The Psychological Well-Being of Content Moderators

The Emotional Labor of Commercial Moderation and Avenues for Improving Support

Miriah Steiger
TaskUs and St. Mary's University
miriah.steiger@taskus.com

Timir J. Bharucha
TaskUs and St. Mary's University
timir.bharucha@taskus.com

Sukrit Venkatagiri
Department of Computer Science
Virginia Tech
sukrit@vt.edu

Martin J. Riedl
School of Journalism and Media
University of Texas at Austin
martin.riedl@utexas.edu

Matthew Lease
School of Information
University of Texas at Austin
ml@utexas.edu

ABSTRACT

An estimated 100,000 people work today as commercial content moderators. These moderators are often exposed to disturbing content, which can lead to lasting psychological and emotional distress. This literature review investigates moderators' psychological symptomatology, drawing on other occupations involving trauma exposure to further guide understanding of both symptoms and support mechanisms. We then introduce wellness interventions and review both programmatic and technological approaches to improving wellness. Additionally, we review methods for evaluating intervention efficacy. Finally, we recommend best practices and important directions for future research. Content Warning: we discuss the intense labor and psychological effects of CCM, including graphic descriptions of mental distress and illness.

CCS CONCEPTS

• **Human-centered computing** → *HCI design and evaluation methods*; *Collaborative and social computing*; • **Applied computing** → **Psychology**.

KEYWORDS

content moderation, wellness, social justice, human computation

ACM Reference Format:
Miriah Steiger, Timir J. Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3411764.3445092

1 INTRODUCTION

Commercial content moderation (CCM) consists of assessing user-generated content (UGC) for compliance with a commercial social media platform's terms of service and community guidelines [59, 67, 129]. While most UGC posts are categorized as benign, large amounts of non-compliant text, image, audio, and video content are also posted. To highlight the scale of this problem [155], 160,000 instances of violent extremism were taken down in one year on Google Drive, Photos, and Blogger [20], and Facebook removed or applied warning labels to approximately 3.5 million items of uncivil or violent content in the first quarter of 2018 [48].

Non-compliant posts range from copyright infringement and infractions of regional speech to obscenity laws, such as profanity or nudity, to disinformation [e.g., 71]. Generally, CCM policies are difficult to universalize and put into hierarchies, as they are context-dependent (e.g., local cultures, languages, countries, and laws) and dependent on individual platforms' cultural norms. Extreme visual content can include depictions or actual acts of gore or lethal violence, such as murder, suicide, violent extremism [42], animal abuse, hate speech [92, 136], sexual abuse, child or revenge pornography [133], and more [26, 87, 131].

Ideally, we could rely on machine learning to automatically detect problematic content. However, human interpretation is often necessary due to high accuracy requirements and costs of errors, the subjective nature of the task, and complex, ever-changing moderation policies and forms of offending content [24, 55, 127, 128]. Chen [26] estimates that over 100,000 paid content moderators are staffed globally, spanning internal reviewers, contract workers from third parties, and outsourcing to online labor [59, 125].

While moderation work might be expected to be unpleasant, there is recognition today that repeated, prolonged exposure to specific content, coupled with limited workplace support, can significantly impair the psychological well-being of human moderators [19, 45, 67, 109, 113]. Along with the risk of continued exposure, juggling interactions or relations with management or platform users [161], and needing to maintain externally prescribed accuracy or throughput quotas for acceptable job performance exacerbates psychological discomfort [26, 56, 161]. An oft-noted concern is that moderation has led to a form of *posttraumatic stress disorder* (PTSD) known as *vicarious trauma* for some moderators [19].

Moderators, often limited in their possibilities to publicly speak about issues they face in their jobs, have articulated themselves through open letters some of their most pressing concerns: high accuracy and throughput targets, exposure effects, insufficient counseling, low wages, and lack of hazard pay, the sense that they should

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445092

be directly employed instead of through outsourcing companies, as well as concerns over the impossibility to speak about the job publicly, among other issues [45, 50, 53, 119, 135]. News articles have further shed light on alleged pressure on counselors to share the content of counseling sessions [14]. Moderators have articulated in open letters that companies create a "Big Brother environment" which is taking away their "sense of humanity" [119], and demanded from Facebook to "keep moderators and their families safe," to "maximize at-home working," to "offer hazard pay," "end outsourcing," and "offer real healthcare and psychiatric care" [53].

Contributions. Building on our earlier technical report [89], this article makes three primary contributions. Firstly, we discuss moderators' psychological symptomatology, drawing on other occupations involving trauma exposure to further guide understanding of both symptoms and support mechanisms. Secondly, we introduce wellness interventions for confronting these challenges through a combination of psychological and technological interventions. We further discuss how to measure the effectiveness of such interventions. Finally, we recommend additional best practices, as well as important directions for future research.

Our authorship spans academia and industry, bridging expertise in technology, social science, and clinical mental health. We come together through a strong alignment of our shared values and interests in finding and developing opportunities to support the well-being of content moderators' betterment. We advocate for the continued growth of academic-industry partnerships engaging with these issues, advancing both scholarship and practice.

2 THE HUMAN COST OF MODERATION

2.1 About Content Moderation

Moderation work varies greatly in practice. A moderator may act behind the scenes, reviewing a content queue [109] to approve/reject without any interaction with platform users. Alternatively, a moderator may be a participant in an online community such as Reddit, and thus known to users and held accountable by these users for moderation decisions. Degrees of agency of moderators diverge, depending on the approach towards moderation a company is taking - between a more laissez-faire approach at Reddit and the rather rigorous regimes of the likes of Facebook or Google, as well as on whether moderation is done on a volunteer basis, or as a job.

A variety of tools and processes are typically employed by moderators to manage the volume of UGC. As Myers West [102] notes, ML tools and filters are increasingly utilized to increase efficiency. Platform flagging tools can also allow users to report questionable content, cued up for human moderators to adjudicate. Thus, moderation can be carried out either as content gets uploaded or after problematic content is reported by users [85].

Artisanal, community-reliant, and commercial moderation work differ in their nature and demands [21, 76]. Social media companies such as Facebook, Google, and Twitter would fall into the paradigm of commercial moderation. In contrast, smaller companies may pursue artisanal or community-reliant strategies that may allow higher agency for volunteer moderators. Such volunteer moderation has been key to many online communities and groups and can be understood as a form of *civic labor* [93, 94]. Flagging mechanisms utilize volunteers to assess which content should be taken down [34, 103]. A broader definition of moderation may also include collective corrective action, in which users bond together to drown out negative speech [164]. As content moderation pervades all aspects of digital life, it is imperative for companies, governments, and citizens to develop collaborative strategies of governance in which responsibilities for the policing of content are shared among multiple stakeholders in a framework of "cooperative responsibility" [70]. We distinguish unpaid volunteer work vs. paid work when exploring the emotional labor of moderation, the workers involved, and the situational dynamics and triggers at play.

2.2 A Social Justice Perspective

According to Kranzberg's first law of technology [86], "Technology is neither good nor bad; nor is it neutral." An important stream of HCI research interrogates the many ways technology can impact social justice [43] issues, directly and indirectly, and how new technologies can be designed to advance social justice through advocacy or reduce inequality and injustice [143].

Consideration of content moderators and advocating on their behalf is also consistent with work on giving voice to underrepresented or vulnerable populations which might be adversely impacted by new technologies [41, 132, 146, 160], along with research promoting worker-centered design [41, 52]. Various research has sought to support online workers in particular [12, 63, 64, 73]. In these various spaces, HCI research can actively pursue and contribute to positive social change by striving to understand those at risk of exclusion, adopting a social justice orientation via appropriate design strategies, and putting this into practice [143].

While social justice themes are often invoked in calls to support moderators [9], we should also consider the framing of work underlying such calls, shaping personal and popular perception about what is just and whether assistance is truly needed. On one hand, CCM has been portrayed as stigmatized work of last resort [159], a"dirty job" [116], and 'the worst job in technology" [158]. On the other hand, there is a competing portrayal of noble work to keep the internet safe for others, which workers can take pride in [1, 45, 112, 162]. When work is stigmatized, calls for aid may conversely provoke apathy or concern, depending on the audience. A framing of helplessness may bolster such calls, while a frame of autonomy may suggest external intervention is unnecessary.

Similarly, when work is invisible [73] or perceived as unimportant, there may be less awareness or concern about social justice. This reflects another dichotomy in perceptions today of work and its value for social media platforms. Regarding invisibility of moderation work, as Newton [109] writes, "...so many people have written to me just to say that they didn't know that human beings were actually doing this work. They assumed it was all automated."

Perception of work importance is often focused on required skills rather than the necessity of the work itself. While skilled and creative IT work is often respected and well-compensated, relatively unskilled and rote data processing is typically afforded less stature and reward. While moderation work is relatively unskilled, it is simultaneously "one of the most crucial jobs created by the internet economy" [106] and lies at the very essence of what social media platforms offer today [59, 129], yet it is largely taken for granted.

When Covid-19 struck, Facebook's Zuckerberg [165] discussed: "...a small percent of our critical employees who can't work remotely, like content reviewers working on counter-terrorism or suicide and self-harm prevention..." Calls for social justice for moderators often contrast the critical role moderators play in social media platforms' functioning – and the greater health hazards of the work – vs. their seemingly lower status and benefits provided vs. other platform personnel, who are directly employed [9].

2.3 Human Computation and Crowdsourcing

Human abilities still exceed state-of-the-art AI for many data processing tasks, including content moderation. While AI will continue to improve, the use of human computation enables companies to deliver greater system capabilities today [8, 19, 90, 121], what Gray and Suri call AI's "last mile" [64]. Growth in human computation also creates new economic mobility opportunities, earning crowdsourcing the moniker of "The New Sewing Machine" [115].

A strength and weakness of crowd work [84] is that it is often invisible to consumers [73]. Terminology such as "Human Processing Units (HPUs)" [38], "Remote Person Calls (RPCs)" [10], or "the Human API" [73] highlight how human labor can be obscured and mediated by opaque APIs. Such APIs are central to integrating human work into these systems. Still, the technical jargon may also hide or diminish the crucial role of human workers in powering these systems, perpetuating invisibility of a global workforce that is by its very distributed nature difficult to put a face on [5, 64].

