The Online Identity Help Center: Designing and Developing a Content Moderation Policy Resource for Marginalized Social Media Users

SAMUEL MAYWORM, University of Michigan, USA
SHANNON LI, University of Michigan, USA
HIBBY THACH, University of Michigan, USA
DANIEL DELMONACO, University of Michigan, USA
CHRISTIAN PANEDA, University of Michigan, USA
ANDREA WEGNER, University of Michigan, USA
OLIVER L. HAIMSON, University of Michigan, USA

Marginalized social media users struggle to navigate inequitable content moderation they experience online. We developed the Online Identity Help Center (OIHC) to confront this challenge by providing information on social media users' rights, summarizing platforms' policies, and providing instructions to appeal moderation decisions. We discuss our findings from interviews (n = 24) and surveys (n = 75) which informed the OIHC's design, along with interviews about and usability tests of the site (n = 12). We found that the OIHC's resources made it easier for participants to understand platforms' policies and access appeal resources. Participants expressed increased willingness to read platforms' policies after reading the OIHC's summarized versions, but expressed mistrust of platforms after reading them. We discuss the study's implications, such as the benefits of providing summarized policies to encourage digital literacy, and how doing so may enable users to express skepticism of platforms' policies after reading them.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; **Collaborative and social computing**.

Additional Key Words and Phrases: content moderation, social media, digital literacy, marginalization, marginalized identity

ACM Reference Format:
Samuel Mayworm, Shannon Li, Hibby Thach, Daniel Delmonaco, Christian Paneda, Andrea Wegner, and Oliver L. Haimson. 2024. The Online Identity Help Center: Designing and Developing a Content Moderation Policy Resource for Marginalized Social Media Users. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 129 (April 2024), 30 pages. https://doi.org/10.1145/3637406

Authors' addresses: Samuel Mayworm, mayworms@umich.edu, University of Michigan, Ann Arbor, Michigan, USA; Shannon Li, lishann@umich.edu, University of Michigan, Ann Arbor, Michigan, USA; Hibby Thach, hibby@umich.edu, University of Michigan, Ann Arbor, Michigan, USA; Daniel Delmonaco, delmonac@umich.edu, University of Michigan, Ann Arbor, Michigan, USA; Christian Paneda, ccpaneda@umich.edu, University of Michigan, Ann Arbor, Michigan, USA; Andrea Wegner, amweg@umich.edu, University of Michigan, Ann Arbor, Michigan, USA; Oliver L. Haimson, haimson@umich.edu, University of Michigan, Ann Arbor, Michigan, USA.

ACM 2573-0142/2024/4-ART129
https://doi.org/10.1145/3637406

1 INTRODUCTION

Marginalized communities[1] depend on social media and digital technologies to meet their unique information and social needs, such as creating community support systems, exploring their identities, and finding identity-related information for their own well-being [16, 23, 33, 43, 53, 57, 61, 65, 68, 89]. Despite how marginalized communities uniquely rely on social media for their own well-being, they also experience (in general and as specific social groups) disproportionate rates of content moderation and removals on social media platforms, even in instances where their content does not violate platforms' community guidelines [23, 47]. Examples of these disproportionate removals include social media platforms incorrectly flagging images of transgender and nonbinary users' bodies as "not safe for work" [47], or the incorrect removals of Black users' content that openly discusses racial justice or the Black Lives Matter movement [46, 47]. The disproportionate moderation of marginalized users' content harms marginalized social media users by restricting their freedom of self-expression on platforms and their ability to safely use platforms to meet their general or identity-related needs [23, 47].

Marginalized users struggle with the excessive removals of their content, which is compounded by the difficulty social media users in general experience when navigating the text of platforms' policies in attempts to understand whether their content should have been removed at all [73, 76]. The significant majority of social media users are reluctant to read platforms' policies prior to using their services, typically citing policies' overwhelming length, the excessive amount of time necessary to read platforms' policies, and the opacity of platforms' policies, as factors that discourage user readership [73, 76]. Marginalized users are particularly likely to not trust social media platforms to enforce their own policies correctly while moderating marginalized users' content [16, 29, 63], often relying on personal or community-built folk theories to guide their social media behaviors instead of reading and trusting policies to accurately inform them of what they can post online [15, 16, 63]. As a result, marginalized users are often left in a position where they can neither trust social media platforms to moderate their content fairly nor to accurately inform them of what they can post to begin with.

We designed and developed the Online Identity Help Center (OIHC, www.oihc.org), an online content moderation and social media policy resource designed to highlight and center marginalized users' content moderation experiences, as a response to the challenges marginalized users face. We provide marginalized social media users with social media moderation-related resources and digital literacy resources that can improve their understanding of how to navigate platforms' policies and moderation practices. Such resources include summarized versions of platforms' policies, explanations of social media users' rights, and instructions to appeal moderation decisions across multiple different platforms. As a hub for online resources related to marginalization and content moderation, the OIHC cannot directly intervene in individual users' moderation experiences or appeal processes, nor can it provide services that require live support (such as 24/7 chat support or community forums requiring content moderation of their own). Instead, the goal of the OIHC is to address the inequities faced by marginalized social media users by providing moderation-related resources that make it easier for users to engage with platforms' policies, to understand their rights on social media platforms, and to navigate their experiences with social media content moderation and removals.

We address three research questions in this work:

[1]Marginalized communities are defined in our paper as communities that experience systemic exclusion, discrimination, and social inequities based on factors such as, but not limited to, race, ethnicity, religion, gender identity, sexuality, or disability.

- **RQ1:** What resources and topics would marginalized users like to have included on a comprehensive web resource that centers marginalized social media users' online rights and content moderation experiences?
- **RQ2:** How might we develop a comprehensive web resource that provides accurate information about social media policies to help marginalized users navigate their online rights and content moderation experiences?
- **RQ3:** How might marginalized users use and perceive the resources included on this comprehensive web resource?

This paper addresses RQ1 by describing the findings of our qualitative analysis of (n = 24) interviews with marginalized social media users who have previously experienced content removals on social media platforms, whose input was utilized while designing and developing the OIHC. We then address RQ2 by describing the design and development process of the OIHC, including our survey of n = 75 social media users to determine how to prioritize content for the OIHC, and our competitive analysis of existing official social media policy resources and their common shortcomings. We then address RQ3 by describing the findings from qualitative analysis of (n = 12) OIHC user tests and interviews, where marginalized social media users gave feedback on a prototype version of the OIHC and its contents. This paper contributes a description of the design and development of an online digital literacy resource for marginalized populations to learn more about social media content moderation. We contribute findings on how marginalized social media users use such a resource; this includes unanticipated interactions with the resource, such as choosing to read platforms' policies in full, critically reflecting on platforms' policies, and expressing skepticism toward platforms' policies. We then contribute a discussion on how digital literacy resources can serve as starting points for social media users to learn more about social media platforms' policies, along with a discussion of marginalized users' skepticism of platforms' policies after reading them.

2 RELATED WORK

2.1 Social Media Content Moderation and Guidelines

Content moderation is a necessary part of the work platforms must do to protect their users from abusive user behavior and to remove illegal or potentially harmful content [39, 40, 98]. Grimmelman defines social media content moderation as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse" [45]. Content moderation can take the form of "top-down" moderation enacted by platform administrators and their algorithmic moderation tools [41, 44, 60, 71] or "bottom-up" moderation enacted by the platform's own users, often voluntarily and unpaid [6]; some platforms employ a combination of both top-down and bottom-up content moderation structures [92]. Several major social media platforms prominently featured throughout the OIHC (Facebook, Instagram, Twitter, and TikTok) employ "top-down" moderation on their platform, while only one (Reddit) employs a "hybrid" moderation model, utilizing a combination of subreddit-specific volunteer moderators, sitewide administrators, and algorithmic moderation tools [82, 83]. US-based social media platforms often outsource human moderation through third-party companies located outside of the United States; these human moderators are then responsible for assessing and removing content posted by users and communities across the entire world [98]. Content moderation is an integral part of the work on the OIHC and this study, as the methods employed by platforms to moderate content affects the visibility and rights of individual social media users and user groups [40].

Though social media platforms generally employ some form of content moderation while requiring users to adhere to their platforms' policies (commonly referred to as "community guidelines"),

the significant majority of social media users do not fully read platforms' policies and rules before using their services [2, 73, 74, 76]. Social media users typically perceive platforms' policies as being too long, confusing, or time-consuming to read, or "irrelevant" to their overall experiences as social media users [73, 76, 85]. The difficulty of reading and understanding platforms' policies can be particularly problematic for users who wish to appeal the removal of their social media content, as the appeal instructions may be so difficult to understand that users feel discouraged from appealing content removals at all [94]. Users may also experience frustration regarding platforms' algorithmic moderation tools, which may struggle to correctly moderate users' content, including "gray area" content that is not clearly addressed by platforms' policies [47]; this reflects a broader trend of automated systems struggling to moderate behaviors that may require human judgment and intervention [59, 79].

Social media users often respond to challenging platform policies and appeal processes by relying on folk theories (defined by French and Hancock as a "person's intuitive, causal explanation about a system") about social media policies and content moderation to guide their behavior on social media instead of reading platforms' rules [15, 17, 27, 34, 70]. Folk theories are typically created by ordinary users to help them understand how a system works in practice and to inform their behavior and decision-making related to that system [36]. However, the difficulty of understanding platforms' terms of service as they are written can pose threats to social media users who may be unable to fully understand their platforms' terms of service [31, 58]. Luger argues that the poor readability of platforms' terms of service is "fundamentally an issue of inclusion and accessibility," arguing that the overly inaccessible language and length of these policies can make them functionally unreadable for many users [58].

Past literature has evaluated the readability of online policies presented to users; Fabian et al. found that readable online policies can simplify online decision-making for users, resulting in improved online user experiences [28]. Past literature has also explored different formats for presenting policies to users in more readable formats; Kelley et al. developed a "Nutrition Label for Privacy," a visually-friendly format resembling a food nutrition label, as a way to present privacy policies to users in a readable, easily digested way [56]. Our study builds off of similar information presentation goals, with the goal of designing the OIHC to present short, summarized versions of social media platforms' policies that emphasize clarity and readability for social media users, helping social media users more easily digest platforms' policies even when platforms themselves do not present their policies in "readable" formats.

2.2 Content Moderation and Marginalized Users

Content moderation practices occur both algorithmically and by human content moderators, both of which can detrimentally impact social media users and their online experiences [37, 62]. Marginalized social media users are disproportionately likely to experience unique harms and challenges related to content moderation on social media platforms, such as the disproportionate removals of their content or accounts, or the suppression of identity-related speech [5, 19, 23, 30, 32, 42, 47, 86]. Certain groups of users, such as Black social media users, LGBTQIA+ users (particularly transgender users), and women (particularly women of color) are especially likely to experience content suppression and removal [14, 19, 21, 23, 30, 38, 47, 62, 86, 100]. Marginalized users may experience feeling pressured to behave or present themselves in specific ways due to the types of bodies and behaviors content moderation practices emphasize as "normal" [30]. Marginalized users' content that includes images of their bodies may also face particular scrutiny and moderation in attempts to police sexuality via nudity and sexual content regulations and bans [88]; Gerrard and Thornham situated this aspect of content moderation as platforms' "sexist assemblages" which can perpetuate harmful gender roles [38]. Marginalized users' content that is removed often falls

into content moderation "gray areas," in which both algorithmic and human moderator methods cannot easily categorize content as "right" or "wrong"; Haimson et al. argued that platforms should embrace these gray areas in their moderation practices rather than forcing content to fit into strict permissible or removable categories [47].

