Classification and Its Consequences for Online Harassment: Design Insights from HeartMob

LINDSAY BLACKWELL, University of Michigan School of Information
JILL DIMOND, Sassafras Tech Collective
SARITA SCHOENEBECK, University of Michigan School of Information
CLIFF LAMPE, University of Michigan School of Information

Online harassment is a pervasive and pernicious problem. Techniques like natural language processing and machine learning are promising approaches for identifying abusive language, but they fail to address structural power imbalances perpetuated by automated labeling and classification. Similarly, platform policies and reporting tools are designed for a seemingly homogenous user base and do not account for individual experiences and systems of social oppression. This paper describes the design and evaluation of HeartMob, a platform built by and for people who are disproportionately affected by the most severe forms of online harassment. We conducted interviews with 18 HeartMob users, both targets and supporters, about their harassment experiences and their use of the site. We examine systems of classification enacted by technical systems, platform policies, and users to demonstrate how 1) labeling serves to validate (or invalidate) harassment experiences; 2) labeling motivates bystanders to provide support; and 3) labeling content as harassment is critical for surfacing community norms around appropriate user behavior. We discuss these results through the lens of Bowker and Star's classification theories and describe implications for labeling and classifying online abuse. Finally, informed by intersectional feminist theory, we argue that fully addressing online harassment requires the ongoing integration of vulnerable users' needs into the design and moderation of online platforms.

CCS Concepts:

• Human-centered computing: Human computer interaction (HCI) • Human-centered computing: Collaborative and social computing • Social and professional topics: Computing / technology policy

KEYWORDS

Online harassment; classification; labeling; intersectionality; moderation; support; bystanders; social norms.

ACM Reference format:

Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proc. ACM Hum.-Comput. Interact.*, Vol. 1, No. 2, Article 24 (November 2017), 19 pages. https://doi.org/10.1145/3134659

Authors' addresses: Lindsay Blackwell (lindsay.blackwell@umich.edu), University of Michigan School of Information, Ann Arbor, Michigan, United States. Jill Dimond (jill@sassafras.coop), Sassafras Tech Collective, Ann Arbor, Michigan, United States. Sarita Schoenebeck (sarita.schoenebeck@umich.edu), University of Michigan School of Information, Ann Arbor, Michigan, United States. Cliff Lampe (cacl@umich.edu), University of Michigan School of Information, Ann Arbor, Michigan, United States.

2573-0142/2017/11-ART24 $15.00

https://doi.org/10.1145/3134659

1 INTRODUCTION

Roughly four in ten American internet users have personally experienced harassment online, including name-calling, embarrassment, physical threats, stalking, sexual harassment, and sustained harassment [20,21,30]. Online harassment can be particularly devastating for marginalized populations, including women of color and lesbian, gay, bisexual, and transgender (LGBT) people [20,30], limiting their ability to participate safely and equitably in online spaces.

We conducted semi-structured interviews with 18 users of HeartMob, an online platform built by and for people who are disproportionately affected by the most severe forms of online harassment. HeartMob was developed as a grassroots platform under its parent organization, Hollaback!, a non-profit organization dedicated to ending harassment in public spaces. A cornerstone of HeartMob is its community-based approach, which seeks to combat harassment through bystander support—with a focus on amplifying the voices of marginalized internet users.

In this work, we build on extensive prior research that has explored misbehavior in online communities [17,19,40] and mechanisms for moderating and mitigating it [8,27,29]. Our results contribute three insights: first, for harassment targets, labeling experiences as 'online harassment' provides powerful validation of their experiences. Second, for bystanders, labeling abusive behaviors as 'online harassment' enables bystanders to grasp the scope of this problem. Third, for online spaces, visibly labeling harassment as unacceptable is critical for surfacing norms and expectations around appropriate user behavior.

We use Bowker and Star's [7] classification theories and Becker's [5] labeling theory of deviant behavior to better explicate the role of power and social oppression in the classification—whether by technical systems, platform policies, or users—of harassment behaviors as normative or non-normative. We discuss the implications of our results for technical approaches to labeling online abuse, which often fail to address structural power imbalances perpetuated by classification.

Finally, we turn to intersectional feminist theory [1,15,25] to further elucidate the limitations of traditional classification systems and to inform potential alternatives. When a classification system is created with dominant values and morals in mind, the needs of marginalized users are neglected. This framework allows us to bring a CSCW argument to bear on the limitations of current efforts to prevent, manage, or detect online harassment, including platform policies, reporting tools, and machine learning-based approaches. We conclude by advocating for more democratic and user-driven processes in the generation of values that underpin technology systems. Ultimately, centering those who are most vulnerable results in technologies that better address the needs of all users.

2 RELATED WORK

2.1 Online Harassment

Online harassment refers to a broad spectrum of abusive behaviors enabled by technology platforms and used to target a specific user or users, including but not limited to flaming (or the use of inflammatory language, name calling, or insults); doxing (or the public release of personally identifiable information, such as a home address or phone number); impersonation (or the use of another person's name or likeness without their consent); and public shaming (or the use of social media sites to humiliate a target or damage their reputation). These tactics are often employed concurrently, particularly when many individuals, acting collectively, target just one individual (sometimes referred to as "dogpiling"). One individual may also harass another, as is often the case in instances of cyberbullying [3,39] and non-consensual image sharing (also known as "revenge porn"), a form of doxing in which sexually explicit images or videos are distributed without their subject's consent, often by a former romantic partner [12].

Despite a rich history of research exploring online misbehavior [17,19,40] and community moderation [8,29], harassment and other forms of abuse remain a persistent problem online. A Pew survey conducted in 2017 revealed that 66% of adult internet users have seen someone be harassed online, and 41% of users

have personally experienced online harassment [21]. Although these behaviors are instantiated using technology (such as social media sites, text messages, or emails), targets of online harassment frequently report disruptions to their offline lives, including emotional and physical distress, changes to technology use, and increased safety and privacy concerns [21]. Online harassment is also disruptive to everyday life and requires physical and emotional labor from targets, who must spend time reporting abuse in order for platforms to potentially intervene. Indeed, some targets report distractions from personal responsibilities, work obligations, or sleep [23,35]. In addition, online harassment has a chilling effect on future disclosures: Lenhart et al. [30] found that 27% of American internet users self-censor what they post online due to fear of harassment. A 2017 poll of industry and academic experts revealed fears that harassment and other uncivil online behavior will only get worse, resulting in what many fear could be devastating consequences for free speech and privacy [36].

Although men and women both experience harassment online, women "experience a wider variety of online abuse" [30] and are disproportionately affected by more serious violations, including being stalked, sexually harassed, or physically threatened [20]. Young women are particularly vulnerable to more severe forms of harassment: among all young women surveyed by Pew in 2014 [20], 25% had been sexually harassed online (compared with 6% of all internet users), and 26% had been physically threatened (compared with 8% of all users). People of color are also more susceptible to online abuse: 59% of Black internet users have experienced online harassment. 25% of Blacks and 10% of Hispanics have been targeted with harassment online because of their race, compared with just 3% of white respondents [21]. Lesbian, gay, and bisexual (LGB) persons also disproportionately experience more serious forms of online abuse: 38% of LGB individuals had experienced intimate partner digital abuse, compared with 10% of heterosexual individuals [30]. LGB persons are more likely "to feel scared or worried as a result of harassment, experience personal or professional harms, or take protective measures to avoid future abuse" [30].