In addition, whenever work becomes less publicly visible, working conditions do as well. Irani and Silberman [73] opine that "by hiding workers behind web forms and APIs... employers see themselves as builders of innovative technologies... unconcerned with working conditions." Gray and Suri [64] describe as "ghost work" the unseen labor needed to maintain online content infrastructures and services, such as annotation and database work necessary to fuel AI. Similarly, Ekbia and Nardi [47] suggest intelligent systems in practice nearly always embody man-machine *heteromation*, yet the human contribution is often overlooked or diminished vs. the narrative of technological determinism and advancement.

Content moderation fits a natural niche for human computation since the content is difficult to automatically moderate. Perversely, while human computation mechanisms now enable us to call on human workers in such cases, this also seems like precisely the sort of task that one might most wish to automate since exposure to disturbing content cannot harm an algorithm. Dwoskin [45] reports that "It's the first job where I interviewed people where several people told me they would be happy if A.I. took over their job."

As noted earlier (Section 2.2), moderation work also remains largely invisible today. Some moderation work is, in fact, done via online crowdsourcing. Much is performed in secure offices with non-disclosure agreements, preventing moderators from discussing work or inviting others into their workplace [45, 57]. Like moderation, there are also competing narratives about worker empowerment vs. exploitation. The portrayal of crowd workers as victims is common in popular press [107], while the narrative of agency emerges more in first-hand accounts in worker forums [153]. Deng et al. [41] pursue a value-sensitive design approach to address the "duality of empowerment and marginalization" of such work.

2.4 Emotional Labor in Volunteer Moderation

Whereas commercial platforms can be challenging to study, a growing body of work has studied volunteer moderation on Reddit and Twitch [23, 81, 161].

Wohn [161] investigates the emotional labor of volunteer moderators on the Twitch live-streaming platform. Since moderation work involves great repetition of negative experience, Wohn notes this can trigger *secondary trauma* and eventual burnout. "Secondary trauma is the acute response to being exposed to someone else's traumatic experience" [161], while burnout results from "continual exposure to traumatic material" [95]. Another emotional toll found by Wohn [161] was a simple lack of appreciation where moderators did not feel sufficiently valued for their work contributions.

Dosono and Semaan [44] investigate the emotional labor of volunteer moderators on Reddit, noting that little was known about the experiences of the people doing this largely invisible yet necessary work to sustain the online community for others. They find that "the work moderators engage in is personally emotional, and they encounter threats to their personal privacy and well-being. The longevity of online communities rests on the backs of the moderators... [r]isking personal safety and wellness for the social good... constantly exposed to disturbing content that may have long-term effects on their mental health."

Jhaver et al. [75] discuss how automated content regulation on Reddit can reduce moderator exposure to offensive content but at the emotional cost and anxiety around real or potential automation mistakes. Regarding emotional labor on another Reddit forum, Gilbert [58] describes one moderator's crisis of conscience over whether to remove certain content: "we just felt so shitty as moderators... our community... is meant to be giving people answers about the past, but ... it's become a platform for these poor women to become humiliated again..." To cope, moderators were found to band together behind the scenes to share their feelings, be they annoyance, frustration, or more significant emotional burdens.

2.5 Emotional Labor in a Commercial Setting

The emotional cost of CCM work has been reported in popular press for nearly a decade [24–26, 69]. Sarah Roberts [124, 129] can be attributed as observing and describing the emotional toll of content moderation for the longest time in academic research, and has also coined the term 'commercial content moderation' for industrial-style moderation. In recent years, reporting by Casey Newton [108–113] has shone publicity onto connected issues.

In one example, Newton [111] reports moderators finding their belief systems altered and psychological wellness impaired by continual exposure to disturbing content. Newton [109] describes a content queue at one company dedicated entirely to violent and extremist content, such as rapes, be-headings, or murders, with another on child abuse videos. Newton [111] even reports a moderator dying at his desk from a heart attack, presumably triggered by the work. In the documentary film "The Cleaners", a moderator suicide occurs after repeated requests for a transfer are denied [116].

Moderators may triage content anywhere from 9-10 hours a day, 4-5 days a week [141]. Dwoskin [45] reported that one of five counselors supporting 450 moderators in Austin, TX stated the work could cause a form of PTSD known as *vicarious trauma*.

"They have to pause the video, they have to rewind the video. They have to zoom in on the video to see what's really happening. They have to see it, and they say they can't unsee it" [46].

We do not know how prevalent PTSD is among moderators. In 2019, Cambridge Consultants [19], commissioned by Ofcom (the UK's communications regulator), reported that "Moderating harmful content can cause significant psychological damage to moderators... The psychological effects of viewing harmful content is well documented, with reports of moderators experiencing post-traumatic stress disorder (PTSD) symptoms and other mental health issues as a result of the disturbing content they are exposed to." Newton [113] writes, "From my own interviews with more than 100 moderators over the past year, it appears to be a significant number [get PTSD]. And many other employees develop long-lasting mental health symptoms that stop short of full-blown PTSD, including depression, anxiety, and insomnia."

To date, no scientific studies have been conducted quantifying the prevalence of PTSD among moderators. We know that within the broader population of people exposed to secondary trauma, 7.8% experience lifelong symptoms, whereas 3.6% will have a 12 month period at which they exhibit full criteria for PTSD. Notably, such secondary exposure to trauma yields far reduced rates of symptomology vs. those who experience trauma directly themselves [104]. We must also consider sampling bias, especially in non-scientific journalistic or anecdotal accounts, in which the population of participants (e.g., those interviewed) may disproportionately reflect symptoms. Ultimately, more research is needed.

In scholarly work, the *emotional labor* of CM work is attracting increasing attention [36, 44, 75, 80, 129, 161], including related conferences: Roberts' "All Things in Moderation" (2017) at UCLA [154], as well as events at Santa Clara (2018) [134] and at USC (2018) [27, 151]. The Santa Clara event [134] included a recorded session on mental well-being, and Roberts' [126] essay highlights challenges and opportunities for worker wellness. Given recent litigation [54, 56], Roberts speculates that "...there may be liability for firms and platforms that do not take sufficient measures to shield... workers from damaging content whenever possible and to offer them adequate psychological support when it is not."

It is important to note that many factors contribute to the stress of CCM work beyond exposure. For example, volume quotas (akin to a call center) increase pressure on moderators, and moderators reported that "constant measurement for accuracy is as pressurizing as a quota" [45]. Roberts [129] has discussed how the factory-like nature of CM can cause burnout for many workers, including growing reluctant to discuss their work with those close to them so as not to burden them. During Covid-19, another stressor was being required to come into the office and risk exposure, while most platform employees were allowed to work from home [89, 165].

Psychologist Stefania Pifer runs the *Workplace Wellness Project* [120], which advises technology companies on moderator care. According to Pifer [45], a key challenge with companies today is "a clash between a call center model designed for low-cost labor and mechanized tasks and a feeling among workers that the burdens placed on them go well beyond that of a traditional call center employee." In her words, companies "might provide Zumba and yoga and access to a counselor, but they aren't thinking about how not being able to get up at any point in the day [162] might be increasing the psychological impact — or how holding people to a rigid number of tickets, or accuracy counts could be adding more harm." Facebook has similarly noted [45] that "Finding the right balance between content reviewer well-being and resiliency, quality, and productivity... is very challenging at the scale we operate in. We are continually working to get this balance right..."

The adverse effects of moderation can be better understood using Karasek and Theorell [79]'s *demand-control model* of risk factors. Job requirements such as volume quotas and accuracy goals increase demands while limiting autonomy, sense of control, and recovery time, increasing the probability of the employee struggling with heightened psychological stress [2]. High job demands and a low sense of control result in high-stress workplaces [2]. Employees with a greater understanding of power possess increased self-efficiency, mental health, self-esteem, and productivity.

3 LESSONS FROM RELATED PROFESSIONS

Given the limited research on the impact of graphic content on moderators, we draw on related occupations that have an extended body of prior work studying distress due to direct or indirect exposure to disturbing material. These findings can inform understanding of moderators' potential experience and symptomatology when similarly exposed to graphic content and potential prevention and treatment options that have been tested. Of course, exposure alone is not the exclusive determinant of what renders content moderation emotionally taxing. We can relate moderation to these other professions via both exposure and broader job stressors.

Regarding exposure, when drawing comparisons with other professions, it is essential to differentiate between primary and secondary exposure and active versus passive engagement. Content moderation involves second-hand exposure, with actions more passive vs. these other frontline professions. However, these actions – though farther removed from constituents – can still have a consequential impact. Moderators may rightfully feel as though they are shielding others from content exposure, a predisposition that is not passive. Overall, findings from these three related occupations suggest that content moderators may develop similar symptomatology and be at risk of developing anxiety, depression, or STS.

Journalists. Reporters often witness traumatic events, such as natural disasters, murders, mass casualties, etc. Monteiro et al. [100] conclude that exposure to traumatic events and media can induce or exacerbate feelings of psychological distress in journalists, leading to an unwillingness to work on certain types of investigations and a reduction in trust, morale, and job satisfaction. Feinstein et al. [51] explored the effect of uncensored UGC on the psychological health of 116 journalists and found that the frequency of exposure to UGC was a predictor of *psychopathologies* such as anxiety, depression, or PTSD. Journalists may be in situations where they encounter distressing material in the field first-hand and second-hand in online research. The connection between PTSD and journalism has manifested itself in research programs, such as the *Dart Center for Journalism and Trauma* at Columbia's Journalism School, which is interested in not only the ethical coverage of trauma but also the effects on journalists from covering traumatic events.

Emergency Dispatchers. Dispatchers are also exposed to significant traumatic material in the form of distressed callers. Just as content moderators, dispatchers are not on the scene of the incident but are expected to decipher graphic details of traumatic incidents in a timely manner [61]. As a result of the work, dispatchers report *peritraumatic distress*, or distress development immediately following the trauma, due to at least one call while on duty [152]. Pierce and Lilly [118] assessed trauma exposure, peritraumatic distress, and PTSD symptomatology in 911 dispatchers and found, as, in Troxell [152], a high prevalence reported peritraumatic distress. Additionally, researchers state that even though dispatchers are physically isolated from the traumatic event, exposure to the critical event can reach the threshold to produce PTSD symptomatology. Similar to content moderators, emergency dispatchers are not on the scene. However, they directly interact with those experiencing trauma, unlike the commercial moderation setting.