Marginalized users can find disproportionate moderation and content removals frustrating and painful, particularly if they perceive their removed content as not violating the platforms' guidelines [47, 86] or as incorrectly removed by platforms' algorithmic moderation tools [9]. Instances of user disagreements with platforms' moderation decisions highlight the mismatch between user experiences on platforms and content moderation policies in the form of contested platform governance [90]. The disproportionate removals of marginalized users' content can prevent marginalized users from using social media as freely as non-marginalized users, and may result in marginalized users not trusting social media platforms to moderate their content correctly or to protect them from harm [54, 87]. These disproportionate removals can also have major human rights implications, such as in the case of Palestinian Facebook and Instagram users, whose posts documenting airstrikes in the Gaza Strip were disproportionately removed from both platforms [3]. The OIHC was developed to address the challenges faced by marginalized social media users by centering marginalized users and their unique challenges regarding social media content moderation, with the goal of creating an online social media policy resource that can help marginalized social media users navigate the disproportionate content moderation and removal experiences that they face. This study builds on the literature by developing insight into the moderation-related informational resources that marginalized users on the OIHC found helpful, along with exploring marginalized users' engagement with (and perceptions of) social media platforms' community guidelines and moderation practices.

2.3 The Digital Divide and Online Digital Literacy Resources

The "digital divide" is an ambiguous term with various definitions, but broadly refers to the gap between those who do and do not have access to information technology [96]. Multiple divides emerge as technology access and usage increases, including differences in technology usage and digital literacy skills between marginalized and privileged groups [69]. Digital literacy is broadly described by Reddy et al. as the "necessary skills and competencies to perform tasks and solve problems in digital environments" [84]. Young marginalized people, such as those who are Indigenous, culturally and linguistically diverse, or living with a disability, are less likely to use information communication technologies, and more likely to experience lower digital literacy levels, compared to other youths [4, 81]. Online agency diminishes without adequate digital literacy skills; users without these skills are less likely to reap the full benefits of what technology has to offer [8, 18, 24]. Barriers to digital literacy for marginalized individuals can make them more vulnerable to hate and discrimination online, as they may not possess the resources and preparedness necessary to safely respond to these interactions [7]. Thus, digital divide discussions widen in scope beyond access, becoming not only a technological problem, but a social justice problem as well [95].

Online digital literacy resources are often centered around mainstream identities, failing to account for the background and experiences of marginalized communities. This failure is prevalent within education systems that rely on digital technologies for teaching [75]. Decentering whiteness within education pedagogies to center around marginalized communities contributes towards a safe digital learning environment for communities of color to share their experiences and incur active engagement [49]. The failure to bring narratives from marginalized communities further reify unequal power structures, continuously disempowering marginalized people from becoming digitally literate [75]. Hourcade explains that the challenges of educating marginalized youth on digital skills is a result of the lack of digital infrastructure and qualified teachers within their

communities, exacerbated by difficulties to effectively engage marginalized youth to learn and find value in digital literacy skills [50].

Establishing adequate access to digital devices and technologies is a significant first step, but only addresses some digital literacy and online safety concerns faced by marginalized users. For example, Epstein and Quinn emphasized that socioeconomic inequality contributes towards and worsens online privacy marginalization that puts vulnerable groups who are less informed on online privacy protection at great risk of privacy loss to be most impacted by systematic disparities in privacy literacy [26, 35]. Existing technologies for digital literacy may also propel already-powerful groups forward without a comprehensive framework guiding design and development [64]. Design principles of online digital literary tools have implications on marginalized youth's technology use by leaving out important considerations that hinder user experience [50]. Closing the gap between what is currently understood about marginalized people's digital literacy skills and the reality of their experience provides a valuable opportunity to recenter digital literacy resources for marginalized people [49]. In navigating the digital space, existing tools for online digital literacy have been created to address marginalized people's resource needs. The Santa Clara Principles, co-written by organizations such as the Electronic Frontier Foundation and the Center for Democracy and Technology, provides a published set of general guidelines outlining standards for users' free speech rights and moderation transparency on social media platforms; social media users may read through these principles to better familiarize themselves with their rights on social media platforms [25]. Some social media platforms also provide tools to help social media users appeal the removals of their content; for example, Facebook allows users to appeal their content removals to the Oversight Board [66], while Twitter offers an appeal submission form for users to appeal account suspensions [93].

Several digital literacy resources help provide guidance for marginalized users whose income and contributions toward society have been impacted by content removals. The Syrian Archive supports activists by archiving records of human rights violations removed by social media platforms that are vital for human rights research, working with social media companies to redefine their content moderation policies and reinstate many records, and providing contact support for content removals [91]. The Internet Freedom Foundation provides information and support for Indian social media users who experience political censorship on social media [52]. The Don't Delete Art campaign launched an online gallery of art by LGBTQ+ artists who were algorithmically removed by social media platforms' moderation tools, an appeal process absent of concrete steps, and unclear community guidelines as a statement to demand for social media platforms create guidelines that democratizes art to be shared freely online [20]. The campaign provides artists with guidance for posting work, resources for navigating the appeal process for multiple social media platforms, information regarding community guidelines, and artworks that have been removed as an opportunity to have them displayed.

Online digital literacy tools have also been created to counteract privacy concerns; these tools can be used by marginalized users to equip themselves with knowledge and awareness of privacy protections to reduce social marginalization [26]. Consumer Reports Security Planner's platform equips people with expert-reviewed personalized guides for staying safe online, including topics such as safeguarding online accounts and protecting mobile data [11]. Data cooperatives are another kind of collective data resource built to enable greater autonomy over personal data and to restore trust between users and the organizations who utilize users' personal data, which can address social media users' distrust of social media platforms regarding data privacy and use [67]. Resources that provide educational materials for communities and individuals on security practices and secure sensitive information online can also be useful for marginalized social media users; examples

include the Digital Security Helpline [1], which provides immediate emergency assistance and 24/7 real time services of any digital risks and concerns.

The digital literacy resources discussed above are often not specific to content moderation or marginalized social media users' moderation-related needs, or are only built to address specific moderation-related contexts; as a result, many of these resources will not meet the needs of marginalized social media users broadly. We developed the OIHC to address this gap by presenting a range of moderation and digital literacy-related content, with the goal of providing relevant, helpful moderation-related information to a broad range of marginalized social media user communities.

In what follows, we describe the three phases of our research: an initial interview study to determine what to include on the site, the design and development of the OIHC site, and then our evaluation study to determine potential users' reactions to the site. In each of these sections, we first describe our methods and then our results.

3 INTERVIEW STUDY

3.1 Interview Study Methods

3.1.1 Data Collection. To answer our research questions, we first conducted a structured interview study with n = 24 participants. A goal of the interview study was to identify topics related to content moderation and social media policies that are relevant and important to marginalized social media users; this information was later used to determine what kinds of content to feature on the OIHC. This interview study was reviewed and deemed exempt from oversight by our university's Institutional Review Board (IRB). We contacted 26 participants to schedule interviews, with 24 participating in their scheduled interview. Recruiting these participants happened in three ways: (1) participants from our prior survey study on social media content moderation and removals who indicated interest in participating in a follow-up interview were invited to participate (n = 6); (2) participants who filled out a screening survey we promoted via our social media accounts on Twitter were invited to participate (n = 6); (3) participants who were a part of a research recruiting service and matched our internal screening survey process were invited to participate (n = 12). We screened for adult social media users from marginalized groups (i.e., racial/ethnic minorities, gender and sexual minorities) who stated that their content or accounts were removed from a social media platform in the past year for reasons they disagreed with. To ensure that our sample was diverse and included people from marginalized groups, the screening surveys asked participants for their age, gender, race/ethnicity, LGBTQIA+ status, and whether they specifically are transgender, nonbinary, or both. We used open text in our recruiting surveys for gender and sexuality in order to respect and capture the diversity of terminology and self-identification within the queer and trans population.

Of the 24 interviews, 23 were conducted remotely over Zoom and recorded for audio transcription, while one interview with a deaf participant was conducted through text over email. Interviews lasted for an average of 51.65 minutes (sd = 11.06 minutes, range: 38 - 84 minutes). During interviews, participants completed the informed consent process and multiple interviewers were typically present. The interview presented a series of questions about participants' content or account removals, asking them to describe the removals, whether they thought the removals were incorrect, and how the removal experience may have related to their marginalized identities. They were asked further questions about their perceptions of content moderation and community guidelines on the platforms they use, and how content moderation and community guidelines on platforms could be improved for marginalized users. Participants received $30 for participating in the interview study. The leftmost column of Table 1 details interview participant demographics, showing age,

gender, if participants self-identified as LGBTQ+, and participants' self-identified race/ethnicity. The structured interview study participants are referred to as P1 through P24 to differentiate from the user test interview participants, who are referred to as P25 through P36.

3.1.2 Data Analysis. The three authors who completed the interviews conducted open coding [12] using Atlas.ti. The first three authors first coded the same transcript to begin developing a codebook; after agreeing on all codes and their meanings, they then coded the remaining interviews separately. New codes were identified during the subsequent coding and were discussed by the research team; those that the team agreed on after discussion were added to the codebook. We used axial coding [12] to organize codes into themes. Themes that emerged in our analysis include: community guidelines, general content moderation practices, algorithmic content moderation, unequal moderation of marginalized users and their content, abusive user behaviors and moderation, content visibility and suppression, the relationship between content moderation and social media platforms/corporations, and user behavior changes in response to moderation. After transcribing interviews, we used FigJam to conduct affinity diagramming to organize insights across large amounts of qualitative data [48] and prioritize usability issues to be addressed [57]. Authors consolidated themes until the most salient educational topics that encapsulated users' needs were revealed. These themes included:

(1) Social media users' rights
(2) Privacy and data collection on social media sites
(3) Social media sites' policies and guidelines and their differences
(4) How to contact social media sites for help after a content/account removal
(5) Social media algorithms and how they work
(6) How to file an appeal for a content/account removal
(7) What shadowbanning is, and how to avoid it
(8) Ways to connect with others to engage in activism related to content/account removal

3.2 Interview Study Results: What to Include on the Site

During the initial round of interviews, we asked participants about the content moderation issues they encountered on social media platforms, and to envision and describe an informational resource that they would personally use that might address their concerns; while we did not explicitly present a list of potential formats for the resource to the participants, the majority of participants described their envisioned resource as a website. As described in section 3.1.2, interview participants identified eight major content areas that informed the OIHC's development and design. During the later survey study described in section 4.1.1, participants rated the eight major content areas that they most desired to have featured on the OIHC. Out of the eight content areas, the top four highly rated content areas were:

(1) User Rights
(2) Privacy and Data Collection
(3) Social Media Site Policies and Guidelines
(4) Contacting Social Media Sites.

In this section, we present participants' views on each of these four priority areas. We limit this section to the top four highly ranked content areas as, due to the scope and timeline of the project, we narrowed the OIHC's focus to those four topics instead of all eight (a process described in greater detail in section 4.1.1) [2]

[2]Though User Rights could be interpreted as a sub-topic of Social Media Site Policies and Guidelines, interview study participants described these topics as similar but separate concerns. We designed User Rights and Social Media Site Policies

3.2.1 User Rights. Participants expressed the desire to educate themselves on how their rights (such as freedom of speech or self-expression) apply to social media use, while also asking platforms to describe user rights with clear and concise language. P7 noted that it could be difficult to understand whether *"laws are actually there for online technologies to protect people's [speech],"* or *"[whether] 'freedom of speech' extends to online technologies"* in the first place. P15 shared similar sentiments, stating that she *"would very much use"* an informational web resource on mainstream social media platforms' policies, because *"in this day and age, social media is so powerful... so it makes sense to know what [your] rights are."* P15 then compared her experiences posting on platforms to in-person interactions with police, saying *"before you get arrested, they read you your rights, and although that is something you should still know, on Instagram, they take down your things, but they didn't let you know your rights before they took it [down]."* Participants described user rights conflicts specifically related to marginalization that they witnessed on platforms; P1, P6, and P20 described witnessing the disproportionate removal of posts related to Indigenous activism and the #BlackLivesMatter movement across multiple platforms, while P3 described witnessing the inordinate removal of transgender and nonbinary users' selfies and activist content on Instagram. The participants' examples highlight the particular difficulty for marginalized users, subjected to the excessive removal of their identity-related speech and self-expression, to know how their rights will apply to social media use. Participants' sentiments echo the desire for platforms to keep users in the know about their rights, and expresses the distress users may experience when their content is removed in a way they perceive as violating their rights – a distress disproportionately experienced by marginalized users.

