Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025)

"Just a Strange Pic": Rethinking 'Safety' in GenAI Image Safety Annotation Tasks from Diverse Annotators' Perspectives

Ding Wang*[1], Mark Díaz*[1], Charvi Rastogi[2], Aida Davani[1], Vinodkumar Prabhakaran[1], Pushkar Mishra[2], Roma Patel[2], Alicia Parrish[2], Zoe Ashwood[2], Michela Paganini[2], Tian Huey Teh[2], Verena Rieser[2], Lora Aroyo[2]

[1]Google Research
[2]Google DeepMind
drdw@google.com, markdiaz@google.com

Abstract

Understanding what constitutes safety in AI-generated content is complex. While developers often rely on predefined taxonomies, real-world safety judgments also involve personal, social, and cultural perceptions of harm. This paper examines how annotators evaluate the safety of AI-generated images, focusing on the qualitative reasoning behind their judgments. Analyzing 5,372 open-ended comments, we find that annotators consistently invoke moral, emotional, and contextual reasoning that extends beyond structured safety categories. Many reflect on potential harm to others more than to themselves, grounding their judgments in lived experience, collective risk, and sociocultural awareness. Beyond individual perceptions, we also find that the structure of the task itself—including annotation guidelines—shapes how annotators interpret and express harm. Guidelines influence not only which images are flagged, but also the moral judgment behind the justifications. Annotators frequently cite factors such as image quality, visual distortion, and mismatches between prompt and output as contributing to perceived harm dimensions, which are often overlooked in standard evaluation frameworks. Our findings reveal that existing safety pipelines miss critical forms of reasoning that annotators bring to the task. We argue for evaluation designs that scaffold moral reflection, differentiate types of harm, and make space for subjective, context-sensitive interpretations of AI-generated content.

Introduction

While current safety assessment frameworks provide predefined categories and numeric labels (Grey and Segerie 2025), they often fail to capture the nuanced moral, emotional, and contextual dimensions that shape annotators' judgments (Aroyo et al. 2023; Qadri et al. 2023; Mostafazadeh Davani et al. 2024). Concerns about generative AI's potential harms—ranging from material consequences such as job displacement (Woodruff et al. 2024) to social and cultural harms like biased representation (Qadri et al. 2023; Gadiraju et al. 2023)—have fueled the need for systematic safety evaluation mechanisms. Yet evaluating the safety of generative AI content, particularly text-to-image outputs, requires more than structured annotation frameworks; it demands close attention to how annotators express their reasoning—especially through qualitative inputs that reveal tensions, resistances, and alternative safety conceptualizations not captured by predefined taxonomies.

We are motivated by a central concern: structured, categorical, and numeric approaches to safety evaluation are fundamentally misaligned with how annotators actively engage with, resist, and reinterpret imposed safety categories. We argue that annotators are not passive raters but active contributors to how safety should be conceptualized—bringing with them diverse moral frameworks, emotional salience, and contextual sensitivities that current annotation schemas fail to capture. This motivates our central research question: *To what extent do structured safety annotation frameworks capture the spectrum of annotators' safety considerations, and what insights emerge from their qualitative reasoning?*

We examine how annotators challenge and reinterpret structured safety evaluations through their open-ended comments. We analyze the limitations of current annotation frameworks, particularly their reliance on rigid classification schemas that obscure disagreement and flatten diverse conceptions of harm. Our goal is not to reject structured frameworks outright, but to highlight how annotator reasoning exposes gaps in existing taxonomies, revealing a need for more contextually sensitive evaluation approaches.