Sex-trafficking Detectives. Brady [16] surveyed 433 sex-trafficking detectives to explore the impact of job-related factors with the risk of secondary traumatic stress (STS), burnout, and compassion satisfaction. One in four detectives reported low compassion and a high prevalence of STS and burnout. Similarly, Perez et al. [117] found that officers who investigated Internet child pornography cases reported higher secondary traumatic stress disorder as they were exposed to more disturbing media. Repeated exposure to highly stressful traumatic situations was found to negatively impact some officers' cognitive abilities, memory, mental health, and overall well-being [3]. Detectives may encounter trauma in the field (primary) and online (secondary), with frequent exposure. Responsibility to help victims may increase both stress and pride.

Burns et al. [17] interviewed fourteen internet child exploitation (ICE) investigators on their strategies and best practices when processing disturbing material. Initially, many felt well-adapted to manage the graphic content, but upon review of material, quickly felt overburdened with the amount and extreme graphic nature of the content [17]. When discussing their job requirements, they reported the urgency to understand the breadth of work truly.

4 MODERATOR WORKPLACE WELLNESS

While there is extensive literature on the health benefits of well-structured workplace wellness programs [88], moderation work differs from traditional corporate positions with repeated exposure to disturbing content, challenging working conditions, and stringent performance metrics [22, 88, 123]. This repeated exposure increases moderators' risk of developing anxiety, depression, stress disorders, heart disease, interpersonal conflict, and substance abuse – similar to other occupations involving exposure to traumatic events (Section 3). If left untreated, this can lead to absenteeism, lower quality of life, burnout, and work dissatisfaction [4].

Cognitive health protection seeks to mitigate such risks by providing holistic wellness offerings, including mental health care [147], by addressing the impact of workplace conditions along with performance demands and their relationship with deterioration in mental health [88]. Given vast literature in occupational health, safety psychology, and clinical psychology addressing distinguished risk factors, employers can aid in prevention and mediation of developing adverse symptomology through two types of interventions – *programmatic* (Section 5) and *technological* (Section 6) – spanning three levels. As the first line of defense, *primary* prevention seeks to reduce stress in the work environment itself. Next, *secondary* interventions aim to bolster workers' resilience to stressors. Finally, should a disorder fully manifest, *tertiary* care reactively provides individualized treatment [88]. Providing comprehensive support spanning all three levels is crucial to "prevent and control the impacts of job stress" [88].

Programmatic and technological interventions are necessary to provide top-notch resources and assistance for mental health. Programmatic interventions assist developing of resilience through the acquisition of coping skills and techniques gained from training or sessions with a mental health professional. Technological interventions can be used while reviewing content to reduce exposure to graphic content and provide easier and faster access to mental health support or tools to dampen the impact. Both approaches work hand-in-hand to support the content moderator, and neither serves as a replacement for the other.

4.1 Primary Interventions: Risk Mitigation

Primary interventions apply strategies that limit the risk of the onset of mental health symptoms or prepare an individual to manage potentially adverse situation before the occurrence.

When hearing of preventative or primary measures, many people think of the medical community in the form of vaccinations or interventions one can take to increase overall physical health [83]. However, the concept spans multiple fields to include mental health and technology communities as well. Primary interventions seek to prevent harm by preventing exposure before it occurs, teaching skills, enhancing resiliency, and increasing tolerance prior to exposure. Overall, primary interventions foster a supportive work environment for workers to flourish [77, 88].

Joyce et al. [77] notes most common mental health conditions are treatable and in some cases preventable" (p. 683). Notably, primary interventions are considered more effective than other levels and the most inclusive, providing equal preventative care to all [88].

4.2 Secondary Interventions: More Resilience

Secondary interventions create innovative tools or psychological training to reduce negative symptoms after their onset, thereby helping them to return to a state of stability. Secondary interventions are primarily classified as preventative but are implemented to address risk factors following exposure to an environment or event. They provide skills training or coping material to address the ongoing stressor, thereby once again working on building the individual's resiliency. Tetrick and Winslow [149] conducted a meta-analysis evaluating the effectiveness of secondary intervention programs and found that those focused on stress management were effective, mediated by the type of intervention provided.

With persistent exposure to graphic content, moderators risk shifts in worldview [110]. Some such shifts are particularly maladaptive, such as changes following trauma exposure related to an individual's perception of safety for themselves and others [15]. Workers viewing disturbing material are thus classified as at risk even though they may not meet the disorder criteria. For this reason,

secondary interventions apply reoccurring clinical skills training or technological tooling to counteract such changes in perception.

4.3 Tertiary Interventions: Clinical Support

Finally, tertiary care triages severe cases and provides interventions following manifestation of a psychological disorder that meets full criteria, seeking to assist individuals in distress, and needing interventions focused on recovery or diminishing symptoms [149], e.g., via psychotherapy. Tertiary programs notably consist of psychotherapeutic interventions, providing individual, reactionary care.

Research supports the effectiveness of therapy in the workplace, but limitations on its extent must be noted. Individuals diagnosed with severe depression or anxiety require elevated care that may fall out of the workplace's scope of support. Such individuals are then recommended for referral to outside resources [77].

5 PROGRAMMATIC INTERVENTIONS

5.1 Primary Level: Wellness On-boarding

Occupations that encounter high levels of stress or trauma exposure can incorporate resilience programs to help workers develop foundational skills for stress management and self-soothing techniques to utilize during work at their desks. A well-known example is the *Resilience and Activity for Everyday* (READY) program [18]. Because depressive symptoms can manifest without proper training strategies to manage taxing work environment or job duties [18], such programs focus on strategies for stress reduction to mitigate the risk for developing such symptoms. Programs (like READY) encompass a variety of resiliency protective factors, such as positive emotions, cognitive flexibility (acceptance), life meaning, social support, and active coping [18].

Research has shown that resilience may be nurtured and developed in the workplace through guidance [82]. Because of the malleable nature of resilience, many corporations are moving towards offering resilience training to their workers. According to Seibert [138], resiliency encompasses multiple attributes, including coping, learned optimism, self-efficacy, hardiness, stress resistance, post-traumatic growth, internal locus of control, emotional intelligence, and the survivor personality. Resiliency programs seek to teach the skills necessary for managing stressful environments or situations. They also allow for preemptive care through skill development, instead of traditional reactionary interventions [32, 138].

Arnetz et al. [4]'s two-year longitudinal study provided supportive findings when evaluating program effectiveness with training police officers (Section 3) by measuring scores of physical symptoms, coping, mental well-being, sleep quality, and exhaustion. The program included relaxation training and guided imagery to mirror real-world situations with the application of their tactical skills.

Established resiliency programs have been applied to a multitude of other professions, including education, business, and medicine. However, there are variations in program duration,

By integrating the resiliency program literature components, companies establish an initial on-boarding program structured around primary interventions prior to first exposure of content. This is accomplished by teaching skills associated with highly resilient individuals to content moderators, similar to those mentioned in [157]. Employees can apply teachings to workplace occurrences and identify social supports within their workplaces, such as teammates or leaders, to better manage their unpredictable job tasks or changing environment. Resilience programs are most effective when incorporating organizational, and individual care [149] within the structure.

5.2 Secondary Level: Resiliency Training

Resilience training goes beyond on-boarding prevention programs like READY (Section 5.1), with ongoing training to facilitate individual and team success when encountering challenges [130]. Such ongoing training has been shown to positively impact employees' mental health and well-being, as well as performance and productivity [130], potentially aiding in the alleviation of stress associated with metrics with training focused on job performance.

Programmatic secondary interventions provide skills training led by a mental health professional or information to assist with coping strategies following exposure to an event. Tetrick and Winslow's [149] meta-analysis on the effectiveness of secondary clinical intervention programs related to the general and unwell population found that interventions focused on stress management were effective, mediated by the type/quality of intervention provided. Programs that incorporated cognitive behavioral therapy interventions were found to increase well-being and alleviate stress by measures of physiological and psychological assessments, attrition, absenteeism, and self-report. The authors attributed this to changing thoughts and behaviors instead of using simpler distraction techniques [149], which are seemingly the most prevalent form of care within the current content moderation literature [9]. Results indicate that training programs are beneficial to implement within the workplace if they are purposeful and grounded in theoretical practice.

5.3 Tertiary Level: Interpersonal Clinical Care

Numerous corporations follow the movement of implementing mental wellness counselor support on-site for employees, allowing employees to process both work and personal issues [31]. Previous studies have shown that therapy in the workplace is an effective intervention in reducing anxiety, stress, and depression for the majority of employees who utilize the services. Additionally, prior research reveals that common organizational interventions, for example, training or team meetings outside of those with clinical instruction, had no impact on employee psychological or physical well-being [31].

McLeod [96] conducted a meta-analysis assessing the effectiveness of workplace counseling, including internal services, offsite amenities paid by the employer, or offering of an employee assistance program. Findings from article reviews included 80% employee satisfaction with clinical services offered, a positive impact on employee psychological symptoms and stress reduction, psychological concerns, and work performance. The meta-analysis results specified intervention effectiveness to relate directly with acute, immediate onset of symptoms and low well-being, as well as "specific work-related psychological and behavioural problems, such as anxiety, low self-esteem, emotional burnout, occupational PTSD and substance abuse" [96] (p. 241).

6 TECHNOLOGICAL INTERVENTIONS

6.1 Primary Level: Preventing Exposure

Various algorithms have been proposed to automatically detect problematic content in an effort to minimize the amount of human moderation work required [42, 67, 78, 92, 123, 136, 156].

While supervised machine learning (ML) can be used to detect problematic content automatically, it is important to understand that there is no free lunch: human moderation work is still required. Firstly, ML algorithms require human-annotated training data (and usually vast amounts [68]) to perform accurately. Secondly, some level of ongoing human moderation is continually required to verify algorithmic performance or update training data for changing policies for acceptable and new forms of offending content. Thirdly, when algorithmic accuracy falls short of required targets, human moderators will be called in to close "the last mile" [64].

Companies already engaged in human moderation can "recycle" human moderation decisions into annotated training data [67]. Still, many companies do not have existing moderation work, nor are the few existing public data sets necessarily pertinent to the differing types of content relevant to each company. It also bears note that human annotators are more widely susceptible to exposure effects than is popularly recognized. For example, the Linguistic Data Consortium has reported that intensive exposure to regular news articles for a relatively benign annotation task-induced nightmares and overwhelming feelings for some annotators [144].

The most significant issue is that state-of-the-art algorithms are typically still insufficient to meet practical needs, predominantly due to high accuracy requirements and cost of errors, coupled with complex moderation policies often requiring human interpretation [55]. Today, practical solutions often adopt a human-in-the-loop approach [19, 90] spanning human and algorithmic detection.

