AURA: Amplifying Understanding, Resilience, and Awareness for Responsible AI Content Work

ALICE QIAN ZHANG[*], Carnegie Mellon University, USA
JUDITH AMORES, Microsoft Research, USA
HONG SHEN, Carnegie Mellon University, USA
MARY CZERWINSKI, Microsoft Research, USA
MARY L. GRAY, Microsoft Research, USA
JINA SUH, Microsoft Research, USA

Behind the scenes of maintaining the safety of technology products from harmful and illegal digital content lies unrecognized human labor. The recent rise in the use of generative AI technologies and the accelerating demands to meet responsible AI (RAI) aims necessitates an increased focus on the labor behind such efforts in the age of AI. This study investigates the nature and challenges of content work that supports RAI efforts, or "RAI content work," that spans content moderation, data labeling, and red teaming – through the lived experiences of content workers. We conduct a formative survey and semi-structured interview studies to develop a conceptualization of RAI content work and a subsequent framework of recommendations for providing holistic support for content workers. We validate our recommendations through a series of workshops with content workers and derive considerations for and examples of implementing such recommendations. We discuss how our framework may guide future innovation to support the well-being and professional development of the RAI content workforce.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; • **Security and privacy**; • **Social and professional topics** → **Computing / technology policy**;

Additional Key Words and Phrases: Responsible AI, worker well-being, red teaming, content moderation, data labeling, data work

ACM Reference Format:
Alice Qian Zhang[*], Judith Amores, Hong Shen, Mary Czerwinski, Mary L. Gray, and Jina Suh. 2025. AURA: Amplifying Understanding, Resilience, and Awareness for Responsible AI Content Work. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW033 (April 2025), 45 pages. https://doi.org/10.1145/3710931

1 Introduction

On July 21, 2023, the United States White House released a statement detailing the voluntary commitments of companies leading in developing artificial intelligence (AI) [65]. These commitments include promises to ensure AI systems are safe through "internal and external testing" before their introduction to the public. Such promises subsequently raise concerns about how human expertise is being recruited and supported in this type of testing. Thus in this paper,

[*]This work was completed while the author was an intern at Microsoft Research

Authors' Contact Information: Alice Qian Zhang[*], aqzhang@andrew.cmu.edu, Carnegie Mellon University, USA ; Judith Amores, judithamores@microsoft.com, Microsoft Research, USA; Hong Shen, hongs@andrew.cmu.edu, Carnegie Mellon University, USA; Mary Czerwinski, marycz@microsoft.com, Microsoft Research, USA; Mary L. Gray, mlg@microsoft.com, Microsoft Research, USA; Jina Suh, jinsuh@microsoft.com, Microsoft Research, USA.

ACM 2573-0142/2025/4-ARTCSCW033
https://doi.org/10.1145/3710931

we explore how to best support people engaging in work practices that ensure ethical and safe AI products.

We define those practices as **Responsible AI (RAI) content work**, which involves generating, reviewing, or reasoning about digital content with the goal of ensuring safety and ethical standards in AI systems [106]. In this paper, we focus on three key aspects of RAI content work to scope our study: content moderation, data labeling, and the emerging practice of red teaming. These areas are critical to ensuring the ethical and responsible development of contemporary AI systems. It is important to note, however, that individual RAI content workers may engage in a multitude of these activities, reflecting the multifaceted nature of their role in supporting responsible AI development. Regardless of the specific activities workers engage with, the support for human efforts behind these initiatives is often overlooked despite the importance of the work conducted [33, 52, 114, 127, 143, 152]. Without a comprehensive understanding of these efforts, we may see history repeat itself with content work facing challenges of invisibility of the workforce and a lack of well-being support crucial to workers.

Prior human-computer interaction (HCI) literature on harmful content exposure within content moderation has surfaced key challenges of developing psychological symptoms such as anxiety, depression, and burnout within populations [33, 114, 127, 143, 152]. However, empirical data on how these challenges manifest in other types of content work and factors unique to RAI (e.g., sudden increases in content volume due to interest in AI integration) remains limited. Studies have also explored using technologies to mitigate harmful content exposure and treat symptoms [30, 62, 73, 142], but were limited to primarily image and video-based content that does not cover the full spectrum of types of exposure in all types of content work. Recent calls within Computer-Supported Cooperative Work (CSCW) advocate examining the transformation of human labor within AI systems [20, 130]. In this context, we investigate the emergence of RAI content work as a new form of digital labor and the potential disruptions generative AI may bring to the digital content ecosystem, raising uncertainty about the impact on those maintaining AI system safety.

Previous studies within CSCW have examined content moderation challenges specific to end-user communities [17, 123] and platforms [59, 117]. However, the challenges related to the well-being and work quality of content workers employed and working with AI systems have yet to be explored in depth. To address this gap, we ground our study in the lived experiences of self-identifying content workers engaged in various activities with and around content. We aim to highlight the need to evaluate challenges content workers face amidst growing AI-related content demands and to inform future practices of content work in the age of AI.

We take a comprehensive approach, examining all types of content work from content moderation to red teaming through a two-phase study (see Figure 1 for the study flow). In the first phase, we provide empirical insights from surveys and interviews on the nature of content work (**RQ1**) and the challenges content workers face (**RQ2**). We illustrate the multi-faceted nature of content work, detailing findings from the main factors that constitute it: workers' roles, types of content that workers are exposed to, protective tools they use, impacts of engaging with content, and practices for collaboration. Building off of these insights, we surface challenges about misconceptions about the realities of content work, shortcomings of tools and metrics, failures of workplace support, and barriers to career growth. These challenges informed our proposal of a framework for amplifying understanding, resilience, and awareness (AURA) for RAI Content Workers comprised of four categories: *recruitment, tooling, adaptive wellness, and retention*. In the second phase, we further revise our recommendations through validation workshops that surface challenges and considerations for the applications of these recommendations in our framework. Overall, our study informs future improvements in the design of content work, developments that can support the

Fig. 1. Flow of our two-phase study. In the first phase, we conducted a survey study (N=67) and an interview study (N=22) to understand the nature of content work. From these insights, we developed a set of recommendations to improve content worker well-being. In the second phase, we validated the challenges we discovered and our recommendations to address those challenges, within the AURA framework that organize those recommendations, through interactive workshops (N=14).

well-being of workers, and progress in defining the professional identity and growth of the RAI workforce.

2 Related Work

In the following section, we delve into relevant existing literature to provide context for our study. We first provide a background of how RAI content work is conceptualized, tracing the origins of content-related endeavors from annotation and moderation of harmful online content to the formulation of our definition of content work within RAI, as many forms of content work are dedicated to shielding users from potentially harmful AI content and behaviors. Subsequently, we summarize the well-documented challenges that content workers dealing with harmful content or behaviors may encounter, setting the stage for exploring the new obstacles and difficulties content workers face during RAI efforts. Lastly, we highlight prior work on targeted support measures for content workers, from tooling and automation to organizational interventions to motivate renewed emphasis on holistic examination of support for all types of content-related work necessitated by the AI boom.

2.1 Conceptualizing Content Work for Responsible AI

Within the landscape of RAI, "content work" (or "data enrichment work" [108]) encompasses any form of labor that requires human review, judgment, or intelligence on a digital content or data that may be used to ensure the safety and responsibility during AI or ML model development. This includes data labeling, annotations, or validation as well as content moderation, human feedback, or corrections. The labor of annotating, reviewing, or moderating potentially harmful content or behaviors from within technology-facilitated spaces or interactions could be traced back to early online communities of unpaid *volunteers* [127] that enforced community norms and rules [90, 121] across platforms like Discord [128], Twitch [152], Reddit [81], and Facebook [49]. In addition, *end-users* who are not formal community members may also engage in content moderation by reporting violations [26, 69], adjusting moderation preferences [48, 68, 76]. As the popularity of social media platforms grew, this labor evolved into a structured and industrialized form with human workers in various ways. *Crowdsourced content workers* are those who are recruited to complete small human intelligence tasks, such as labeling content on Amazon MTurk [30, 61, 97].

Commercial content workers (i.e., commercial content moderators or CCMs [121]) are a dedicated workforce employed in some full- or part-time capacity. Recently, the demand for commercialized content work seems to be on the rise. For example, content moderation, as an already booming market [40], is now met with an increased demand for moderating AI-generated content [6].

However, the traditional data labeling and content moderation are not the only forms of content work on the rise. Originating as a military strategy [85, 160], red teaming has since been applied to the field of security [1, 75, 102, 153, 159] and more recently within RAI as a mandate [65] and a method [14, 47] to safeguard AI deployments. Similarly to the traditional forms of content work, RAI red teaming has been seen in the public domain with volunteers [22, 46, 107] and crowdworkers [47], and seen in commercial domains with full-time operators [95, 105, 106]. RAI red teaming also shares similar goals as other forms of content work where it broadly seeks to identify potentially harmful capabilities or outputs from AI systems [10, 135], with the exception that it often involves a structured and systematic adversarial testing [95, 105]. Despite such recent popularity, there is limited research about best practices for RAI red teaming or the experiences of people who identify as conducting RAI red teaming [39, 43, 79, 95, 133]. Because of the adversarial nature of RAI red teaming work, it should not be prematurely assumed as equal to or distinct from the other forms of content work.

Our work proposes a renewed and holistic examination that is inclusive of data labeling, content moderation, and red teaming as an emerging form of content work. Such examination is important for the design a supportive framework for RAI content work and timely because all such forms of content work may increase in demand and become further commercialized as AI becomes prevalent. In doing so, we leverage past learning from well-known professions, such as content moderation, and incorporate the lived experiences of all types of content workers to inform recommendations for existing and emerging RAI practices.

2.2 Well-being Challenges for Content Workers from Harmful Content Exposure

The labor of annotating, moderating, and testing data for RAI poses specific challenges concerning the well-being of content workers, as they consistently face the risk fo encountering potentially harmful content. Prior research highlighted the emotional toll on content workers along a wide spectrum of psychological impacts. Impacts include symptoms of secondary trauma, burnout, a sense of undervaluation for their contributions, feelings of privacy infringement [33, 114, 127, 152] as well as alterations to their belief systems [34, 99, 140] and the development of post-traumatic stress disorder (PTSD) [4, 36, 94, 124]. Unfortunately, much of the prior research on content workers remain in the volunteer or crowdwork domain. On the commercial domain, there is limited knowledge about content workers' challenges and the types of support they typically receive in organizations [13, 121]. For example, CCM research often relies on end-user perceptions due to limited transparency from platforms [98, 145]. Some journalists have reported on their challenging working conditions as working up to nine hours a day in what they described as crowded workspaces [36, 134].

Given the limited empirical data on the impact of reviewing harmful content on content workers in the commercialized or professional setting, prior research has turned to analogous fields to anticipate challenges RAI content workers face. Steiger et al. [143] noted similarities between the psychological impact experienced by content moderators and professionals like journalists, emergency dispatchers, and sex-trafficking detectives who witness traumatic scenes, potentially leading to PTSD, peritraumatic distress, secondary traumatic stress, and burnout [9, 42, 143, 147]. Expanding on this analysis and recognizing the cognitive dissonance that RAI content workers my face of holding adversarial or harmful values along with their personal and societal values, we look to numerous other occupations that confront similar moral dilemmas and psychological

distress. For instance, military soldiers and healthcare professionals often grapple with moral injury when making decisions that conflict with their personal beliefs [31, 53, 84, 91]. Similarly, military interrogators, actors, and ethical hackers may face moral conflicts when required to adopt roles contrary to their values [3, 70, 78]. Social workers navigate bureaucracy and systemic issues, paralleling content workers who must confront distressing content they cannot prevent [56]. Our work aims to highlight lived experiences and well-being challenges of RAI content workers in professional and organizational settings and to explore the unique moral challenges encountered by content workers.

2.3 Workplace Support for Well-being and Harmful Content Exposure

In line with the plethora of challenges identified for content workers, emerging literature has examined how content work may be supported from multiple angles. Prior studies have explored the benefits and implementations of reducing exposure by limiting the visual field [82], limiting color display [138], limiting amygdala activation from viewing facial emotions [25, 45], reducing intrusions [64], displaying content in monochrome greyscale, blocking faces, blurring visuals, muting audio, controlling the speed of videos, and allowing content workers to play Tetris immediately after viewing content [30, 63, 73]. Other well-being approaches include monitoring the short- and long-term impact of viewing content [150], programs to improve resilience skills [142], evidence-based psychotherapy [24, 29, 139, 151], and a suite of workplace, clinical, and technological interventions [143].

While much work on content workers we outlined has focused on content exposure and mitigation strategies, our work examines the broader workplace context surrounding a content worker because of the intricate relationship between work, individuals, and their well-being [8, 18, 77, 122, 136]: work can simultaneously be a source of well-being support via Employee Assistance Programs (EAPs) [89], a source of meaning [154], and a source of well-being challenges [23, 54, 103]. Therefore, in addition to the utilization of exposure reduction tools in their day-to-day work, we examine content workers' physical work setup, organizational support, and its utilization, as well as their perception toward work and place in the broader context of society. We also investigate how workers access or incorporate treatment resources, such as therapy, to cope with symptoms of exposure with organizational support.

Automated content regulation, which reduces moderator exposure to harmful content, has also been explored through various algorithmic approaches in images, videos [32, 119, 149], and text [50, 57, 58, 72, 86, 109, 125, 126]. These methods often target specific categories of harm, including pornography, pro-eating disorder content, mental health content, personal attacks, and hate speech [16, 58, 86, 132, 155, 161].

Recently, tools have been developed to increase automation in RAI red teaming [7, 92], raising concerns about their robustness and coverage of discovering harms and effectiveness in reducing harm exposure. In general, the introduction of automated content work can lead to concerns regarding potential automation errors [67] or exacerbating the problem they aim to solve [51], even acknowledged by platform leadership grappling with nuances [12]. In our work, we explore the benefits and challenges of using automated tools from workers' perspectives and the organizational support for such tools.

3 Phase 1: Understanding the Nature and Challenges of Content Work

To contextualize content work and the well-being of content workers within the space of generative AI deployment and safety, we examined content moderation and data labeling, alongside the emerging practices of responsible AI (RAI) red teaming. Our research questions are as follows:

Phase 1: Survey	
Activity	*Count*
CM only	27
DL Only	11
RT only	4
CM and DL	20
DL and RT	1
All	4

Phase 1: Interview	
Activity	*Count*
CM only	5
DL Only	2
RT only	1
CM and DL	8
DL and RT	3
All	3

Phase 2: Workshop	
Self-identified role	*Count*
Content moderator/analyst	7
Content designer	3
Data annotator	3
Operations or enforcement manager	2
Volunteer AI tester	1

Table 1. Number of participants in various self-reported content activities or roles across three studies (survey, interview, workshop) within two phases.

RQ1 What is the nature of content work, in terms of professional roles, the types of content handled, the work environment, its impact on well-being, and collaborative practices, as experienced by content workers?
RQ2 What are the main challenges faced by content workers in their well-being?
RQ3 How do we best support the well-being of content workers?