2.2 Technical Approaches for Addressing Online Harassment

Recent efforts for addressing online harassment have largely revolved around natural language processing and machine learning techniques for the classification of abusive language, enabling automatic detection and prevention. In one of the earliest published machine learning approaches to harassment detection, Yin et al. [51] focus on detecting "intentional annoyance" in discussion-style (e.g., MySpace and Slashdot) and chat-style (e.g., Kongregate) communities. Yin et al. [51] are able to improve a basic supervised model for identifying harassing posts through the addition of contextual features (the context in which a given post occurs) and semantic features (for example, foul language combined with the use of second-person pronouns). More recently, Chandrasekharan et al. [9] draw on large-scale, preexisting data from other online communities (4chan, Reddit, Voat, and MetaFilter) to generate a computational model that, when applied to an unrelated community, can identify abusive behavior with 75% accuracy (and 92% accuracy after the model is trained on 100,000 human-moderated posts from the target community). Similarly, Wulczyn, Thain, and Dixon [50] use a supervised classifier trained on 100,000 human-labeled Wikipedia comments to automatically identify over 63 million personal attacks across the platform. Other automated attempts to detect abusive language online include Google and Jigsaw's Perspective project, an API that uses machine learning models to assign a "toxicity score" to a string of input text [26].

2.2.1 Limitations of Automated Approaches

Though harassment detection approaches have improved dramatically, fundamental limitations remain. As Hosseini, Kannan, Zhang, and Poovendran [26] demonstrate with Google's Perspective, automatic detection efforts are easily outmaneuvered by "subtly modifying" an otherwise highly toxic phrase in such a way that an automated system will assign it "a significantly lower toxicity score." Yin et al. [51] also note that too many spelling errors render the sentiment features of their detection model ineffective. Automated approaches to sanctioning, such as Twitter's 'time out,' frustrate users who perceive them as opaque or unfairly applied. In one case, a transgender woman was sanctioned for including the phrase "Fuck you" in a tweet directed @VP, the government account of Vice President Pence [46]. This event, which some users perceived as a limitation of a citizen's right to criticize government administrations, drew outrage from

users who had unsuccessfully reported racist content from white supremacy groups using the same platform—demonstrating the difficulty of imparting social nuance into the classification of harassing behaviors at scale.

2.3 Classification Has Consequences

Fundamentally, problems that arise from technical approaches to the detection and categorization of harassing behaviors online are a consequence of *classification.* Classification—the foundation of information infrastructures—describes how information is sorted according to recognized patterns to facilitate some improved understanding. However, as Bowker and Star [7] describe, classification is an inherently human process—and as such, it requires human decision-making about what does and does not belong in a given category. Classification becomes a concern when labeling decisions are made with little consideration of the biases inherent in—and thus, risks associated with—those decisions. Still, technology systems rely on classification as a necessary and robust mechanism for scalability: database systems, user accounts, and social media profiles all require the labeling and categorization of people into a series of digital bits.

Bowker and Star [7] emphasize that classification systems embody moral choices that reflect greater societal values. They cite 1950's South Africa, in which the Population Registration Act and Group Areas Act required that people be classified by racial group and constrained as to where they could live and work—the precursor for the brutality of South African apartheid. While extreme, we see how classification systems can valorize prevailing or dominant points of view while silencing others [7]. Power, then, is held by those who are creating the labels.

More recently, we see how classification is inherent in everyday technological systems and that classifications are largely socially constructed. For example, in 2014, Facebook revised its gender field from "male" and "female" to allow 56 additional options, including transgender, gender non-binary, and gender questioning identities. As of 2017, Facebook offered 71 total gender categories for users to choose from, including a custom field for users to define themselves [49]. From a social justice perspective, this change is welcome and embraced. However, practically, Facebook still needs to serve advertisements as part of its business model, for which gender is inevitably one feature. As a result, users may not in fact know how their chosen gender categories are subsequently categorized by Facebook's algorithms in order for Facebook to more effectively target advertisements.

In this work, we explore the opportunities and limitations of classification in the domain of online harassment. Using HeartMob as a case study, we suggest that classification can motivate bystanders to provide support, both by demonstrating the breadth and scale of harassment experienced online and by surfacing clear avenues of support. We also show how classification can invalidate harassment experiences, particularly for users whose experiences do not fall within the bounds of a 'typical' harassment experience. As argued by Bowker and Star [7], classification systems embody the morals and values of their creators—whether in the context of a technical system, such as Facebook's automatic detection of harassing language, or a platform policy, such as Twitter's categorization of what constitutes abusive behavior. When a classification system is created with dominant values and morals in mind, the needs of marginalized users are neglected. To reconcile these tensions, we turn to intersectional feminist theory to further elucidate the limitations of traditional classification systems and to inform potential alternatives.

2.4 Intersectional Feminist Theory and Practice

Intersectional feminist theory [1,15,25] holds that various identities (such as race, gender, class, sexuality, religion, disability, and nationality) are inextricably bound in systems of entrenched structural oppression (such as racism, sexism, cissexism, classism, heterosexism, ableism, and colonialism). These systems of oppression "intersect" and cannot be understood independently of each other [15]. For those who have many identities that have been historically dominated, the effect of various oppressions is intensified [15,25]. Crenshaw—who coined the term intersectionality—illustrates that under United States

law, women are assumed to be white, and Black people are assumed to be men [16]. As a result, there are severe gaps in legal protections for women of color, which white women do not experience [16]. Thus, the experiences of Black women cannot be understood as additive (e.g., Black + woman), but must instead be understood as intersecting, interdependent, and mutually constitutive [16].

For example, two of HeartMob's co-founders, Courtney Young and Debjani Roy, are women of color who reside in the United States. As a result, they experience oppression both offline and online based on their gender as women and as being African American and South Asian, respectively. Roy's transnational experience as an immigrant from the U.K. to the U.S. has also shaped her experiences. In contrast, three of the present work's authors are women and are white or pass as white. They experience oppression based on their gender identity as women, but not based on their perceived race or immigration status. Intersectionality offers a lens through which to understand how the oppression women of color experience can be both different from and more acute than that of white women. In the context of the current work, intersectionality helps to position individuals' experiences of online harassment as both reflective of and inextricable from systems of structural oppression, such as racism, sexism, cissexism, and so on. This framework allows us to 1) better understand the limitations of current approaches to online harassment and 2) consider user-driven alternatives.

2.4.1 HeartMob

HeartMob (iHeartMob.org), launched in January 2016, is a private online community designed to provide targets of online harassment with access to social and instrumental support. HeartMob was created by leaders of Hollaback!, an advocacy organization dedicated to ending harassment in public spaces. Hollaback! leaders and their colleagues experienced consistent and often severe harassment online as a result of their work. Given their collective expertise in intersectional feminist practice, social movement framing, and bystander intervention to combat harassment in physical spaces [18], the Hollaback! team sought to translate these practices online to support others with similar experiences. Their goals were 1) to understand if, and to what extent, online tools could help create communities of accountability, and 2) to organize people who witness online harassment (i.e., bystanders) to provide support to harassment targets.

HeartMob was designed with an intersectional feminist underpinning: it was built by and for people who are disproportionately affected by the most severe forms of online harassment due to their intersecting oppressions. In late 2014, the team convened a diverse group of journalists, academics, and feminist activists who had all experienced severe online harassment. In the workshop, the second author organized speculative design activities so that participants were directly informing the system's ultimate design. The design of the system went through several design iterations, with continuous feedback from the convening community. As a consequence, the design team learned the ways in which actual targets of harassment wanted to be helped, including supportive (and moderated) messages, a space to document their experiences, and assistance reporting harassment to the platforms on which it occurs. Additionally, participants wanted the option to make their harassment experience public, including a description of the perceived motive.