Finally, not all useful automation requires ML. The *Global Internet Forum to Counter Terrorism* (GIFCT.org), founded by Facebook, Microsoft, Twitter and YouTube, maintains a shared database of extremist content [19]. Similarly, Microsoft's PhotoDNA [98] database, catalogues known child exploitation material. Such "remembering" allows known disturbing material, when recirculated, to be automatically removed without any additional exposure to moderators by way of hashing techniques [62]. However, content is often modified to circumvent such detection by exact match. To address this, Gorwa et al. [62] suggest approximate matching techniques. Such near-duplicate detection techniques can automatically filter content or group related content to ease moderation work [67].

6.2 Secondary Level: Reducing Exposure

One way to further reduce exposure to disturbing content and its emotional impacts is to investigate affective interface design for how content is presented to moderators.

In early work, De Cesarei and Codispoti [39] investigated the effects of image size and blur on 40 university students' emotional response. The authors found a significant reduction in emotional response when image blur and size reduction were applied. Participants found the altered images to be less vivid, less arousing, and less pleasant than the original pictures. In subsequent work, De Cesarei and Codispoti [40] further measured the electrodermal activity of skin conductance, which expresses the arousal and engagement in behavior in response to a critical event such as exposure to graphic content. They found that picture blurring reduced skin conductance, which was otherwise greater when viewing emotionally arousing images. Taken together, the two studies suggest that altering images can inhibit *action preparation*, helping to suppress the autonomic systems heightened arousal state.

Bekhtereva and Müller [11] showed momentary emotional distractors to participants to study how color may impact the time of attentional bias in the visual cortex. As the participants completed visual tasks, the researchers briefly projected unpleasant and neutral images in the background either in color or grayscale. They found that the photos in color produced a more significant distraction effect compared to grayscale. Moreover, the effect was lengthier with unpleasant photos in color. Participants rated the unpleasant scenes higher in emotional negativity. Additionally, greater arousal was observed for images shown in color.

In 2018, Dang et al. [36] proposed an interactive image blurring interface to reduce moderator exposure to disturbing material. The tool offered three modalities: 1) a slider for varying the blur level of the image; 2) a mouse-over option that temporarily unblurred the image while the mouse cursor remains over it, and 3) a mouse-click option required an explicit click to permanently unblur the image. The authors proposed an evaluation design for psychological well-being via the Positive Affect Negative Affect Scale (PANAS) and job performance (e.g., impacts on accuracy or speed). However, no evaluation was actually performed.

Since at least January 2019, Microsoft's video moderation tool has supported black-and-white and moderator controlled variable blurring transformations [148]. That May, Facebook "quietly announced it would be giving moderators new controls to help shield themselves from the ill effects of continually watching disturbing content" [45, 148]. Moderators were provided a preference pane to control image/video blurring, or audio muting [49]. In July 2019, Cambridge Consultants [19], in a report commissioned by Ofcom (the UK's communications regulator), suggested that blurring could allow moderators to perform their jobs while reducing exposure. Chris Harrison, a psychologist on Facebook's global resiliency team, stated, "shielding moderators from harm begins with giving them more control of what they're seeing and how they're seeing it, so just the existence of ...preferences helps" [148]. Moderators could sometimes classify content using only the associated text [45, 148].

Karunakaran and Ramakrishan [80] studied the impact of grayscaling and blurring images with Google moderators. The grayscale conversion could be manually activated, and moderators could mouse-hover over an image to view it in its original form. As in Dang et al. [36], psychological well-being was measured using PANAS. Viewing content in grayscale improved the reviewers' positive affect while still allowing effective identification of the most extreme and violent images. On the other hand, moderators were irritated by the blurring treatment as a negative outcome.

In 2020, Das et al. [37] extended Dang et al. [36]'s study by evaluating their interactive blurring tool. Using the Scale of Positive and Negative Experience (SPANE), the authors found that negative emotion was the highest for the unblurred baseline and at a minimum for fixed blur and hover. The Positive and Negative Affect Scale (I-PANAS-SF) showed that all three interfaces produced higher positive affect score vs. the baseline, with the slider option being

significantly superior to the baseline. In contrast with Karunakaran and Ramakrishan [80]'s findings regarding static blurring, Das et al. [37] report interactive blurring not only diminished the emotional impact of moderation but without compromising accuracy or speed.

6.3 Tertiary Level: Supporting Treatment

Technology can also aid in clinical care for content moderators. For example, when an individual experiences a flashback or views content in which they can connect to a prior traumatic memory [13], virtual reality (VR) technology can assist through the use of *dialectical behavioral mindfulness therapy* to redirect the mind to the present moment, as reported by Navarro-Haro et al. [105]. Participants within the study reported significantly higher state mindfulness levels after the VR intervention compared to those in the mindfulness group without VR. Individuals who received the VR intervention exhibited significantly lower levels of sadness, anger, anxiety, and increased relaxation from pre to post-test analysis. Other uses for VR include relaxation, increases in positive emotion, and stress reduction [7, 139]. Strassmann et al.[145] found that the VR environment allowed individuals to quickly achieve a relaxed state through immersion in the technology and separation from the current environment leading to an increase in well-being.

O'Leary et al. [114] suggest that tooling development can also significantly improve peer-support and mental health among those experiencing distress, specifically tooling that introduces peers based on similarities such as characteristics, beliefs, and needs. Moreover, they urge variety in the tools for engagement to include peer-support through text, audio, and visual formats. Finally, the researchers found that the tool entails resources in managing and reducing risk and adverse symptomology. Participants in their study spoke to the exclusion of clinicians in the tool forum to preserve agency, however, others spoke to the risks of using peers only for support. The researchers propose training to better equip peers to assist with mental health. However, limitations exist as trained peers lack experience and knowledge to sufficiently handle ongoing mental health needs at the moderate or severe level, calling for a need to maintain trained clinical staff.

7 EVALUATING INTERVENTIONS

For any type of intervention – programmatic (Section 5) or technological (Section 6) – we must be able to measure the outcome of the intervention in order to assess: 1) its actual benefit (if any); and 2) given cost of intervention, the return-on-investment.

There are several key challenges when measuring the impact of well-being and resilience interventions. There is no universal definition today for the constructs that direct the choice of measurement. Furthermore, while emotions are quite transient, research shows that short-term emotions can have a long-term impact. Based on symptomology from related occupations (Section 3) – depression, anxiety, secondary traumatic stress, compassion fatigue, and burnout [3, 16, 51, 117, 118, 152] – we suggest the following assessments to measure the effectiveness of programmatic and technological interventions on moderator well-being.

The General Health Questionnaire (GHQ). Given the limited empirical evidence linked to moderation work, it is imperative to baseline mental health functioning. The Mental Well-being: GHQ [6, 60] is a self-screening tool for psychiatric disorders within the domains of "depression, anxiety, somatic symptoms, and social withdrawal" [74] (p.57). GHQ has prior application in work environments by evaluating those most at risk for developing acute symptoms that do not meet the criteria for psychotic disorders. The measurement therefore, can be utilized to distinguish between healthy populations versus psychological illness, as well as track onset of endorsement of symptomatology [22]. Companies employing content moderators can administer the GHQ before training and initial exposure, then re-administer the assessment in one month intervals through the first year to monitor the progression of any potential risk to employees. Jackson [74] stresses the necessity to combine the GHQ with other forms of behavioral measurement that share a relationship with psychological distress, including absenteeism, low productivity, or increased turnover in the workplace.

The Perceived Stress Scale (PSS) Chronic stress is correlated to low workplace satisfaction, negative attitudes or behaviors within the workplace, and poor physical health [99]. The PSS [29] is thus an ideal measurement to assess the need for integration of additional program options or re-direction based on average company scores. The scale measures a person's interpretation of their environment within the categories of unpredictable, uncontrollable, and overloaded [30]. Recommendations for administration include monthly or quarterly to reassess the impact of the program.

The Oldenburg Burnout Inventory (OLBI). Unattended chronic work stress can lead to burnout [66], with severe levels of exhaustion and diminished attitudes towards work. While the Maslach Burnout Inventory is the most widely used assessment to measure burnout, it best suits the human services field. Halbesleben and Demerouti [66] developed the OLBI, which is widely applied to a range of occupations, from physical labor to information processing. Unlike previous burnout inventories, the OLBI uses both positively and negatively structured items to assess burnout's key characteristics: exhaustion and disengagement.

The Connor-Davidson Resilience Scale (CD-RISC). The CD-RISC measures resilience [33]. It is comprised of 25 items scored from 0-100, with a higher score indicating increased resilience. Connor and Davidson [33] applied the assessment to both the general and clinical samples. Results show that it has good internal consistency and test-retest reliability. Finally, the assessment's validity is comparable to other widely used measures of stress [33].

8 FURTHER RECOMMENDATIONS

Facebook researchers have recommended "providing access to licensed counselors, providing group therapy sessions, and screening applicants for suitability for the role as part of the recruiting process." [67]. Similarly, the recommendations below complement our earlier discussion of programmatic and technological interventions.

8.1 Disclose Risks to Prospective Workers

Newton [113] reports a moderator telling him, "If I knew from the beginning how this job would impact our mental health, I would never have taken it." We support Newton's [108, 112] call for firms to fully-disclose risks of exposure, up-front when advertising openings, to let prospective hires make a fully-informed decision whether to accept work. Consider further examples. Sex-trafficking

detectives (Section 3) also reported the urgency to truly understand the breadth of work when discussing their job requirements [17]. A *Technology Coalition* [150] of companies fighting online child sexual exploitation created an Employee Resilience Guidebook [28] to support workers interacting with such content, including that workers should be informed what their role will entail from the beginning. University Institutional Review Boards (IRBs) similarly stress the importance of informed consent for research subjects.

8.2 Limit Exposure

Limiting exposure to extreme content was a key finding from Burns et al. [17] interviews of detectives. Novice investigators reported that slowly increasing the amount of exposure to the disturbing content, led to greater comfort with it. They also conveyed the importance of setting limitations on the amount of material processed daily [17]. This finding also aligns with the results of Meischke et al. [97] who found that over-commitment to the completion of work tasks was positively associated with increased symptoms of stress. One way to limit exposure was by alternating the viewing of material with other tasks. This strategy allowed for a much-needed break [17]. The ERG [28] further advocates for the use of effective tooling that can assist in limiting the amount of human interaction required which partially mitigates adverse effects of exposure.

When asked how much content is safe to view per day, Facebook's Chris Harrison, a clinical psychologist, responded that, "Scientifically, do we know how much is too much? Do we know what those thresholds are? The answer is no, we don't... If there's something that were to keep me up at night... it's that question" [108].