We broadly define "content work" as any work activity related to anticipating, generating, reviewing, reasoning about, or making decisions on digital content. To examine the diversity of content-related work experiences, we divided content work into three role categories of content-related activities, which we used throughout our study: (1) *Content Moderation* involves reviewing various forms of online content (e.g., text, photos, audio, and video) with the intent to flag or identify any content that potentially violates the platform's policy or guidelines. (2) *Data Labeling* involves reviewing, labeling, or categorizing various types of content (e.g., text, photos, audio, and video) according to specific labeling or sorting guidelines, which aids in data analysis and training machine learning models. (3) *Red Teaming* involves critical assessments of product or platform features by simulating the actions of potential bad actors or testing system vulnerabilities to identify if these features can inadvertently generate or promote content that violates policy or guidelines, thus enhancing the product's security and safety measures. In our study, we specifically refer to RAI red teaming, which evaluates AI system outputs for potential harm.

We conducted a two-phase study (Figure 1), focusing on content workers engaging in the three categories of activities above, who were employed in some capacity (i.e., full-time, part-time, contractor) to conduct content-related activities. The first phase was formative and aimed at answering our research questions via a survey and interviews with various content workers. In this section, we present our survey and interview protocols, our findings from both studies and a set of preliminary recommendations aimed at supporting the well-being of content workers. The second phase involved validating our findings and recommendations with content workers in a series of workshops. In Section 4, we present our workshop protocol and a set of refined recommendations. The distribution of self-reported content activities and roles across our phases are found in Table 1.

3.1 Methods

Our mixed-methods formative study aimed to answer **RQ1** and **RQ2** by conducting a survey and semi-structured interviews in parallel. We used the survey to get a broad understanding of the diversity of experiences and the interviews to obtain a detailed view of how content work was carried out and the challenges associated with it. Our study was conducted between June and August of 2023.

3.2 Survey Study

Survey participants were recruited through a sample of manually validated contacts and referrals via a snowball sampling method. We reached out to several technology companies that either conducted content work in-house or hired vendor companies, as well as vendor companies who conducted content work for these technology companies. These companies helped expand our reach by sharing the study information with other organizations and appropriate individuals involved in the content-related activities defined above. This 15-20 minute survey was anonymous, and participation was voluntary, with a strong reminder to avoid accidental employer disclosure to protect participant anonymity (see Appendix A.1). To encourage participation, we did not collect employer data, acknowledging this introduces some opacity as a limitation. Participants received a $10 gift card via an independent survey that only captured their contact information for compensation and interview follow-up. The lead institution's Institutional Review Board (IRB) approved our survey study protocol.

We received a total of 67 complete responses (see Table 1 for activity distribution). The median age of our survey population was 30. All reported the highest level of education to be at least some degree post-high school diploma (i.e., vocational training, Bachelor's degree, postgraduate degree, and beyond). 47.8% of our population identified as female, 49.2% as male, and 3.0% as non-binary. 77.6% reported as being full-time employees, 20.9% as full-time contractors, and one as a part-time employee. Most content workers had at least two years of experience in their roles, with 22.4% having more than five years, 31.3% having 2-5 years, and the rest having less than two years. In this paper, we note our survey participants with the prefix S.

Our survey questions included several sections aimed at understanding the nature and impact of content work: (1) *Background:* We obtained basic demographic information (e.g., age, gender, education), employment status, involvement in different content activities described above, and tenure in these activities. We asked participants to describe their motivations for conducting current work. (2) *Work Description:* We asked participants to report the modality (e.g., text, image, audio, video) and categories (e.g., child abuse, graphic violence, sexual content) of content they worked with, the average weekly hours, and average daily contiguous hours of content exposure. (3) *Work Tools and Strategies:* To understand tools and strategies used for work, we enumerated 17 work tools and 23 coping strategies based on existing resources outlined in Section 2.2 potentially used during or after reviewing or generating content and asked whether they had access to them, whether they used them, and how useful they were. All tools and coping strategies listed in the survey can be found in Figures 2 and 3. (4) *Work Impact:* To understand how content work impacts well-being, we asked participants to describe the positive and negative impact of work, challenges related to content work, opportunities for supporting their work, and the impact of the recent rise in the use of generative AI technologies. We asked participants to subjectively assess the quality of sleep (5-points) and the frequency of nightmares, flashbacks, or intrusive thoughts (7-points), as they are relevant symptoms often associated with PTSD [5]. We provide our full survey questionnaire in Appendix A.2.

3.3 Interview Study

Interview participants were recruited from those who expressed interest in follow-up interviews in the survey or through referrals. These participants were then consented to participate in a 1-hour interview study. Each interview was recorded and transcribed over video-conferencing software. We asked participants to report which of the three content-related activities they conducted in their role to ensure diversity in sampling. Participants were compensated with a $50 gift card. Our interview study protocol was approved by the lead institution's IRB.

We conducted 22 semi-structured interviews where 16 reported engaging in content moderation, 16 for data labeling, 7 for red teaming, and 8 who identified with just one activity. The median age of our interview population was 30. 12 participants identified as female, 9 as male, and 1 as gender nonbinary. 18 were full-time employees, and 4 were full-time contractors. In this paper, we note our interview participants with the prefix P.

Following the aspects of the nature of content work we are interested in (**RQ1**), we structured interviews into four main sections: (1) *Work Setup:* We asked participants to describe their roles, work activities, and the types of content they encountered, followed by their work location, physical setup (e.g., equipment), environment (e.g., noise, comfort), social setup (e.g., access to colleagues, collaboration), and the ideal work setup for them. (2) *Conducting Work:* We asked participants to describe how they conduct their content-related work activities, by sharing how they interact with their tools. We explored how automation or generative AI technologies is or could be helpful in their work as a tool used for work (e.g., filtering, labeling, prompt generation) and as materials for work (e.g., AI-generated images). (3) *Coping:* We asked participants to describe how their well-being was impacted by their content work, including their current and ideal strategies, resources, or support they leverage to improve their well-being. (4) *Collaboration:* Finally, we asked participants to describe how they collaborated with others for work, including getting help for content-related work or for social support. We provide a full list of questions in Appendix A.3.

3.4 Survey and Interview Data Analysis

For qualitative data from open-ended survey responses and interview transcripts, we applied reflexive thematic analysis [93] and open-coded [19] qualitative texts. The survey responses were affinity mapped using FigJam[1] by two researchers. The interview transcripts were uploaded to Marvin[2] and open-coded using the tool by the first author on a granular, line-by-line basis and then with higher-level themes in mind in the following coding iterations. At the end of the coding process, three researchers iteratively reviewed codes, resolved disagreements, and refined or grouped codes to identify overarching themes.

For quantitative data from survey responses, we computed percentages of participants that reported different roles, amount of work, exposure to various content types and categories as well as access to tools and coping strategies. We conducted Pearson's correlation where appropriate. All correlation results reported in this paper are statistically significant with $p<0.05$.

3.5 Privacy and Positionality Statement

Our study focused on the well-being of individuals who self-identify as employed in content work roles. As such, we placed great emphasis on acknowledging and addressing potential participant concerns about the privacy and sensitivity of the data with respect to their employment. To conduct our investigation, we gathered perspectives from participants across various workplaces. To ensure that participants were comfortable during the study, we emphasized that participation was voluntary at any point. Additionally, participants were told their responses would remain anonymous, and they were given the option to remove any data they provided to us afterward. We especially encouraged participants not to disclose their place of employment or personally identifiable information throughout our survey, interviews, and workshops. After collecting all of our data, we carefully de-identified all accidental disclosures of information that could result in identifying a participant. Therefore, we did not disclose specific organizations or roles of individuals and instead reported demographic data in the aggregate.

[1]https://figma.com
[2]https://heymarvin.com/

We recognize that the perspectives and biases we hold as a research team play a role in our study. Our research team brings interdisciplinary research expertise in HCI, CSCW, critical computing, affective computing, and psychology and comprises individuals with diverse gender, racial, and cultural backgrounds, including people of color and immigrants. We acknowledge that, as researchers, we may have influenced the system under study by observing and probing it. We provided participants with a space in which to reflect on aspects of their work experiences and ways to improve them, and such an opportunity to voice their perspectives may not have been readily available. It is in this way that our research and biases may, in turn, have provided opportunities for understanding and empathizing with participants. On the other hand, we note that we bring biases in our interpretations of what content work is because, as researchers, we cannot assume that we understand it as well as those who conduct such work on a daily basis.

3.6 Survey and Interview Findings

In the following section, we detail our findings addressing our research questions focusing on the nature (**RQ1**) and challenges (**RQ2**) of content work.

3.6.1 RQ1: What is the nature of content work? In exploring the nature of content work, we aim to provide a detailed understanding of the main factors that constitute this field. Content work involves a variety of roles, types of content, impacts, protective tools, and collaboration practices. By examining these elements, we clarify what content work entails and the various factors that influence it. This context serves as a foundation for understanding the complexities and scope of content work.

How content workers define their profession Our participants reported performing various content-related activities, including labeling text for harm categorizations, triaging user feedback, deciding on content removal or reporting it as illegal, and generating prompts to assess harmful content and model restrictions. Participants performing a variety of these activities in both our survey population and interviews described doing so specifically for AI systems or anticipating their work to shift to focus more on AI-related harms. Our study found that 37.3% (25/67) of survey participants reported performing activities spanning multiple role categories – content moderation, data labeling, and red teaming – indicating that the boundaries between these roles are not always distinct. For instance, 39.2% (20/51) of content moderators also did data labeling, and 4 out of 9 red teamers participated in content moderation and data labeling, pointing to a need for a broader categorization of content work as a diverse set of roles.

In addition to the diversity and blurring of roles within content work, we found that content workers collectively differentiated their roles through professional requirements. These include self-awareness of personal comfort boundaries with content (e.g., P5, P11), resilience to cope with psychological demands (e.g., P4, P22), subject matter expertise in harmful content (e.g., P6), analytical and investigative skills to interpret human language nuances (e.g., P10, P13), and empathy. For instance, P11 emphasized the need for self-awareness to recruits: *"I want you to be honest with yourself about how willing you are to talk about sexual content, about profanity about religion, about political beliefs, and not only that but to understand the opposing views of those subject matters."* Participants highlighted the need to adopt the perspective of harmful content creators (e.g., a white supremacist) to identify *"coded language"* (P10) used to perpetuate bias or to understand how *"people operate differently in those two environments online versus at home"* (P16). Red teamers described their investigative processes as intuitive, adaptive, and developed over time, with P21 emphasizing the need for *"a team of people who are looking forward and anticipating and adapting to the changing circumstances."* Overall, our participants highlighted the diverse activities and unique skills required for content work, indicating that this profession may not be suited for everyone.

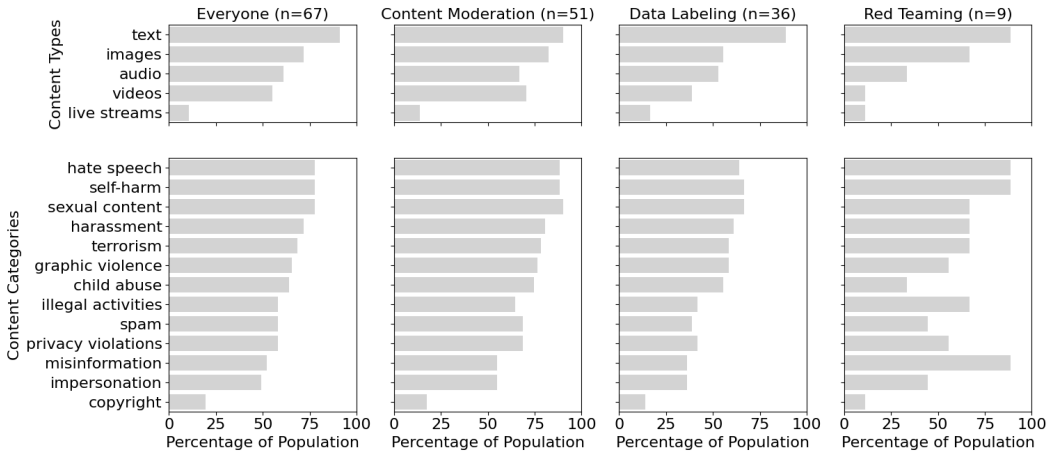

Fig. 2. Percentages of our survey population exposed to various content types (e.g., text, images, audio) and categories (e.g., hate speech, self-harm) across everyone (N=67), content moderation (N=51), data labeling (N=36), and red teaming (N=9).

Types and modalities of content that workers engage with We found significant overlap in the types and modalities of content our participants reasoned about, with most reviewing multiple modalities and categories of impactful content. 34.3% (23/67) of survey participants reviewed all four modalities (text, images, videos, and audio), while 20.9% (14/67) reviewed two modalities. Text and images were the most commonly reviewed, with 90.2% (46/51) of content moderators, 88.9% (32/36) of data labelers, and 8 out of 9 red teamers engaging with text, and similarly high percentages for images (82.4%, 55.6%, and 6 out of 9, respectively). The content source also varied, as several interview participants (9 red teamers, P18, and P19) reported working with AI-generated content or content within AI systems. Moreover, interview participants reported exposure to highly impactful content, such as abusive or hate speech, child sexual abuse material, and terrorist and violent extremist content. Those involved in red teaming described their work as involving both *"generating"* and *"processing"* such content. For instance, P6 described their role as dealing with *"explicit content that is either sexual content or child endangerment content. Violence, racism, the bevy of the worst of the worst of the Internet is what I have to generate and also test, moderate, and sift through."* Figure 2 shows that half of our population was exposed to all categories of content except for copyright (19.4%; 13/67) and others (15.0%; 10/67). Red teaming participants reported higher exposure to hate speech, self-harm, and misinformation (all 8 out of 9) compared to the general survey population (77.6%, 77.6%, and 52.2%, respectively). Content moderators encountered sexual content more frequently (90.2%; 46/51) than the population (77.6%; 52/67). These findings suggest that content workers are frequently exposed to highly impactful content regardless of their specific roles.

Impacts of performing content work Prior research has shown the significant negative impact of exposure to harmful content (Section 2.2). Our findings confirm this but show that symptoms vary by individuals and specific activities. We found that the negative psychological impacts persisted long after exposure, resulting in residual effects such as moral injury, lower sleep quality, intrusive thoughts, and hypervigilance. We found that moral injury [129] stemmed not only from viewing content conflicting with one's moral values or beliefs but also from the adversarial red teaming. For example, P6 stated that they had to *"dive into white supremacist blogs...absorb the information of*

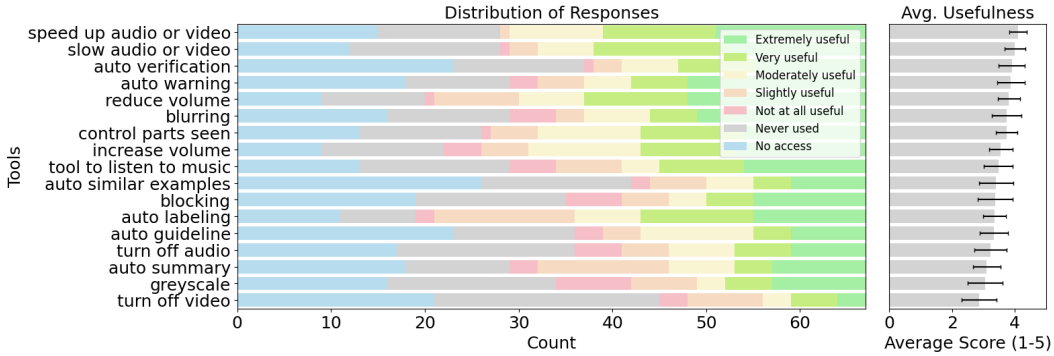

Fig. 3. The access and usefulness of tools that help during content work which are sorted in descending order of usefulness with 95% confidence intervals.

how people talk there, how people utilize language, and then bring that back to the team and process that together." S11 described their work as *"scarring [their] brain for money"* since content work activities *"can be stressful and emotionally impactful,"* highlighting the emotional toll. We found that the better a content worker red teamed a model, the worse they could feel about themselves for making that model generate harmful output.