Users also expressed the value of simple summaries and definitions for platform policies and their associated jargon; P14 spoke on this, stating that *"the most important thing is to make [policies] 'bite-sized' [and] very easy to consume,"* compared to the *"words that you see a lot in legal papers and legal guidelines."* P14 also suggested that users may be more likely to read summarized, less overwhelming versions of policies than their full-text counterparts, stating that *"I do have the power to go through and read the [full] guidelines, but... I just don't care enough. I want to, but I don't. But presenting them in a very easy, digestible format would probably spread more awareness [of platforms' policies]."* Users also suggested presenting this information in a list format; P8 suggested a user-friendly list format *"like LinkTree"* providing users with a checklist *"of things can be posted [on social media platforms]"* or *"for appealing a [moderation] decision."* With our interviewees' suggestions in mind, we made sure to include links and information regarding user rights on various social media platforms, easily accessible to website visitors. We also use clear and concise language when discussing user rights, attempting to make the website's language accessible and not overwhelming to a general audience (see Figure 1).

3.2.2 Data Privacy and Collection. Another important topic participants spoke on was the issue of privacy and data collection on social media platforms. A fear amongst participants was what happens to their information while using a platform. P21 wondered *"how much Instagram actually looks at what you post [or] pays attention to it, stores it, [etc.],"* while sharing his perception of how social media platforms treat users' personal data: *"I know that's how social media functions, [it] sells*

and Guidelines as two separate features of the OIHC to reflect participants' perspectives, with the former focused on informing users of how their basic rights apply to social media use (e.g. how "freedom of speech" is interpreted on social media platforms), and the latter focused on informing users of individual websites' policies (e.g. whether individual platforms allow certain kinds of "nude" or "graphic" content). We also acknowledge that some important themes from our affinity diagramming were determined to be outside the OIHC's scope, such as social media algorithms and shadowbanning. Additionally, the OIHC excludes some content moderation topics that participants occasionally mentioned during the interviews but were not among their most salient educational topics, such as shadow bans, the sale of users' data, or marginalized users reporting other users' abusive content.

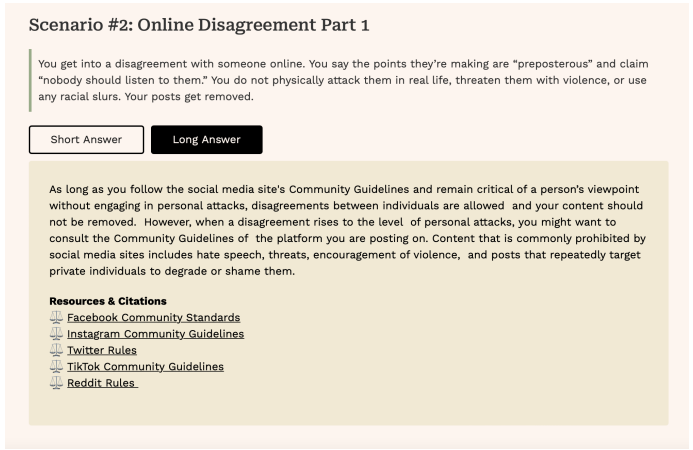

Fig. 1. A hypothetical moderation scenario and explanation featured on the OIHC Social Media Rights page.

your information." This informed us that participants want clear knowledge on who or what has access to user information when agreeing to use an app or platform, and that there is a popular belief that social media profits by selling user information. Participants discussed a wide range of online privacy and safety concerns affecting marginalized social media users specifically, such as the surveillance of marginalized activists' social media activity or the personal safety threats marginalized users experienced during everyday social media use. P18 shared his concerns related to online surveillance of marginalized activists, stating that *"social media platforms accumulating all that data,"* could *"later use [user data] against members of marginalized communities, either in the US or in other parts of the world"* by creating and weaponizing *"profiles... of marginalized activist community members"*. P14 voiced concerns related to online privacy and safety during everyday social media use; experiencing misogynistic harassment from strangers on Instagram resulted in P14 feeling *"insecure"* posting about their experiences with misogyny on the platform, stating that Instagram's *"lack of privacy"* could expose them to further harassment. So, not only should platforms be clear about how they use user information, they should also attempt to minimize harms and risks as much as possible, as the most marginalized may be affected more greatly. These concerns are reflected in our approach to designing the OIHC, as we include resources about how platforms use user information and provide guidance for how people can protect their data and online privacy (see Figure 2).

3.2.3 Social Media Site Policies and Guidelines. Participants found it important that the OIHC not only describe their general rights as social media users (described in the User Rights section), but also detail various social media platforms' current policies and guidelines. P21 noted that it can *"be helpful for people to understand what not to post and... why certain things have been taken down."* P21's comments reflect users' need for platform policies that clearly guide what they can and cannot post on platforms, which P21 found lacking in the social media platforms that they use. P12 spoke on this, stating that she *"would like to learn more about the specific guidelines that each social media site enforces, as well as examples for each guideline of what is considered right and what is considered wrong."* Relatedly, P10 said, *"I think that the comparative details about how the rules and guidelines have been changing would help people to understand in what direction exactly that social media companies are going."* This means laying out clearly how social media platforms enforce their guidelines, while also keeping a history of how these guidelines and their

Protect your account with built-in features

- Use more complex and different passwords than you have used in the past
- Use two-factor authentication when possible

Resources

Protecting Your Internet Accounts Keeps Getting Easier

Avoid logging into social media accounts from public devices, if possible

- Check if your connection is secure. Having https in the web address means it's secure, but doesn't necessarily mean it's legitamate.
- Public devices that are not your own is susceptible to data breaches.
- Avoid logging into any of your social media accounts when possible

Resources

How to Protect Your Digital Privacy

Fig. 2. Examples of data privacy guidance and resources featured on the OIHC Privacy and Data Collection page.

enforcement have changed. Some participants described how clear explanations of social media site policy and enforcement could specifically benefit marginalized social media users. P6, who spoke about guidelines for showcasing trans bodies, stated that they *"would love to see [stuff] about trans people posting their bodies, both pre-op and post-op, because obviously those guidelines are going to be different,"* reflecting transgender and nonbinary social media users' experiences with platforms disproportionately and inconsistently moderating content featuring their bodies, even when their content does not violate platforms' guidelines [47, 78]. P12 expressed frustration with *"[platform] guidelines that allude to discrimination against certain groups... but don't give very good descriptions of what each guideline is,"* making it unclear to what extent marginalized users are protected from discriminatory treatment on different platforms. P9 mentioned how it would be helpful to include *"guidance on... what words to avoid"* or how to phrase social media posts to avoid disproportionate content moderation; P9's example reflects marginalized social media users' use of alternative phrases (such as such as "yt" instead of white or "seggs" instead of sex [10, 80]) to avoid disproportionate algorithmic moderation while discussing identity-related topics online. Overall, users suggested they wanted a clear history of social media platforms' policies and guidelines and specifics related to particular identity groups, as well as how to avoid content moderation through alternative language. To help with this, the OIHC details policies and guidelines for five specific social media platforms, and provides details regarding how content moderation may impact marginalized users (see Figure 3).

3.2.4 Contacting Social Media Sites. Many participants spoke about social media appeals systems' vagueness, or the seemingly impossible task of reaching out to speak to someone about their account ban or content removal. P5 said they would appreciate OIHC telling users *"how to find communication pathways to dispute moderation or to contact moderators or owners."* P12 agreed, also

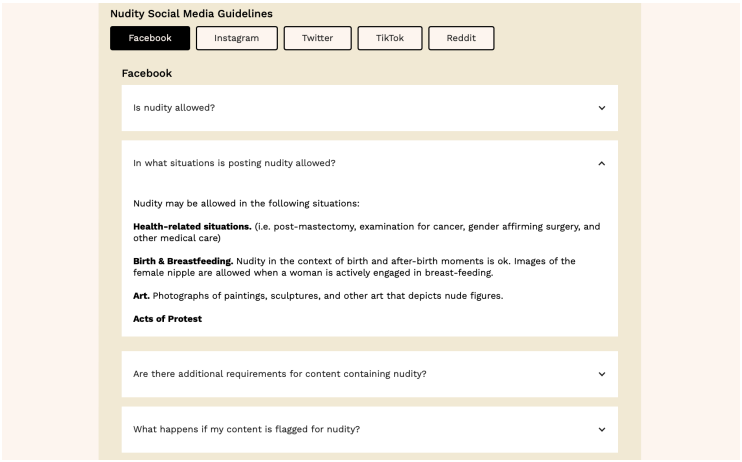

Fig. 3. Examples of exceptions to Facebook's nudity policies featured on the OIHC Social Media Guidelines page.

saying they *"would like to see contact information for specific people you can reach out to if you have questions [about moderation], whether that's an email or phone number. I think that's important for people to have."* Users like P13 described the frustrations of appealing content moderation decisions, mentioning a platform that didn't *"have a contact number or a live operator. I've never seen anything like that. It's always just an email."* Similarly, P22 had to *"[do] some Google searches [to] find out how to get a hold of service representatives,"* stating that the contact information was *"difficult to find"* even while using a search engine. As detailed by our participants, content moderation appeals processes are unclear and labyrinthine at times. Thus, we designed the OIHC to assist users in contacting social media platforms to appeal account bans or content removal. The site lists links to platforms' official policy pages, online appeal forms, and information about how to contact social media platforms, consolidating this information into one place for users to easily contact social media platforms regarding their content moderation decisions (see Figure 4).

4 DESIGN AND DEVELOPMENT OF THE ONLINE IDENTITY HELP CENTER

4.1 Requirements Gathering

4.1.1 Surveys. As previously mentioned in section 3.2, after completing the interview study, we conducted a survey (n = 75) asking participants from marginalized groups (i.e. racial/ethnic minorities, gender and sexual minorities) to rank the eight different educational topics identified during the interview analysis in order of importance on a scale of 1 (most important) to 8 (least important). Our goal was to use the survey findings to help prioritize relevant resource topics for the OIHC. We conducted this survey through online survey recruitment platform Prolific, receiving 75 eligible responses within the 6 days that the survey was live. The research team determined that the pool of 75 participants provided an appropriate representative sample of the OIHC's target population while also fitting the timeline and scope of the project. Of the survey participants, 25.33% responded that they had content taken down from a social media site for reasons they disagreed with within the last year. We compiled the average rank among each topic to determine which were the most important. We focused the OIHC on the top four topics: social media user rights, privacy and data collection, social media sites policies and guidelines, and how to contact social media sites for help after content or account removal. The center column of Table 1 details

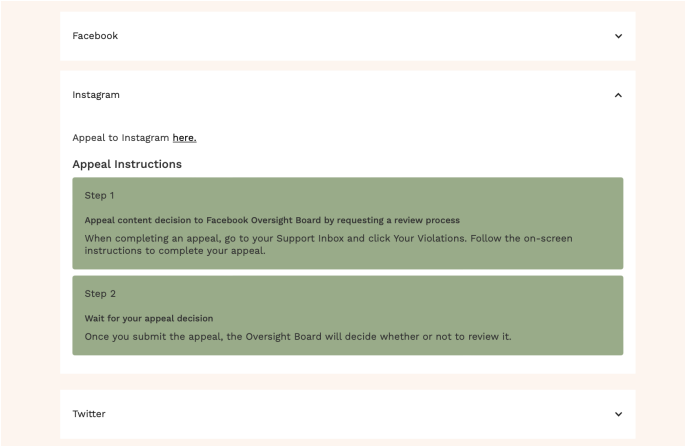

Fig. 4. Instructions to file an Instagram content removal appeal featured on the OIHC Social Media Appeals page.

survey participant demographics, showing age, gender, if participants self-identified as LGBTQ+, and participants' self-identified race/ethnicity.