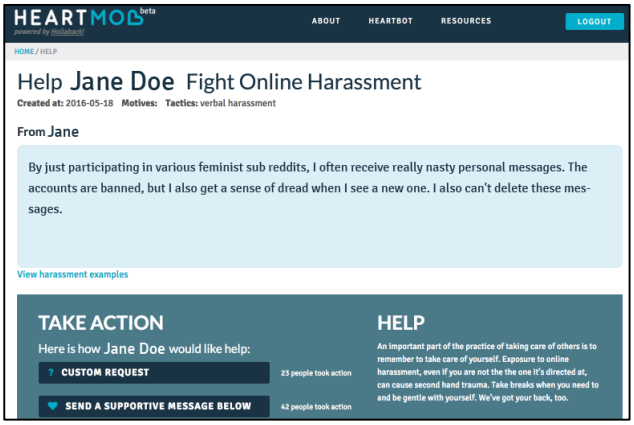
2.4.2 How HeartMob Works

The resulting HeartMob system was designed around the specific needs of people who had experienced online harassment. After creating an account, a person experiencing harassment can submit a harassment case, which includes a description of their experience, the type of harassment (e.g., stalking or doxing), the perceived motive (e.g., racist or misogynist harassment), and any screenshots or additional documentation (see Figures 1 and 2). The user can optionally create a help request and specify the type of help they would like (for example, supportive messages or assistance reporting harassment). The user account, harassment case, and help request are all moderated by trained HeartMob employees. The target can optionally make their harassment case public outside of the HeartMob community to make online harassment more visible.

Users can also apply to be a "HeartMobber," or a bystander who supports targets of harassment by fulfilling their requests for help. HeartMobbers are vetted using a tiered trust system. Level 1 HeartMobbers are only required to verify one social media account but are still moderated. As a result, they are limited to

the information they can see about a particular help request until they are accepted as a "Trusted HeartMobber." In order to submit an application to be a "Trusted HeartMobber," users must choose a combination of disclosing additional data and meeting certain criteria based on account duration and approved community participation.

As of February 2017, there were a total of 1,455 approved HeartMobbers and 98 approved target accounts, producing a total of 86 approved harassment cases and 71 associated help requests. Having a help request is not required, as some targets simply want to share their story. Targets may also request help from individual people instead of (or in addition to) the HeartMobber community. There have been a total of 4,555 actions taken by HeartMobbers on targets' help requests. Supportive messages are the most popular type of help provided on HeartMob, with a median of 41.5 messages and 46.5 "Got Backs" (similar to the Facebook 'Like' button) per help request (see Table 1). Although targets most frequently request help reporting abuse to platforms, there are not as many actions taken by HeartMobbers as supportive messages. This is likely due to the various challenges of reporting content to platforms, which will be discussed further in the results.



Figures 1 and 2: HeartMob Categories for Perceived Harassment Motives; Anonymized HeartMob Help Request.

	Percentage of cases with a request for this action	Median number of actions taken per case
Supportive Messages & 'Got Backs'	67.6%	41.5, 46.5
Reporting Abuse	70.4%	2.5
Documenting Abuse	46.4%	1
Other	0.7%	9

Table 1: HeartMob Supportive Actions.

3 METHODS

We conducted semi-structured interviews with 18 users of the HeartMob system. As of February 2017, HeartMob had 1,455 total users. Participants were recruited via an email blast sent to current HeartMob users (e.g., anyone who had created an account) beginning in September 2016. All HeartMob users aged 18 and older were invited to participate, whether or not they had personally experienced online harassment. Four email blasts were sent in total, with the last recruitment email sent in January 2017. 25 HeartMob users expressed interested in participating; 18 users ultimately completed an interview. All participants were located in the United States or western Europe, and interviews were conducted in English.

Before the interview, each participant was asked to review and sign an online consent form, which included a short demographic survey. Participants were asked their age (open response); gender (open response); whether they identify as transgender (yes or no); sexual orientation (Heterosexual or straight; Gay or lesbian; Bisexual; or Please specify); and race (White; Hispanic, Latino/a/x, or Spanish origin; Black or African American; Asian; American Indian or Alaska Native; Native Hawaiian or Other Pacific Islander; or Please specify). Participants ranged in age from 28 to 71 years, and the median age was 40.5. 10 of our 18 participants identified as heterosexual; one participant identified as hetero-poly (for *polyamorous,* or someone who consensually pursues multiple sexual or romantic relationships). Four participants identified as gay or lesbian, one as queer, one as bisexual, and one as pansexual. The high number of non-heterosexual participants in our sample may be partially explained by the increased likelihood of LGB people to experience harassment online [30].

	Age	***Gender Identity***	***Sexual Orientation***	***Race***
P1	28	Male	Gay or Lesbian	White
P2	64	Female	Heterosexual	White
P3	43	Male	Heterosexual	White
P4	34	Female	Bisexual	White, Hispanic
P5	36	Male	Gay or Lesbian	White
P6	65	Female	Gay or Lesbian	White
P7	45	*	Gay or Lesbian	White
P8	28	Male/They	Hetero-poly	Italian, Irish
P9	30	Female	Heterosexual	White
P10	42	Female	Heterosexual	White
P11	31	Female	Queer	White
P12	57	Male	Heterosexual	White
P13	51	Female	Heterosexual	White
P14	48	Female	Pansexual	White
P15	71	Male	Heterosexual	White
P16	35	Male	Heterosexual	White
P17	31	Female	Heterosexual	White, Black
P18	39	Female	Heterosexual	Asian

Table 2: Participant Demographics (Self-identified).

We provided an open text field for participants to identify their gender. Ten participants self-identified as female; 7 participants self-identified as male. One self-identified male participant, P8, included a pronoun preference, and will be identified in the paper as they/them. The remaining participant (P7) chose not to

identify a gender*, saying that "gender is a system of oppression, not an identity." Instead, she identified her sex as female. None of our participants identified as transgender. Although transgender and gender non-conforming people are likely to endure severe harassment online due to systemic transphobia, misogyny, and homophobia [13], these experiences remain unrepresented in recent online harassment research (e.g., [30]), including in the present study. This is an important limitation that should be explored in future research.

Interviews were conducted between September 2016 and January 2017. Interviews lasted an average of 52 minutes; the longest interview lasted 120 minutes and the shortest 19 minutes. Generally, interviews with participants who had experienced online harassment themselves were longer than interviews with participants who hadn't, due to the personal nature of online harassment experiences and the emotional labor of recounting them. All interviews were conducted and recorded using BlueJeans, a secure online video service. Participants were not compensated. This study was approved by an accredited Institutional Review Board.

Most users had experienced some form of online harassment themselves (n=11; P2, P3, P4, P5, P6, P7, P8, P11, P13, P17, P18), but not all of these participants had submitted a case to HeartMob (some were motivated by their own harassment experiences to join HeartMob as a supporter). We asked these users about their harassment experiences, what support they received (if any), and the process of using HeartMob to seek support (if they had done so). Other participants joined HeartMob purely in a supportive role (n=7; P1, P9, P10, P12, P14, P15, P16); we asked these users about their motivation to join HeartMob, whether they felt a sense of community with other HeartMob users, and about the process of offering support to people experiencing online harassment. We asked all users what the phrase *online harassment* means to them and about their impressions of the HeartMob site. We sought to elicit specific experiential narratives from our participants through the use of general questions centered on specific emotions (e.g., "Tell me about your most recent experience with online harassment, or about an experience that was particularly difficult" or "Tell me about a time where you felt the support you provided was helpful"). Last, participants were asked what additional support HeartMob and other platforms (for example, social media sites such as Facebook and Twitter) could provide to users like them.

Interview recordings were transcribed by Rev.com. We used an inductive approach to develop codes [42]. One member of the research team individually read through interview transcripts and noted codes by hand. After discussing these initial codes as a research team, we created a more comprehensive list of codes (61 codes in total). Resulting codes were organized around several themes, including but not limited to defining online harassment, impacts of harassment experiences, changes in technology or privacy use, seeking or providing support, and visibility and audience. Two researchers each coded four interview transcripts in a pilot coding process to test and refine the codebook. We coded interviews using Atlas.TI, frequently discussing codes to maintain agreement. Each interview transcript was coded by two members of the research team. Quotations have been lightly edited for readability.