Sleep quality was on average "fair" to "good" ($\bar{x}$=3.46 out of 5, σ=1.03), but lower sleep quality was significantly correlated with weekly exposure (Pearson r=0.357) and contiguous hours worked per day (Pearson r=0.430). Nightmares or intrusive thoughts were rare ($\bar{x}$=2.56 out of 7, σ=1.61) and negatively correlated with sleep quality (Pearson r=-0.274). S9 and S19 frequently revisited past choices, experiencing intrusive thoughts and guilt about their work outside work hours. Exposure to harmful content increases sensitivity and hypervigilance in daily life. For example, S47 mentioned being *"unable to watch certain TV shows, movies or even listen to stories or podcasts that involve child endangerment, child crimes, gore, severe violence, mutilation."* S2 reported feeling less empathy and compassion, while P12 felt both hardened and hypersensitive.

Participants experienced exhaustion from viewing high quantities of content for extended periods. Specifically, 61.2% (41/67) reviewed or generated content for 30-40 hours a week, with 44.8% (30/67) spending more than four contiguous hours per day (median = 3.5 hours per day). Red teamers, on the other hand, were exposed to content for a median of 15 hours per week, with a median of 1.5 contiguous hours per day. Several interviewees, especially those in content moderation, lamented the physical and psychological distress from long exposure to harmful content. For example, P13 mentioned that, *"Sometimes it's nearly a full day's work just going through toxic reports ...it's incredibly draining."* Such findings indicate that the long work hours compounded the other impacts of content work described earlier.

Tools content workers use for protection In terms of the tools participants used for protection while performing content work activities, we found that the most useful tools were video and audio speed controls (4.1), automatic verification (3.9), and automatic warning (3.9). We observed a significant positive correlation between a content worker's contiguous exposure to content and the usefulness of blurring (Pearson r=0.340) and blocking faces (Pearson r=0.476). As illustrated in Figure 3, these tools received high average usefulness scores.

Interview participants who red-teamed disclosed access to a differing subset of tools specific to their role, such as those for prompting models. For instance, red teamers used a prepared list of prompts that workers would then send as input to models and wait to evaluate model outputs (P19,

P20). P20 found this tool useful, saying *"the advantage of using this data set would be [reassuring of] 'okay, I didn't come up with this.' I just use whatever inputs are there are run them."* Thus, we find that tools used by red teamers may aim to alleviate the internal dissonance they face when generating harmful prompts that go against their beliefs.

How content workers collaborate as a team The final aspect of content work we explored was its collaborative nature. Participants benefited from diverse teams, which provided a diversity of perspectives and allowed them to tailor their work experiences to their personal preferences. P5 stated that each team member had unique domain expertise, such as vendor management, user interface design, and video editing to *"spot fakes"*, recognizing generated harmful content. This diversity enabled P20 to divide responsibilities according to team strengths when red teaming, while P6 typically processed content collaboratively with their team. However, P20 cautioned against relying solely on domain expertise, emphasizing the importance of lived experience: *"you don't want a team full of white guys testing the feature. You want everybody there...But that's not the only way, especially if...seeing that repeated offensive information about it can have a long-term effect on you."* Ultimately, some content work activities benefit from collaboration, but the collaborative nature also opened up challenges in questions about how diversity should be incorporated in a team without overburdening individuals who may have more underrepresented skills or experiences with needing to do more work.

3.6.2 RQ2: What are the main challenges faced by content workers in their well-being? In this section, we illustrate key challenge our participants experienced as content workers, which informed our recommendations listed in Table 2. Each subsection details the identified challenges and the corresponding recommendation categories we developed.

Misconceptions and realities of content work Several participants found their initial perceptions of content work to be inaccurate after starting their roles. While some viewed it as *"just a job"* (P4, P7, P17), this view was often challenged by the realities of the work's impact. P2 reflected: *"I was like, oh, this is going to be a really easy job...And it wasn't easy."* Participants used various methods to cope with the "not easy" parts of content work, such as putting up figurative shields (P9), but these often backfired. As P13 noted, becoming *"hardened to [harmful content] in the sense that it doesn't impact [them] as much...can cloud [their] judgment"* due to frequent exposure. Many participants expressed the need for adequate warnings about the nature of content work, including the activities performed (P2, P17), types of content (P2, P12, P17), and potential impacts (P17). P19 recommended surveying new workers about their comfort level and offering alternative assignments. Training resources, such as video curricula with coping strategies, were also found helpful (P1).

From these challenges, we developed our first category of recommendations: *Recruitment.* We urge that potential RAI content workers receive comprehensive education about the impact of content exposure. This education helps them assess their suitability for content work, which demands a specialized skillset honed through experience. Early, personalized training should equip new hires to cope with the highly individual naure of exposure-related symptoms, potentially utilizing tools like a generative training dataset for controlled exposure.

Shortcomings of existing tools and metrics We identified several challenges with the tools available to content workers. Many tools were inaccessible to some participants, including automatic clustering (38.8%; 26/67), automatic verification or guidelines (34.3%; 23/67), turning off video (31.3%; 17/67), and blocking (28.4%; 19/67). 35.8% (24/67) and 28.4% (19/67) of participants who had access to turning off video or audio never used these features. This is concerning as these tools were particularly useful with greater contiguous exposure to impactful content (see section 3.6.1). Participants who regularly used tools recognized the need to address variability in tooling

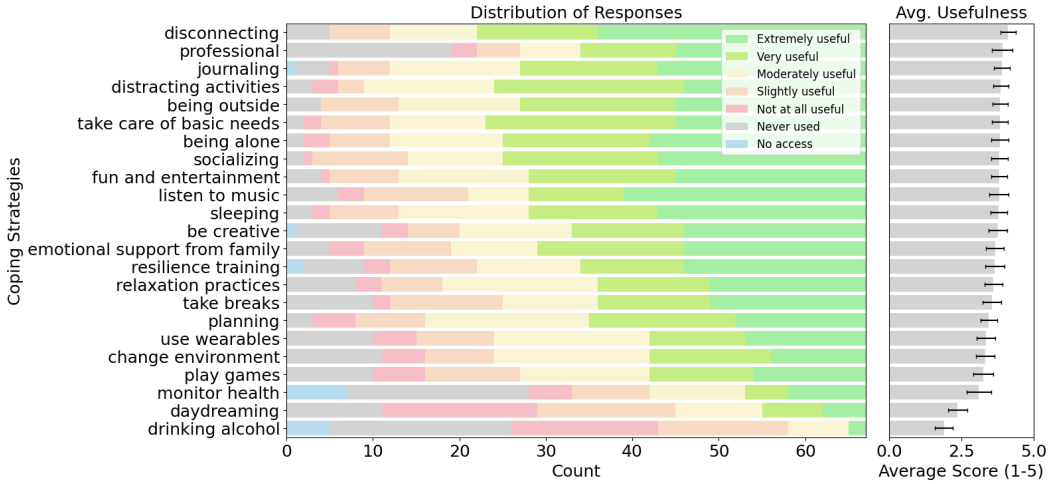

Fig. 4. The access and usefulness of coping strategies used to manage the demands of content work which are sorted in descending order of usefulness with 95% confidence intervals.

needs. For instance, while P13 viewed images unfiltered to understand the context, P7 needed the black-and-white filtering feature. Generative AI was mentioned as an emerging tool, but has many areas for improvement in application. For instance, P6 emphasized the need for human involvement due to the focus of content work being on *"the bleeding edge of technology"*, making it difficult to automate red teaming fully. As such, challenges remain in reconciling the benefits of generative AI with the drawbacks of it potentially exacerbating the burden on workers by creating more content for workers to engage with. P11 anticipates this being a key challenge: *"I have a list of 100 terms, and I use a prompt against it to identify which of these terms is sexual content [...] I run that prompt four times. If each instance of that prompt results in a different result, how do we reconcile that?"*

Participants also highlighted challenges with how productivity was measured, sometimes facilitated by the tools they use for work. Many felt pressured to work long hours because their productivity was measured by time worked or content volume reviewed. Instead, they preferred metrics that helped them manage when to take breaks (P1) and recognize the impact of their work. P7 noted that being told to view content often worked against their productivity on days with insufficient content to review. Participants were more receptive to a combination of metrics, including content severity or the severity of impact as well as the quantity and time spent viewing content.

Based on these findings, we recommend the category of *Tooling*. Tools should be customized to reduce content exposure based on individual preferences. Continuous feedback from content workers should be incorporated into tool improvement and creation to ensure relevance to their needs.

Failures in workplace support and coping mechanisms We found that content workers use highly individualized and specialized methods to cope with the psychological demands of their work. Most coping strategies were accessible, but some lacked access to health monitoring (10.4%; 7/67) and resilience training (3.0%; 2/67). Despite all participants reporting having access to professional support services (e.g., therapy), only 71.6% (48/67) of participants used them, with only 5 out of 9 red teamers participating. The most useful coping strategies were disconnecting ($\bar{x}$=4.1 out of 5,

σ=1.1), professional support ($\bar{x}$=3.9 , σ=1.3), journaling ($\bar{x}$=3.9 , σ=1.1), distracting activities ($\bar{x}$=3.9 , σ=1.1), and being outside ($\bar{x}$=3.8 , σ=1.1), as shown in Figure 4.

Interview data highlighted variabilities in individual preferences. For example, P7 found work therapy sessions unhelpful and *"[felt] like a waste"* because they were *"not allowed to talk about issues that are bugging [them] from outside of work."* In contrast, P3 benefited from workplace coaching. P20 found group therapy helpful for shared experiences and found it to be *"a really nice space because [they] get to see what everybody else and going through and [they] realize that sort of a lot of [them are] having the same or similar experiences."* However, we found an under-utilization of professional support services among red teamers, partly due to the nascent nature of their roles. For example, P21, who was satisfied with playing Tetris or taking walks for occasional red teaming and seeing a disturbing image, conjectured that *"if somebody was doing it for six and a half hour days, I would probably say mandatory. Like, I might literally force people to play Tetris if they had to do that much content moderation or something akin to it because, boy, it'd be hard."* P19 did not feel affected enough to join support services, and P18 highlighted that *"[well-being sessions] are not for vendors."*

Access to coping strategies was also limited by the work environment. P9 found home-based strategies, like having pets or listening to music outloud, effective, while others preferred access to colleagues in the workplace for fostering stronger interpersonal relationships. P5 found regular walks with colleagues beneficial for well-being *"check-in"* and venting. P7 noted challenges of working in the office as well as working from home. They noted being *"blown away in the face"* by the content their colleagues were viewing when they pass by their screens in the office and causing *"second-hand smoke"* when seeking support from their loved ones at home. These findings underscore the need for *Adaptive Wellness*. We recommend flexible and individualized support for content workers, recognizing the varied impacts of their work. Providing diverse well-being resources and fostering workplace connections are crucial for validating their experiences and offering support through both informal and formal interactions.

Barriers to career growth and support We found that participants faced challenges regarding access to career opportunities and resources, as many lacked an understanding of career growth in content work. Most participants viewed content work as an entry-level position with opportunities to advance in fields like user experience design, project management, or software development (P6, P9, S33, S60). For example, P16 remained in their role due to advancement opportunities: *"I started showing that skill set to want to dive deeper and to ask more questions...when the interview process came up for the tier two position, I was asked if I was interested in interviewing."* Conversely, P5 felt there were no growth opportunities: *"We don't have networking opportunities, and I don't see any growth. The way I grow in this role is I get a new skill set and then leave this role. There's no promotion. There's no career pathway."* The perception of content work as temporary and replaceable thus contributed to the lack of well-known career opportunities.

To remain motivated, some participants sought closure by learning about the positive impacts of their work through reports of the number of children they save (P5) or progress on investigations (P13), but accessing such data was challenging due to the sensitive nature of the content (P2, P4, and P16). S39 found it rewarding: *"While it is hard seeing some of the content I am exposed to, it can be a bit rewarding, knowing that you are keeping more vulnerable users safe."* Even then, several participants expressed frustration that the public did not understand the importance of content work. P5 recounted a discouraging remark: *"Man, your job sounds great. you just look at porn all day."* P21 highlighted the lack of positive recognition: *"You only notice [red teaming work] when something goes wrong. You don't notice it when it's going well."*

Participants expressed concerns about growing workloads due to the rapid development of AI systems and the insufficient resources for content workers. P20 described the constant pressure on their team, noting, *"given how much we're shipping, how many features we've had to review, and how much [time in] our schedule, we allow for a certain number of [sessions] in a week, and we're always overbooked. "* Anticipating similar challenges, P18 advocated for hiring more content workers but worries about the lack of resources for current workers. Generative AI and automation have been introduced to reduce workloads and exposure. Some welcomed having automated tools for the purpose of reducing exposure and potentially lessening the negative impacts of their work (P12 and P15). While we anticipated a concern about automation replacing workers, our participants did not share this sentiment. Several participants were confident that advances in automation would not drastically alter their work, believing that such attempts could not replace their roles, saying that *"no amount of automation is going to be able to do the nuanced work that humans can do in this thing"* (P11). Crucially, participants knew that several automated methods do not even apply to activities such as red teaming for generative AI (P13 and P15). Thus, we confirm that a challenge persists in supporting content work amid growing demands in AI development while utilizing state-of-the-art technologies such as generative AI to complement existing workflows.

Participants also emphasized the need for supportive leadership that understood content work, advocate for and listen to workers (P5, P9). P9 argued that stakeholders need first-hand experience: *"it's just one of those things where unless you're actually in it, it is difficult to understand the gravity and impact of the work."* With increasing workloads, the need for strong, capable leaders is urgent. Specifically, red teamers expressed concerns about rising demands (P8, P20), with P8 expressing nervousness: *"We expect this [type of content's] volume to increase significantly, and keeping up with it will be a real challenge. I do not know how to do that, and I'm nervous about that. Everybody else I'm working with is nervous about that."*

Overall, we highlight the need for clear career pathways and growth resources facilitated by prepared leadership in the face of increasing workloads. Our final category of recommendations,*Retention*, calls for efforts to demonstrate the value of RAI content work and support skill development and retention. Career pathways and transferrable skills learning can facilitate transitions from perceived short-term roles. Networking opportunities, such as internal conferences, can promote personal growth and validation.

4 Phase 2: Evaluating Recommendations

In the second phase of our study, we aimed to address **RQ3**: *"How do we best support the well-being of content workers?"*. To do this, we validate our recommendation framework and surface persisting challenges and consideration in its implementations through a series of small-group workshops with those who identify as conducting content moderation, data labeling, or red teaming activities. Our workshops are designed to facilitate a form of "member checking" [27, 83, 158] as a way to validate our findings from the first phase [28, 83] and to provide an avenue for consensus and empowerment, through highlighting participants' voices on the feasibility and implementation of our recommendations within the complex individual, organizational, and societal contexts our participants navigate daily [21]. In this section, we present our workshop protocol, our findings from these workshops, and revised recommendations.