4.1.2 Competitive Analysis. We performed a competitive analysis of social media policy resources published by various social media platforms, including Facebook, Twitter, Tumblr, Instagram, TikTok, Google, Reddit, Youtube, and Linktree. We compared these tools across the following criterias: information layout, searchability of information, topics, navigation of resources, social media contacts, appeal process, information for marginalized people. We conducted closed coding [13] during the analysis to better understand how different policy resources presented and communicated information compared to one another, such as differences in the resources' navigability and overall visual layouts. We also sought to identify common shortcomings of existing social media policy resources while conducting our analysis to avoid repeating those shortcomings while designing the OIHC. The analysis revealed social media platforms' shortcomings while explaining their policies, along with barriers to finding specific policy-related information on platforms' policy resources.

4.2 Design and Development

The process of deciding what content to include on the OIHC is expanded on in the Results; the following is a brief description of the UX design and development of the OIHC. We created a journey map, a process of visualizing user's needs and experiences over time across their interactions with a system [51], to understand participants' pain points around content moderation and social media. We identified potential design interventions to address these pain points on the OIHC. We developed the OIHC's visual and brand design based on results from the competitive analysis, and ensured that the OIHC is inclusive of and appealing to marginalized user communities; we also maintained consistent and clean styling throughout the OIHC while featuring visible logos and imagery indicating its affiliation with a major university, ensuring the final product appears to be a legitimate resource. The OIHC's information architecture was collaboratively developed by the design team on FigJam; the information architecture was organized to emphasize the information prioritized by the study participants. We followed an iterative design process [99] to make design decisions for the OIHC based on requirements gathering results; the design team used results from the competitive analysis and user interviews to implement design patterns that aligned

with the target population. Design choices prioritized providing a centralized collection of advice about content moderation for marginalized people and reflecting the needs of diverse communities affected by wrongful content removals or account bans. See Figures 1, 2, 3, and 4 for examples of the site design. We created the OIHC as an educational resource website because this was the format most requested by participants. We employed mobile first design so that the site would be accessible to those without computers. We considered including additional affordances suggested during the interview study, such as live chat support and user forums, but determined that these were beyond the scope of the OIHC's capabilities because they would require constant availability and active content moderation.

5 EVALUATION STUDY

5.1 Evaluation Study Methods

5.1.1 Data Collection: User Tests and Interviews. After designing the OIHC, we conducted two rounds of usability tests with a total of n = 12 participants (n = 5 participants in the first round and n = 7 participants in the second) to learn more about how users would use the OIHC, and about the site's effectiveness, as well as to inform design iterations. This research was reviewed and deemed exempt from oversight by our university's Institutional Review Board (IRB). We recruited participants by contacting participants from prior study stages who indicated they may be open to participating in a follow-up interview (n = 16) and by posting about the study on Twitter and Instagram. Interested participants filled out a screening survey to ensure a diverse group of participants from marginalized groups. Of the prior study participants, n = 5 agreed to participate in the first round of usability tests. We recruited second-round participants through online recruitment service User Interviews; n = 7 eligible User Interviews respondents participated in the second-round of usability tests. All interviews/usability tests were conducted remotely over Zoom and recorded for audio transcription. Participants viewed a link to the OIHC prototype on Figma and were presented with a series of tasks to accomplish using the prototype. They were asked to communicate their thought processes out loud and were asked follow up questions about their experiences. The second round participants were asked additional questions about their experiences on social media as marginalized social media users, including their experiences with content removals on social media platforms. Participants received $30 for participating in the interviews/usability tests. The rightmost column of Table 1 details user test participant demographics, showing age, gender, if participants self-identified as LGBTQ+, and participants' self-identified race/ethnicity. The user test interview participants are referred to as P25 through P36 to differentiate from the initial interview study participants.

5.1.2 Data Analysis. We collaboratively analyzed the interviews/usability test data. We recorded notes from the usability tests using Figjam. The qualitative data analysis followed the same process described in section 3.1.2, drawing from Corbin and Strauss's qualitative analysis methods [12]. This analysis resulted in the several themes which we discuss in Results.

5.2 Evaluation Study Results: Engagement with Social Media Platforms and Policies

5.2.1 Participants Critically Reflected on Platform Policy Resources Linked in Social Media Guidelines Page. The Social Media Site Policies and Guidelines page includes embedded links to social media platforms' official community guidelines, including Facebook's Community Standards page for Adult Nudity. Interview study participants were asked to navigate the OIHC to locate and read a summary of Facebook's official guidelines on artistic nudity. After finding and reading the summarized policy, several participants chose to open the embedded link to Facebook's Community

Table 1. Participant Demographics

	Initial interviews (n = 24)	**Survey** (n = 75)	**Evaluation study** (n = 12)
Gender			
Woman	11 (45.8%)	34 (45.3%)	6 (50.0%)
Nonbinary	7 (29.2%)	13 (17.3%)	4 (33.3%)
Man	6 (25.0%)	29 (38.7%)	2 (16.7%)
Race/Ethnicity			
Asian	10 (41.7%)	9 (9.9%)	4 (33.3%)
Latinx/e	6 (25.0%)	10 (11.0%)	2 (16.7%)
Black or African American	5 (20.8%)	20 (22.0%)	6 (50.0%)
White	2 (8.3%)	42 (46.2%)	0 (0.0%)
Middle Eastern	1 (4.2%)	2 (2.2%)	0 (0.0%)
Native Hawaiian or Pacific Islander	1 (4.2%)	0 (0.0%)	0 (0.0%)
American Indian	0 (0.0%)	4 (4.4%)	0 (0.0%)
Alaska Native	0 (0.0%)	4 (4.4%)	0 (0.0%)
LGBTQ+			
Yes	15 (62.5%)	57 (76.0%)	7 (58.3%)
No	8 (33.3%)	15 (20.0%)	5 (41.7%)
Did Not Disclose	1 (4.2%)	3 (4.0%)	0 (0.0%)
Age			
18-24	10 (41.7%)	23 (30.7%)	3 (25.0%)
25-34	10 (41.7%)	34 (45.3%)	8 (66.7%)
35-44	4 (16.7%)	11 (14.7%)	0 (0.0%)
45-54	0 (0.0%)	7 (9.3%)	1 (8.3%)

Participants could choose multiple gender and race/ethnicity options, so percentages add up to greater than 100%.

Standards page to explore their nudity policies in greater detail. P6 stated that the OIHC provided *"all the information [I] need on the [summarized policy] page"* to navigate Facebook's nudity guidelines, while adding that the embedded link to Facebook's full guideline *"is also nice to have... just for my own further reading."* P26 stated that users *"who really want to deeply engage with the subject would want to 'read more' on an 'official' [platform policy] website."* P35 agreed, though he would *"have to actually read the guidelines themselves"* to understand the policies and their nuances in greater detail.

The participants who read Facebook's full artistic nudity guidelines critiqued the perceived ambiguity of the policies; P34 stated that the policies were *"a little gray"* and that she *"still has questions"* after reading them, while P29 stated that *"Facebook's Terms of Service are ambiguous,"* making it difficult to *"figure out what is actually allowed to be posted."* P36 agreed, stating that it *"felt odd"* that Facebook would allow or forbid different kinds of nude content without explaining how their moderation system would distinguish between exempt and non-exempt nude imagery. She connected her lack of trust that Facebook would enforce their policies properly with her frustrations with social media content moderation as a whole:

> *"It's not just Facebook... I feel like rules on many social media [platforms] are really vague. And I think that's done purposely, because when things go wrong, that vagueness excuses [platforms] from culpability. It's just odd to me because... I feel like, with nudity in protest or art, or just most nudity in general, I would say about 75% to 80% is probably okay, not too bad. I've seen some violent things on Facebook, and like... how is that okay, but then a naked body is wrong? I don't know. It's just odd to me."*

Participants like P36 not only found and read Facebook's full policy on nudity, but they critically engaged with the policy and critiqued platform rules that they found "ambiguous" or difficult to enforce in practice; in P36's case, she connected her mistrust of Facebook's policies and rule enforcement to her mistrust of social media platforms' content moderation practices in general. Overall, the participants engaged deeply with Facebook's full nudity policies, developing further insight into the policies, critiquing policies that they perceived as flawed, and even reinforcing their overall negative perceptions of the platform itself.

Several participants reflected on their past experiences with social media content removals as they browsed through Facebook's policies, particularly participants whose content was removed from platforms other than Facebook. P35 shared that he previously had content removed from Reddit, and initially stated that the Social Media Guidelines resource could have potentially been a *"big help... in combating that moderation [decision]."* But as P35 continued to reflect on Facebook's policies, he stated that most moderation decisions on Reddit *"depend on who the mod is for the subreddit,"* and that reading through Reddit's sitewide moderation policies may not be as helpful as it would be to a user of Facebook, Twitter, or other platforms with a *"more centralized team for moderation and removal."* Though P35's Reddit content removal was not due to nudity, navigating Facebook's nudity policies still resulted in him critically assessing the differences he perceived between Facebook and Reddit's moderation systems. P35's reflection also revealed a potential limitation of the OIHC itself; his comments highlighted how the OIHC's summarized policies may be less helpful to users of social media platforms that employ a "hybrid" model of moderation (such as Reddit) compared to platforms employing more centralized, "top-down" moderation (such as Facebook), as policies may be enforced unequally across "hybrid" platforms based on human moderators' individual decisions [6, 41, 44, 60, 71].

P32 also reflected on their past content removal experiences while reading through Facebook's nudity guidelines; they stated that the summarized policies are *"really helpful... because rules are a thing that different social media sites vary on broadly,"* especially *"about nudity specifically, which is a point of tension for many people."* As P32 explored Facebook's guidelines on nudity, they reflected on their own past experience of having content incorrectly removed for "nudity" on Tumblr:

> *"On Tumblr, I've had stuff removed because of 'nudity,' even though there was none. I don't know if Tumblr has an appeals process... their [moderation process] seems very sporadic. So I'm not sure if anything like [this page] could have been helpful for my experiences."*

Like P35, P32 reflected on their content removal experience on another platform while reading through Facebook's guidelines, questioning whether reading Tumblr's community guidelines would have helped them during their removal experience and expressing skepticism that Tumblr's moderation system would enforce its own guidelines correctly. When participants like P35 and P32 were introduced to Facebook's policies through the OIHC, they not only learned more about Facebook's policies themselves, but they also reflected on moderation practices that they previously witnessed on other platforms, critically comparing the perceived differences in moderation practices between social media platforms, including comparing the perceived effectiveness of different platforms' moderation systems. However, experiences like P32's also made visible some of the OIHC's limitations, such as the limited number of social media platforms that it provides policy information about.

5.2.2 Participants Expressed Skepticism that Platforms Would Respond to Their Appeals or Restore Their Content. Participants described appreciating the OIHC's links to content moderation appeal resources across a variety of platforms. However, as participants engaged with the platforms' appeal pages, many expressed skepticism that filing content moderation appeals would result in their content being restored. Some of this doubt was due to a lack of information provided by platforms

regarding the turnaround time for content moderation appeals. P27 suggested that the Social Media Appeals page include *"some information on the typical turnaround time for an appeal"* to clarify *"how long[users are] supposed to wait"* to hear back from the platform. P27 stated that adding this information could benefit users who may not know what to expect from the appeal process, which she perceived to be *"very confusing"* to ordinary social media users.

Some participants also questioned whether social media platforms would respond to their moderation appeals at all. P27 shared that she felt *"hesitant [to file an appeal] because... it's not guaranteed that Twitter will actually give your account or posts back."* P32 shared similar frustrations, stating that *"sometimes, even if you do follow the [appeal] steps, Twitter will still suspend your account or not give it back."* P32's lack of trust that social media platforms would act on users' appeals also made them feel less confident in the Social Media Appeals page's instructions, stating that there is *"not really much of a guarantee"* that following the instructions correctly would result in a platform restoring their content. Comments such as P27's and P32's highlighted participants' mistrust of platforms' moderation practices; even in a hypothetical situation where their content was incorrectly removed, participants still expressed doubt that platforms would acknowledge their potential error or restore their content in response to moderation appeals. The doubt expressed by the participants may reflect an underlying perception that the OIHC's linked appeal resources may not be equally helpful for all users appealing a removal due to the perceived unreliability of platforms' appeal processes themselves.