3.1 Position statement

Per feminist methodology, the authors recognize the importance of positioning the research team in relation to the current work and our analysis [4,48]. All of the authors have been close with someone who has experienced online harassment, and two of the authors have experienced it personally. Three of the authors are women, and one is a man. All of the authors identify as white. As our sample is also majority white, the absence of experiences from or interpretations by people of color is a significant limitation: because of cultural racism, people of color face different kinds of harassment online than do white people. Further, women of color experience a unique intersection of racist and misogynist online harassment, which is different from the experience of racist harassment or the experience of misogynist harassment. This experiential understanding is notably absent from the current work. Similarly, all authors are cisgender, and none of our interview participants identified as transgender. More research should explore and amplify the online harassment experiences of trans people.

4 RESULTS

Results are organized around three major themes: first, for targets, labeling experiences as 'online harassment' provides powerful validation of their experiences. Second, for bystanders, labeling abusive behaviors as 'online harassment' enables bystanders to grasp the scope of this problem. Third, for online spaces, visibly labeling harassment as unacceptable is critical for surfacing norms and expectations around appropriate user behavior.

4.1 Labeling Validates Targets' Harassment Experiences

When a user submits their harassment experience to HeartMob, they select the type of harassment (e.g., stalking or doxing) and the perceived motive (e.g., racist harassment or misogynist harassment) from drop-down lists. HeartMob moderators then individually approve each case of online harassment an individual user submits. Participants felt validated when their experiences were accepted and labeled as 'online harassment.' P5 said:

> "It's the safety net. Right now, the worst that can happen is somebody experiences harassment and they have nowhere to go—that's what's normal in online communities. With HeartMob, if someone says they're experiencing harassment, then at least they get heard… at least they have an opportunity to have other people sympathize with them."

P11 referred to HeartMob as a means of "harm reduction"—she said it "doesn't have the capacity to single-handedly solve the problem, but it makes being online bearable." P9 said that even though she had not experienced harassment online, offering support to others on HeartMob made her feel safer: "Personally, it makes me feel a little safer. Even though I've never used this as someone experiencing harassment, I know that it's there—it's comforting to know there's this network." Similarly, P8 said that supporting other users made him feel like HeartMob is "something I can turn to."

Many participants expressed a preference for support from users who empathized with their unique experiences. Some participants felt that HeartMob's system of labeling enabled them to find other users with similar experiences or shared identities. This finding underscores the importance of understanding harassment through an intersectional lens, as different groups of people experience oppression differently. P13 felt that HeartMob provided immediate access to social support from people who understood the impacts of her specific experiences:

> "I feel like what HeartMob is doing is so instrumental in keeping people from going off the deep end—from feeling alone in this. Most people don't quite understand… how it invades every aspect of your life, basically, when this happens to you. Even my friends—they knew on a daily basis what was going on, and they still couldn't really grasp it."

Other participants felt that more targeted support could help them better prepare and protect themselves in the future. P11, who had several experiences with sustained, high-volume attacks (e.g., dogpiling), said that during her first experience with harassment, she had no idea "how scary it is to see hundreds and hundreds of people wishing death upon you." During subsequent experiences, knowing what she could expect helped P11 to "rally her troops" and preemptively seek support for potential harassment:

> "I was able to tell the folks around me, 'Hey, this is gonna be really rough, so I'm gonna need you to send me some love. If you see an article posted on someone's Facebook and you see nasty comments, I'm gonna need you to come to my defense.' People really came to my defense, which was incredible to me… I didn't have that [the first time], because I had no idea what was coming."

P13, whose former partner created a number of defamatory websites that had disrupted her personal and professional life for years, felt isolated from her friends and family and had a difficult time accessing relevant information and legal resources. After joining HeartMob to support other targets of online harassment, P13 suggested that a system for categorizing submitted cases according to specific harassment behaviors and their impacts could help connect isolated users with the support they need most: "I think the person would feel a little more connected. I feel like we're all in silos."

4.1.1 Classification Privileges Dominant Experiences

Participants like P13, who was not able to use HeartMob to identify others who shared her circumstances, were likely to minimize the impact of their own experiences. Often, these participants made comparisons to harassment experiences they had observed or perceived others to be experiencing. P3, a 43-year-old white man, had lost several employment opportunities due to defamatory information posted about him online by a former associate—and had spent thousands of dollars trying to expunge the defamatory information from his online record. P3 joined HeartMob specifically to access other users who could directly empathize with his experiences: "When I registered with HeartMob, my only intention was to get some kind of support network... I hoped to identify with someone that's dealing with something similar to me. I don't think there are a lot of people out there that experience what I experienced."

P3 said that first joining HeartMob had been difficult, in that he didn't know how to best categorize his experience using the system's available labeling tools. Ultimately, P3 felt that his inability to use the HeartMob system to identify users experiencing similar abuse signaled that his experiences might not 'count' as online harassment:

> "Trying to find the right checkbox to categorize yourself is tough sometimes… when someone thinks of online harassment, they don't think of what I've been going through. I don't even think people would define my case as online harassment. I know there are people out there that just don't care what I'm going through."

P3 ultimately had difficulty soliciting the unique support he needed from the HeartMob community, and he suggested—based on other cases he had seen on HeartMob—that other, more marginalized users were more deserving of the community's support:

> "The lesbian and gay community, how they are discriminated against online and harassed—that's more important than my situation. My case probably seems very small and insignificant, considering. I don't think the world would feel sorry for someone like me."

For P8, whose middle school classmates had once impersonated him online to harass a favorite teacher, reading other users' harassment cases on HeartMob made him doubt the severity of his own experiences: "I've never experienced anything where I have felt threatened for my safety… nothing like what I see on HeartMob." P13 similarly minimized her own harassment experiences. When P13 ended a romantic relationship, her former partner created a defamatory website suggesting P13 had participated in prostitution. He included P13's contact information and professional history, and circulated the website to over 300 of P13's friends, family members, colleagues, and clients. P13 had been working with local and federal law enforcement and with Google for more than four years to have the defamatory sites removed. Still, P13 said: "At one point there were 14 websites. I feel like 14 is a lot—but as I read other people's stories [on HeartMob], I realize it could be a lot worse."

4.1.2 Labeling Reflects Community Norms

Some participants did not see their own identities reflected in the cases ultimately posted to HeartMob (i.e., approved by moderators), and as a consequence, they questioned whether or not their experiences belonged—and by extension, whether or not they belonged—on the site. When asked whether she had sought support on HeartMob for her own harassment experiences, P2, who is 64 years old, said she hadn't considered it because the interactions she had witnessed on the site left her with the impression that most users were significantly younger: "It didn't occur to me… I'm there to provide support. The bulk of people I [am supporting] are much younger."

Some participants felt that system labels on HeartMob privileged certain perspectives over others. P7 had applied to become a trusted HeartMobber, but had been denied—which P7 suspected was a direct result of writing publicly about her controversial views on gender, which others had characterized as abusive. P7 felt that she had been unfairly characterized by HeartMob, and as a result, she felt she would not be welcome to solicit support on HeartMob for her own harassment experiences:

> "I think that HeartMob comes from this leftist liberal mindset. When people who have political views that are supported by a community engage in that type of behavior, they're praised. When

> people like me—who have unpopular views—engage in that type of behavior, we're accused of abuse. I would hazard a guess that if someone like me went to [HeartMob] and was like, 'I'm really being harassed,' they probably wouldn't help me, honestly."

Participants also sought ways to more easily locate users who shared their social identities: P5 joined HeartMob specifically hoping to support other LGBT users experiencing harassment online, but he could not locate experiences similar to his own experiences of harassment and exclusion on Wikipedia: "I could browse other people's stories, and there were no LGBT stories in the queue. I thought I might have wanted to put my story in an LGBT queue, but there was no such thing at the time." Most participants felt that support services for people experiencing online harassment should welcome a diversity of users and perspectives. Said P7: "I think they should make it very clear that they are agnostic about who they support, and that it doesn't matter what you believe. If you've been abused, they'll support you. I think that would be great." However, this perspective may be at odds with HeartMob's original design goals, which prioritize marginalized users and intentionally script these values into the design of the system itself.