4.1 Workshop Methods and Data Analysis

The recruitment method for workshops was similar to that of the interviews. Consented participants attended a 1-hour workshop session conducted remotely over video-conferencing software and

FigJam[3]. We recruited a total of 14 participants across 2 workshops (Table 1). 11 participants reported as being full-time employees and 3 as being full-time contractors. 5 participants identified as female, 8 identified as male, and 1 identified as transgender. The median age range of our workshop participants was 36-45 years old. All reported the highest level of education to be at least some degree post-high school diploma (i.e., vocational training, Bachelor's degree, postgraduate degree, and beyond). Participants were compensated with a $50 gift card. Our workshop study protocol was approved by the lead institution's Institutional Review Board (IRB).

We kicked off each workshop with a quick introduction and tutorial on FigJam (10 minutes). In the first portion of the workshop (20 minutes), we asked participants to review challenge themes we discovered in phase 1 and vote on those that they agree or disagree with, comment on whether they think the challenge is accurate, react to each other's notes, and discuss emerging themes. In the second portion of the workshop (30 minutes), we asked participants to review the AURA framework with its preliminary recommendations from phase 2, comment on what they liked, what they didn't like, what they would change, how the recommendations could be improved, how they would want the recommendation implemented, and discuss emerging themes. Detailed structure of the workshop, including the challenges and recommendation cards and screenshots of the interactive boards presented to the participants, can be found in the Appendix A.4.

We applied reflexive thematic analysis [93] on workshop responses, both written and spoken. Participant reactions were systematically collated to assess the accuracy of challenges by ascertaining the extent of consensus or divergence among participants through voting. We further analyzed the remaining qualitative data by surfacing themes in responses to our proposed recommendations. In this paper, we note our workshop participants with the prefix W.

4.2 Workshop Findings

In the following section, we present our participants' discussions on the identified challenges and concerns regarding our recommendations. Overall, our findings validate the AURA framework and its four pillars. They emphasize the importance of our recommendations and highlight participants' concerns about their implementation. These insights are integrated into our revised table of recommendations (Table 2) to enhance the comprehensiveness and impact of our guidance.

4.2.1 Agreement with Discovered Challenges As we delved into the challenges unearthed during the initial phase of our study, we found participants actively engaged with these challenges and envisioned their personal relevance. Participants most strongly agreed that six of the eleven challenges we identified were accurate: psychological and physical symptoms from generating, analyzing, or researching harmful content, lack of career opportunities to grow in content work, an inaccurate measure of productivity or exposure, lack of specialized tools for content work, lack of appreciation for content work, and lack of closure on the positive impact of work. W3, reflected on how accurate the challenge of lack of career opportunities was: *"the skills [they] develop does not feel like it opens many opportunities nor are there many closely adjacent roles. As such, there's a limited number of options to progress, mostly into people management positions"*. On the other hand, participants least strongly agreed with the accuracy of three challenges: the double-edged sword of heroism, moral injury from working with harmful content, and inaccurate expectations about the cost of content work. For instance, while W10 felt *"[they were] 100% aware of what [they were] signing up for,"* they still noted the inaccuracy in job descriptions for new hires and advocated for the implementation of a recommendation to address this. As such, even when participants did not experience the challenges themselves, they valued the recommendations for others facing these issues.

[3]https://figma.com

	Recommendation		Example Applications of Recommendation
Recruitment	R1	Ensure potential participants of RAI content work are informed about the psychological impact of content exposure.	Providing descriptions of types of content that are present in the datasets within content work job descriptions
	R2	Educate content workers with basic and ongoing training to minimize exposure and ease psychological burdens associated with content work.	Providing AI-generated examples of each type of content a new hire will be exposed to
Tooling	R3	Allow configuration of tools to accommodate individual sensitivity to content exposure.	Limiting content workers to viewing one case at a time (e.g., one image at a time)
	R4	Perpetually integrate feedback into tools to stay responsive to the evolving demands of content.	Providing a platform for bug and feature requests
Adaptive Wellness	R5	Provide flexible well-being support to accommodate highly variable responses to content exposure.	Providing a craft table for content workers to engage in creative hobbies (e.g., origami)
	R6	Foster strong interpersonal work connections that can act as first-aid to the psychological burdens of content work.	Organizing optional team lunches with food paid for on a monthly basis
Retainment	R7	Enable content workers to observe the beneficial effects of their work, mitigating psychological stress resulting from feedback absence and demonstrating the value of their contributions.	Sending a weekly newsletter with a digest of a team or individual's contribution to keeping a platform safe
	R8	Design well-defined career pathways for content workers to foster the retention of domain experts.	Organizing a conference on responsible AI for content workers to network with peers across teams and organizations

Table 2. Revised AURA framework and its eight recommendations to support RAI content workers, categorized by the framework's four key categories of holistic support.

4.2.2 Implementation of Recommendations Through discussions with participants, we surfaced many concerns regarding how the recommendations should be implemented in practice. We integrated these perspectives into our revised recommendations, providing detailed considerations in the subsequent section and updated examples illustrating our recommendations' application. We also note that participants did not propose additional recommendations beyond those we presented; instead, they expressed concerns with a critical focus on how our recommendations may be implemented. When considering hosting a conference for learning and networking, W14 exclaimed that *"If this could be shared from [their] managers to the team, it would be helpful."* Conversely, W13 expressed concern that attending such events would take away from their assigned production hours–current productivity metric–to attend such an event, making them less likely to go. From

another perspective, W1 suggested making such meetings *"a peer-to-peer meeting style that allows for a collective group of content moderations/designers/researchers to nominate an agenda."* These concerns highlight the need for organizational and managerial support because, as our participants reported, content workers often lack the resources to initiate such changes. Further from W1's insights, we find that organizers should carefully account for workplace dynamics to ensure content workers are in spaces they deem productive for growth. However, there were also recommendations that content workers could support independently. For instance, W10 stated that *"[their] team is currently interviewing candidates for an open role. The job description is absurdly vague and doesn't accurately describe the work or its impact,"* highlighting an opportunity for content workers to improve job descriptions. Similarly, W11 shared how they provided feedback on the tools they used, demonstrating content workers' agency in implementing our recommendations.

4.2.3 Considerations for Recommendations In the next section, we synthesize participants' insights into considerations for each pillar of the framework, along with our revised recommendations. We include examples from workshop participants to illustrate each consideration.

Recruitment: We found that careful consideration should be placed to ensure that recruitment involves more than a single training instance. W3 stated that this recommendation *"should be more than recruitment but ongoing training with new information."* In fact, W3's point highlights how new training should be provided continuously so content workers' skills for well-being and their work may be kept up to date with the evolving demands of content work. We emphasize in our recommendation (R2) that content workers are educated with not only basic training at the initial recruitment stage but provided*ongoing* training throughout their work.

Tooling: Participants predicted that implementing tooling guidelines may distribute the responsibility of ensuring the effectiveness of such tools to engineering teams (W13). As such, engineering teams need the proper support (e.g., allocating a portion of their time to maintaining tools) to integrate the feedback they receive from content workers about performance issues and feature requests. Per W2:

> *"[doing so] could allow for quick resolution of technical issues with content moderation tools, as well as minimize the psychological burden of simultaneously dealing with gruesome content and managing the stress of hitting [their] production hours–which may arise due to issues with tooling."*

As per our participants' advice, we advocate for engineering teams to be viewed as crucial for content work and efforts for AI testing. We urge decision-makers to allocate sufficient resources to support these teams and, by extension, the content workers who depend on them.

Additionally, workshop participants emphasized that AI tools assist content workers within their existing workflows. W1 noted that tools allowing free-form user input for unique content cases to report *"could limit the work [they] do, pushing it onto their co-workers."* As such, we advise that tool improvements involve direct collaboration with content workers, as they are the primary end-users.

Adaptive wellness: Concerns appeared about how flexibility in well-being support may unfairly distribute team responsibility. W2 worried that it *"leaves it up to [the content workers] to create the space rather than it being provided as part of the job."* W13 added that *"[we] need guidelines on what is acceptable among teams so as to minimize opportunities for bias among employees"* while W3 pointed out the need to specify management responsibilities. We advise that applications of our recommendations for adaptive wellness are implemented with clarifications on where the responsibilities lie for each team member. For instance, it should be clear if content workers must provide materials for a craft table, as in the R5 example.

Additionally, W3 expressed concern about the cost of more flexible well-being as a limitation. To address this, we recommend that such support be part of workplace provisions and not replace external benefits. Participants also worried about activities becoming "mundane" and desired alone time (W2). Thus, fostering strong interpersonal work connections through optional team events may be beneficial when offered in variation.

Retention: Several workshop participants expressed concerns about lacking time and support for skill development within the workday (W13, W14). For instance, W13 worried: *" This suggestion would take away time from my assigned production hours, in which case I would be less likely to attend [career development events] even if I felt like it was beneficial to my personal growth."* Thus, we revised our recommendations to suggest that content workers have allocated time to focus on career growth on at least a monthly basis. Ultimately, our participants' immediate focus on applying our recommendations surfaces a sense of urgency they feel as professionals in the RAI content work space to receive such support.

5 Discussion

In this paper, we explored content work involved in the development of RAI technologies. We identified recent challenges faced by content workers to inform the future design and support for content work. We confirmed that the exposure to harmful content that is demanded in these professions has a profound and long-lasting psychological impact on workers, as described by prior studies of content moderation on social media platforms [114, 120, 124]. In addition, our work revealed nuanced challenges about the profession, including misconceptions about the work, lack of adequate workplace tools and support, and barriers to career growth. Our work encompassed the examination of various roles that may be exposed to harmful digital content, including the nascent RAI red teaming role, where we discovered nuanced differences. For example, the generative and adversarial aspect of RAI red teaming added another layer to the moral injury and introduced the need for prompting tools. We posit that the nascency of red teaming as a role (i.e., it has not been firmly established as a full-time profession at the time of our study) could attribute to the fewer number of content hours or the under-utilization of support services in comparison to other content roles. At the same time, we discovered heightened nervousness around rising demands for red teaming. As the workload of all content workers increases amid growing demand within the tech sector and particularly for that of generative AI capabilities [10, 65, 113, 135], we urge the community to heed the warnings from prior research on content moderation [114, 124, 142] in light of new and old forms for content work. We must develop a comprehensive strategy for establishing resilient human infrastructure that safeguards and supports these content workers, and our recommendation framework is just an initial step.

5.1 Practical considerations

Prior research has focused on the psychological impact of content work, recommending strategies to prevent, reduce, and treat exposure to harmful content [30, 62, 73, 142, 143, 150]. Our findings corroborated these strategies, leading us to incorporate *Adaptive wellness* and *Tooling* to address individualized psychological impacts. In these pillars, our findings revealed that organizations should indeed provide benefits such as psychological support services, but this organizational support should go beyond traditional Employee Assistance Programs (EAPs). We found that simply having resources available is inadequate; organizations should encourage utilization, provide personalized support, and maintain feedback loops. Additionally, we found that organizational and professional support are crucial for recruitment and maintaining healthy career trajectories. For these reasons, our recommendations included *Recruitment* and *Retention* as two of the four key pillars, emphasized by our workshop participants.

Informed by the lived experiences of content workers, our AURA framework encompasses these four pillars, designed to holistically support content workers' psychological and professional well-being. Therefore, we recommend that future work on content worker support or future ideation of recommendations cover all four pillars covered in the AURA framework. Here, we discuss practical considerations based on our findings for implementing our recommendations.

5.1.1 Importance of RAI content worker involvement Our validation workshops helped discover practical considerations for implementing our recommendations. We identified the need for learning and networking opportunities and the importance of integrating these into workers' responsibilities through organizational leadership support. The workshops also fostered discussions of "who" should participate in implementing the recommendations and "how", with workers identifying their own roles in promoting awareness and transparency of their work, such as writing job descriptions. This exercise helped content workers reflect on both the organization's and their roles in improving well-being of their prospective hires, their team, and their organization. We posit that there may be an opportunity for worker empowerment through participation where employees can exercise direct and indirect control over their work environments [88].

Based on this experience, we recommend that researchers and organizational leaders include worker perspectives in the iterative design and discussion of organizational processes and policies, as demonstrated by prior work using codesigning methods for integrating RAI practices [87]. A recent study reported that organizational benefits and practices should be viewed as investments in employee well-being, not just business costs [131] since work-related stress has a profound impact on the business [60]. This perspective is crucial for content workers motivated by heroism mentality but lacking adequate support and recognition. In this regard, organizations must consider employee involvement and participation in evaluating human resource management practices, as well as in the organizational design and decision-making processes, to promote greater acceptance of change [148] and work satisfaction [55]. Although our study focused on workers' perspectives, implementing these recommendations is a design exercise [71] that involves multiple stakeholders (e.g., workers, managers, organizational leaders, human resources). As our findings suggest, individual differences in content work and well-being needs introduce complexities of interpreting and supporting well-being [74], that top-down policies cannot address. Careful stakeholder involvement can lead to ideal implementations by "deliberately eliciting potential tensions that occur when stakeholders' values conflict [110]."

5.1.2 Importance of socio-ecological perspective on RAI content work Our study also revealed the importance of considering workers' environment in understanding and supporting them. We found that content workers' physical setup, such as having access to musical instruments or outdoor spaces, and social environments, like team support and managers who advocated for them, directly influence their ability to cope. Beyond work, we found that content workers carefully crafted boundaries around their loved ones and society to manage work's impact, considering both themselves and others, based on prior interactions involving the content and their work experiences. We also saw content workers struggle with heightened attention on mitigation failures and less celebration of successes, which impacted their perception toward work and well-being.

Throughout our study, we found that content workers' well-being was never just about the exposure to potentially harmful content; external factors also play a role. In fact, many of the decisions made outside of work have a trickle-down effect on the workers themselves [141]. In relation, prior research shows that workplace well-being involves interaction between individual characteristics and the surrounding environment [8, 38, 118, 144]. Therefore, future research should adopt a socio-ecological approach to analyze worker well-being and the effects of program implementations.

5.1.3 Importance of holistic tooling for RAI content work The four pillars in the AURA framework provide a structure for ideating and categorizing new and existing technology innovation opportunities, including automated tools, to support content workers holistically. Unlike prior research focusing on algorithmic decision support or content filtering [86, 109, 125], our participants suggested technologies extending beyond daily tasks to include orientation, work management, and analytics. Examples include using generative AI technologies to simulate content exposure to increase understanding of what the job entails for new or potential hires, dynamic adjustment of breaks using affective and ubiquitous computing technologies, and tracking the positive impact of their work via a digital dashboard to have a sense of closure in their work. In designing such tools, however, prior research has urged for incorporating contextual factors surrounding the work to avoid being a nuisance [157]. Therefore, we urge future technology innovation research to be inclusive of all aspects surrounding the content work, not just the content exposure or the direct impact of tools.

5.2 Labor considerations

Our study examined the human labor involved in the review and refinement of potentially harmful content generated by AI and the digital ecosystem. With a renewed interest necessitated by the proliferation of generative AI technologies, we also examined an emerging form of labor we called "RAI red teaming", which rapidly evolved and materialized right before our eyes as we were preparing this paper [44, 105]. Our work highlighted workers' perspectives on defining what this work is, who should be doing this work, and how to support that work. However, many questions still remain around the blurred roles within content work that includes RAI red teaming, all embedded within the persistent invisibility of such necessary content work. Here, we discuss our findings that may guide future research in understanding and supporting RAI content work.