Some participants also expressed skepticism that the appeal resources provided in the Social Media Appeals section would apply to all content removal situations that they may experience. P27 stated that she had previously struggled to complete a content removal appeal on Reddit as it was *"not very clear which [removal reason] would align with [her] content that was taken down."* P31 shared similar concerns, stating that *"people whose [content removal] situations are dramatic, like being discriminated against,"* may feel that their situations have escalated into *"something legal,"* and may prefer legal help regarding the removal rather than trusting the platform itself to handle the appeal correctly.

Overall, the Social Media Appeals page was praised for providing different platforms' appeal resources to the OIHC's users, which participants found particularly beneficial due to the difficulty of finding appeal resources on social media platforms themselves. However, multiple participants stated that platforms' appeal resources may not meet their specific needs in the first place, and that they may have to explore alternative options to challenge their content takedowns instead. The difficulty of finding platforms' appeal resources without the help of a resource like the OIHC, as well as the possibility that platforms' appeal resources may not account for all users' moderation situations, led multiple participants to reinforce their frustration and negative perceptions of social media platforms. These frustrations also highlighted several of the OIHC's limitations in helping users navigate content moderation and appeals, such as its inability to guarantee that platforms will reliably moderate content or appeal removed content in line with their own policies.

5.3 Evaluation Study Results: Engagement with the Online Identity Help Center

5.3.1 Social Media Rights. The Social Media Rights page includes a series of educational scenarios describing realistic content moderation and removal situations, including scenarios where a social media user may face offline consequences for their social media content (such as being fired from a job for posting objectionable content online). These scenarios include summarized information on social media users' rights in each scenario, long-form explanations as to why the content removal (or real-world consequence) may or may not be correct, and links to external resources (such as platforms' official policy pages) for users who may wish to learn more. Participants described the Social Media Rights page to be informative, presenting realistic moderation-related scenarios

that engage users while informing them of their user rights in different moderation contexts. P25 stated that user rights descriptions throughout the page were *"precise and easy to understand,"* while P29 praised the vocabulary of the page for being *"short and simple"* while avoiding *"lots of legal language or jargon."* Participants praised the hypothetical moderation scenarios presented throughout the page, particularly "gray area" scenarios where the correct removal decision may be unclear. P29 stated that the gray area scenarios *"did a good job of bringing up questionable situations where whether [users] have freedom of speech, or whether [removals] would be justified or not, depends on other factors."* Overall, participants expressed that the Social Media Rights page served as an accessible, informative starting point to begin learning more about their rights as social media users.

Many participants stated that they appreciated the links to external informational resources about social media rights provided throughout the page at the end of each hypothetical moderation scenario. P29 stated that the external resources were helpful when addressing the legal aspects of users' social media rights, stating that *"legal texts can be really difficult [to understand] if you're not a lawyer... so it's nice to have links to websites that translate laws for a general audience so they're easier to understand."* P35 appreciated that *"the resources could give [him] more information"* about his social media rights than he could find on the OIHC itself; P26 agreed, stating that *"it's nice that [external resources] are included"* to supplement the content of the Social Media Rights page. Other participants expressed interest in exploring the external resources even after leaving the page; P26, P27, and P29 suggested that the OIHC provide a full *"list of [external] resources"* near the end of the Social Media Rights page for users to explore in their own time. Overall, participants expressed that the Social Media Rights page provided helpful "basic" information about social media users' rights and how those rights may apply to everyday social media use. However, many participants primarily expressed their perception of the Social Media Rights page as a "hub" for external social media rights resources instead, where users can find links to platforms' official policies, legal resources provided by the U.S. government, and other "official" resources that may help them navigate future social media content or account removals. Though the users praised the page's original content as informative and useful, it is possible that users primarily value the Social Media Rights page as an entry point for finding official platform policies and legal resources.

Participants offered their thoughts on the hypothetical moderation scenarios that they encountered as they progressed through the Social Media Rights page. Several participants initially expressed skepticism toward the page's explanations for moderation decisions, but challenged their own reactions after reading the explanations in depth. P29 expressed that their *"gut reaction"* to the Snapchat removal scenario (where a student was expelled for using profanity in a Snapchat post) was to disagree with the student's expulsion. Though the page's explanation affirmed P29's reaction by stating that the expulsion would be incorrect, it also provided example contexts in which a student may *not* be protected from disciplinary action against their social media content, such as if the content was posted during school hours or if it relates to school activities. After reading the explanation, P29 somewhat reevaluated their stance, stating that they now *"understood that there are instances where something like [the Snapchat post] would not be appropriate."* P31 also reevaluated her perceptions of moderation and social media users' rights after reading the page's explanations in depth. While reading a scenario about an employee being fired from their job due to their social media content, P31 initially answered that their firing was *"probably justified,"* as the employee's social media posts may have been inappropriate. However, the explanation stated that their firing was a *"gray area"* decision that could vary based on workplace policies and other external factors. Like P29, P31 reevaluated her initial answer after reading the explanation, concluding that her answer *"was not wrong, but was not right either."*

Overall, participants critically engaged with the page's explanations for the "correct" conclusions to moderation and user rights scenarios. This critical engagement led several participants to openly reassess their personal perceptions of how moderation and users' rights work, developing a more nuanced understanding of how users' social media rights can operate in various contexts.

5.3.2 Data Privacy and Collection. The Data Privacy and Collection section of the OIHC present short-form advice on a range of personal data and privacy-protection principles, such as advice on protecting one's social media account information or using public wireless connections. Participants found the Data Privacy and Collection section to provide a range of broadly applicable information about protecting their personal data; P27 stated that *"many people would benefit from this information and find it useful,"* while P25 stated that it offers *"very nice advice in terms of security,"* particularly regarding two-factor authentication and safely using public internet connections. Several participants stated that the Data Privacy and Collection section could be particularly useful for users who may not be familiar with general data privacy principles; P36 stated that the data privacy advice within the section is *"not obvious to most users,"* while P32 stated that the information in the section includes *"many basic things [that] most users forget a lot of the time."* Overall, participants expressed confidence that the page could be a useful resource for users who are new to the concept of data privacy and collection.

Like other sections of the OIHC, participants also praised the external informational resources provided throughout the Data Privacy and Collection section. P32 stated that the linked resources made them *"curious to know"* more about personal data privacy practices; P36 agreed, stating that the external resources can *"help people understand"* data privacy in greater detail than what is included in the section itself. Some participants, such as P30, stated that the contents of the Data Privacy and Collection page itself were *"helpful... but mostly things that [she] already knew."* However, P30 found the page's external resources to be beneficial by offering an *"in-depth look"* at the data privacy concepts described throughout the page, allowing P30 to progress from reviewing familiar data privacy principles to learning new ways to protect her privacy online. As a result, P30 perceived the external resources as *"increasing the usefulness of [the page],"* particularly for social media users who may already have a basic understanding of how to protect their data and privacy online, but may wish to understand those principles in greater detail.

Overall, participants praised the Data Privacy and Collection section for presenting an accessible starting point for everyday users to learn about protecting their data online. Like the Social Media Rights section, participants were particularly likely to discuss the external resources provided on the Data Privacy and Collection page; it is possible that these external resources were the most valuable feature of the page for participants who were already familiar with basic data privacy principles and were interested in finding more "in-depth" information.

5.3.3 Social Media Appeals. The Social Media Appeals section of the OIHC provides users with direct links to social media platforms' official content moderation appeal resources, along with summarized instructions for how to appeal content or account removals on different platforms. The participants praised the Social Media Appeals section for simplifying the process of finding and filing content moderation appeals with social media platforms. P33 stated that the instructions were *"clear and short,"* and praised the use of screenshots to visually guide participants through the appeal process. P27 agreed that the page's instructions for filing appeals were *"very straightforward and helpful,"* stating that the simplicity of the page could particularly benefit *"ordinary users"* navigating their *"first time experiencing the [appeal] process."*

P29 remarked on the availability of appeal resources on the page that may be difficult for users to find on social media platforms themselves, stating that *"it's nice that [the OIHC] gives you the steps that are involved."* P29 then shared their personal experience of having an account removed

from Twitter: *"My Twitter account was locked... without any information about how to appeal it. I didn't even know that there was an option to appeal!"* This experience resulted in P29 perceiving social media platforms as deliberately *"making [appeal resources] really obscure to find,"* concluding that this happens because *"social media sites don't want to deal with [appeals]."* Afterwards, P29 stated that the Social Media Appeals page could directly benefit the OIHC's users by *"providing the links to where you go [to appeal]"* instead of leaving users to find appeal resources (perceived by P29 to be deliberately obscured) on their own.

Overall, participants reacted positively to the appeal resources and instructions provided on the Social Media Appeals page. However, users like P29 expressed frustration with the difficulty of finding appeal resources on their respective platforms in the first place, reinforcing their negative perceptions of social media platforms as deliberately suppressing users' ability to appeal content removals.

5.3.4 Share Your Story. The Share Your Story section provides an online form where OIHC users may submit testimonies of their past content or account removal experiences that they felt were unfair or incorrect. With the user's explicit consent, the OIHC's staff can read these submissions and, when applicable, follow-up with the user via email with further resources that may help them navigate their content or account removals. Users are also welcome to use the form simply to express their feelings (or "vent") about their content removal experiences; with the user's consent, anonymized versions of these testimonies may also be featured on the OIHC for other users to read. Participants expressed appreciation that the Share Your Story page offered the OIHC's users the opportunity to share their content removal experiences for other OIHC users to read. P34 stated that the Share Your Story page makes it *"easy for [users] to make connections and hear one another's stories,"* and that the OIHC's users would benefit from reading these stories and knowing that others have faced similar moderation experiences. P36 stated that she mostly envisioned herself using the OIHC to "vent" about her past instances with content removals; she then shared her past experience of having content removed from Reddit in a way she perceived to be incorrect, and how she would have used the OIHC to air her frustrations about the removal:

> *"I probably would have used [the OIHC] as a venting experience... because it was such a shock when my content was removed. There was nothing explicit about [my post]! I didn't even use any inappropriate language! So I contacted the moderator to say, "I don't understand," because I was genuinely curious whether I had missed something. Then the moderator insinuated that [my post] was removed for "soliciting!" And I thought, "whoa... what?" So I think I would have used the OIHC to vent about that."*

Participants like P34 and P36 discussed why they perceive a space for social media users to "vent" about their removal and appeal experiences to be important, and why some users may even primarily use the OIHC to "vent"about their negative experiences via the Share Your Story feature. Even participants who have already learned how to navigate moderation and appeals processes expressed wanting to "share their story" with other users; like P36, these users may even find the Share Your Story feature to be the most appealing element of the OIHC.

Some participants also suggested that the Share Your Story page integrate other community-oriented features, such as a forum or chat feature, that could allow its users to directly communicate with other OIHC users who have experienced content removals or are undergoing appeals. P34 stated that *"[the OIHC] can't go wrong with making a little community on [the website] for users to reach out to each other";* P30 agreed, stating that the OIHC could foster *"a community... for people who don't understand why their content was removed"* by allowing its users to speak with one another *"about the online censorship issues they face."* P30 described her ideal communication tool as *"a bit like a Reddit thread where people can reach out to other [users] who had their content taken*

down," arguing that this feature would benefit the OIHC's users by allowing them to *"discuss their removal experiences and stories"* with one another. P30 also expressed her belief that integrating a communication tool could *"increase [users'] chances of using the OIHC more in the future,"* as the feature would help the OIHC *"foster a community... [where] we could all speak about the moderation issues we face and our grievances with the system."*

Overall, participants like P34 and P30 valued the ability to share their removal experiences with the OIHC's users via the Share Your Story form. However, users also desired the ability to use the OIHC to directly communicate with others who have also experienced removals (a feature that the OIHC currently does not provide). Like P36, P34 and P30 indicated that the ability to express the emotional toll of content removal experiences, as well as finding support from other users who can relate to those experiences, is a similarly high priority for users as finding information to help navigate their removals and appeals.