4.1.3 System Labeling Invalidates Harassment Impacts

Participants also felt invalidated when their experiences were outright rejected by the system, or otherwise labeled as *not* harassment. This was particularly salient for social media users, as most major social media companies rely on scripted responses that do not acknowledge individual experiences or the impacts of being harassed. P17, who endured large volumes of harassment as a result of her work as a writer, felt that the labor required to report ongoing harassment was "completely useless." She continued: "There's really no point in reporting stuff on social media. Last time, I spent several hours going through and reporting tweets. It felt like maybe less than 10% were considered threatening—and either they had their account indefinitely suspended, or just suspended until they took the tweet down." Reporting was particularly frustrating if reported content was found not to be in violation of existing policies, which many participants felt was a frequent occurrence. P17 went on to say:

> "What I think was really frustrating was the level of what people could say and not be considered a violation of Twitter or Facebook policies. That was actually really scary to me—if they're just like, 'You should shut up and keep your legs together, whore,' that's not a violation because they're not actually threatening me. It's really complicated and frustrating, and it makes me not interested in using those platforms."

Even when social media platforms removed abusive content, participants felt the process by which a verdict was reached (or a sanction determined) was obscure. After P18 and her colleagues had publicly launched a new product, some users threatened their physical workplace. When P18 was sent an image on Twitter of a man pointing a sniper from a rooftop, she reported the tweet: "We did ask Twitter to take that down, and they did—but I don't know what they did with the person who posted it. I don't remember what happened with that." The challenges of reporting abuse directly to the platforms where it occurs were similarly frustrating for participants who wished to support harassment targets, which may explain why so few HeartMobbers provide this type of support—despite it being the type of help most frequently requested by targets (See Table 1). P9 had tried to help a target report abuse, but the content was no longer available: "I went online to the link the person had posted, but all of the posts were already gone. I guess somebody had—somebody must've already reported it. You could see that two people had responded to the request, but that was kind of the end of it." This gap between system labels and user experiences led many participants to wonder whether targets were receiving the support they requested, or whether they were doing enough to support them. P14 said: "There was somebody who had some really significant, terrible comments. I did worry that my words of hope and encouragement in that situation weren't enough. I felt I could have been far more helpful."

4.2 Labeling Helps Bystanders Understand Targets' Harassment Experiences

We find that for bystanders, labeling the variety of abusive experiences enabled by technology as 'online harassment' helps them understand the breadth and impact of this problem. For P14, participating as a HeartMobber changed her perceptions about the severity of online harassment. P14 felt that providing support on HeartMob helped her better understand the breadth of online harassment experiences:

> "In the work that I do, I help people understand how society plays a role in the violent culture that we have—but I've never had too much opportunity to actually see the evidence on the internet. I knew it was there. I talk about it, present about it, but actually seeing the horrific things that people are seeing and doing to others online really brought that to a whole different place for me."

P11 suggested that visible disclosures of harassment experiences, like the collection of cases available on HeartMob, could encourage targets to seek support:

> "The second someone's vulnerable about their experience with violence online, it creates this sense of community amongst others who've also had that experience. It's like, oh my gosh, you too—we're some of the only people talking about this, so we have to have each other's back."

Still, participants worried that targets may make themselves vulnerable to additional abuse by publicly disclosing their experiences. P9 felt that even a protected space like HeartMob could jeopardize targets' ultimate safety and comfort online: "If somebody hacks in, it's so much more traumatizing if you get harassed from within that space."

HeartMob also exposed users to the diversity of abusive behaviors used to target people online. P17 expressed that online harassment could include a range of potential experiences: "It's anything—email, direct messages, in the comments. It's the doxing of people's information. It's sustained threats, and sometimes just little ones here and there. It can happen either once or over a sustained period of time." P3, who had experienced an atypical form of online harassment, acknowledged that it was difficult for others to empathize with his experiences: "I think until you live this nightmare I've been living with, you just don't know."

P11 suggested that the problem of online harassment is particularly insidious because bystanders—particularly those who have not experienced harassment themselves—feel powerless to help. By classifying different harassment experiences and providing specific, labeled ways in which to provide support, P11 felt HeartMob made it easier for bystanders to offer support: "HeartMob is a brilliant way of addressing a problem that I think immobilizes most people, because it seems so big and daunting—so they don't do anything at all." P1, who worried that his family members and friends did not consider online harassment to be a 'real' problem, felt his participation in HeartMob was a good first step toward speaking out against online harassment in more public spaces: "I think it's very good participating in these cases while not exposing yourself to abusers in public spaces. I think of HeartMob as the stepping stone to participating in the public spaces."

4.3 Labeling Helps Define Responses to Harassment

Last, we find that visible resistance to harassment in online spaces is important not only for targets, but for surfacing norms about what is and is not acceptable behavior. P11 said that online harassment is not taken seriously as a problem, and that in particular, people do not recognize the full breadth of the problem:

> "When I think of the phrase online harassment, I think about death by a thousand cuts. I think about how we either don't take it seriously as a society, and we think it's just the internet and turn it off, or we view individual tweets or emails or comments in isolation—we don't view the breadth of it, recognize the avalanche."

4.3.1 Resisting Normalization

For many participants, seeing other users support harassment targets on HeartMob provided visual evidence that demonstrated the full scale of user resistance to the problem of online harassment. This community opposition to online abuse was not always apparent on other platforms. P9 noted that online harassment can be particularly isolating when targets are exposed to significant volumes of abuse, but receive comparably limited support:

> "It's something that's very isolating, because it can make you feel—especially if there are multiple people doing the harassing—like everybody would be against you… like they're representing society."

P9 went on to say that harassment perpetrated by anonymous or pseudonymous users can seem representative of society at large: "Because it's anonymous, it can give you this sense that they're representative. It makes you lose faith in humanity. There are just so many people devoting their time to bringing other people down."

Participants who regularly experienced harassment and abuse online admitted to feeling like their experiences had become normative. P7 said that she had become indifferent to online harassment over time, due to the volume of threats she had experienced:

> "It's like exposure therapy. After time, it's alright. You're able to put boundaries around what's a real threat, what isn't actually a real threat—even though it might hurt my feelings, and even though I might feel my body react or my heart raise."

P17 agreed: "It's annoying, because I'm getting used to it. It's just becoming the norm for how our society conducts itself, how people discuss things and interact." P11, a writer, said that after witnessing so many others experience online abuse, her first experience felt familiar: "There was this familiarity… it was like I had been initiated into the experience." Her friends were unsurprised or even apathetic, P11 said: "I was met with a lot of like, yeah, welcome to being a woman online. Welcome to engaging with American media. Welcome to the club. There was this apathetic, cynical response that I found very discouraging." Participants felt that more visible resistance to harassing behaviors online would discourage these types of experiences from becoming normalized.

Participants also worried that the normalization of harassing behaviors online impacted other users. P17 worried about the impact of her public harassment on other users: "I definitely get messages from people who are like, 'I want to share my story, but I see what you go through.'"

4.3.2 Reclaiming Space through Labeling

Traditionally, social movements have taken to public spaces, such as streets, to reclaim space as an act of resistance. Similarly, participants sought to reclaim online spaces, where abusive behaviors were perceived to occur frequently. For participants who had experienced harassment, social media site use became undesirable or even frightening. P17, who had taken a hiatus from social media sites as a way to avoid ongoing harassment as a result of her work as a writer, said: "The work that I do, I have to use [social media]. It is not a tool that is fun for me anymore. It is a tool for work." P17 discussed the need for a tool to mitigate high volumes of public abuse, "to break up the monotony of the hatred." Specifically, P17 felt overwhelmed by the sheer amount of abuse visible in her social media notifications:

> "You know that scene [in Harry Potter] where Harry uses his Patronus for the first time, and it knocks out all the Dementors? It's this white light that pushes everybody back. I wish there were—I don't know the social media equivalent of that. I need a Patronus right in the moment to just push everybody back."

