values Reflections by Engaging Privacy Futures Using Design Workbooks. *Proc. ACM HCI* CSCW (2017). https://doi.org/10.1145/3134746
[119] Allison Woodruff. 2019. 10 things you should know about algorithmic fairness. *interactions* 26, 4 (2019), 47–51.
[120] Peter Wright and John McCarthy. 2008. Empathy and experience in HCI. In *Proc. CHI.*

Received January 2021; revised April 2021; accepted May 2021