6 DISCUSSION

6.1 Accessible policy "starting points" encourage users to read policies in full

Though social media users typically do not read social media platforms' terms of service prior to using social media platforms [2, 73, 76], the behavior of the OIHC's user test participants suggests that users may be more willing to read platforms' policies when presented in a simplified, accessible format. The study participants easily found and read through the summarized platform policies on the OIHC, and overwhelmingly expressed that the summarized policy content was both informative and easy for them to understand. Participants also shared similar sentiments about sections of the OIHC that were related to social media policies but were not exclusively about content removals, such as the Social Media Rights Page. Most participants indicated that the OIHC's summarized policy information sufficiently informed them of how to navigate content moderation on social media platforms, including participants like P35 who acknowledged that a summarized policy by nature cannot address every nuance or exception to a rule. Past research has explored and developed visually friendly formats for communicating policy information to users in a way that is easier to understand than long, impenetrable legal documents [22, 56]. Examples of user-friendly policy formats include Kelley et al.'s "Nutrition Label for Privacy" and Drodz and Kirrane's "Consent Request User Interface (CURE) Prototype," both of which were designed to improve users' comprehension of privacy policies so they may make more informed choices about their personal data and how it is collected and used [22, 56]. Both the Nutrition Label and the CURE Prototype were found to improve users' comprehension of privacy policies; Kelley et al. also found that their study participants rated the visually friendly Nutrition Label format as "more enjoyable" to read than pre-existing privacy policies [22, 56]. We argue that a resource like the OIHC is valuable to social media users for similar reasons as Nutrition Label for Privacy [56] and the CURE prototype [22]: because it presents an accessible introduction to platforms' policies in a convenient, user-friendly format that encourages everyday users to begin familiarizing themselves with platforms' policies, directly confronting the trend of users either skimming platforms' policies or not reading them at all [2, 73, 76].

However, OIHC goes beyond prior work by not only encouraging users to read summarized versions of social media platforms' policies but by successfully encouraging users to read the full versions of platforms' policies *after* reading the OIHC's summarized version. As described in Section 5.2.1, after reading the summarized descriptions of platforms' policies, multiple participants then chose to read the full text versions of those policies that were directly linked within the OIHC. Though these resources were located outside of the OIHC and were not presented in a "summarized" format, multiple participants still expressed their desire to read the full versions of

platforms' policies after first reading their summarized versions. The participants did not indicate that they read the full versions of policies due to feeling insufficiently informed by the OIHC's policy summaries; instead, participants like P6 indicated that the summarized policies sufficiently informed them, and that their decision to explore the full versions of policies was more motivated by personal interest. The participants who read the full versions of policies critically engaged while reading them, such as by voicing their opinions on the policies and whether they can be enforced, indicating a rich level of engagement with policies that they ordinarily would not have read in the first place. This level of engagement may indicate that social media users can develop confidence and willingness to read full, "complex" versions of platforms' policies, and can more easily understand complex versions of platforms' policies after being introduced to a summarized version. We argue that the OIHC reflects the strength of "readable" policy presentation formats by presenting "readable" versions of policies when platforms themselves fail to do so, helping sufficiently inform users of how platforms work while encouraging them to feel confident enough to read full policies afterward. This is a significant development considering that social media users typically do not feel incentivized to read or critically engage with platforms' full policies in the first place [73, 76], particularly policies that users do not perceive to be "readable" [28]. This summarized policy format can benefit users not only by directly resolving the barriers to initially familiarizing themselves with policies, but also by encouraging them to read the full text of policies after developing an initial awareness about how the platform works – thus making it easier for social media users to understand platforms' policies at increasing levels of complexity and nuance.

We note, however, that while the summarized policy format was generally well-received and perceived as sufficiently informative by most participants, the summarized policies were not necessarily perceived as helpful by every study participant, particularly those who were skeptical that social media platforms would enforce their own policies as written; this signals that the summary format will not be helpful for all users. We also acknowledge that our findings related to users going from reading summarized policies to full policies were unanticipated findings of our study, and that our insights related to these findings were limited by the number of participants in our evaluation study; future research can specifically explore social media users' engagement patterns with summarized and non-summarized policy information on a larger scale.

6.2 Users may remain skeptical of policies even after reading them in full

Though multiple participants expressed a desire to read the official versions of platforms' policies after first reading a summarized version on the OIHC, they typically did not respond to the official platform policies by expressing trust that the platforms would enforce their own rules properly. Instead, participants typically responded with skepticism while reading social media platforms' official policy pages, contrasting participants' positive receptions of summarized policies presented on the OIHC itself. Multiple participants who read platforms' official policies began expressing negative sentiments toward social media platforms' content moderation practices. For example, after reading Facebook's artistic nudity policies on Facebook's Community Standards page, participants expressed frustration with what they perceived to be the policies' ambiguities, while questioning whether Facebook's algorithmic moderation tools can accurately distinguish between permissible and impermissible nude content. Similar patterns emerged as participants explored the external appeal resources in the OIHC's Social Media Appeals page. Participants who engaged with platforms' appeal resources expressed doubt that the platforms would respond to content moderation appeals promptly (if at all), or that appealing content removals would be equally effective on different platforms. Though many participants chose to read the official guidelines, they did not necessarily trust that the guidelines would be enforced as written.

Instead of trusting that platforms will moderate content in line with their own rules, some participants shared their personal folk theories as to how platforms moderate their content instead. Folk theorization related to social media content moderation and removals is particularly common among marginalized social media users [15, 16, 55], reflecting the disproportionately high rates of incorrect or gray-area content removals that marginalized users face on social media platforms [47]. Marginalized users who have experienced or witnessed these disproportionate removals often distrust platforms' rules and content moderation systems, and choose to develop their own theories as to how platforms' content moderation systems operate, relying on these theories to guide their behavior and decision-making on social media platforms [15, 16, 63]. P36's voiced skepticism of Facebook's content moderation practices serves as an example of folk theorization in practice; while P36 critiqued Facebook's artistic nudity policies as overly ambiguous, she connected her critique to her perception that social media platforms' policies are overly vague in general. She then took her perception a step further, theorizing that platforms deliberately develop vague platform policies to avoid culpability when their content moderation practices attract backlash from social media users. Like many social media users before her, P36 deferred to her folk theories about social media platforms' policies to guide her perception of Facebook's moderation practices, even after reading Facebook's full artistic nudity policy.

Social media folk theorization is sometimes framed as social media users developing theories about platforms' moderation practices primarily as a substitute for reading platforms' rules [63]. However, we found that even when our participants read platforms' official policies, they still continued to develop theories as to how platforms would enforce their rules instead of trusting that platforms would enforce their rules correctly. Past literature has explored the role of marginalized users' distrust of platforms in both their reluctance to read platforms' guidelines [63] and their use of folk theories to guide their decision-making on social media platforms [15, 16, 63]. Our study tied these ideas together by showing that, even after reading platforms' policies in detail, marginalized social media users may continue to rely on their folk theories to guide their behavior on platforms instead of the platforms' guidelines themselves. We argue that marginalized users' folk theories about platforms' guidelines may have less to do with whether they've read and understood a platforms' policies, and more to do with their pre-existing distrust of platforms themselves, often derived from their negative experiences with identity-related content moderation and removals.

We also argue that platforms themselves can take proactive steps that may reduce marginalized users' perceived need to theorize about how guidelines are enforced. One possible approach for platforms could involve developing content moderation transparency resources that publicly clarify how guidelines are enforced in cases related to marginalization. For instance, the Oversight Board website is an example of an existing transparency resource. The Oversight Board is an independent body overseeing the governance of Meta platforms like Facebook and Instagram, and provides information on its website describing its decisions to uphold or overturn content moderation decisions that were appealed by Meta platforms users, including decisions related to marginalized users' content [77]. One example includes the Oversight Board's decision to overturn Meta's removal of two transgender and nonbinary Instagram users' posts related to top surgery [78]. The Oversight Board detailed how it determined that the trans users' posts were incorrectly identified by Meta as violating its Adult Nudity and Sexual Activity Community Standard, ultimately ruling that the two trans users' posts be restored. The Oversight Board also published public comments submitted by social media users related to the appeal case, which prominently featured submissions by trans and nonbinary users. These public comments overwhelmingly expressed disagreement with Meta's initial removal of the two trans users' posts, advocating for the posts to be restored while describing the potentially harmful implications of the removals for trans users of Meta's platforms [78]. Though the existence of transparency resources like the Oversight Board's website does not

guarantee that marginalized users will trust platforms to enforce their guidelines appropriately, similar resources could at least provide marginalized users with greater insight into how platforms' content moderation decisions related to marginalization are made, potentially reducing marginalized users' perceived need to theorize about how platforms moderate content related to marginalization.

6.3 Implications for Design

In a sense, developing a site like the OIHC is our implication for design; because people often do not trust social media platforms and are unlikely to read their guidelines [74], it is necessary to create and maintain external sites that summarize and engage people about social media site policies. We developed the OIHC directly in response to user needs that we identified while interviewing marginalized social media users, such as their need for accessible versions of social media platforms' policies. Resources addressing social media users' rights, social media policies and guidelines, user privacy and data protection, and contacting social media platforms were four major content areas that participants in our study identified as necessary in order for the OIHC to meet their needs. By addressing these content areas on the OIHC, we aim to provide marginalized social media users with easy access to the resources necessary for them to better understand their rights as social media users while more easily navigating their experiences with content moderation and removals. Participants also expressed their underlying mistrust of social media platforms' guidelines and content moderation practices, reflecting the broader trend of marginalized social media users distrusting platforms after experiencing disproportionate, inequitable patterns of content moderation and removals [47, 63]. By presenting platforms' policy and moderation resources in an easily digested format, we aim for the OIHC to make platforms' policies more transparent for its users. By visibly displaying the OIHC's major research university affiliation on the service, we aim for the OIHC's summarized policy information to be perceived as trustworthy by its users. Making platforms' policies more transparent can help marginalized social media users more easily interpret platforms' rules and values, allowing them to make more informed decisions about what they post, say, or do on social media platforms.

One challenge in providing a resource like the OIHC is keeping it maintained and updated as social media platforms' guidelines continually change. The OIHC is currently manually maintained and updated by the research team; we acknowledge that manually updating the OIHC's social media policy and informational resources could pose several challenges, such as the difficulties of knowing when a platform has updated their policies and monitoring policy changes across multiple platforms. Future work on the OIHC could determine efficient ways to automate updates to the site's policy content or resources, ensuring that the OIHC remains a relevant resource for marginalized social media users into the future. Additionally, we acknowledge upcoming EU regulatory changes, namely through the Digital Services Act (DSA), that will require platforms to clearly explain how their content sorting and recommendation algorithms operate, and how they make removal decisions for illegal content [72, 97]. Future developments in this area, namely increased clarity about platforms' algorithms and removal decisions, may change the perceived helpfulness or relevance of the OIHC's "simplified" policies.

6.4 Limitations

User tests of the OIHC were split into two rounds (n = 5 participants in the first round, n = 7 in the second). Though all user test participants were marginalized social media users, the second round participants were asked more questions about their personal experiences with social media content or account removals, along with specific questions about identity-related content removals. The lack of identity-related questions asked to the first round user test participants limited our insight into the identity-related social media experiences of our user test participants, and whether

the participants feel the OIHC fully meets their specific needs as marginalized social media users. We also acknowledge that this study was also conducted in a U.S. context and primarily centered the experiences of marginalized social media users from the U.S. Future research can explore the most appropriate ways to meet the online policy, content moderation, and digital literacy-related needs of marginalized social media users from outside of the U.S. or of Western cultural contexts. Additionally, we acknowledge that the OIHC has not been heavily promoted, thus limiting the number of users using the service. Future promotion and search engine optimization could increase traffic to the OIHC. Increased traffic could result in more insight on users' perceptions of the OIHC, allowing us to further improve the service in the future.