Participants perceived online spaces as being overrun by harassment and abuse, with little visible resistance. P13 said: "The internet has become the town square where people are taken to be embarrassed and punished. The only difference is the whole world can see it, not just the town. Then it stays forever—it doesn't go away." P7 felt that it was important to define the line between abuse and disagreement on social media sites, particularly in a divisive political climate:

> "In order to have a free and fair and just society, all people need to feel empowered to speak. It gets tricky on social media, when people are designating certain things as abusive speech that are actually political disagreement. For example, I'm a Democrat. Someone who's a hardcore Trump supporter, I might not like them, I might think that they're stupid, but they have every right to their opinion, and it's not hate speech for them to post that they think Trump should build a wall. I might think it's racist, I might think it's wrong, but it should be permissible for people to say that."

P12 felt that it was important to emphasize civility online: "We've got to start taking back the internet. It's got to come back to what it was—a place of information, a place of sharing ideas. If I go on a website and I don't like what they're talking about, I can leave. I don't have to attack them." P13 felt that seeing greater resistance to harassment in online spaces would encourage her to express visible support for targets: "If I go online to support a person, and then I look at the other messages to that person—I think, 'Oh, that's neat. There are other people out there doing it, too.'" One way to reclaim online spaces is through making norm violations visible and labeling them as harassment. Many participants reflected on the relationship between a perceived increase in harassing behaviors online and the current political climate at the time of data collection. Several participants (n=5; P13, P17, P2, P145, and P10) specifically mentioned Donald Trump as having participated in or incited online harassment, including his criticism of Chuck Jones on popular social network site Twitter. Jones, the president of United Steelworkers 1999 (a labor union), suffered substantial harassment following Trump's tweet [38]. P17 felt President Trump's comportment would normalize similar behaviors:

> "I just had this moment of, 'Oh my god; this is not going to end.' I had this realization… it's happening from the top, and we now have a president who condones this. That terrifies me. There are millions of Twitter users who see that this behavior he's doing is okay—that this is how you conduct yourselves when talking to people who disagree with you."

P5 felt that social media platforms could also benefit from exposure to a diverse corpus of documented harassment experiences. P5 was concerned that internet companies may not be equipped to fully understand the experiences of their diverse users: "For example, the Wikimedia Foundation—they're based in San Francisco, but Wikipedia gets international harassment problems. If harassment happens in, say, India, white people in San Francisco don't really know what it's like to be a teenage gay boy in India." P5 felt that companies could be doing more to partner with activist platforms or support organizations where diverse harassment experiences can be visibly aggregated. This desire to reclaim civility in online spaces suggests that labeling behavior as abusive and unwelcome plays an important role in defining normative responses to harassment within a given community.

5 DISCUSSION

Classification—whether instantiated by technical systems, platform policies, or users themselves—plays a critical role in validating and supporting online harassment experiences, as well as enabling bystanders to intervene. We discuss our results through the lens of Bowker and Star's classification theories and Becker's labeling theory of deviant behavior. We discuss the implications of our results for current approaches to labeling and classifying online abuse, such as platform reporting tools and machine learning techniques for the automatic identification of harassing language. Finally, we present alternatives for representing diverse experiences in online systems.

5.1 Surfacing Social Norms through Visible Classification

Many of our participants discussed feeling uneasy about or apathetic toward the ways in which abusive behaviors are seemingly becoming normalized online. Participants who were harassed publicly (e.g., dogpiling on social media sites or defamation in the media) often desired more public demonstrations of support than the HeartMob system could realistically provide. This result suggests that visibly labeling harassing behaviors as inappropriate—whether by users or by the system itself—can make opaque, system-driven classification systems more visible, and consequently may help users to identify and define the boundaries of appropriateness in online spaces.

5.1.1 Visible labels create powerful descriptive norms

Empirical evidence suggests that injunctive norms—expectations for how you *should* behave, such as those articulated in a platform policy—are often less powerful than descriptive norms (how most others behave) in encouraging behavior change [11,21]. Cialdini [11] argues that descriptive norms offer "an information-processing advantage," in that by understanding how most people behave in a given situation, an individual can more quickly decide how to behave themselves. Similarly, Marwick [32] argues that

social media users "monitor each other by consuming user-generated content, and in doing so formulate a view of what is normal, accepted, or unaccepted in the community." Users who are uncertain about how to behave will adapt to visible descriptive norms, particularly in cue-sparse online environments [45].

Cheng et al. [10] found that perpetrators of online harassment were not, contrary to popular narratives, anti-social actors dedicated to violating norms of civility. Rather, people could be primed to harass others online in an experimental setting when earlier instances of harassing behavior were made visible in a comment thread. Where harassment is visible—and any sanctioning of that harassment is not—then the descriptive norm could easily become that harassment is an appropriate activity. If a new Twitter user is exposed to high volumes of harassing content on the platform with little visible resistance (e.g., platform- or user-issued sanctions), for example, the user may determine that harassing behavior is appropriate on Twitter. Our results suggest not only that visible sanctions serve as important validation for harassment targets, but also that public demonstrations of support may help other users determine what behaviors are and are not appropriate in a given space.

5.1.2 Visible classification penetrates the "fog of audience"

Computer-mediated communication can be cue-sparse and persistent, making it difficult for users to assess who has seen what and when [6,37]. In discussing the implications of using technology to engage in social surveillance, Marwick [32] argues that networked online platforms give users an ambient awareness of others, but obscure critical details such as social context and differences in power—a phenomenon which Lampe [28] refers to as the "fog of audience." A digital audience is "large, unknown and distant" [34], and users must continually negotiate their own privacy and impression management practices to satisfy both their own expectations and those of a functionally invisible audience [31,34,44]. Any one user cannot reliably determine the experiences or expectations of all others in the network, making it difficult for users to assess the norms of a given community.

The difficulty in detecting the norms around harassment on a social media platform is further complicated by policy-driven classifications, which are not typically made visible to users. Many online platforms rely on policies to enforce formal (i.e., codified) norms for what is and is not appropriate behavior when using their services. Although the policies themselves are accessible to users, how and why those policies are actually *enforced* is more opaque—and any one user is typically unaware of how other users expect policies to be applied, or how they experience policy enforcement. In the absence of transparency, users are left to decide for themselves why platforms make certain choices, and may consequently ascribe values to a system that its creators did not intend. For example, in June 2017, ProPublica [2] revealed that U.S. congressman Clay Higgins' Facebook post calling for the slaughter of "radicalized" Muslims was not removed by the site's content moderators. Conversely, a post by Black Lives Matter activist Didi Delgado—"All white people are racist. Start from this reference point, or you've already failed"—was removed, and resulted in a week-long account suspension. Although it is unlikely Facebook's content moderators made any direct comparisons between these two posts when deciding how best to enforce their policies, the obscurity of Facebook's classification system leaves users to draw their own conclusions about what is or is not appropriate on Facebook and why.

Importantly, Bowker and Star [7] argue that classification systems—though ubiquitous in everyday lives—are typically invisible, and thus people remain largely unaware of the social and moral order they create. Classification systems are typically made visible only "when they break down or become objects of contention" [7]. As documented by ProPublica, Facebook content moderators are trained to recognize hate speech as curses, slurs, calls for violence, or other attacks directed at "protected categories," which are rooted largely in Western legal definitions (i.e., sex, race, ethnicity, sexual orientation, gender identity, religious affiliation, national origin, and serious disability or disease). However, factors such as social class, age, and occupation are *not* protected under U.S. law, nor are categorical factors such as religions or countries (in other words: attacking an individual's religious affiliation is not allowed, but attacking a specific religion is). As a result, ProPublica argues, Facebook's algorithm is designed to "defend all races and genders equally"—an approach which Danielle Citron, quoted within, argues will "protect the people who least need it and take it away from those who really need it." Though Facebook was the focus of this particular report, most online platforms—including Twitter, Craigslist, and others—must present publicly available policies (e.g., community standards) while also engaging in often "invisible" classifications to effectively enforce these policies at scale. This obscurity not only contributes to uncertainty among users about what is and is not considered acceptable behavior in online spaces, but also creates distrust about which values these technologies privilege.