5.2.1 Defining the work In our study, we used three categories of content work – data labeling, content moderation, and red teaming – to define our research scope. We found that while these categories fit within RAI practices, individual workers may participate in a multitude of activities, making strict categorization less meaningful. Such diversity in activities, coupled with individual characteristics, was observed alongside diverse work experience, work structures, resources availability and utilization, and personalized coping strategies. We saw correlations between well-being outcomes and the types and amounts of exposure. There was an overlap between traditional content moderation or data labeling and RAI red teaming, potentially due to the expanding scope of work required by generative AI deployments, as some of our participants eluded to.

So, what is becoming of RAI content work, and where is it going? Such blurring of content work boundaries is important to monitor, especially when it involves workers who are not prepare them for potentially harmful content exposure. We hope these boundaries will stabilize over time to protect worker well-being and urge organizations to conduct longitudinal studies to observe how RAI-related content work, particularly RAI red teaming, takes shape within and around RAI practices. It is important to monitor the placement of these activities within the overall AI lifecycle (e.g., design, development, or deployment phases) and their long-term effects. Since our study uncovered subjective experiences of moral dissonance, future research should systematically monitor moral injury and negative self-appraisal among those actively engaging in generative adversarial activities compared to those only viewing AI-generated content.

5.2.2 Invisibility of work Our findings confirm the prior observation of the invisibility of content work [13, 52, 121, 137, 142] that reported de-humanizing the work by reducing workers' role into "a human cleansing device" [13, 124]. Unfortunately, our participants reported that workers are made invisible through others underestimating their importance (i.e., not noticing red teaming until a

model has harmful output) and overlooking what content work actually involves (i.e., thinking content moderation for high-risk content is simply watching porn all day). While some suggest that the misperception of content work may be attributable to its relatively unskilled and rote nature [143], our findings contrarily highlight that the content work requires highly specialized skills of resilience, analytical thinking, and domain expertise. Content workers' "pride" [128] arises from their noble mission of safeguarding others and their exceptional abilities to manage harmful content while balancing their own well-being, and it should not be attempted without the proper preparation or qualifications. The lack of public awareness, stigmatization of the work, and measures to increase productivity deeply impact the psychological well-being of content workers [143]. This concern extends to red teamers, who face similar challenges due to the nascence of their activity and the lack of visibility into what their work entails [15, 66, 105]. Thurs, future research should examine *different ways in which we make content work further invisible.*

In the context of generative AI technologies, we cannot ignore the continuous hype of "automating away" content work [52, 80]. Our participants welcomed automated tools that can protect them from harmful content and integrate seamlessly into their workflows in a way that preserves control over their tools and their decision-making [51]. However, while automation efforts are often motivated by the desire to protect workers from harmful content, it is crucial to examine whether these efforts legitimize the value of content work or undermine human skills. Some assume that using people is a temporary stop-gap solution until automation can take over entirely: at first, domain experts might label and generate harmful content [106], then crowd workers [47, 112, 146, 156] and offshore vendors might take over [113], with the eventual goal of full automation. However, many scholars argue that human involvement will remain essential even with advancements in automation [51, 52, 142]. Emerging tools for red teaming, such as those generating datasets of red teaming prompts [115], fine-tuning prompts through various approaches (e.g., search-based) [41], or assisting RAI practices [11, 116], must be promoted to transform and elevate content work rather than replace and obscure human contributions [124].

5.2.3 Supporting the workforce Our current study focused on content workers employed by organizations whose responsibilities included RAI-related content activities. The challenge of fostering and growing these highly skilled professionals still remains, as the content profession is sometimes considered a "dead-end career" [100]. For example, none of our workshop participants, including subject matter experts, could imagine career progression beyond managing other content workers. In anticipation of increased demand for RAI content work, *how should we think about the content work profession and their career growth?* Some suggest that content work should be time-bound [101], serving as a launching point for *"better job opportunities"* (P16). Many participants reported that their specialized skills are difficult to translate to other professions, suggesting the need to expand their skill sets to those transferrable to other fields. However, viewing the content profession as temporary may reinforce its perception as unskilled and rote [143]. Therefore, future research must focus on developing a variety of career pathways within and beyond content work.

Formally employed content workers are more likely to receive EAP benefits, as all of our survey participants reported access to professional support services. However, the workforce ensuring responsible AI deployment extends beyond formal employees to expert volunteers [104] and crowdsource and gig workers [47, 61], who may lack access to support services. Public events (e.g., [15, 22, 66]) may offer mental health services, but barriers like stigma and lack of awareness can impede meaningful utilization to available resources [35, 37, 96]. These examples show that efforts to expand participation in RAI content work have begun, and it is urgent to understand and provide the support this new population of content workers needs.

The push to conduct RAI activities and "crowdsource a diverse set of failure modes" [111] raises the questions: *who should be doing this work, and are we providing adequate support?* Considering that our findings highlight well-being challenges despite available tools and resources, we must carefully design recruitment and inclusion strategies. This includes considering the sustained engagement of and support for workers without professional support services or short gigs without adequate training and preparation. Organizers of content work events involving emerging activities, where labor implications have yet to be fully understood, should not ignore the impacts of RAI content work and the need for upfront and ongoing support for the workforce.

5.3 Limitations

Our study has several limitations that need acknowledgment. We primarily focused on analyzing employed content workers, so readers should be cautious about applying our findings to volunteer and crowdsourced workers, as issues like retention might not be relevant to gig workers with one-off tasks, and well-being benefits may not apply to non-full-time employees. As our participants alluded to, there is a general perception of content work being outsourceable, and we have already seen it being crowdsourced [47, 112, 146, 156]. Future research should aim to provide holistic support for the entire content workforce. Additionally, our study relied on the experiences of 96 participants, a small subset compared to an estimated 100,000 commercial content moderators [143]. Our sample may be biased toward those who have persevered in content work, so further research should strive to amplify the voices of those who have left the field. Our recommendations are validated via workshops but were not explicitly implemented in an organization to evaluate its efficacy. Lastly, our study is limited due to the field of RAI red teaming rapidly evolving, even within months of our research and writing. The definition put out by The Frontier Model Forum [105] was not available when we first launched our study in June of 2023. We hypothesize many changes from the submission of this paper to the eventual publication that may impact the interpretation of our results. Regardless of these changes, we recommend integrating research on improving content workers' well-being into AI deployment safety efforts from the outset.

6 Conclusion

As national leadership and industries increasingly call for safer AI systems, it becomes crucial to consider how the humans behind such efforts are supported. Through this study, we establish that RAI content work requires expertise that cannot be developed by anyone or instantly. Individualized support is necessary for those engaging in RAI content work, and this support should cover aspects of *adaptive wellness, tooling, recruitment, and retention*. We consolidate our recommendations for holistic support into a framework for amplifying understanding, resilience, and awareness for RAI content workers (AURA). Our recommendations are validated and enhanced with examples of potential applications from a series of workshops. Ultimately, our approach surfaces a critical need to address existing challenges with content work amid increasing demands for it, particularly given the increased interest in AI-related advancements. We urge the reader to consider that offering such support alone is inadequate; support should be actively promoted among every level of organizational structure, from content work teams to leadership. In turn, we call for these support structures to be in place as part of RAI deployment safety efforts.

References

[1] Hussein Abbass, Axel Bender, Svetoslav Gaidow, and Paul Whitbread. 2011. Computational red teaming: Past, present and future. *IEEE Computational Intelligence Magazine* 6, 1 (2011), 30–42.

[2] Jackie Andrade, Jon May, Catherine Deeprose, Sarah-Jane Baugh, and Giorgio Ganis. 2014. Assessing vividness of mental imagery: the Plymouth Sensory Imagery Questionnaire. *British Journal of Psychology* 105, 4 (2014), 547–563.

[3] Jean Maria Arrigo. 2004. A utilitarian argument against torture interrogation of terrorists. *Science and Engineering Ethics* 10, 3 (2004), 543–572.

[4] Andrew Arsht and Daniel Etcovitch. 2018. The human cost of online content moderation. *Harvard Journal of Law and Technology* 2 (2018).

[5] American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*. Vol. 5. American psychiatric association Washington, DC.

[6] Avasant. 2024. https://avasant.com/report/impact-of-generative-ai-on-content-moderation/

[7] Azure. 2024. PyRIT: Python Risk Intelligence Toolkit. https://github.com/Azure/PyRIT Accessed: 2024-07-15.

[8] Gianluca Biggio and ClaudioG Cortese. 2013. Well-being in the workplace through interaction between individual characteristics and organizational context. *International journal of qualitative studies on health and well-being* 8, 1 (2013), 19823.

[9] 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 (2017), 305–318.

[10] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, J. B. Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. 2020. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. arXiv:2004.07213 [cs] http://arxiv.org/abs/2004.07213

[11] Zana Buccinca, Chau Minh Pham, Maurice Jakesch, Marco Tulio Ribeiro, Alexandra Olteanu, and Saleema Amershi. 2023. AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms. *ArXiv* abs/2306.03280 (2023). https://api.semanticscholar.org/CorpusID:259088554

[12] Katie Canales. 2021. Mark Zuckerberg said content moderation requires "nuances" that consider the intent behind a post, but also highlighted Facebook's reliance on ai to do that job. https://www.businessinsider.com/zuckerberg-nuances-content-moderation-ai-misinformation-hearing-2021-3?r=US&IR=T

[13] Elinor Carmi. 2019. The hidden listeners: regulating the line from telephone operators to content moderators. *International Journal of Communication* 13 (2019), 440–458.

[14] Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. 2023. Explore, Establish, Exploit: Red Teaming Language Models from Scratch. *ArXiv* abs/2306.09442 (2023). https://api.semanticscholar.org/CorpusID:259187620

[15] Sven Cattell. 2023. Generative Red Team Recap. https://aivillage.org/defcon%2031/generative-recap/

[16] Stevie Chancellor, Yannis Kalantidis, Jessica A Pater, Munmun De Choudhury, and David A Shamma. 2017. Multimodal classification of moderated online pro-eating disorder content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3213–3226.

[17] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 1201–1213.

[18] Ramya Chari, Chia chia Chang, Steven L. Sauter, Elizabeth L. Petrun Sayers, Jennifer L. Cerully, Paul A. Schulte, Anita L. Schill, and Lori Uscher-Pines. 2018. Expanding the Paradigm of Occupational Safety and Health: A New Framework for Worker Well-Being. *Journal of Occupational and Environmental Medicine* 60 (2018), 589–593. https://api.semanticscholar.org/CorpusID:4561492

[19] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. sage.

[20] EunJeong Cheon. 2023. Powerful Futures: How a Big Tech Company Envisions Humans and Technologies in the Workplace of the Future. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–35.

[21] Jeasik Cho and Allen Trent. 2006. Validity in qualitative research revisited. *Qualitative research* 6, 3 (2006), 319–340.

[22] Rumman Chowdhury. 2023. Redefining Red Teaming. https://www.hackthefuture.com/news/redefining-red-teaming

[23] Thomas W Colligan and Eileen M Higgins. 2006. Workplace stress: Etiology and consequences. *Journal of workplace behavioral health* (2006).

[24] Christine L Cook, Jie Cai, and Donghee Yvette Wohn. 2022. Awe Versus Aww: The Effectiveness of Two Kinds of Positive Emotional Stimulation on Stress Reduction for Online Content Moderators. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–19.

[25] Sergi G Costafreda, Michael J Brammer, Anthony S David, and Cynthia HY Fu. 2008. Predictors of amygdala activation during the processing of emotional stimuli: a meta-analysis of 385 PET and fMRI studies. *Brain research reviews* 58, 1 (2008), 57–70.
[26] 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.
[27] John W Creswell. 2012. *Educational research: Planning, conducting, and evaluating quantitative and qualitative research.* Pearson Education, Inc.
[28] John W Creswell and Dana L Miller. 2000. Determining validity in qualitative inquiry. *Theory into practice* 39, 3 (2000), 124–130.
[29] Karen Cusack, Daniel E Jonas, Catherine A Forneris, Candi Wines, Jeffrey Sonis, Jennifer Cook Middleton, Cynthia Feltner, Kimberly A Brownley, Kristine Rae Olmsted, Amy Greenblatt, et al. 2016. Psychological treatments for adults with posttraumatic stress disorder: A systematic review and meta-analysis. *Clinical psychology review* 43 (2016), 128–141.
[30] 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 AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 33–42.
[31] Wendy Dean, Simon Talbot, and Austin Dean. 2019. Reframing clinician distress: moral injury not burnout. *Federal Practitioner* 36, 9 (2019), 400.
[32] 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.
[33] 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.
[34] Evelyn Douek. 2021. More content moderation is not always better. https://www.wired.com/story/more-content-moderation-not-always-better/
[35] Benjamin G Druss, Thomas Bornemann, Yvonne W Fry-Johnson, Harriet G McCombs, Robert M Politzer, and George Rust. 2008. Trends in Mental Health and Substance Abuse Services at the Nation's Community Health Centers: 1998–2003. *American Journal of Public Health* 98, Supplement_1 (2008), S126–S131. doi:10.2105/ajph.2005.076943
[36] Elizabeth Dwoskin. 2019. Inside facebook, the second-class workers who do the hardest job are waging a quiet battle. https://www.washingtonpost.com/technology/2019/05/08/inside-facebook-second-class-workers-who-do-hardest-job-are-waging-quiet-battle/
[37] Daniel Eisenberg, Marilyn F Downs, Ezra Golberstein, and Kara Zivin. 2009. Stigma and help seeking for mental health among college students. *Medical Care Research and Review* 66, 5 (2009), 522–541.
[38] Susan L Ettner and Joseph G. Grzywacz. 2001. Workers' perceptions of how jobs affect health: a social ecological perspective. *Journal of occupational health psychology* 6 2 (2001), 101–13. https://api.semanticscholar.org/CorpusID:5102747
[39] Daniel Fabian. 2023. Google's AI Red Team: The ethical hackers making ai safer. https://blog.google/technology/safety-security/googles-ai-red-team-the-ethical-hackers-making-ai-safer/
[40] Fact.MR. 2024. Content Moderation Solution Market. https://www.factmr.com/report/4522/content-moderation-solutions-market
[41] Michael Feffer, Anusha Sinha, Zachary C Lipton, and Hoda Heidari. 2024. Red-Teaming for Generative AI: Silver Bullet or Security Theater? *arXiv preprint arXiv:2401.15897* (2024).
[42] 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.
[43] Sorelle Friedler, Ranjit Singh, Borhane Blili-Hamelin, Jacob Metcalf, and Brian Chen. 2023. Ai Red-teaming is not a one-stop solution to ai harms. https://datasociety.net/wp-content/uploads/2023/10/Recommendations-for-Using-Red-Teaming-for-AI-Accountability-PolicyBrief.pdf
[44] Sorelle Friedler, Ranjit Singh, Borhane Blili-Hamelin, Jacob Metcalf, and Brian J Chen. 2023. AI Red-Teaming Is Not a One-Stop Solution to AI Harms.
[45] Paolo Fusar-Poli, Anna Placentino, Francesco Carletti, Paola Landi, Paul Allen, Simon Surguladze, Francesco Benedetti, Marta Abbamonte, Roberto Gasparotti, Francesco Barale, et al. 2009. Functional atlas of emotional faces processing: a voxel-based meta-analysis of 105 functional magnetic resonance imaging studies. *Journal of psychiatry and neuroscience* 34, 6 (2009), 418–432.
[46] Hack The Future. 2023. https://www.airedteam.org/news/ai-village-at-def-con-announces-largest-ever-public-generative-ai-red-team
[47] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma,

Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. arXiv:2209.07858 [cs] http://arxiv.org/abs/2209.07858

[48] R Stuart Geiger. 2016. Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society* 19, 6 (2016), 787–803.