7 CONCLUSION

We designed and developed the OIHC to confront the challenge of inequitable content moderation faced by marginalized social media users, providing marginalized users with information about their rights on social media platforms, social media platforms' policies, and directions for appealing content and account removals. We share the interview participants' priorities for topics to feature on the OIHC, such as resources to help them better understand social media policies, how to appeal content removals, what their rights are as social media users, and how to protect their data online. We also share findings from user testing of the OIHC, such as user test participants' decisions to read platforms' full policies after reading summarized versions on the OIHC, as well as participants' skepticism that platforms will restore their removed content after they have filed an appeal. We then discuss the implications of our findings, such as the use of summarized platform policies as "starting points" to encourage users to read platforms' policies in full. We also discuss participants' continued distrust of social media platforms after reading their policies, including participants' use of folk theories to help interpret how social media platforms may moderate their content in practice. It may be the case that marginalized users will continue to primarily defer to folk theories to guide their behavior on social media platforms even after reading platforms' policies – an important area to explore more in future research.

ACKNOWLEDGMENTS

We thank our study participants for sharing their insights and experiences with us. We also thank the members of the Community Research on Identity and Technology (CRIT) Lab at the University of Michigan School of Information (UMSI) for their helpful feedback and comments on our work. We also thank our anonymous reviewers for their constructive comments that improved this work. This work was supported by the National Science Foundation grant #1942125.

REFERENCES

[1] Access Now. 2022. About. https://www.accessnow.org/help/

[2] Yannis Bakos, Florencia Marotta-Wurgler, and David R. Trossen. 2014. Does Anyone Read the Fine Print? Consumer Attention to Standard-Form Contracts. *The Journal of Legal Studies* 43, 1 (Jan. 2014), 1–35. https://doi.org/10.1086/674424

[3] Sam Biddle. 2022. Facebook Report Concludes Company Censorship Violated Palestinian Human Rights. https://theintercept.com/2022/09/21/facebook-censorship-palestine-israel-algorithm/

[4] Michelle Blanchard, Atari Metcalf, Jo Degney, Helen Herrman, and Jane Burns. 2008. Rethinking the digital divide: Findings from a study of marginalised young people's information communication technology (ICT) use. *Youth Studies Australia* 27, 4 (2008), 35–42. https://academics.hamilton.edu/ebs/pdf/rtdd.pdf

[5] 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/

[6] Ben Bradford, Florian Grisel, Tracey L. Meares, Emily Owens, Baron L. Pineda, Jacob N. Shapiro, Tom R. Tyler, and Danieli Evans Peterman. 2019. *Report of the Facebook Data Transparency Advisory Group.* Technical Report. The

Justice Collaboratory - Yale Law School. https://law.yale.edu/system/files/area/center/justice/document/dtag_report_5.22.2019.pdf

[7] Kara Brisson-Boivin and Samantha McAleese. 2021. How digital literacy can help close the digital divide. *Institute for Research on Public Policy* (April 2021). https://policyoptions.irpp.org/magazines/april-2021/how-digital-literacy-can-help-close-the-digital-divide/

[8] Erik P. Bucy. 2000. Social Access to the Internet. *Harvard International Journal of Press/Politics* 5, 1 (Jan. 2000), 50–61. https://doi.org/10.1177/1081180X00005001005

[9] Erik Calleberg. 2021. Making Content Moderation Less Frustrating: How Do Users Experience Explanatory Human and AI Moderation Messages. http://sh.diva-portal.org/smash/record.jsf?pid=diva2%3A1576614&dswid=6239

[10] Anna Woorim Chung. 2020. *Subverting the algorithm: Examining anti-algorithmic tactics on social media.* Ph. D. Dissertation. Massachusetts Institute of Technology. https://hdl.handle.net/1721.1/127453

[11] Consumer Reports. 2022. Keep Your Data Secure With a Personalized Plan. https://securityplanner.consumerreports.org

[12] Juliet Corbin and Anselm Strauss. 2008. *Basics of Qualitative Research (3rd ed.): Techniques and Procedures for Developing Grounded Theory.* SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States. https://doi.org/10.4135/9781452230153

[13] Benjamin F. Crabtree and William L. Miller (Eds.). 1999. *Doing qualitative research* (2nd ed ed.). Sage Publications, Thousand Oaks, Calif.

[14] Amanda L. L. Cullen and Bonnie Ruberg. 2019. Necklines and 'naughty bits': constructing and regulating bodies in live streaming community guidelines. In *Proceedings of the 14th International Conference on the Foundations of Digital Games.* ACM, San Luis Obispo California USA, 1–8. https://doi.org/10.1145/3337722.3337754

[15] Michael Ann DeVito. 2021. Adaptive Folk Theorization as a Path to Algorithmic Literacy on Changing Platforms. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–38. https://doi.org/10.1145/3476080

[16] Michael Ann DeVito. 2022. How Transfeminine TikTok Creators Navigate the Algorithmic Trap of Visibility Via Folk Theorization. *Proceedings of the ACM on Human-Computer Interaction* (2022). https://michaelanndevito.files.wordpress.com/2022/05/transfemmetiktok_cscw2022_preview.pdf

[17] Michael A. DeVito, Jeremy Birnholtz, Jeffery T. Hancock, Megan French, and Sunny Liu. 2018. How People Form Folk Theories of Social Media Feeds and What it Means for How We Study Self-Presentation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM, Montreal QC Canada, 1–12. https://doi.org/10.1145/3173574.3173694

[18] Jan A. G. M. Dijk. 2017. Digital Divide: Impact of Access. In *The International Encyclopedia of Media Effects* (1 ed.), Patrick Rössler, Cynthia A. Hoffner, and Liesbet Zoonen (Eds.). Wiley, 1–11. https://doi.org/10.1002/9781118783764.wbieme0043

[19] Christina Dinar. 2021. *The state of content moderation for the LGBTIQA+ community and the role of the EU Digital Services Act.* Technical Report. Heinrich-Böll-Stiftung. 23 pages.

[20] Don't Delete Art. 2022. The Campaign. https://www.dontdelete.art/thecampaign

[21] Emily Dreyfuss. 2018. Twitter Is Indeed Toxic for Women, Amnesty Report Says. https://www.wired.com/story/amnesty-report-twitter-abuse-women/

[22] Olha Drozd and Sabrina Kirrane. 2020. Privacy CURE: Consent Comprehension Made Easy. In *ICT Systems Security and Privacy Protection*, Marko Hölbl, Kai Rannenberg, and Tatjana Welzer (Eds.). Vol. 580. Springer International Publishing, Cham, 124–139. https://doi.org/10.1007/978-3-030-58201-2_9 Series Title: IFIP Advances in Information and Communication Technology.

[23] Ángel Díaz and Laura Hecht-Felella. 2021. Double Standards in Social Media Content Moderation. *Brennan Center for Justice at New York University School of Law* (2021). https://www.brennancenter.org/our-work/research-reports/double-standards-socialmedia-content-moderation

[24] Matthew S. Eastin and Robert LaRose. 2006. Internet Self-Efficacy and the Psychology of the Digital Divide. *Journal of Computer-Mediated Communication* 6, 1 (June 2006), 0–0. https://doi.org/10.1111/j.1083-6101.2000.tb00110.x

[25] Electronic Frontier Foundation. 2021. The Santa Clara Principles on Transparency and Accountability in Content Moderation. https://santaclaraprinciples.org

[26] Dmitry Epstein and Kelly Quinn. 2020. Markers of Online Privacy Marginalization: Empirical Examination of Socioeconomic Disparities in Social Media Privacy Attitudes, Literacy, and Behavior. *Social Media + Society* 6, 2 (April 2020), 205630512091685. https://doi.org/10.1177/2056305120916853

[27] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I "like" it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.* ACM, San Jose California USA, 2371–2382. https://doi.org/10.1145/2858036.2858494

[28] Benjamin Fabian, Tatiana Ermakova, and Tino Lentz. 2017. Large-scale readability analysis of privacy policies. In *Proceedings of the International Conference on Web Intelligence*. ACM, Leipzig Germany, 18–25. https://doi.org/10.1145/3106426.3106427

[29] Margaret Fernandes. 2022. *Making Sense of Digital Content Moderation from the Margins*. Ph. D. Dissertation. http://hdl.handle.net/10919/110747

[30] 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 (May 2020), 1–28. https://doi.org/10.1145/3392845

[31] Casey Fiesler, Cliff Lampe, and Amy S. Bruckman. 2016. Reality and Perception of Copyright Terms of Service for Online Content Creation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, San Francisco California USA, 1450–1461. https://doi.org/10.1145/2818048.2819931

[32] Electronic Frontier Foundation. 2019. Who Owns a Word? https://www.eff.org/tossedout/who-owns-word

[33] Jesse Fox and Rachel Ralston. 2016. Queer identity online: Informal learning and teaching experiences of LGBTQ individuals on social media. *Computers in Human Behavior* 65 (Dec. 2016), 635–642. https://doi.org/10.1016/j.chb.2016.06.009

[34] Megan French and Jeff Hancock. 2017. What's the Folk Theory? Reasoning About Cyber-Social Systems. *SSRN Electronic Journal* (2017). https://doi.org/10.2139/ssrn.2910571

[35] Seeta Peña Gangadharan. 2017. The downside of digital inclusion: Expectations and experiences of privacy and surveillance among marginal Internet users. *New Media & Society* 19, 4 (April 2017), 597–615. https://doi.org/10.1177/1461444815614053

[36] Susan A. Gelman and Cristine H. Legare. 2011. Concepts and Folk Theories. *Annual Review of Anthropology* 40, 1 (Oct. 2011), 379–398. https://doi.org/10.1146/annurev-anthro-081309-145822

[37] Ysabel Gerrard. 2018. Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society* 20, 12 (Dec. 2018), 4492–4511. https://doi.org/10.1177/1461444818776611

[38] Ysabel Gerrard and Helen Thornham. 2020. Content moderation: Social media's sexist assemblages. *New Media & Society* 22, 7 (July 2020), 1266–1286. https://doi.org/10.1177/1461444820912540 Publisher: SAGE Publications.

[39] Tarleton Gillespie. 2017. Governance of and by platforms. In *The SAGE Handbook of Social Media*. SAGE, New York, 30.

[40] Tarleton Gillespie. 2018. *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, New Haven London.

[41] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7, 2 (July 2020), 205395172094323. https://doi.org/10.1177/2053951720943234

[42] GLAAD. 2021. GLAAD's Social Media Safety Index. https://www.glaad.org/blog/glaads-social-media-safety-index

[43] Amy L. Gonzales. 2017. Disadvantaged Minorities' Use of the Internet to Expand Their Social Networks. *Communication Research* 44, 4 (June 2017), 467–486. https://doi.org/10.1177/0093650214565925

[44] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (Jan. 2020), 205395171989794. https://doi.org/10.1177/2053951719897945

[45] James Grimmelmann. 2017. *The Virtues of Moderation*. preprint. LawArXiv. https://doi.org/10.31228/osf.io/qwxf5

[46] Jessica Guynn. 2019. Facebook while black: Users call it getting 'Zucked,' say talking about racism is censored as hate speech. *USA Today* (April 2019). https://www.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/

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

[48] H. Rex Hartson and Pardha S. Pyla. 2019. *The UX book: Agile UX design for a quality user experience* (second edition ed.). Morgan Kaufmann, an imprint of Elsevier, Amsterdam. OCLC: 1076548565.

[49] Jennifer Higgs, Steven Athanases, Alexis P. Williams, Danny Martinez, and Sergio Sanchez. 2021. Amplifying Historically Marginalized Voices Through Text Choice and Play With Digital Tools: Toward Decentering Whiteness in English Teacher Education. *Contemporary Issues in Technology and Teacher Education* 21, 3 (Sept. 2021), 583–612. https://www.learntechlib.org/p/217723/ Society for Information Technology \& Teacher Education.