5.2 Classification Reifies Oppression

Classification systems impact the ways in which social norms are created and enforced—for example, the medical profession's classification of homosexuality as an "illness" during the nineteenth century led to stigmatization which persists even today [47]. It is important to note the relationship between social deviance and social oppression: what is considered "deviant" in a given society is defined by dominant social forces, and thus deviant labels may unjustly malign members of marginalized groups. The importance of social labeling in the emergence and persistence of social norms [47] is critical to understanding not only societal perceptions of behaviors, but also the ways in which labels may be leveraged in the oppression of non-dominant groups.

As Bowker and Star [4] emphasize, classification systems emphasize the concerns of dominant groups, and are often created specifically to impose dominant norms upon oppressed persons. Marxist social conflict theory [33] similarly defines deviant behaviors as those which conflict with the goals of social institutions and the ruling class. Rooted in the recognition of structural differences in power and social class within capitalist societies, conflict theory [33] asserts that more powerful social groups are motivated to retain their power over oppressed groups, and as such, they assert that power through the application of laws and other classifications for behavior designed to oppress less powerful groups.

A promising direction for addressing online harassment is through the automated detection of abusive words and phrases (e.g., [9,50,51]); however, our findings show that, like in other cases where descriptive norms are being constantly negotiated, accounting for fluid social nuance may prove challenging for automated approaches. Similarly, Crawford and Gillespie [11] surface critical limitations of technical tools that enable users to flag content: for example, Facebook's removal of a photograph of two men kissing following its flagging by several users as 'graphic sexual content.' Although Facebook ultimately reinstated the image and apologized, this incident illustrates the limitations of sociotechnical classification systems for labeling non-normative—or non-dominant—behavior. Crawford and Gillespie [11] argue that flagged content is not a proxy for non-normative or deviant behavior, but instead represents complex negotiations between platforms, individual users, and broader regulatory forces. As such, automatic classifiers for detecting harassing language or behaviors are far from a simple solution, and indeed could exacerbate existing structural inequities.

5.2.1 System classifications exclude outsiders

Bowker and Star describe classification systems as "artifacts embodying moral and aesthetic choices that in turn craft people's identities, aspirations, and dignity." Our participants reported that they sometimes felt their harassment experiences did not fall within 'typical' expectations of what online harassment looks like, or even who is harassed online, especially those who were not able to accurately categorize their experience when submitting their case using HeartMob's checkbox system. This is reminiscent of Becker's labeling theory [5], which posits that socially applied labels can affect individuals' self-conceptions and behaviors, making those labeled as 'deviant' more likely to see themselves as outsiders.

In the language of Becker [5], our participants saw themselves as outsiders as a result of HeartMob's classification of their experiences. In this way, the HeartMob system itself acts as a boundary object [41], moving the enactment of power from an automated approach (e.g., Twitter and Facebook's automatic detection of and response to potentially abusive behavior) to human moderators. HeartMob moderators must still make classification decisions about what experiences should and should not be prioritized in the HeartMob community, which inevitably leads to the exclusion of some users who may belong to more privileged groups, or whose experiences do not fit within typical categories.

As Bowker and Star [7] emphasize, classification systems emphasize the concerns of dominant groups, and are often created specifically to impose dominant norms upon oppressed persons. Marxist social conflict theory [33] similarly defines deviant behaviors as those which conflict with the goals of social institutions and the ruling class. Rooted in the recognition of structural differences in power and social class within capitalist societies, conflict theory [33] asserts that more powerful social groups are motivated to retain their power over oppressed groups, and as such, they assert that power through the application of laws and other classifications for behavior designed to oppress less powerful groups.

These concerns resurface when corporations—who are managing policies at an enormous scale and who are accountable to their shareholders—must create content moderation policies that reflect the values

of the company, yet are enforceable at scale. To return to the Facebook example, ProPublica's article accuses the company of favoring "elites and governments over grassroots activists and racial minorities," based on the examples given (a U.S. congressman and a Black Lives Matter activist). From a social conflict theory perspective, the values and worldviews embedded in a company—driven by its leaders, shareholders, and employees—are likely to be reflected in its formal policies, which classify particular behaviors as appropriate or inappropriate. Instead, a more effective system for classifying appropriate behavior may be one that is co-governed between platforms and their users, which would allow for the introduction of additional social nuance in platform policies while fostering a greater sense of accountability among users.

5.3 Implications for Social Change

There cannot be a neutral or value-free approach to harassment categorization [7]. Intersectional feminist theory expands theories of power and oppression by linking different categorical identities—such as gender and race—to systems of structural oppression, such as sexism and racism. We argue that social media platforms and others working to prevent, detect, or manage online harassment must consider power and oppression when creating classification systems, including reporting tools, moderator guidelines, and platform policies.

First, platforms should make visible and disclose the categories, criteria, and process by which harassment is categorized. By rejecting or accepting incidents of online harassment without full disclosure around the values by which such decisions are made, targets may feel invalidated, causing additional harm and potentially affecting their future technology behaviors (e.g., chilling effects). Our data shows that when users receive scripted responses from platforms that do not directly address their specific experiences—or worse, when users receive responses that indicate there has been no violation of a specific site policy, as many of our participants did—targets experience increased social isolation and may subsequently minimize the impacts of their harassment experiences. Thus, system validation for harassment experiences through labeling is critical for targets' emotional health and recovery.

5.3.1 Addressing online harassment by centering vulnerable users

Ultimately, our results suggest a need for more democratic, user-driven processes in the generation of values that underpin technology systems. One movement towards this goal can be observed in the platform cooperativism and worker cooperative movements, which seek to turn users and employees into cooperative owners of—and participants in—such systems. Future research should further explore innovative democratic practices around online harassment management and support, particularly efforts driven by users.

It is important to emphasize that platform-driven design often privileges the experiences and concerns of socially dominant groups. Of 11,445 developers in the United States surveyed by Stack Overflow in 2017 [14], 85.5% were men, a majority of whom were also white. According to the United States Bureau of Labor [43], women held 25% of computing-related occupations in the United States in 2015—a percentage that has been steadily declining since 1991, when it reached a high of 36 percent. Nearly two-thirds of women who held computing occupations in 2015 were white: only 5% were held by Asian women, 3% by Black or African American women, and 1% by Latina or Hispanic women [43]. Thus, white men—who in the United States possess the most structural power [33]—are largely responsible for the ideation and development of policies, moderation guidelines, reporting tools, and other technologies aimed at preventing or managing harassment online: a problem disproportionately experienced by marginalized people [20,21,30].

In applying intersectional feminist theory, we argue that abuse mitigation practices must ultimately protect and be informed by those who are most vulnerable, or the people who historically experience structural oppression. Centering the oppressed in the ideation and development of technology creates stronger objectivity [24] in the categorization of online harassment. A "bottom-up" approach to system design allows us to start from the experiences of those who have traditionally been left out of the production of knowledge, ultimately resulting in technologies that better address the needs of all users. Best addressing online harassment requires the ongoing integration of vulnerable users' needs into the design and moderation of online platforms.

6 CONCLUSION

Harassment and abuse remain a pernicious problem for modern online communities. Through interviews with 18 users of HeartMob, a system designed by and for targets of online harassment, we find that classification is critical in validating and supporting harassment experiences. We also find that labeling abusive behaviors as 'online harassment' enables bystanders to grasp the true scope of this problem, and that visibly labeling harassment as inappropriate is critical for surfacing community norms and expectations for appropriate behavior. We discuss these results through the lens of Bowker and Star's classification theories and Becker's labeling theory of deviant behavior, and we caution that visible classification systems can also marginalize users whose harassment experiences are not typical, or whose experiences are not accounted for in the system's development. We surface significant challenges for better incorporating social context into existing technical systems, which often fail to address structural power imbalances perpetuated by automated labeling and classification. Similarly, platform policies and reporting tools are designed for a seemingly homogenous user base and do not account for individual experiences and systems of social oppression. Finally, informed by intersectional feminist theory, we argue that fully addressing online harassment requires the ongoing integration of vulnerable users' needs into the design and moderation of online platforms. Centering the oppressed in the ideation and development of technology ultimately results in technologies that better address the needs of all users.