Newton [112] offers related recommendations. Because so little is known, he calls for greater investment in research. Beyond limiting exposure on the job, he recommends setting a lifetime cap on moderator exposure to disturbing content. If moderation is understood as a time-bounded profession, he recommends defining real career paths for moderators beyond moderation work. Finally, while health care and counseling support are needed on the job, he further advocates providing these benefits after moderators have left the job, when adverse effects may persist or newly appear.

8.3 Create Space

Dosono and Semaan [44] report that "distancing away from drama" helped moderator coping. When the Linguistic Data Consortium learned its annotators were experiencing nightmares and other overwhelming feelings, "...to reduce the negative psychological impact... project managers implemented downtime. Every other week, approximately one hour was set aside for the team to do something other than annotation, as a stress reliever. The team chose the downtime activities each week, ranging from painting to a walk in the park to going to the local art museum" [144].

There has been a divergence between the level of moderator support claimed by some firms vs. the moderators [45, 162]. Wong [162] reports CCM work environments without sufficient time for breaks or resilience, and moderators not being allowed to use phones at their desk. Dwoskin [45] reported Accenture claiming unlimited wellness time for its moderators including access to counselors on-demand. However, moderators provided documentation from Accenture management showing wellness time limited to 45 minutes/week, or 9 minutes/day. Moderators disliked having to take a formal break to make even a brief personal call. One moderator commented that not being allowed to leave the building during breaks "made me feel like there is no escape from the content..."

8.4 Build Connection

Investigators also spoke to the necessity of strong social support outside of work, such as family and friends [17]. This support system offered the investigators an alternative avenue to express themselves and participate in activities outside of work. Social support led to feelings of belonging and understanding, which was reported as essential by the interviewees [17].

This is further evident in Idås and Backholm [72]'s study of exposure effects on crisis and war journalists: 80% of the 375 participants reported seeking social support following a taxing assignment. The findings were interesting, though as the type of support received was far more relevant than the amount, indicating a need to develop purposeful support systems for those within these professions.

Developing a social support network may not be as simple to implement for content moderators. Roberts [127] notes that industrial stratification and geographic distribution make it difficult for workers to build community beyond the local office. For volunteer moderators on Reddit, Dosono and Semaan [44] describe several strategies used by moderators to manage the emotional stress manifest in their moderation work. This included "empowering moderators through visible social support" and "building solidarity from shared struggles." Similarly, a study of live-streaming Twitch moderators also found that moderators felt better when they made time to commiserate and bond with one another [161].

9 SOCIOTECHNICAL DESIGN CHALLENGES

While various technological tooling has been proposed or implemented, we are still in the infancy of prototyping and evaluating such technological interventions to better support moderator well-being. While some commercial implementations of such technology are now known (e.g., earlier Microsoft and Facebook examples), it is unclear to what extent social media companies, or companies developing supporting technology, are investing in technological development, evaluation, and establishment of best practices.

More accurate detection algorithms should translate into less exposure for human moderators. Moreover, even if overall accuracy is insufficient, some specific types of content might be predicted accurately enough for automatic review. Better understanding and reasoning about algorithmic predictions can help us better know when, where, and why algorithmic predictions can be trusted without human review. Such automated detection is an area in which companies are already investing since decreasing manual labor is a win-win for cost savings, speed, scalability, and reducing exposure.

When human moderators are needed, how can technology better support them to both improve production outputs (e.g., accuracy and speed) and mitigate exposure effects? Since automation is unlikely to replace human moderators anytime soon [47, 64], we might think beyond automation to decision-support technology [140], to aid human moderators via an algorithm-in-the-loop [65]. This coincides with a broader trend toward developing predictive

algorithms that are accurate and interpretable [91] to the human decision makers. As algorithms provide better explanations, human moderators might be able to make oversight decisions based on those explanations, without viewing the actual content [148].

If we could predict the severity of disturbing content, such predicatability could enable an early warning system and/or a healthier, more balanced distribution of the workload across moderators. Other transformations of content could also reduce sensitivity. Karunakaran and Ramakrishan [80] suggest masking specific colors (for example, changing all red to green) or performing more artistic renderings of content, e.g., via Google's Deep Dream [101].

As discussed earlier, it can help to create time and space away from difficult content (Section 8.3). Strassel et al. [144] discussed the value of introducing breaks, not merely as pauses in work, but with diverting activities. Could it be similarly useful to introduce micro-breaks or micro-diversions [35] into moderation work on a more frequent and on-going basis? As an analogy, *WorkRave* [163] lets computer users impose a schedule of micro and macro forced breaks to reduce the risk of incurring a repetitive stress injury. Could setting some similar schedule of breaks into moderation work reduce the risk of incurring emotional injury?

While interfaces to reduce exposure are good, this treats symptoms and not the underlying social problems. To aid prevention, research should investigate sociotechnical designs to discourage the production and sharing of the harmful content itself.

Another intriguing technical challenge is to accurately model each worker's risk of adverse reactions that may manifest into post-traumatic stress disorder over time – or to a lesser severity, a stress-related disorder? Given the diverse risk factors and complexities involved, this is a challenging prediction task. Recent research has proposed machine learning to compile data from outcome assessments and confounding variables to predict individuals most susceptible to developing a stress-related disorder [122]. As with other sensitive patient data in digital health care records, strong safeguards must be put in place for safe use and stewardship of such data and model predictions.

To truly understand the challenges moderators face and develop workable solutions, researchers might engage with moderators in participatory design [137]. Rather than assume, that we researchers or designers know what is best, we engage stakeholders to identify key problems and co-develop realistic solutions.

Finally, researchers are not immune either. Regarding her lab's work on crisis social media, Starbird [142] comments that, "We haven't used external annotators for this data and probably wouldn't due to concerns about potential psychological effects of exposure to this kind of content. It's affected us." Another researcher who wished to remain anonymous shared, "We've historically had problems getting enough ground truth information from annotators because of how emotionally taxing the task of annotation is. Even for experts like me who've built up an emotional shield, I still am really negatively affected by this work."

10 CONCLUSION

In this paper, we contextualize commercial content moderation (CCM) and the implications of persistent exposure to graphic content for the human moderators who provide the last line of defense for social media platforms. While commercial moderation has received attention from journalists and policymakers, relatively little research has been directed toward the types and levels of care at which workers must be supported, or workplace wellness mechanisms that companies can put in place to help moderators.

We have outlined a comprehensive approach to promoting moderator wellness, including programmatic and technological interventions spanning three levels: prevention, mitigation, and treatment. Our discussion of wellness and resilience programming via the occupational health and safety model points to the importance of developing, sustaining, and strengthening such programs.

Ultimately, society must grapple with how governance of user-generated content should unfold. As people continue to use technology platforms to upload content, policing this will remain a necessity. We call upon others to enhance moderator well-being through more collaborative research and implementation of recognized best practices of care. Moreover, we urge companies to prioritize well-being alongside productivity through the transparency of work hazards and supporting worker autonomy once hired.

Finally, it is important to challenge the predominant invisibility and social framing of moderation work today. It is typically seen only when moderation fails, scandals surface, or companies disclose information voluntarily. Narratives that disparage moderation as a 'dirty' job do not recognize or honor the essential contribution of this profession in safeguarding the internet for the rest of us. Research assessing the moderators' sense of purpose and contribution to protecting society could help. Research on public attitudes regarding content moderation also has the potential to help shift perceptions towards a shared acknowledgment of its importance. Beyond gestures, such a shift could have important practical ramifications for the status and benefits provided to moderators within technology companies, moving toward parity with other tech jobs that are already recognized as crucial to the core business.

ACKNOWLEDGMENTS

We thank the reviewers for their valuable feedback and the many content moderators who help safeguard social media platforms for the rest of us. Sukrit Venkatagiri is supported by the Center for Human—Computer Interaction at Virginia Tech. Matthew Lease is supported in part by the Knight Foundation, the Micron Foundation, and Good Systems[1], a UT Austin Grand Challenge to develop responsible AI technologies. The statements made herein are solely the authors' own opinions and not those of the sponsoring agencies.

REFERENCES

[1] Accenture. 2019. Content Moderation Associate. https://www.linkedin.com/jobs/view/content-moderation-associate-at-accenture-1647075188/
[2] Jafar Akbari, Rouhollah Akbari, Mahnaz Shakerian, and Behzad Mahaki. 2017. Job demand-control and job stress at work: A cross-sectional study among prison staff. *Journal of education and health promotion* 6, 15 (2017).
[3] Judith P Andersen, Konstantinos Papazoglou, Markku Nyman, Mari Koskelainen, and Harri Gustafsberg. 2015. Fostering resilience among police. *Journal of Law Enforcement* 5, 1 (2015), 1–13.
[4] Bengt B Arnetz, Eamonn Arble, Lena Backman, Adam Lynch, and Ake Lublin. 2013. Assessment of a prevention program for work-related stress among urban police officers. *International archives of occupational and environmental health* 86, 1 (2013), 79–88.
[5] Andy Baio. 2008. The Faces of Mechanical Turk. November 20. waxy.org/2008/11/the_faces_of_mechanical_turk.

[1] https://goodsystems.utexas.edu

[6] Michael H Banks, Chris W Clegg, Paul R Jackson, Nigel J Kemp, Elizabeth M Stafford, and Toby D Wall. 1980. The use of the General Health Questionnaire as an indicator of mental health in occupational studies. *Journal of occupational psychology* 53, 3 (1980), 187–194.

[7] Rosa María Baños, Ernestina Etchemendy, Diana Castilla, Azucena García-Palacios, Soledad Quero, and Cristina Botella. 2012. Positive mood induction procedures for virtual environments designed for elderly people. *Interacting with Computers* 24, 3 (2012), 131–138.

[8] Jeff Barr and Luis Felipe Cabrera. 2006. AI Gets a Brain. *Queue* 4, 4 (2006), 24–29.

[9] Paul M. Barrett. 2020. Who Moderates the Social Media Giants? *Center for Business* (2020).

[10] Benjamin B Bederson and Alexander J Quinn. 2011. Web workers unite! Addressing challenges of online laborers. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. 97–106.

[11] Valeria Bekhtereva and Matthias M Müller. 2017. Bringing color to emotion: the influence of color on attentional bias to briefly presented emotional images. *Cognitive, Affective, & Behavioral Neuroscience* 17, 5 (2017), 1028–1047.