[50] Juan Pablo Hourcade, Natasha E. Bullock-Rest, and Heidi Schelhowe. 2010. Digital Technologies and Marginalized Youth. In *Proceedings of the 9th International Conference on Interaction Design and Children - IDC '10*. ACM Press, Barcelona, Spain, 360. https://doi.org/10.1145/1810543.1810614

[51] Tharon Howard. 2014. Journey mapping: a brief overview. *Communication Design Quarterly* 2, 3 (May 2014), 10–13. https://doi.org/10.1145/2644448.2644451

[52] Internet Freedom Foundation. 2022. About IFF. https://internetfreedom.in/about/

[53] Irmi Jenzen, Olu; Karl. 2014. Make, Share, Care: Social Media and LGBTQ Youth Engagement. (2014). https://doi.org/10.7264/N39P2ZX3 Publisher: University of Oregon Libraries.

[54] Olu Jenzen. 2017. Trans youth and social media: moving between counterpublics and the wider web. *Gender, Place & Culture* 24, 11 (Nov. 2017), 1626–1641. https://doi.org/10.1080/0966369X.2017.1396204

[55] Nadia Karizat, Dan Delmonaco, Motahhare Eslami, and Nazanin Andalibi. 2021. Algorithmic Folk Theories and Identity: How TikTok Users Co-Produce Knowledge of Identity and Engage in Algorithmic Resistance. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–44. https://doi.org/10.1145/3476046

[56] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. 2009. A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security - SOUPS '09*. ACM Press, Mountain View, California, 1. https://doi.org/10.1145/1572532.1572538

[57] Leanna Lucero. 2017. Safe spaces in online places: social media and LGBTQ youth. *Multicultural Education Review* 9, 2 (April 2017), 117–128. https://doi.org/10.1080/2005615X.2017.1313482

[58] Ewa Luger, Stuart Moran, and Tom Rodden. 2013. Consent for all: revealing the hidden complexity of terms and conditions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Paris France, 2687–2696. https://doi.org/10.1145/2470654.2481371

[59] Henrietta Lyons, Senuri Wijenayake, Tim Miller, and Eduardo Velloso. 2022. What's the Appeal? Perceptions of Review Processes for Algorithmic Decisions. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–15. https://doi.org/10.1145/3491102.3517606

[60] João Carlos Magalhães and Christian Katzenbach. 2020. Coronavirus and the frailness of platform governance. *Internet Policy Review* 9 (March 2020). https://nbn-resolving.org/urn:nbn:de:0168-ssoar-68143-2

[61] Aida E Manduley, Andrea Mertens, Iradele Plante, and Anjum Sultana. 2018. The role of social media in sex education: Dispatches from queer, trans, and racialized communities. *Feminism & Psychology* 28, 1 (Feb. 2018), 152–170. https://doi.org/10.1177/0959353517717751

[62] Brandeis Marshall. 2021. *Algorithmic misogynoir in content moderation practice.* Technical Report. Heinrich-Böll-Stiftung. 17 pages.

[63] Samuel Mayworm, Michael Ann DeVito, Dan Delmonaco, Hibby Thach, and Oliver L. Haimson. 2023. Content Moderation Folk Theories Among Marginalized Social Media Users. *Under Review for CSCW '23* (2023), 21.

[64] Kerry McNamara. 2003. Information and Communication Technologies, Poverty and Development: Learning from Experience. *The World Bank* (Jan. 2003). https://core.ac.uk/download/pdf/48025103.pdf

[65] Bharat Mehra, Cecelia Merkel, and Ann Peterson Bishop. 2004. The internet for empowerment of minority and marginalized users. *New Media & Society* 6, 6 (Dec. 2004), 781–802. https://doi.org/10.1177/146144804047513

[66] Meta. 2022. Appealed content. https://transparency.fb.com/policies/improving/appealed-content-metric/

[67] Katharine Miller. 2021. Radical Proposal: Data Cooperatives Could Give Us More Power Over Our Data. https://hai.stanford.edu/news/radical-proposal-data-cooperatives-could-give-us-more-power-over-our-data

[68] Ryan A. Miller. 2017. "My Voice Is Definitely Strongest in Online Communities": Students Using Social Media for Queer and Disability Identity-Making. *Journal of College Student Development* 58, 4 (2017), 509–525. https://doi.org/10.1353/csd.2017.0040

[69] Karen Mossberger, Caroline J. Tolbert, and Mary Stansbury. 2003. *Virtual inequality: beyond the digital divide.* Georgetown University Press, Washington, D.C.

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

[71] Jibon Naher, An Taehyeon, and Kim Juho. 2019. Improving Users' Algorithmic Understandability and Trust in Content Moderation. Association for Computing Machinery. https://kixlab.github.io/website-files/2019/cscw2019-workshop-ContestabilityDesign-paper.pdf

[72] David Nosák. 2021. Overview of Transparency Obligations for Digital Services in the DSA. https://cdt.org/insights/overview-of-transparency-obligations-for-digital-services-in-the-dsa/

[73] Jonathan A. Obar and Anne Oeldorf-Hirsch. 2018. The Clickwrap: A Political Economic Mechanism for Manufacturing Consent on Social Media. *Social Media + Society* 4, 3 (July 2018), 205630511878477. https://doi.org/10.1177/2056305118784770

[74] Jonathan A. Obar and Anne Oeldorf-Hirsch. 2020. The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society* 23, 1 (Jan. 2020), 128–147. https://doi.org/10.1080/1369118X.2018.1486870

[75] W. Ian O'Byrne. 2019. Educate, empower, advocate: Amplifying marginalized voices in a digital society. *Contemporary Issues in Technology and Teacher Education* 19 (Dec. 2019), 640–669. https://www.learntechlib.org/p/188279/ Society for Information Technology \& Teacher Education.

[76] Anne Oeldorf-Hirsch and Jonathan A. Obar. 2019. Overwhelming, Important, Irrelevant: Terms of Service and Privacy Policy Reading among Older Adults. In *Proceedings of the 10th International Conference on Social Media and Society*. ACM, Toronto ON Canada, 166–173. https://doi.org/10.1145/3328529.3328557

[77] Oversight Board. 2020. Oversight Board. https://www.oversightboard.com

[78] Oversight Board. 2023. Gender identity and nudity. https://www.oversightboard.com/decision/BUN-IH313ZHJ/

[79] Anette C. M. Petersen, Lars Rune Christensen, and Thomas T. Hildebrandt. 2020. The Role of Discretion in the Age of Automation. *Computer Supported Cooperative Work (CSCW)* 29, 3 (June 2020), 303–333. https://doi.org/10.1007/s10606-020-09371-3

[80] Mikaela Pitcan, Alice E Marwick, and danah boyd. 2018. Performing a Vanilla Self: Respectability Politics, Social Class, and the Digital World. *Journal of Computer-Mediated Communication* 23, 3 (May 2018), 163–179. https://doi.org/10.1093/jcmc/zmy008

[81] Prabha Prayaga, Ellie Rennie, Ekaterina Pechenkina, and Arnhem Hunter. 2017. Digital Literacy and Other Factors Influencing the Success of Online Courses in Remote Indigenous Communities. In *Indigenous Pathways, Transitions and Participation in Higher Education*, Jack Frawley, Steve Larkin, and James A. Smith (Eds.). Springer Singapore, Singapore, 189–210. https://doi.org/10.1007/978-981-10-4062-7_12

[82] Reddit. 2022. AutoModerator. https://mods.reddithelp.com/hc/en-us/articles/360002561632-AutoModerator

[83] Reddit. 2022. Contacting the Admins. https://mods.reddithelp.com/hc/en-us/articles/360002337171-Contacting-the-admins

[84] Pritika Reddy, Bibhya Sharma, and Kaylash Chaudhary. 2020. Digital Literacy: A Review of Literature. *International Journal of Technoethics* 11, 2 (July 2020), 65–94. https://doi.org/10.4018/IJT.20200701.oa1

[85] Joel R. Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James Graves, Fei Liu, Aleecia McDonald, Thomas Norton, Rohan Ramanath, N. Cameron Russell, Norman Sadeh, and Florian Schaub. 2014. Disagreeable Privacy Policies: Mismatches between Meaning and Userss Understanding. *SSRN Electronic Journal* (2014). https://doi.org/10.2139/ssrn.2418297

[86] Shakira Smith, Oliver L Haimson, and Claire Fitzsimmons. 2021. Exclusive: An Investigation into Algorithmic Bias in Content Policing on Instagram | Salty. https://saltyworld.net/algorithmicbiasreport-2/ Section: #MeToo.

[87] Shakira Smith, Oliver L Haimson, Claire Fitzsimmons, and Nikki Echarte Brown. 2021. Censorship of Marginalized Communities on Instagram, 2021. *Salty* (Sept. 2021). https://saltyworld.net/product/exclusive-report-censorship-of-marginalized-communities-on-instagram-2021-pdf-download/

[88] Sanna Spišák, Elina Pirjatanniemi, Tommi Paalanen, Susanna Paasonen, and Maria Vihlman. 2021. Social Networking Sites' Gag Order: Commercial Content Moderation's Adverse Implications for Fundamental Sexual Rights and Wellbeing. *Social Media + Society* 7, 2 (April 2021), 20563051211024962. https://doi.org/10.1177/20563051211024962 Publisher: SAGE Publications Ltd.

[89] Robin Stevens, Stacia Gilliard-Matthews, Jamie Dunaev, Marcus K Woods, and Bridgette M Brawner. 2017. The digital hood: Social media use among youth in disadvantaged neighborhoods. *New Media & Society* 19, 6 (June 2017), 950–967. https://doi.org/10.1177/1461444815625941

[90] Jeanna Sybert. 2021. The demise of #NSFW: Contested platform governance and Tumblr's 2018 adult content ban. *New Media & Society* (Feb. 2021), 1–21. https://doi.org/10.1177/1461444821996715 Publisher: SAGE Publications.

[91] Syrian Archive. 2022. About. https://syrianarchive.org/en/about

[92] Hibby Thach, Samuel Mayworm, Daniel Delmonaco, and Oliver Haimson. 2022. (In)visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit. *New Media & Society* (July 2022), 146144482211098. https://doi.org/10.1177/14614448221109804

[93] Twitter. 2022. Our range of enforcement options. https://help.twitter.com/en/rules-and-policies/enforcement-options

[94] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "At the End of the Day Facebook Does What ItWants": How Users Experience Contesting Algorithmic Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–22. https://doi.org/10.1145/3415238

[95] Jan van Dijk. 2020. *The digital divide*. Polity, Cambridge, UK ; Medford, MA.

[96] Jan A.G.M. van Dijk. 2006. Digital divide research, achievements and shortcomings. *Poetics* 34, 4-5 (Aug. 2006), 221–235. https://doi.org/10.1016/j.poetic.2006.05.004

[97] James Vincent. 2022. Google, Meta, and others will have to explain their algorithms under new EU legislation. https://www.theverge.com/2022/4/23/23036976/eu-digital-services-act-finalized-algorithms-targeted-advertising

[98] Mark Wilson. 2018. The hardest job in Silicon Valley is a living nightmare. https://www.fastcompany.com/90263921/the-hardest-job-in-silicon-valley-is-a-living-nightmare

[99] Anson S. T. Wong and Chul B. Park. 2010. A Case Study of a Systematic Iterative Design Methodology and its Application in Engineering Education. *Proceedings of the Canadian Engineering Education Association (CEEA)* (July 2010). https://doi.org/10.24908/pceea.v0i0.3148

[100] Andrew Zolides. 2021. Gender moderation and moderating gender: Sexual content policies in Twitch's community guidelines. *New Media & Society* 23, 10 (Oct. 2021), 2999–3015. https://doi.org/10.1177/1461444820942483

Received January 2023; revised July 2023; accepted November 2023