[49] Anna D Gibson. 2023. What Teams Do: Exploring Volunteer Content Moderation Team Labor on Facebook. *Social Media+ Society* 9, 3 (2023), 20563051231186109.

[50] Kirsten Gollatz, Felix Beer, and Christian Katzenbach. 2018. The turn to artificial intelligence in governing communication online. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-59528-6.

[51] 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 (2020). https://api.semanticscholar.org/CorpusID:213703822

[52] Mary L Gray and Siddharth Suri. 2019. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.

[53] Neil Greenberg, Mary Docherty, Sam Gnanapragasam, and Simon Wessely. 2020. Managing mental health challenges faced by healthcare workers during covid-19 pandemic. *bmj* 368 (2020).

[54] Joseph G Grzywacz, David M Almeida, and Daniel A McDonald. 2002. Work–family spillover and daily reports of work and family stress in the adult labor force. *Family relations* 51, 1 (2002), 28–36.

[55] David E. Guest. 2002. Human Resource Management, Corporate Performance and Employee Wellbeing: Building the Worker into HRM. *Journal of Industrial Relations* 44 (2002), 335 – 358. https://api.semanticscholar.org/CorpusID:154421694

[56] Wendy Haight, Erin Sugrue, Molly Calhoun, and James Black. 2016. A scoping study of moral injury: Identifying directions for social work research. *Children and youth services review* 70 (2016), 190–200.

[57] 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:2009.10311* (2020).

[58] Hugo Lewi Hammer. 2016. Automatic detection of hateful comments in online discussion. In *Proceedings of the International Conference on Industrial Networks and Intelligent Systems*. Springer, 164–173.

[59] Catherine Han, Joseph Seering, Deepak Kumar, Jeffrey T Hancock, and Zakir Durumeric. 2023. Hate raids on Twitch: Echoes of the past, new modalities, and implications for platform governance. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–28.

[60] Juliet Hassard, Kevin R. H. Teoh, Gintare Visockaite, Philip J. Dewe, and Tom Cox. 2018. The Cost of Work-Related Stress to Society: A Systematic Review. *Journal of Occupational Health Psychology* 23 (2018), 1–17. https://api.semanticscholar.org/CorpusID:26097668

[61] Danula Hettiachchi and Jorge Goncalves. 2019. Towards effective crowd-powered online content moderation. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*. 342–346.

[62] E Alison Holman, Dana Rose Garfin, and Roxane Cohen Silver. 2014. Media's role in broadcasting acute stress following the Boston Marathon bombings. 111, 1 (2014), 93–98. Publisher: National Acad Sciences.

[63] Emily A. Holmes, Ella L. James, Thomas Coode-Bate, and Catherine Deeprose. 2009. Can Playing the Computer Game "Tetris" Reduce the Build-Up of Flashbacks for Trauma? A Proposal from Cognitive Science. 4, 1 (2009), e4153. doi:10.1371/journal.pone.0004153

[64] Emily A. Holmes, Ella L. James, Emma J. Kilford, and Catherine Deeprose. 2010. Key Steps in Developing a Cognitive Vaccine against Traumatic Flashbacks: Visuospatial Tetris versus Verbal Pub Quiz. 5, 11 (2010), e13706. doi:10.1371/journal.pone.0013706

[65] The White House. 2023. FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI. https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

[66] Human Intelligence, SeedAI, and DEFCON AI Village. 2023. *Generative AI Red Teaming Challenge: Transparency Report*. Technical Report. Humane Intelligence. 1–41 pages. Retrieved July 13, 2024, from https://drive.google.com/file/d/1JqpbIP6DNomkb32umLoiEPombK2-0Rc-/view.

[67] 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.

[68] Shagun Jhaver and Amy Zhang. 2023. Do Users Want Platform Moderation or Individual Control? Examining the Role of Third-Person Effects and Free Speech Support in Shaping Moderation Preferences. *arXiv preprint arXiv:2301.02208*

(2023).
[69] Shagun Jhaver, Alice Qian Zhang, Quanze Chen, Nikhila Natarajan, Ruotong Wang, and Amy Zhang. 2023. Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor. *arXiv preprint arXiv:2305.10374* (2023).
[70] Tim Jordan and Paul Taylor. 2017. A sociology of hackers. In *Cyberspace Crime*. Routledge, 163–186.
[71] Sabine Junginger. 2013. Design and Innovation in the Public Sector: Matters of Design in Policy-Making and Policy Implementation. https://api.semanticscholar.org/CorpusID:154113223
[72] David Jurgens, Eshwar Chandrasekharan, and Libby Hemphill. 2019. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. *arXiv:1906.01738* (2019).
[73] 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.
[74] Anna Kawakami, Shreyans Chowdhary, Shamsi T. Iqbal, Qingzi Vera Liao, Alexandra Olteanu, Jin hwan Suh, and Koustuv Saha. 2023. Sensing Wellbeing in the Workplace, Why and For Whom? Envisioning Impacts with Organizational Stakeholders. *Proceedings of the ACM on Human-Computer Interaction* 7 (2023), 1 – 33. https://api.semanticscholar.org/CorpusID:257496602
[75] Emre Kazim and Adriano Koshiyama. 2020. AI assurance processes. *Available at SSRN 3685087* (2020).
[76] Daphne Keller. 2021. The future of platform power: making middleware work. *Journal of Democracy* 32, 3 (2021), 168–172.
[77] Dee K Knight, Christy Crutsinger, and HaeJung Kim. 2006. The impact of retail work experience, career expectation, and job satisfaction on retail career intention. *Clothing and Textiles Research Journal* 24, 1 (2006), 1–14.
[78] Elly Konijn. 2000. *Acting emotions*. Amsterdam University Press.
[79] Ram Shankar Siva Kumar. 2023. Microsoft AI Red Team Building Future of Safer Ai. https://www.microsoft.com/en-us/security/blog/2023/08/07/microsoft-ai-red-team-building-future-of-safer-ai/
[80] Jan Leike and Ilya Sutskever. 2023. https://openai.com/blog/introducing-superalignment
[81] Hanlin Li, Brent Hecht, and Stevie Chancellor. 2022. All that's happening behind the scenes: Putting the spotlight on volunteer moderator labor in Reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 584–595.
[82] Jeffrey Y Lin, Scott O Murray, and Geoffrey M Boynton. 2009. Capture of attention to threatening stimuli without perceptual awareness. *Current Biology* 19, 13 (2009), 1118–1122.
[83] Yvonna S Lincoln and Egon G Guba. 1985. *Naturalistic inquiry*. sage.
[84] Brett T Litz, Nathan Stein, Eileen Delaney, Leslie Lebowitz, William P Nash, Caroline Silva, and Shira Maguen. 2009. Moral injury and moral repair in war veterans: A preliminary model and intervention strategy. *Clinical psychology review* 29, 8 (2009), 695–706.
[85] David F Longbine. 2008. Red teaming: past and present. *School of Advanced Military Studies, Army Command and General Staff College* (2008).
[86] 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).
[87] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020). https://api.semanticscholar.org/CorpusID:210156214
[88] Mick Marchington, J. F. B. Goodman, Adrian Wilkinson, and Peter Ackers. 1991. NEW DEVELOPMENTS IN EMPLOYEE INVOLVEMENT. *Management Research News* 14 (1991), 34–37. https://api.semanticscholar.org/CorpusID:154439260
[89] Dale A Masi. 2020. The history of employee assistance programs in the United States.
[90] J Nathan Matias. 2019. The civic labor of volunteer moderators online. *Social Media+ Society* 5, 2 (2019), 2056305119836778.
[91] Robert G Maunder, William J Lancee, Kenneth E Balderson, Jocelyn P Bennett, Bjug Borgundvaag, Susan Evans, Christopher MB Fernandes, David S Goldbloom, Mona Gupta, Jonathan J Hunter, et al. 2006. Long-term psychological and occupational effects of providing hospital healthcare during SARS outbreak. *Emerging infectious diseases* 12, 12 (2006), 1924.
[92] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249* (2024).
[93] Sharan B Merriam et al. 2002. Introduction to qualitative research. *Qualitative research in practice: Examples for discussion and analysis* 1, 1 (2002), 1–17.

[94] Michel. 2018. Ex-Content Moderator Sues Facebook, Saying Violent Images Caused Her PTSD - e-traces. https://api.semanticscholar.org/CorpusID:150309904

[95] Microsoft. 2023. Introduction to red teaming large language models (llms). https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/red-teaming

[96] Ramin Mojtabai, Mark Olfson, Nancy a Sampson, Benjamin Druss, Philip S Wang, Kenneth B Wells, Harold a Pincus, and Ronald C Kessler. 2011. Barriers to mental health treatment: results from the WHO World Mental Health surveys. *Psychological Medicine* 41, 8 (2011), 1751–1761. doi:10.1017/S0033291710002291.Barriers

[97] Garrett Morrow, Briony Swire-Thompson, Jessica Montgomery Polny, Matthew Kopec, and John P Wihbey. 2022. The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology* 73, 10 (2022), 1365–1386.

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

[99] Casey Newton. 2019. The trauma floor. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona Publication Title: The Verge.

[100] Casey Newton. 2020. What tech companies should do about their content moderators' PTSD. https://www.theverge.com/interface/2020/1/28/21082642/content-moderator-ptsd-facebook-youtube-accenture-solutions

[101] Casey Newton. 2020. YouTube moderators are being forced to sign a statement acknowledging the job can give them PTSD. https://www.theverge.com/2020/1/24/21075830/youtube-moderators-ptsd-accenture-statement-lawsuits-mental-health

[102] Chuyen Nguyen, Caleb Morgan, and Sudip Mittal. 2022. CTI4AI: Threat Intelligence Generation and Sharing after Red Teaming AI Models. *arXiv preprint arXiv:2208.07476* (2022).

[103] The American Institute of Stress. 2021. Workplace Stress. https://www.stress.org/workplace-stress

[104] OpenAI. 2023. https://openai.com/blog/red-teaming-network

[105] OpenAI. 2023. Frontier Model Forum: What is Red Teaming? https://www.frontiermodelforum.org/uploads/2023/10/FMF-AI-Red-Teaming.pdf

[106] OpenAI. 2023. GPT-4 System Card. https://cdn.openai.com/papers/gpt-4-system-card.pdf

[107] Will Oremus. 2023. Meet the hackers who are trying to make AI go rogue. The Washington Post. https://www.washingtonpost.com/technology/2023/08/08/ai-red-team-defcon/

[108] PAI. 2021. Responsible sourcing of data enrichment services. https://partnershiponai.org/wp-content/uploads/2021/08/PAI-Responsible-Sourcing-of-Data-Enrichment-Services.pdf

[109] Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting comment moderators in identifying high quality online news comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 1114–1125.

[110] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. 2022. Designing Fair AI in Human Resource Management: Understanding Tensions Surrounding Algorithmic Evaluation and Envisioning Stakeholder-Centered Solutions. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022). https://api.semanticscholar.org/CorpusID:248419785

[111] Alicia Parrish, Hannah Rose Kirk, Jessica Quaye, Charvi Rastogi, Max Bartolo, Oana Inel, Juan Ciro, Rafael Mosquera, Addison Howard, William J. Cukierski, D. Sculley, Vijay Janapa Reddi, and Lora Aroyo. 2023. Adversarial Nibbler: A Data-Centric Challenge for Improving the Safety of Text-to-Image Models. *ArXiv* abs/2305.14384 (2023). https://api.semanticscholar.org/CorpusID:258865543

[112] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red Teaming Language Models with Language Models. arXiv:2202.03286 [cs] http://arxiv.org/abs/2202.03286

[113] Billy Perrigo. 2023. OpenAI used Kenyan workers on less than $2 per hour: Exclusive. https://time.com/6247678/openai-chatgpt-kenya-workers/

[114] Amit Pinchevski. 2023. Social media's canaries: content moderators between digital labor and mediated trauma. *Media, Culture & Society* 45, 1 (2023), 212–221.

[115] Bhaktipriya Radharapu, Kevin Robinson, Lora Aroyo, and Preethi Lahoti. 2023. Aart: Ai-assisted red-teaming with diverse data generation for new llm-powered applications. *arXiv preprint arXiv:2311.08592* (2023).

[116] Bhaktipriya Radharapu, Kevin Robinson, Lora Aroyo, and Preethi Lahoti. 2023. AART: AI-Assisted Red-Teaming with Diverse Data Generation for New LLM-powered Applications. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:265213125

[117] Harita Reddy and Eshwar Chandrasekharan. 2023. Evolution of Rules in Reddit Communities. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 278–282.

[118] Lucie Richard, Louise Potvin, Natalie A Kishchuk, Helen Prlic, and L. W. Green. 1996. Assessment of the Integration of the Ecological Approach in Health Promotion Programs. *American Journal of Health Promotion* 10 (1996), 318 –