[12] Janine Berg, Marianne Furrer, Ellie Harmon, Uma Rani, and M Six Silberman. 2018. *Digital labour platforms and the future of work: Towards decent work in the online world*. International Labour Office Geneva.

[13] Shaw Beth. 2019. When Trauma Gets Stuck in the Body. *Psychology Today* (Oct 2019). https://www.psychologytoday.com/us/blog/in-the-body/201910/when-trauma-gets-stuck-in-the-body

[14] Sam Biddle. 2019. Trauma counselors were pressured to divulge confidential information about Facebook moderators, internal letter claims. *The Intercept* (2019). August 16. https://theintercept.com/2019/08/16/facebook-moderators-mental-health-accenture/. Visited December 23, 2020.

[15] Dinka Corkalo Biruski, Dean Ajdukovic, and Ajana Löw Stanic. 2014. When the world collapses: changed worldview and social reconstruction in a traumatized community. *European Journal of Psychotraumatology* 5, 1 (2014), 24098.

[16] Patrick Q Brady. 2017. Crimes against caring: Exploring the risk of secondary traumatic stress, burnout, and compassion satisfaction among child exploitation investigators. *Journal of police and criminal psychology* 32, 4 (2017), 305–318.

[17] Carolyn M Burns, Jeff Morley, Richard Bradshaw, and José Domene. 2008. The emotional impact on and coping strategies employed by police teams investigating internet child exploitation. *Traumatology* 14, 2 (2008), 20–31.

[18] Nicola W Burton, Ken I Pakenham, and Wendy J Brown. 2010. Feasibility and effectiveness of psychosocial resilience training: a pilot study of the READY program. *Psychology, health & medicine* 15, 3 (2010), 266–277.

[19] Cambridge Consultants. 2019. The Use of AI In Online Content Moderation. https://www.ofcom.org.uk/research-and-data/internet-and-on-demand-research/online-content-moderation

[20] K Canegallo. 2019. Meet the teams keeping our corner of the internet safer. *The Keyword* (2019).

[21] R Caplan. 2018. Content or context moderation. *Data & Society Research Institute* (2018).

[22] Hudson Wander de Carvalho, Christopher J Patrick, Miguel Roberto Jorge, and Sérgio Baxter Andreoli. 2011. Validation of the structural coherency of the General Health Questionnaire. *Brazilian Journal of Psychiatry* 33, 1 (2011), 59–63.

[23] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 32 (Nov. 2018), 25 pages. https://doi.org/10.1145/3274301

[24] Adrian Chen. 2012. Facebook releases new content guidelines, now allows bodily fluids. Gawker.com. http://gawker.com/5885836/facebook-releases-new-content-guidelines-now-allows-bodily-fluids

[25] Adrian Chen. 2012. Inside Facebook's outsourced anti-porn and gore brigade, where 'camel toes' are more offensive than 'crushed heads'. Gawker.com. http://gawker.com/5885714/inside-facebooks-outsourced-anti-porn-and-gore-brigade-where-camel-toes-are-more-offensive-than-crushed-heads

[26] Adrian Chen. 2014. The laborers who keep dick pics and beheadings out of your facebook feed. https://www.wired.com/2014/10/content-moderation/

[27] George Civeris. 2018. The new 'billion-dollar problem' for platforms and publishers. *Columbia Journalism Review* (2018). https://www.cjr.org/tow{_}center/facebook-twitter-tow-usc.php

[28] The Technology Coalition. 2013. Employee Resilience Guidebook for Handling Child Sexual Abuse Images. March. https://www.thorn.org/wp-content/uploads/2015/02/EmployeeResilienceGuidebookFinal7-13-1.pdf. Visited March 2, 2020.

[29] Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. 1983. A global measure of perceived stress. *Journal of health and social behavior* (1983), 385–396.

[30] Sheldon Cohen, T Kamarck, R Mermelstein, et al. 1994. Perceived stress scale. *Measuring stress: A guide for health and social scientists* 10 (1994).

[31] Jill Collins, Alison Gibson, Sarah Parkin, Rosemary Parkinson, Diana Shave, and Colin Dyer. 2012. Counselling in the workplace: How time-limited counselling can effect change in well-being. *Counselling and Psychotherapy Research* 12, 2 (2012), 84–92.

[32] ComPsych Corporation. 2019. More than One-third of Employees Say "People Issues" Cause the Most Stress at Work: Highlights the 2019 StressPulseSM survey. Oct 22. https://www.compsych.com/press-room/press-article?nodeId=5e35641b-dfe3-4e87-9066-66c420b0a234. Visited March 15, 2020.

[33] Kathryn M Connor and Jonathan RT Davidson. 2003. Development of a new resilience scale: The Connor-Davidson resilience scale (CD-RISC). *Depression and anxiety* 18, 2 (2003), 76–82.

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

[35] Peng Dai, Jeffrey M Rzeszotarski, Praveen Paritosh, and Ed H Chi. 2015. And now for something completely different: Improving crowdsourcing workflows with micro-diversions. 628–638.

[36] Brandon Dang, Martin J Riedl, and Matthew Lease. 2018. But who protects the moderators? the case of crowdsourced image moderation. *arXiv preprint arXiv:1804.10999*.

[37] Anubrata Das, Brandon Dang, and Matthew Lease. 2020. Fast, Accurate, and Healthier: Interactive Blurring Helps Moderators Reduce Exposure to Harmful Content. In *Proceedings of the 8th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

[38] J. Davis, J. Arderiu, H. Lin, Z. Nevins, S. Schuon, O. Gallo, and M.H. Yang. 2010. The HPU. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*. 9–16.

[39] Andrea De Cesarei and Maurizio Codispoti. 2008. Fuzzy picture processing: effects of size reduction and blurring on emotional processing. *Emotion* 8, 3 (2008), 352.

[40] Andrea De Cesarei and Maurizio Codispoti. 2010. Effects of picture size reduction and blurring on emotional engagement. *PloS One* 5, 10 (2010), e13399.

[41] Xuefei Nancy Deng, K. D. Joshi, and Robert D. Galliers. 2016. The Duality of Empowerment and Marginalization in Microtask Crowdsourcing: Giving Voice to the Less Powerful through Value Sensitive Design. *MIS Quarterly* 40, 2 (June 2016), 279–302.

[42] Oscar Deniz, Ismael Serrano, Gloria Bueno, and Tae-Kyun Kim. 2014. Fast violence detection in video. In *2014 international conference on computer vision theory and applications (VISAPP)*, Vol. 2. IEEE, 478–485.

[43] Lynn Dombrowski, Ellie Harmon, and Sarah Fox. 2016. Social Justice-Oriented Interaction Design: Outlining Key Design Strategies and Commitments. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*. 656–671.

[44] Bryan Dosono and Bryan Semaan. 2019. Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on Reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[45] Elizabeth Dwoskin. 2019. Inside Facebook, the second-class workers who do the hardest job are waging a quiet battle. *The Washington Post* (2019). https://www.washingtonpost.com/technology/2019/05/08/inside-facebook-second-class-workers-who-do-hardest-job-are-waging-quiet-battle/

[46] Elizabeth Dwoskin. 2019. *Internet gatekeepers pay a psychic toll*. Washington, D.C. https://www.washingtonpost.com/podcasts/post-reports/californias-secret-climate-deal-with-automakers-bypasses-trump-administration-regulations/?itid=lk_interstitial_manual_22

[47] Hamid Ekbia and Bonnie Nardi. 2014. Heteromation and its (dis) contents: The invisible division of labor between humans and machines. *First Monday* 19, 6 (2014).

[48] Facebook. 2018. Facebook Publishes Enforcement Numbers for the First Time. https://about.fb.com/news/2018/05/enforcement-numbers/

[49] FastCompany. 2019. Facebook Preferences Pane for Content Moderators. https://images.fastcompany.net/image/upload/w_596,c_limit,q_auto:best,f_auto/wp-cms/uploads/2019/06/Quick-Settings.png

[50] Fbcontentmods. 2020. This is a message of solidarity from a group of current and former content moderators at Facebook. *Medium.com* (2020). June 8. https://medium.com/@fbcontentmods/this-is-a-message-of-solidarity-from-a-group-of-current-and-former-content-moderators-at-facebook-6af1b3b2a020. Visited December 23, 2020.

[51] Anthony Feinstein, Blair Audet, and Elizabeth Waknine. 2014. Witnessing images of extreme violence: a psychological study of journalists in the newsroom. *JRSM open* 5, 8 (2014), 2054270414533323.

[52] Sarah E Fox, Vera Khovanskaya, Clara Crivellaro, Niloufar Salehi, Lynn Dombrowski, Chinmay Kulkarni, Lilly Irani, and Jodi Forlizzi. 2020. Worker-Centered Design: Expanding HCI Methods for Supporting Labor. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[53] Foxglove. 2020. Open letter from content moderators re: pandemic. *foxglove.org* (2020). November 18. https://www.foxglove.org.uk/news/open-letter-from-content-moderators-re-pandemic. Visited December 23, 2020.

[54] Sandra E. Garcia. 2018. Ex-Content Moderator Sues Facebook, Saying Violent Images Caused Her PTSD. *The New York Times* (2018). https://www.nytimes.com/2018/09/25/technology/facebook-moderator-job-ptsd-lawsuit.html

[55] Arpita Ghosh, Satyen Kale, and Preston McAfee. 2011. Who moderates the moderators? crowdsourcing abuse detection in user-generated content. In *Proceedings of the 12th ACM conference on Electronic commerce*. 167–176.

[56] Abhimanyu Ghoshal. 2017. Microsoft sued by employees who developed PTSD after reviewing disturbing content. The Next Web. *The next web* (2017). https://thenextweb.com/microsoft/2017/01/12/microsoft-sued-by-employees-who-developed-ptsd-after-reviewing-disturbing-content/

[57] David Gilbert. 2020. Bestiality, Stabbings, and Child Porn: Why Facebook Moderators Are Suing the Company for Trauma. *Vice)* (2020). https://www.vice.com/en_us/article/a35xk5/facebook-moderators-are-suing-for-trauma-ptsd

[58] Sarah A Gilbert. 2020. " I run the world's largest historical outreach project and it's on a cesspool of a website." Moderating a Public Scholarship Site on Reddit: A Case Study of r/AskHistorians. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–27.

[59] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.

[60] David P Goldberg and Valerie F Hillier. 1979. A scaled version of the General Health Questionnaire. *Psychological medicine* 9, 1 (1979), 139–145.