ACKNOWLEDGMENTS

This work was partially supported by the Knight Foundation. We thank Emma Gardiner for assistance with data analysis and Vidhya Aravind, Renée Cymry, Jennifer Rubin, and Sarah Vieweg for feedback on drafts.

REFERENCES

[1] Sara Ahmed. 2017. *Living a Feminist Life*. Duke University Press.
[2] Julia Angwin. 2017. Facebook's Secret Censorship Rules Protect White Men from Hate Speech But Not Black Children. *ProPublica*.
[3] Zahra Ashktorab and Jessica Vitak. 2016. Designing Cyberbullying Mitigation and Prevention Solutions Through Participatory Design With Teenagers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16). ACM, New York, NY, USA, 3895–3905.
[4] Shaowen Bardzell and Jeffrey Bardzell. 2011. Towards a feminist HCI methodology: social science, feminism, and HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11). ACM, New York, NY, USA, 675–684.
[5] Howard Becker. 1963. Outsiders. Glencoe. *The Free Press* 9: 1982.
[6] Lindsay Blackwell, Jean Hardy, Tawfiq Ammari, Tiffany Veinot, Cliff Lampe, and Sarita Schoenebeck. 2016. LGBT Parents and Social Media: Advocacy, Privacy, and Disclosure During Shifting Social Movements. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16). ACM, New York, NY, USA, 610–622.
[7] Geoffrey C. Bowker and Susan Leigh Star. 2000. *Sorting things out: Classification and its consequences*. MIT press.
[8] Amy Bruckman, Catalina Danis, Cliff Lampe, Janet Sternberg, and Chris Waldron. 2006. Managing deviant behavior in online communities. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '06). ACM, New York, NY, USA, 21-24.
[9] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17). ACM, New York, NY, USA, 3175-3187.
[10] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '17). ACM, New York, NY, USA, 1217–1230.
[11] Robert B. Cialdini. 2007. Descriptive social norms as underappreciated sources of social control. *Psychometrika* 72, 2: 263.
[12] Danielle Keats Citron and Mary Anne Franks. 2014. *Criminalizing Revenge Porn*. Social Science Research Network, Rochester, NY.
[13] Rep Katherine Clark. 2015. Online Violence Against Trans Women Perpetuates Dangerous Cycle. *Huffington Post*.
[14] Keith Collins. 2017. Tech is overwhelmingly white and male, and white men are just fine with that. *Quartz*.
[15] Patricia Hill Collins. 1993. Black feminist thought in the matrix of domination. *Social theory: The multicultural and classic readings*: 615–625.
[16] Kimberle Crenshaw. 1991. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford law review*: 1241–1299.
[17] Julian Dibbell. 1993. A Rape in Cyberspace. *Village Voice* XXXVIII, 51.
[18] Jill P. Dimond, Michaelanne Dye, Daphne Larose, and Amy S. Bruckman. 2013. Hollaback!: the role of storytelling online in a social movement organization. In *Proceedings of the 2013 conference on Computer supported cooperative work* (CSCW '13). ACM, New York, NY, USA, 477-490.
[19] Judith Donath. 1999. Identity and Deception in the Virtual Community. *Communities in cyberspace*. Psychology Press.
[20] Maeve Duggan. 2014. Online Harassment. Pew Research Center.
[21] Maeve Duggan, 2017. Online Harassment 2017. Pew Research Center.
[22] Noah J. Goldstein, Robert B. Cialdini, and Vladas Griskevicius. 2008. A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of consumer Research* 35, 3: 472–482.
[23] Mark Griffiths. 2002. Occupational health issues concerning Internet use in the workplace. *Work & Stress* 16, 4: 283–286.
[24] Sandra Harding. 1992. Rethinking standpoint epistemology: What is "strong objectivity?" *The Centennial Review* 36, 3: 437–470.

[25] Bell Hooks. 2000. *Feminism is for everybody: Passionate politics*. Pluto Press.
[26] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving Google's Perspective API Built for Detecting Toxic Comments. *arXiv:1702.08138 [cs]*.
[27] Sara Kiesler, Robert E. Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design*. MIT Press, Cambridge, MA.
[28] Cliff Lampe. 2014. Gamification and social media. *The Gameful world: Approaches, issues, applications*. MIT Press, Cambridge, MA.
[29] Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '04). ACM, New York, NY, USA, 543–550.
[30] Amanda Lenhart, Michelle Ybarra, Kathryn Zickuhr, and Myeshia Prive-Feeney. 2016. *Online Harassment, Digital Abuse, and Cyberstalking in America*. Data & Society Institute.
[31] Eden Litt and Eszter Hargittai. 2016. The imagined audience on social network sites. *Social Media + Society* 2, 1: 2056305116633482.
[32] Alice E. Marwick. 2012. The public domain: Social surveillance in everyday life. *Surveillance & Society* 9, 4: 378.
[33] Karl Marx and Friedrich Engels. 1967. *The communist manifesto (1848). Trans. Samuel Moore*. London: Penguin.
[34] Leysia Palen and Paul Dourish. 2003. Unpacking privacy for a networked world. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '03). ACM, New York, NY, USA, 129-136.
[35] Michael L. Pittaro. 2007. Cyber stalking: An analysis of online harassment and intimidation. *International Journal of Cyber Criminology* 1, 2: 180–197.
[36] Lee Rainie, Janna Anderson, and Jonathan Albright. 2017. The Future of Free Speech, Trolls, Anonymity and Fake News Online. Pew Research Center.
[37] Sarita Schoenebeck, Nicole B. Ellison, Lindsay Blackwell, Joseph B. Bayer, and Emily B. Falk. 2016. Playful Backstalking and Serious Impression Management: How Young Adults Reflect on Their Past Identities on Facebook. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (CSCW '16). ACM, New York, NY, USA, 1475-1487.
[38] Michael D. Shear. 2016. Trump as Cyberbully in Chief? Twitter Attack on Union Boss Draws Fire. *The New York Times*.
[39] Peter K. Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. 2008. Cyberbullying: its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry* 49, 4: 376–385.
[40] Lee Sproull, Sara Kiesler, and Sara B. Kiesler. 1992. *Connections: New Ways of Working in the Networked Organization*. MIT Press.
[41] Susan Leigh Star and James R. Griesemer. 1989. Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social studies of science* 19, 3: 387–420.
[42] Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation,* 27(2), 237-246.
[43] U. S. Department of Labor. 2015. Current Population Survey: Detailed occupation by sex and race. Bureau of Labor Statistics.
[44] Jessica Vitak, Stacy Blasiola, Eden Litt, and Sameer Patil. 2015. Balancing Audience and Privacy Tensions on Social Network Sites: Strategies of Highly Engaged Users. *International Journal of Communication* 9: 20.
[45] Joseph B. Walther and Malcolm R. Parks. 2002. Cues filtered out, cues filtered in. *Handbook of interpersonal communication* 3: 529–563.
[46] Christina Warren. 2017. Twitter's New Abuse Filter Works Great, If Your Name Is Mike Pence. *Gizmodo*.
[47] Jeffrey Weeks. 1999. Discourse, desire and sexual deviance: some problems in a history of homosexuality. *Culture, society and sexuality. A reader*: 119–42.
[48] Amanda M. Williams and Lilly Irani. 2010. There's methodology in the madness: toward critical HCI ethnography. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '10). ACM, New York, NY, USA, 2725-2734.
[49] Rhiannon Williams. 2014. Facebook's 71 gender options come to UK users. *The Telegraph*.
[50] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2016. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web* (WWW '17). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1391-1399.
[51] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D. Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB* 2: 1–7.

Received April 2017; revised July 2017; accepted August 2017