328. https://api.semanticscholar.org/CorpusID:46837613
[119] Christian X Ries and Rainer Lienhart. 2014. A survey on visual adult image recognition. *Multimedia tools and applications* 69, 3 (2014), 661–688.
[120] Sarah T Roberts. 2016. Commercial content moderation: Digital laborers' dirty work. (2016).
[121] Sarah T Roberts. 2017. *Content moderation.*
[122] Ivan T Robertson, Cary L Cooper, and Sheena Johnson. 2011. *Well-being: Productivity and happiness at work.* Vol. 3. Springer.
[123] Sabirat Rubya and Svetlana Yarosh. 2017. Video-mediated peer support in an online community for recovery from substance use disorders. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing.* 1454–1469.
[124] Minna Ruckenstein and Linda Lisa Maria Turunen. 2020. Re-humanizing the platform: Content moderators and the logic of care. 22, 6 (2020), 1026–1042. Publisher: Sage Publications Sage UK: London, England.
[125] Marcos Rodrigues Saude, Marcelo de Medeiros Soares, Henrique Gomes Basoni, Patrick Marques Ciarelli, and Elias Oliveira. 2014. A strategy for automatic moderation of a large data set of users comments. In *Proceedings of the 2014 XL Latin American Computing Conference (CLEI'14).* IEEE, 1–7.
[126] 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.
[127] Angela M. Schöpke-Gonzalez, Shubham Atreja, Han Na Shin, Najmin Ahmed, and Libby Hemphill. 2022. Why do volunteer content moderators quit? Burnout, conflict, and harmful behaviors. (2022), 146144482211385. doi:10.1177/14614448221138529
[128] Joseph Seering, Brianna Dym, Geoff Kaufman, and Michael Bernstein. 2022. Pride and Professionalization in Volunteer Moderation: Lessons for Effective Platform-User Collaboration. *Journal of Online Trust and Safety* 1, 2 (2022).
[129] Jonathan Shay. 2014. Moral injury. *Psychoanalytic psychology* 31, 2 (2014), 182.
[130] Alyssa Sheehan and Christopher A Le Dantec. 2023. Making Meaning from the Digitalization of Blue-Collar Work. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–21.
[131] Sara J. Singer, Stacie Vilendrer, Grace Joseph, Jason Kim, and Jeffrey Pfeffer. 2020. Employers' Role in Employee Health: Why They Do What They Do. *Journal of Occupational & Environmental Medicine* (2020). https://api.semanticscholar.org/CorpusID:225410719
[132] Monika Singh, Divya Bansal, and Sanjeev Sofat. 2016. Behavioral analysis and classification of spammers distributing pornographic content in social media. *Social Network Analysis and Mining* 6, 1 (2016), 41.
[133] Ranjit Singh, Borhane Blili-Hamelin, and Jacob Metcalf. 2023. Can we red team our way to AI accountability? https://techpolicy.press/can-we-red-team-our-way-to-ai-accountability/
[134] Daisy Soderberg-Rivkin. 2023. Five myths about online content moderation, from a former content moderator. https://www.rstreet.org/commentary/five-myths-about-online-content-moderation-from-a-former-content-moderator/
[135] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III, Jesse Dodge, Ellie Evans, Sara Hooker, Yacine Jernite, Alexandra Sasha Luccioni, Alberto Lusoli, Margaret Mitchell, Jessica Newman, Marie-Therese Png, Andrew Strait, and Apostol Vassilev. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. arXiv:2306.05949 [cs] http://arxiv.org/abs/2306.05949
[136] Alfonso Sousa-Poza and Andres A Sousa-Poza. 2000. Well-being at work: a cross-national analysis of the levels and determinants of job satisfaction. *The journal of socio-economics* 29, 6 (2000), 517–538.
[137] Franchesca Spektor, Estefania Rodriguez, Samantha Shorey, and Sarah Fox. 2022. AI and essential labor: representing the invisible work of integration. *XRDS: Crossroads, The ACM Magazine for Students* 28, 2 (2022), 16–19.
[138] Ian Spence, Patrick Wong, Maria Rusan, and Naghmeh Rastegar. 2006. How color enhances visual memory for natural scenes. *Psychological science* 17, 1 (2006), 1–6.
[139] Ruth Spence, Amy Harrison, Paula Bradbury, Paul Bleakley, Elena Martellozzo, and Jeffrey DeMarco. 2023. Content moderators' strategies for coping with the stress of moderating content online. *Journal of Online Trust and Safety* 1, 5 (2023).
[140] Thomas Stackpole. 2022. Content moderation is terrible by design. https://hbr.org/2022/11/content-moderation-is-terrible-by-design#:~:text=Their%20work%20is%20largely%20invisible,or%20video%20at%20a%20time.
[141] PAI Staff. 2021. Responsible Sourcing of Data Enrichment Services. https://partnershiponai.org/paper/responsible-sourcing-considerations/
[142] Miriah Steiger, Timir J Bharucha, Wilfredo Torralba, Marlyn Savio, Priyanka Manchanda, and Rachel Lutz-Guevara. 2022. Effects of a Novel Resiliency Training Program for Social Media Content Moderators. In *Proceedings of Seventh International Congress on Information and Communication Technology: ICICT 2022, London, Volume 4.* Springer, 283–298.
[143] 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 *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama Japan, 2021-05-06).

ACM, 1–14. doi:10.1145/3411764.3445092
[144] Daniel Stokols. 1992. Establishing and maintaining healthy environments. Toward a social ecology of health promotion. *The American psychologist* 47 1 (1992), 6–22. https://api.semanticscholar.org/CorpusID:12152556
[145] Nicolas P Suzor, Sarah Myers West, Andrew Quodling, and Jillian York. 2019. What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication* 13 (2019), 18.
[146] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* (2022).
[147] 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 Doctoral dissertation.
[148] André Ullrich, Malte Reißig, Silke Niehoff, and Grischa Beier. 2023. Employee involvement and participation in digital transformation: a combined analysis of literature and practitioners' expertise. *Journal of Organizational Change Management* (2023). https://api.semanticscholar.org/CorpusID:258996239
[149] 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*. IEEE, 54–57.
[150] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 6 (1988), 1063.
[151] Bradley V Watts, Paula P Schnurr, Lorna Mayo, Yinong Young-Xu, William B Weeks, and Matthew J Friedman. 2013. Meta-analysis of the efficacy of treatments for posttraumatic stress disorder. *The Journal of clinical psychiatry* 74, 6 (2013), 11710.
[152] 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*. 1–13.
[153] Bradley J Wood and Ruth A Duggan. 2000. Red teaming of advanced information assurance concepts. In *Proceedings DARPA Information Survivability Conference and Exposition. DISCEX'00*, Vol. 2. IEEE, 112–118.
[154] Amy Wrzesniewski, Clark McCauley, Paul Rozin, and Barry Schwartz. 1997. Jobs, careers, and callings: People's relations to their work. *Journal of research in personality* 31, 1 (1997), 21–33.
[155] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web (WWW'17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1391–1399. doi:10.1145/3038912.3052591
[156] Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2950–2968.
[157] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019). https://api.semanticscholar.org/CorpusID:127989976
[158] Robert K Yin. 2009. *Case study research: Design and methods*. Vol. 5. sage.
[159] Micah Zenko. 2015. *Red Team: How to succeed by thinking like the enemy*. Basic Books.
[160] Micah Zenko and Richard Haass. 2015. "red team: How to succeed by thinking like the enemy". https://www.cfr.org/event/red-team-how-succeed-thinking-enemy
[161] Alice Qian Zhang, Ashlee Milton, and Stevie Chancellor. 2023. # Pragmatic or# Clinical: Analyzing TikTok Mental Health Videos. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 149–153.

7 Appendix

In this appendix, we include frequently asked questions shared during recruitment (Section 7.1), survey questions (Section 7.8), interview questions (Section 7.2), graphics shown during our workshops, and snapshots of our workshop information (Section 7.15).

7.1 Frequently asked questions (FAQ) about the study

The following FAQ was shared with the potential participants during recruitment for our survey and interviews. Please note that we emphasized the anonymity of the responses to protect the identify of our participants and to minimize the risk of retaliation by the employers if identified.

What is Microsoft Research? Microsoft Research is an academic arm of Microsoft. Although Microsoft Research is funded by Microsoft, the research conducted by Microsoft Research follows the same standards that academic institutions follow to conduct their research. Microsoft Research, as a research institution, not only conforms to the privacy and security regulations governed by the company but also follows the federal ethical guidelines. Additionally, Microsoft Research is able to retain academic freedom and discourse without our research topics requiring approval by PR or marketing.

Why are you conducting this study? We are interested in understanding the perspectives and experiences of content moderators, data labelers, and red-team members who deal with various types of online content on a daily basis. The purpose of this research is to develop an understanding of the nature of content moderation, data labeling, and red-teaming work and the tools being used that directly and indirectly impact the work. We hope to identify the challenges and opportunities for improving and supporting the work of content moderators, data labelers, and red-team members.

Who can participate in this study? We are looking for participants that conduct content or feature review tasks as part of their job. As part of your job, you may be exposed to potentially disturbing or impactful content. A content moderator is someone who reviews online content (text, photo, audio, video) to determine its suitability for a platform or a product based on its policy or guidelines. A data labeler reviews, labels, and categorizes various types of content according to specific labeling or sorting guidelines, which aids in data analysis or training machine learning models. A red-team member conducts critical assessments of product or platform features by simulating the actions of potential bad actors or testing system vulnerabilities. While you do not need to be doing these tasks full-time, you need to spend at least 1 hours/week reviewing content or features and have been working as a content moderator, a data labeler, or a red-team member for at least 3 months to be eligible. You must also be at least 18 years old.

Will my employer know that I'm participating? The survey does not ask for or collect any information about your employer. Participating in this study will not affect your employment. Any information we collect about you and as part of study coordination will not be linked to your responses and will not be shared with your employer.

How will you use the data from the study? We will use the data to analyze the themes and patterns that emerge from the responses. We will also use the data to generate recommendations and suggestions for improving and supporting content moderation, data labeling, or red-teaming work. We will publish our findings in academic journals and conferences, as well as share them with relevant stakeholders and organizations.

How will you protect my privacy and confidentiality? The survey is anonymous. We will only collect your name and email address for the purpose of sending you your gift card and information about follow-up interviews, both of which are optional.

Note: we provided an FAQ question with information about the lead research institution, but omit it for purposes of maintaining anonymity.

7.2 Interview questions

The following protocol contains questions we asked during our semi-structured interviews. We structured interview questions into four main sections to examine the aspects of the nature of content work outlined in RQ1. Keeping the semi-structured nature of the interview in mind, the questions we present below include a superset of questions that could asked during the interview

based on the discussion. If screen sharing was involved, we specifically requested our participants not to share any potentially harmful or sensitive content with the researchers.

7.3 Grounding

Please choose one of the following activities as your primary activity:

- **Content Moderation:** I review various forms of online content (including text, photos, audio, and video) with the intent to flag or identify any content that potentially violates the platform's policy or guidelines. This could lead to the content's modification or removal to ensure the platform maintains a safe and respectful environment.
- **Data Labeling:** My work primarily involves reviewing, labeling, or categorizing various types of content (including text, photos, audio, and video). This is done according to specific labeling or sorting guidelines, which aids in data analysis and training machine learning models.
- **Red-Teaming:** I conduct critical assessments of product or platform features by simulating the actions of potential bad actors or testing system vulnerabilities. The objective is to identify if these features can inadvertently generate or promote content that violates policy or guidelines, thus enhancing the product's security and safety measures.

To help us understand your experience with this activity, tell us about this activity. What does it entail? Where does this work come from? What are you trying to achieve with this work? For the following questions, let's use this specific activity that you described as your "work."

7.4 Work Setup

Could you describe your work setting, including where you work, hardware or software you work with, people you work with, your work environment?

- **Location:** Where are you doing the work (desk, cubicle, etc.)?
- **Equipment:** What do you work with (hardware, software)?
- **Environment:** What is the environment like (noise, temperature, comfort level generally)?
- **People:** Who is around you (individual vs. collaborative work)?
- **Ideal Setup:** What would your ideal setup (including location, environment, people) be? What are the barriers to accessing this ideal setup?

7.5 Conducting Work

- What tools do you use to help you with this work? Who develops the tool?
- What support from other people is most helpful for you in your work?
- How helpful are automated (AI) tools for your daily work? If you could request changes to how automated tools work, what changes would you want?
- What is an ideal tool that would help with your work, what would that look like? Why? What are the barriers to accessing this ideal tool?
- How have generative AI tools impacted you doing your work (either/both as a solution and problem)?
- How have generative AI tools impacted *the demands* of your work (either/both as a solution and problem)?

7.6 Coping

Tell me about a time when your well-being was impacted by the content moderation/data labeling/red-teaming work you do.

- What strategies, resources, and support were helpful for improving your well-being? (organizational vs. ad-hoc)
- What other strategies, resources, and support would you want to have access to in an ideal scenario? (organizational vs. ad-hoc)
- What support from other people is helpful for managing these demands (i.e., do you share tips, talk with co-workers, managers, etc. about wellness)? What are the barriers to accessing these strategies, resources, and supports?
- If you were to wave a magic wand to create any kind of technology/software/UX experience that would help with managing the demands of or coping with your work, what would that look like? Why? (e.g., help coping with stress and anxiety, or maladaptive memories) What would a tool like that look like? (e.g., VR immersive interface, or audio-only based interventions, scent and other modalities, an empathic agent, etc.)
- If you were to wave a magic wand to create any kind of resource or activity that would help with managing the demands of or coping with your work, what would that look like? Why?
- What resources, tools, or activities have been most helpful in managing the demands of your work (e.g., training, breaks, etc.)? Which of these resources, tools, or activities do you use at work or outside of work? How does the modality of these resources impact how helpful they are (for remote workers, is it necessary to have in-person access and vice versa)?
- What other strategies do you use to cope with the demands of your work? (e.g., meditation, video games)
- What other support is most helpful for managing emotional and wellbeing demands of your work?
- Which of these supports (resources, strategies, support from other people) is formally provided/encouraged by your organization?
- Which of these are ad hoc?
- How does the organizational and ad hoc support you have differ from your ideal amount of support?

7.7 Collaboration

- How do you decide how to classify content or generate adversarial content?What do you do when you see content or behavior that is more in the "grey" area?
- Who do you work with to make these decisions (co-workers vs managers)? Who needs to approve your decisions? Who supports you in these decisions?
- What about this process is challenging for your work? What is most helpful?

7.8 Survey questions

Our survey was structured to follow our RQ1 and RQ2, ranging from demographics, understanding the nature of content work, challenge and opportunities associated with work, tools used for work, to how workers cope with work demands. In addition to the questions below, we included three validated psychometric scales: (1) the short form of the Plymouth Sensory Imagery Questionnaire [2], (2) the full version of the Burnout Assessment Tool [**?**], and (3) the Emotion Regulation Questionnaire [**?**]. Please refer to the source references for the questions. These questionnaires were included in the survey, but their analysis results were not included in this paper.

7.9 Demographics

Please tell us about yourself.

Q: What is your age?

- 18-25 years old
- 26-35 years old
- 36-45 years old
- 46-55 years old
- 56-65 years old
- 66+ years old
- Prefer not to say

Q: How do you describe your gender identity?

- Female/Women (some examples: cisgender women, female-identified people, transgender women)
- Gender nonbinary (some examples: gender diverse, gender fluid, gender non-conforming, gender questioning, genderqueer, nonbinary, two-spirit)
- Transgender (some examples: transgender female, transgender male, transgender non-conforming, nonbinary)
- Male/Men (some examples: cisgender men, male-identified people, transgender men)
- Unsure
- Not listed or prefer to self-describe
- Prefer not to say

Q: What is the highest level of education you have achieved?

- Less than high school
- High school diploma or equivalent
- Some college or vocational training
- Bachelor's degree
- Some postgraduate degree
- Master's degree
- Doctoral or professional degree
- Prefer not to say

Q: What is your employment status?

- Full-time employee
- Part-time employee
- Full-time contractor
- Part-time contractor
- Volunteer
- Other

Q: Which of the following activities do you perform as part of your job? Please select all that apply.

- Content Moderation: I review various forms of online content (including text, photos, audio, and video) with the intent to flag or identify any content that potentially violates the platform's policy or guidelines. This could lead to the content's modification or removal to ensure the platform maintains a safe and respectful environment.
- Data Labeling: My work primarily involves reviewing, labeling, or categorizing various types of content (including text, photos, audio, and video). This is done according to specific labeling or sorting guidelines, which aids in data analysis and training machine learning models.
- Red-Teaming: I conduct critical assessments of product or platform features by simulating the actions of potential bad actors or testing system vulnerabilities. The objective is to

identify if these features can inadvertently generate or promote content that violates policy or guidelines, thus enhancing the product's security and safety measures.

Q: How would you describe your involvement in content moderation?

- It is my primary job responsibility.
- It is part of my job responsibilities, but not the main focus.
- I volunteer or participate in my free time.
- Other (please specify)

Q: How would you describe your involvement in data labeling?

- It is my primary job responsibility.
- It is part of my job responsibilities, but not the main focus.
- I volunteer or participate in my free time.
- Other (please specify)

Q: How would you describe your involvement in the Red Team?

- It is my primary job responsibility.
- It is part of my job responsibilities, but not the main focus.
- I volunteer or participate in my free time.
- Other (please specify)

Q: How long have you been working as a content moderator? (Please sum up all the months and years of experience as a content moderator, even if there were breaks in between.)