[61] Sarah E Golding, Claire Horsfield, Annette Davies, Bernadette Egan, Martyn Jones, Mary Raleigh, Patricia Schofield, Allison Squires, Kath Start, Tom Quinn, et al. 2017. Exploring the psychological health of emergency dispatch centre operatives: a systematic review and narrative synthesis. *PeerJ* 5 (2017), e3735.

[62] 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 (2020), 2053951719897945.

[63] Mark Graham, Vili Lehdonvirta, Alex Wood, Helena Barnard, Isis Hjorth, and Peter D Simon. 2017. The risks and rewards of online gig work at the global margins. (2017).

[64] Mary L Gray and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass.* Eamon Dolan Books.

[65] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[66] Jonathon RB Halbesleben and Evangelia Demerouti. 2005. The construct validity of an alternative measure of burnout: Investigating the English translation of the Oldenburg Burnout Inventory. *Work & Stress* 19, 3 (2005), 208–220.

[67] Alon Halevy, Cristian Canton Ferrer, Hao Ma, Umut Ozertem, Patrick Pantel, Marzieh Saeidi, Fabrizio Silvestri, and Ves Stoyanov. 2020. Preserving integrity in online social networks. *arXiv preprint arXiv:2009.10311* (2020).

[68] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems* 24, 2 (2009), 8–12.

[69] Reyhan Harmanci. 2012. The Googler Who Looked At The Worst Of The Internet. *BuzzFeed* (2012). August 21. http://www.buzzfeed.com/reyhan/tech-confessional-the-googler-who-looks-at-the-wo.

[70] Natali Helberger, Jo Pierson, and Thomas Poell. 2018. Governing online platforms: From contested to cooperative responsibility. *The information society* 34, 1 (2018), 1–14.

[71] Alex Hern and Dan Sabbagh. 2020. NHS Announces Plan to Combat Coronavirus Fake News. *The Guardian* (2020). https://www.theguardian.com/world/2020/mar/10/nhs-plan-combat-coronavirus-fake-news

[72] Trond Idås and Klas Backholm. 2017. Risk and resilience among journalists covering potentially traumatic events. *The assault on journalism* (2017), 235.

[73] Lilly C Irani and M Six Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 611–620.

[74] Craig Jackson. 2007. The general health questionnaire. *Occupational medicine* 57, 1 (2007), 79–79.

[75] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5 (2019), 1–35.

[76] Jialun Aaron Jiang. 2020. Identifying and Addressing Design and Policy Challenges in Online Content Moderation. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–7.

[77] Sadhbh Joyce, Matthew Modini, Helen Christensen, Arnstein Mykletun, Richard Bryant, Philip B Mitchell, and Samuel B Harvey. 2016. Workplace interventions for common mental disorders: a systematic meta-review. *Psychological medicine* 46, 4 (2016), 683–697.

[78] David Jurgens, Eshwar Chandrasekharan, and Libby Hemphill. 2019. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. *arXiv preprint arXiv:1906.01738* (2019).

[79] R. A. Karasek and Töres Theorell. 1990. *Healthy work, stress, productivity, and the reconstruction of working life.* Basic Books.

[80] Sowmya Karunakaran and Rashmi Ramakrishan. 2019. Testing Stylistic Interventions to Reduce Emotional Impact of Content Moderation Workers. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 50–58.

[81] Charles Kiene, Kenny Shores, Eshwar Chandrasekharan, Shagun Jhaver, Jialun" Aaron" Jiang, Brianna Dym, Joseph Seering, Sarah Gilbert, Kat Lo, Donghee Yvette Wohn, et al. 2019. Volunteer Work: Mapping the Future of Moderation Research. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 492–497.

[82] Danielle D King, Alexander Newman, and Fred Luthans. 2016. Not if, but when we need resilience in the workplace. *Journal of organizational behavior* 37, 5 (2016), 782–786.

[83] Lisa A Kisling and Joe M Das. 2019. Prevention Strategies. *NCBI* (2019).

[84] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The Future of Crowd Work. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*. 1301–1318.

[85] Kate Klonick. 2017. The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev.* 131 (2017), 1598.

[86] Melvin Kranzberg. 1986. Technology and History:" Kranzberg's Laws". *Technology and culture* 27, 3 (1986), 544–560.

[87] Till Krause and Hannes Grassegger. 2016. Inside Facebook. Sueddeutsche Zeitung. http://international.sueddeutsche.de/post/154513473995/inside-facebook

[88] Anthony D LaMontagne, Angela Martin, Kathryn M Page, Nicola J Reavley, Andrew J Noblet, Allison J Milner, Tessa Keegel, and Peter M Smith. 2014. Workplace mental health: developing an integrated intervention approach. *BMC psychiatry* 14, 1 (2014), 131.

[89] Matthew Lease, Miriah Steiger, Timir J. Bharucha, Martin J. Riedl, and Sukrit Venkatagiri. 2020. *Promoting Psychological Wellness of Content Moderators.* Technical Report TR-20-02. University of Texas at Austin, Department of Computer Science. https://apps.cs.utexas.edu/apps/tech-reports/194782

[90] Daniel Link, Bernd Hellingrath, and Jie Ling. 2016. A Human-is-the-Loop Approach for Semi-Automated Content Moderation. In *Proceedings of the Information Systems for Crisis Response and Management (ISCRAM) Conference.*

[91] Zachary C Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 31–57.

[92] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one* 14, 8 (2019).

[93] J Nathan Matias. 2016. The civic labor of online moderators. In *Internet Politics and Policy conference. Oxford, United Kingdom.*

[94] J Nathan Matias. 2019. The civic labor of volunteer moderators online. *Social Media+ Society* 5, 2 (2019), 2056305119836778.

[95] I Lisa McCann and Laurie Anne Pearlman. 1990. Vicarious traumatization: A framework for understanding the psychological effects of working with victims. *Journal of traumatic stress* 3, 1 (1990), 131–149.

[96] John McLeod. 2010. The effectiveness of workplace counselling: A systematic review. *Counselling and Psychotherapy Research* 10, 4 (2010), 238–248.

[97] Hendrika Meischke, Ian Painter, Michelle Lilly, Randal Beaton, Debra Revere, and Becca Calhoun. 2015. An exploration of sources, symptoms and buffers of occupational stress in 9-1-1 emergency call centers. (2015).

[98] Microsoft. 2009. PhotoDNA. https://www.microsoft.com/en-us/PhotoDNA/

[99] Teodor Mihăilă. 2015. Perceived Stress Scale as a Predictor of Professional Behavior and Aspects of Wellbeing. *Romanian Journal of Cognitive Behavioral Therapy and Hypnosis* 2, 2 (2015), 1–14.

[100] Susana Monteiro, Alexandra Marques Pinto, and Magda Sofia Roberto. 2016. Job demands, coping, and impacts of occupational stress among journalists: A systematic review. *European Journal of Work and Organizational Psychology* 25, 5 (2016), 751–772.

[101] A Mordvintsev, M Tyka, and C Olah. 2015. Google deep dream. June 17. https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html. Visited March 3, 2020.

[102] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383.

[103] Teresa K Naab, Anja Kalch, and Tino GK Meitz. 2018. Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society* 20, 2 (2018), 777–795.

[104] National Collaborating Centre for Mental Health (UK and others. 2005. *Post-traumatic stress disorder: the management of PTSD in adults and children in primary and secondary care.* Gaskell.

[105] María V Navarro-Haro, Yolanda López-del Hoyo, Daniel Campos, Marsha M Linehan, Hunter G Hoffman, Azucena García-Palacios, Marta Modrego-Alarcón, Luis Borao, and Javier García-Campayo. 2017. Meditation experts try Virtual Reality Mindfulness: A pilot study evaluation of the feasibility and acceptability of Virtual Reality to facilitate mindfulness practice in people attending a Mindfulness conference. *PloS one* 12, 11 (2017), e0187777.

[106] Annalee Newitz. 2020. We Forgot About the Most Important Job on the Internet. *The New York Times* (2020). March 13. https://www.nytimes.com/2020/03/13/opinion/sunday/online-comment-moderation.html. Visited March 17, 2020.

[107] Andy Newman. 2019. I Found Work on an Amazon Website. I Made 97 Cents an Hour. *The New York Times* (2019). November 15. https://www.nytimes.com/interactive/2019/11/15/nyregion/amazon-mechanical-turk.html. Visited March