- Less than 6 months
- 6 months to 1 year
- 1 to 2 years
- 2 to 5 years
- More than 5 years
- Prefer not to say

Q: How long have you been working as a data labeler? (Please sum up all the months and years of experience as a data labeler, even if there were breaks in between.)

- Less than 6 months
- 6 months to 1 year
- 1 to 2 years
- 2 to 5 years
- More than 5 years
- Prefer not to say

Q: How long have you been working in the Red Team? (Please sum up all the months and years of experience in the Red Team, even if there were breaks in between.)

- Less than 6 months
- 6 months to 1 year
- 1 to 2 years
- 2 to 5 years
- More than 5 years
- Prefer not to say

Q: What motivated you to take on the role of a content moderator, a data labeler, or a red team member? (Please do not include any information in your response that could be used to identify you as an individual or your employer.)

7.10 Nature of work

Please tell us about the work you do. For the following questions, when we refer to reviewing content, we mean all content that is either generated (by you or by product/platform features) or reviewed (by you) as part of your content moderation, data labeling, or red-teaming activities.

Q: What type of content do you review? (Select all that apply)

- Text
- Images
- Videos
- Audio
- Live streams
- Other

Q: On average, how many hours do you spend reviewing content per week?

- Less than 10 hours
- 10 to 20 hours
- 20 to 30 hours
- 30 to 40 hours
- More than 40 hours

Q: On average, how many contiguous hours do you spend reviewing content per day?

- Less than 1 hour
- 1 to 2 hours
- 2 to 3 hours
- 3 to 4 hours
- More than 4 hours

Q: What categories of content do you review? Please select all that apply.

- Child abuse or exploitation: Depictions of child abuse, child pornography, or any content that exploits or endangers minors.
- Terrorism and extremist content: Content promoting or supporting terrorism, violent extremism, or the recruitment of individuals for such activities.
- Hate speech and acts: Language, imagery, or actions that promote hatred, discrimination, or violence against individuals or groups based on factors such as race, ethnicity, religion, gender, sexual orientation, or disability.
- Harassment and bullying: Content that targets, harasses, or bullies individuals, including personal attacks, defamation, or intimidation.
- Graphic violence and gore: Depictions of excessive violence, injuries, or gore, including real-life acts of violence or cruelty, as well as fictional representations in movies, games, or other media.
- Self-harm and suicide: Content that promotes, encourages, or glorifies self-harm, suicide, or other harmful behaviors.
- Sexual content and nudity: Explicit sexual content, pornography, or non-consensual sharing of intimate images or videos (revenge porn), as well as gratuitous nudity or sexually suggestive material.

- Illegal activities: Content that promotes, encourages, or provides information about illegal activities, such as drug use, theft, or hacking.
- Misinformation and disinformation: False, misleading, or deceptive information, including conspiracy theories, deepfake videos, or manipulated content.
- Copyright infringement: Content that violates intellectual property rights, such as sharing copyrighted music, movies, or images without proper authorization.
- Spam and scams: Content that is intended to deceive, manipulate, or exploit users, including phishing, malware, or other fraudulent schemes, as well as spam or unsolicited promotional material.
- Privacy violations: Content that invades the privacy of individuals, such as sharing personal information without consent, stalking, or doxing.
- Impersonation: Content that falsely represents a person or entity, including fake accounts, profiles, or pages created to deceive or mislead others.
- Other

Q: How would you rate the overall quality of your sleep since you started working as a content moderator, a data labeler, or a red-team member?

- Very poor - I consistently have difficulty falling asleep, staying asleep, or experience restless and non-restorative sleep.
- Poor - I frequently have difficulty falling asleep, staying asleep, or experience restless and non-restorative sleep, but there are occasional nights with better sleep quality.
- Fair - I have a mix of good and bad nights, with some difficulty falling asleep, staying asleep, or experiencing restless and non-restorative sleep.
- Good - I generally sleep well, with only occasional difficulties falling asleep, staying asleep, or experiencing restless and non-restorative sleep.
- Very good - I consistently sleep well, with minimal to no difficulties falling asleep, staying asleep, or experiencing restless and non-restorative sleep.

Q: Have you ever experienced nightmares, flashbacks, or intrusive thoughts related to the content you review?

- Never
- Extremely rarely - less than once a year
- Very rarely - less than once per month
- Rarely - less than once per week
- Occasionally - Once to twice per week
- Frequently - 3 to 5 times per week
- Almost always - 6 or more times per week

Q: Has your work as a content moderator, a data labeler, or a red-team member impacted you positively? If so, please describe how. (Please do not include any information in your response that could be used to identify you as an individual or your employer.)

Q: Has your work as a content moderator, a data labeler, or a red-team member impacted you negatively? If so, please describe how. (Please do not include any information in your response that could be used to identify you as an individual or your employer.)

Q: Has your work as a content moderator, a data labeler, or a red-team member been impacted by the recent rise in the use of generative AI technologies (e.g., large language models like OpenAI ChatGPT, image generation models like DALL-E)? If so, please describe how. (Please do not include any information in your response that could be used to identify you as an individual or your employer.)

7.11 Challenges and opportunities

Please tell us about challenges of your work and opportunities for supporting your work.

Q: In your opinion, what are the most significant challenges facing content moderators, data labelers, and red-team members today? (Please do not include any information in your response that could be used to identify you as an individual or your employer.)

Q: What are some ways that the challenges that you outlined above can or should be addressed by technology? (Please do not include any information in your response that could be used to identify you as an individual or your employer.)

Q: If you could change one aspect of your role as a content moderator, a data labeler, and a red-team member, what would it be and why? (Please do not include any information in your response that could be used to identify you as an individual or your employer.)

7.12 Tools

Please tell us about the tools that you use for work.

For the following questions, when we refer to reviewing content, we mean all content that is either generated (by you or by product/platform features) or reviewed (by you) as part of your content moderation, data labeling, or red-teaming activities.

Q: For each of the tools below, please indicate their usefulness while reviewing content, both directly (i.e., it helps me review the content more accurately or efficiently) or indirectly (i.e., it helps me reduce potentially negative impacts associated with reviewing upsetting content). **Scale:**

- I don't have access to this
- I never used this
- Not at all useful
- Slightly useful
- Moderately useful
- Very useful
- Extremely useful

Tools:

- Automated verification or vetting of user accounts
- Automated warning about certain contents as potentially triggering
- Automated content summarization
- Automated content labeling
- Automated assistance on bringing relevant guideline information for content moderation
- Automated assistance on viewing similar or example content for comparison and decision-making
- Applying monochrome, grey scale, or color filter on visual content
- Blurring of visual content (e.g., blurring face)
- Blocking of visual content (e.g., blocking face)
- Turning off audio and viewing the video only
- Reducing audio volume
- Increasing audio volume
- Turning off video and listening to audio only
- Reducing the speed of audio or video
- Increasing the speed of audio or video
- Listening to other music while viewing the video or text only

- Having control over what part of the content is seen or heard
- Having control over how the content is seen or heard

Q: If there are tools that you found useful while reviewing content which are not mentioned above, please describe them here. (Please do not include any information in your response that could be used to identify you as an individual or your employer.)

7.13 Coping

Please tell us about how you manage demands from work.

For the following questions, when we refer to reviewing content, we mean all content that is either generated (by you or by product/platform features) or reviewed (by you) as part of your content moderation, data labeling, or red-teaming activities.

Q: For each of the activities below, please indicate their usefulness for managing the demands of reviewing content and its impact on your overall wellbeing (e.g., positive coping after viewing upsetting content).
Scale:

- I don't have access to this
- I never used this
- Not at all useful
- Slightly useful
- Moderately useful
- Very useful
- Extremely useful

Activities:

- Changing my work environment (e.g., lighting, temperature)
- Listening to music
- Taking breaks from content moderation
- Conducting distracting activities between content moderation
- Being creative
- Playing games (e.g., video games, board games)
- Tracking and monitoring my physical and mental health throughout the day, week, or month
- Using wearables (e.g., Fitbit, Apple Watch, Oura) to track health, sleep, or mood
- Engaging in relaxation or spiritual practices (e.g., mindfulness, meditation, deep breathing, praying)
- Practicing gratitude, positive thinking, reframing negative thoughts
- Journaling, expressive writing, or reflecting on the day
- Disconnecting from work and/or devices
- Being in nature or outside
- Moving your body (e.g., exercising, walking, stretching)
- Socializing or chatting with colleagues
- Being alone or having a quiet moment alone
- Taking care of basic needs (e.g., eating, hydrating)
- Resting, napping, or sleeping
- Drinking alcoholic beverages or using recreational substances
- Engaging in fun or entertaining activities (e.g., reading, watching TV, shopping)
- Daydreaming
- Getting emotional support from others (e.g., friends, family)
- Seeing a professional therapist or a counselor

- Taking resilience or wellness training or classes
- Planning, task management, or problem-solving

Q: If there are tools that you found useful for managing the demands of reviewing content and its impact on your overall wellbeing which are not mentioned above, please describe them here. (Please do not include any information in your response that could be used to identify you as an individual or your employer.)

Q: Based on your experience with any of the tools used during content review and for managing the demands of reviewing content and its impact on your overall wellbeing, what are some high priority challenges with the tools that should be addressed? (Please do not include any information in your response that could be used to identify you as an individual or your employer.)

Q: What other comments or suggestions related to content moderation, data labeling, or red-teaming, its challenges, or potential improvements in the field do you have? (Please do not include any information in your response that could be used to identify you as an individual or your employer.)

7.14 Compensation and Follow-up

Q: Thank you for participating in the survey to help us explore innovative ways to support content moderators, data labelers, and red-team members. Before we let you go, would you like to receive a gift card for your time spent providing your survey responses? Your survey response will still remain anonymous.

- Yes, I would like to receive a gift card.
- No, I do not want a gift card.

Q: We are looking for participants to join a follow-up interview study to share more insights about their experiences or to envision technology support for content moderators, data labelers, and red-team members. The interview will be conducted online and will take about 1 hour. You will receive a $50 gift card as a token of appreciation. Would you like to be contacted with the follow-up study information? Your survey response will still remain anonymous.

- Yes, please contact me with follow-up study information.
- No, I am not interested in a follow-up study at this time.

7.15 Workshop details

The workshops were structured into three main segments:

(1) *Orientation (10 minutes):* We first gave a quick tutorial on FigJam and asked everyone to introduce themselves. We used this as an exercise to post sticky notes and to react to each other's notes using stamps.
(2) *Validating discovered challenges (20 minutes):* We shared 11 challenge cards (**??** left), each describing one challenge theme discovered in phase 1 and a set of example quotes that illustrate that challenge theme. The example quotes were generated by interpreting and paraphrasing qualitative data from the first phase to preserve anonymity and to evaluate our interpretation. We asked participants to independently look through each challenge card, and vote on those that they agree or disagree with, comment on whether they think the challenge is accurate, and react to each other's notes. If there were missing challenges, they were asked to generate a new challenge card. We finished with a quick discussion of emerging themes. A zoomed-in screenshot of the challenge board is shown in Figure 7.
(3) *Reviewing recommendations (30 minutes):* We introduced the AURA framework with its preliminary recommendations from phase 1. Each recommendation card (**??** right) stated

the recommendation and an example of how that recommendation could be implemented. We asked participants to independently review each recommendation card and comment on what they liked, what they didn't like, what they would change, and how they would want the recommendation implemented. Just as before, they can read and react to each other's notes and can add new recommendation cards if any are missing. As a group, we discussed emerging themes, how the recommendations could be improved, and how our recommendations may be implemented. A zoomed-in screenshot of the recommendation board is shown in Figure 8.

The workshop sessions had slight variations from each other because of the makeup of the group and the comfort level of participants to speak up. The facilitator generally asked for volunteers to speak up or called out participants to elaborate on their stickies to encourage active participation from everyone in the session.

Received January 2024; revised July 2024; accepted October 2024

Fig. 5. A full set of challenge cards used in our workshops.

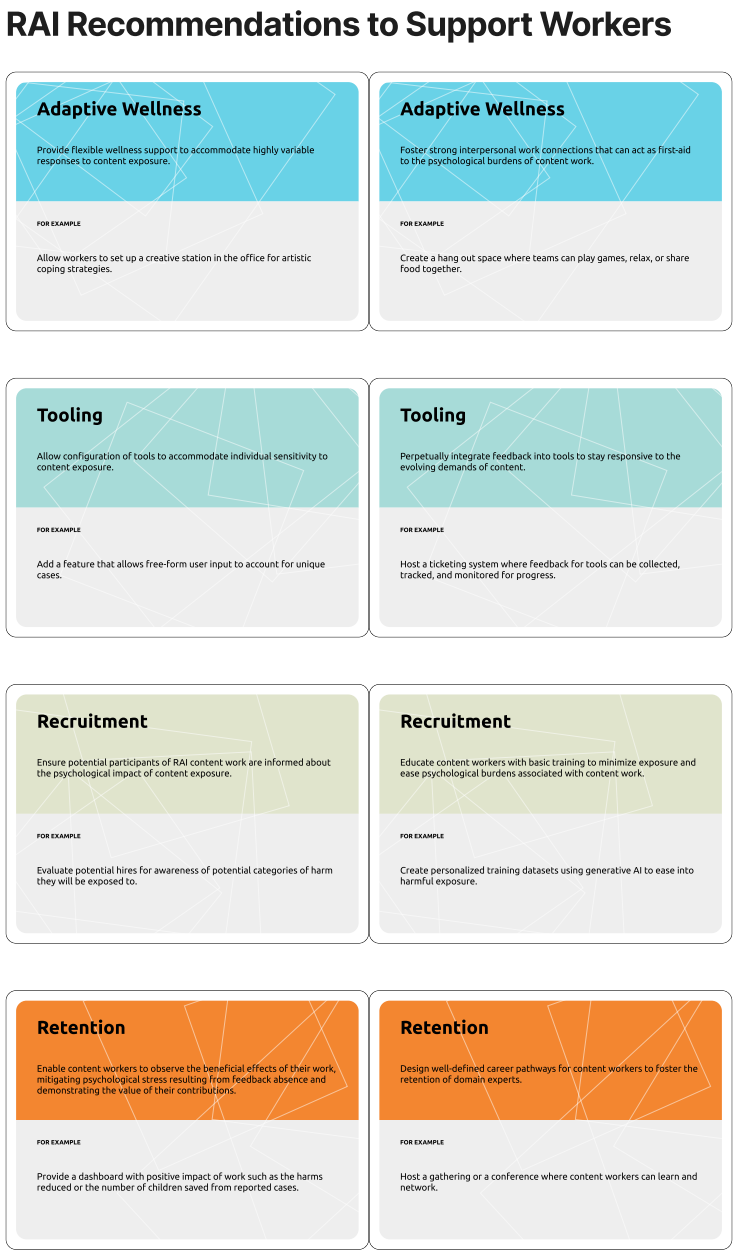

Fig. 6. A full set of recommendation cards used in our workshops.

Fig. 7. A zoomed-in snapshot of the discussion of challenges in our workshop. For each challenge, we asked participants to write about their agreements and examples of what the challenge might look like. Participants reacted to each other's stickies. Participants additionally added thumbs-up stickers for challenges they agreed with the most across all challenges.

Fig. 8. A zoomed-in snapshot of the discussion of our recommendations. For each recommendation, we asked participants to write about their likes, dislikes, and examples of what the recommendation might look like. Participants reacted to each other's stickies.