6, 2020.
[108] Casey Newton. 2019. Bodies in seats. *The Verge* (2019). June 19. https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-tampa. Visited March 2, 2020.
[109] Casey Newton. 2019. The Emotional Toll Of Content Moderation. *All Things Considered* (2019). December 21. https://www.npr.org/2019/12/21/790492548/the-emotional-toll-of-content-moderation. Visited March 6, 2020.
[110] Casey Newton. 2019. The Terror Queue. *The Verge* (2019). December 16. https://www.theverge.com/2019/12/16/21021005/google-youtube-moderators-ptsd-accenture-violent-disturbing-content-interviews-video. Visited 3/2/20.
[111] Casey Newton. 2019. The Trauma Floor. The secret lives of Facebook moderators in America. *The Verge* (2019). February 25. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona. Visited March 2, 2020.
[112] Casey Newton. 2020. What tech companies should do about their content moderators' PTSD. *The Verge* (2020). January 28. https://www.theverge.com/interface/2020/1/28/21082642/content-moderator-ptsd-facebook-youtube-accenture-solutions. Visited March 6, 2020.
[113] Casey Newton. 2020. YouTube Moderators are Being Forced to Sign a Statement Acknowledging the Job Can Give Them PTSD. *The Verge* (2020). January 24. https://www.theverge.com/2020/1/24/21075830/youtube-moderators-ptsd-accenture-statement-lawsuits-mental-health. Visited March 6, 2020.
[114] Kathleen O'Leary, Arpita Bhattacharya, Sean A Munson, Jacob O Wobbrock, and Wanda Pratt. 2017. Design opportunities for mental health peer support technologies. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1470–1484.
[115] Praveen Paritosh, Panos Ipeirotis, Matt Cooper, and Siddharth Suri. 2011. The computer is the new sewing machine: benefits and perils of crowdsourcing. In *Proceedings of the 20th Intl. conference companion on World wide web*. 325–326.
[116] Lisa Parks. 2019. Dirty Data: Content Moderation, Regulatory Outsourcing, and The Cleaners. *Film Quarterly* 73, 1 (2019), 11–18.
[117] Lisa M Perez, Jeremy Jones, David R Englert, and Daniel Sachau. 2010. Secondary traumatic stress and burnout among law enforcement investigators exposed to disturbing media images. *Journal of Police and Criminal Psychology* 25, 2 (2010), 113–124.
[118] Heather Pierce and Michelle M Lilly. 2012. Duty-related trauma exposure in 911 telecommunicators: Considering the risk for posttraumatic stress. *Journal of traumatic stress* 25, 2 (2012), 211–215.
[119] Rob Price. 2019. Facebook moderators are in revolt over 'inhumane' working conditions that they say erodes their 'sense of humanity'. *Business Insider* (2019). February 15. https://www.businessinsider.com/facebook-moderators-complain-big-brother-rules-accenture-austin-2019-2. Visited December 23, 2020.
[120] The Workplace Wellness Project. 2011. The Workplace Wellness Project: Partnering with technology companies to build cultures of resilience. https://theworkplacewellnessproject.com/
[121] Alexander J Quinn and Benjamin B Bederson. 2011. Human computation: a survey and taxonomy of a growing field. In *2011 Annual ACM SIGCHI conference on Human factors in computing systems*. 1403–1412.
[122] Luis Francisco Ramos-Lima, Vitoria Waikamp, Thyago Antonelli-Salgado, Ives Cavalcante Passos, and Lucia Helena Machado Freitas. 2020. The use of machine learning techniques in trauma-related disorders: A systematic review. *Journal of psychiatric research* 121 (2020), 159–172.
[123] Christian X Ries and Rainer Lienhart. 2014. A survey on visual adult image recognition. *Multimedia tools and applications* 69, 3 (2014), 661–688.
[124] Sarah T Roberts. 2014. *Behind the screen: The hidden digital labor of commercial content moderation.* Ph.D. Dissertation. University of Illinois at Urbana-Champaign.
[125] Sarah T. Roberts. 2016. Commercial content moderation: Digital laborers' dirty work. In *The intersectional internet: Race, sex, class and culture online*, Safiya Umoja Noble and Brendesha M. Tynes (Eds.). Peter Lang, 147–160. https://doi.org/10.1007/s13398-014-0173-7.2 arXiv:arXiv:1011.1669v3
[126] Sarah T. Roberts. 2018. Commercial Content Moderation And Worker Wellness: Challenges & Opportunities. *Techdirt* (2018). February 8. https://www.techdirt.com/articles/20180206/10435939168/commercial-content-moderation-worker-wellness-challenges-opportunities.shtml. Visited March 2, 2020.
[127] Sarah T. Roberts. 2018. Content Moderation. In *Encyclopedia of Big Data*. Springer.
[128] Sarah T. Roberts. 2018. Digital detritus: 'Error' and the logic of opacity in social media content moderation. *First Monday* 23, 3 (2018).
[129] Sarah T Roberts. 2019. *Behind the screen: Content moderation in the shadows of social media.* Yale University Press.
[130] Ivan T Robertson, Cary L Cooper, Mustafa Sarkar, and Thomas Curran. 2015. Resilience training in the workplace from 2003 to 2014: A systematic review. *Journal of occupational and organizational psychology* 88, 3 (2015), 533–562.
[131] Rebecca Roe. 2017. Dark shadows, dark web. In *Keynote at All Things in Moderation: The People, Practices and Politics of Online Content Review – Human and Machine | UCLA December 6-7 2017*. Los Angeles, CA. https://atm-ucla2017.net/
[132] Emma J Rose. 2016. Design as advocacy: Using a human-centered approach to investigate the needs of vulnerable populations. *Journal of Technical Writing and Communication* 46, 4 (2016), 427–445.
[133] Napa Sae-Bae, Xiaoxi Sun, Husrev T Sencar, and Nasir D Memon. 2014. Towards automatic detection of child pornography. In *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 5332–5336.
[134] Santa Clara University. 2018. Content moderation and removal at scale, Conference at Santa Clara University School of Law, February 2, 2018, Santa Clara, CA. (2018). http://law.scu.edu/event/content-moderation-removal-at-scale/
[135] Zoe Schiffer. 2020. Facebook content moderators demand better coronavirus protections. *The Verge* (2020). November 18. https://www.theverge.com/2020/11/18/21573526/facebook-content-moderators-open-letter-coronavirus-protections-demands-accenture. Visited December 23, 2020.
[136] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. 1–10.
[137] Douglas Schuler and Aki Namioka. 1993. *Participatory design: Principles and practices.* CRC Press.
[138] Al Seibert. 2005. The resiliency advantage.
[139] Chris Shaw, Diane Gromala, and Meehae Song. 2011. The meditation chamber: towards self-modulation. In *Metaplasticity in virtual worlds: Aesthetics and semantic concepts*. IGI Global, 121–133.
[140] Jung P Shim, Merrill Warkentin, James F Courtney, Daniel J Power, Ramesh Sharda, and Christer Carlsson. 2002. Past, present, and future of decision support technology. *Decision support systems* 33, 2 (2002), 111–126.
[141] Daisy Soderberg-Rivkin. 2019. Five myths about online content moderation, from a former content moderator. *RStreet* (Oct 2019). https://www.rstreet.org/2019/10/30/five-myths-about-online-content-moderation-from-a-former-content-moderator/
[142] Kate Starbird. 2020. Personal Communication.
[143] Constantine Stephanidis, Gavriel Salvendy, Margherita Antona, Jessie YC Chen, Jianming Dong, Vincent G Duffy, Xiaowen Fang, Cali Fidopiastis, Gino Fragomeni, Limin Paul Fu, et al. 2019. Seven HCI grand challenges. *International Journal of Human–Computer Interaction* 35, 14 (2019), 1229–1269.
[144] Stephanie Strassel, David Graff, Nii Martey, and Christopher Cieri. 2000. Quality Control in Large Annotation Projects Involving Multiple Judges: The Case of the TDT Corpora. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*.
[145] Carolin Straßmann, Sabrina C Eimler, Alexander Arntz, Dustin Keßler, Sarah Zielinski, Gabriel Brandenberg, Vanessa Dümpel, and Uwe Handmann. 2019. Relax yourself-using virtual reality to enhance employees' mental health and work performance. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.
[146] Angelika Strohmayer, Mary Laing, and Rob Comber. 2017. Technologies and social justice outcomes in sex work charities: fighting stigma, saving lives. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3352–3364.
[147] Kelley A Strout, Daniel J David, Elizabeth J Dyer, Roberta C Gray, Regula H Robnett, and Elizabeth P Howard. 2016. Behavioral interventions in six dimensions of wellness that protect the cognitive health of community-dwelling older adults: a systematic review. *Journal of the American Geriatrics Society* 64, 5 (2016), 944–958.
[148] Mark Sullivan. 2019. Facebook is expanding its tools to make content moderation less toxic. *FastCompany* (2019). June 24. https://www.fastcompany.com/90367858/facebook-is-expanding-its-tools-to-make-content-moderation-less-toxic. Visited March 2, 2020.
[149] Lois E Tetrick and Carolyn J Winslow. 2015. Workplace stress management interventions and health promotion. *Annu. Rev. Organ. Psychol. Organ. Behav.* 2, 1 (2015), 583–603.
[150] The Technology Coalition. 2021. The Technology Coalition. http://www.technologycoalition.org
[151] Tow Center for Digital Journalism & Annenberg Innovation Lab. 2018. Controlling the conversation: The ethics of social platforms and content moderation, Conference at University of Southern California, Annenberg School of Communication, February 23, 2018, Los Angeles, CA. https://annenberg.usc.edu/events/annenberg-innovation-lab/controlling-conversation-ethics-social-platforms-and-content
[152] Roberta Mary Troxell. 2008. *Indirect exposure to the trauma of others: The experiences of 9-1-1 telecommunicators.* Ph.D. Dissertation. University of Illinois at Chicago.
[153] TurkerView. 2019. Writer Who Never Learned to Drive Works for Uber. Makes $0.97/hr. November 18. https://blog.turkerview.com/writer-who-never-learned-to-drive-works-for-uber/. Visited March 6, 2020.

[154] University of California Los Angeles. 2018. All things in moderation: The people, practices and politics of online content review - human and machine, Conference at the University of California, December 6-7, 2017, Los Angeles, CA. , 18 pages. https://atm-ucla2017.net/

[155] Bertie Vidgen, Helen Margetts, and Alex Harris. 2019. How much online abuse is there? (2019). https://www.turing.ac.uk/sites/default/files/2019-11/online_abuse_prevalence_full_24.11.2019_-_formatted_0.pdf Nobember 27.

[156] Dong Wang, Zhang Zhang, Wei Wang, Liang Wang, and Tieniu Tan. 2012. Baseline results for violence detection in still images. In *2012 IEEE 9th International Conference on Advanced Video and Signal-Based Surveillance*. 54–57.

[157] Brad Waters. 10. Traits of emotionally resilient people. *Psychology Today, May* (10).

[158] Lauren Weber and Deepa Seetharaman. 2017. The Worst Job in Technology: Staring at Human Depravity to Keep It Off Facebook. *The Wall Street Journal* (2017). December 27. https://www.wsj.com/articles/the-worst-job-in-technology-staring-at-human-depravity-to-keep-it-off-facebook-1514398398. Visited March 17, 2020.

[159] Wikipedia. 2021. Employer of Last Resport. https://en.wikipedia.org/wiki/Employer_of_last_resort

[160] Cara Wilson, Roisin McNaney, Abi Roper, Tara Capel, Laura Scheepmaker, Margot Brereton, Stephanie Wilson, David Philip Green, and Jayne Wallace. 2020. Rethinking Notions of'Giving Voice'in Design. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[161] Donghee Yvette Wohn. 2019. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article 160, 13 pages. https://doi.org/10.1145/3290605.3300390

[162] Queenie Wong. 2019. Murders and suicides: Here's who keeps them off your Facebook feed. *CNET* (2019). June 19. https://www.cnet.com/news/facebook-content-moderation-is-an-ugly-business-heres-who-does-it/. Visited 3/6/20.

[163] Workrave. 2021. Workrave: Take a Break and Relax. http://www.workrave.org/

[164] Marc Ziegele, Teresa K Naab, and Pablo Jost. 2019. Lonely together? Identifying the determinants of collective corrective action against uncivil comments. *New Media & Society* (2019), 1461444819870130.

[165] Mark Zuckerberg. 2020. April 16. https://www.facebook.com/zuck/posts/as-we-start-to-think-about-what-it-will-look-like-to-re-open-society-i-wanted-to/10111807251999141/. Visited April 18, 2020.