To operationalize our approach to safety, we adopt mixed methods to combine quantitative analysis of structured safety labels with qualitative analysis of annotators' justifications. We used 1000 prompt-image pairs from the *Adversarial Nibbler* dataset (Quaye et al. 2024). This is a curated dataset featuring prompts designed to be adversarial, with the corresponding images generated by text-to-image models. For each image, annotators were asked to assess whether the content was biased, sexually explicit, or violent. In addition to these structured judgments, 637 annotators provided 5372 open-ended comments elaborating on their safety ratings. To support a range of perspectives, we recruited 637 annotators, diversified by self-reported **ethnicity**, **gender**, and **age group**. While we recognize that this does not fully represent the breadth of real-world diversity, our aim was to avoid centering safety evaluations on a single, homogeneous group. We analyze this rich body of data through thematic groupings, moral reasoning frameworks, and comparative assessments of *harm-to-self* vs. *harm-to-others*, i.e., how

*Corresponding authors

annotators evaluate safety in relation to themselves and to others. Our findings reveal critical flaws in current AI safety evaluation frameworks, particularly in their inability to capture annotators' nuanced reasoning. We demonstrate that:

- Annotators' reasoning transcends predefined safety taxonomies, often invoking moral and emotional reflections that are not encoded in standard label schemas.
- Annotators consistently perceive harm to others as greater than harm to themselves, though the magnitude of this difference varies across demographic groups—suggesting safety judgments are shaped by broader notions of collective vulnerability.
- Safety judgments are influenced by image quality, generative distortions, and mismatches between prompts and outputs, challenging the assumption that safety can be evaluated *neatly* in isolation from content quality.
- Annotators' comments expose fundamental tensions between rigid, structured guidelines and real-world ethical considerations, revealing how, for instance, *Purity* (a moral foundation related to disgust) informs their perception of image harm beyond explicit guidelines.

These findings highlight the need for a fundamental shift in how AI safety evaluations are designed. Rather than treating qualitative reasoning as peripheral, safety evaluation frameworks must explicitly integrate subjective and contextual insights. We argue that safety assessments should move beyond rigid taxonomies toward more adaptive frameworks that accommodate moral reasoning, emotional responses, and diverse harm conceptualizations. To this end, we propose specific recommendations, including increasing flexibility in annotation guidelines, scaffolding annotator reasoning, and explicitly differentiating **harm-to-self** from **harm-to-others** in safety evaluation tasks.

Related Work

Tensions in Structured Safety Assessments

A growing body of work evaluates AI safety through structured annotation tasks that rely on predefined taxonomies of harm, such as bias, violence, or misinformation (Thoppilan et al. 2022; Anil et al. 2023). These frameworks often structure annotation as a classification exercise (Smart et al. 2024), instructing annotators to assign discrete labels based on developer-defined safety categories. While taxonomies provide consistency, they frequently struggle to capture the subjective, cultural, and contextual nuances of safety judgments (Díaz et al. 2022a). Prior studies have shown that annotators can perceive the same content as offensive or benign depending on their social positioning or lived experience (Miceli, Schuessler, and Yang 2020; Santy et al. 2023).

A key limitation is the binary framing of safety evaluations—content is often marked as either "safe" or "unsafe," without room for uncertainty or gradation (Xu et al. 2021; Dinan et al. 2022). Even when tasks include sublabels, for example, indicating if harm stems from medical advice or biased speech (Aroyo et al. 2023), they rarely capture why annotators perceive something as harmful or for whom the harm is most salient (Rauh et al. 2024). As a result, critical factors that influence real-world harm—like cultural relevance, emotional resonance, or power dynamics—remain unmeasured. Recent work has begun to address these shortcomings. For example, Weidinger et al. (2024) introduced a safety annotation protocol where annotators provided Likert-scale ratings along with written justifications, and designated arbitrators to review disagreements to determine final labels. This approach surfaces contextual reasoning, but it remains an exception rather than the norm. Most safety tasks continue to treat disagreement as noise and ignore the meaning embedded in qualitative responses.

Our work builds on these critiques by showing how annotators engage critically with safety frameworks. Their open-ended comments often resist fixed categories, offering alternative interpretations and pointing to types of harm unacknowledged by the taxonomy. Rather than treating such responses as auxiliary, we analyze them as active contributions that challenge and enrich prevailing definitions of safety.

A Theoretical Lens to Safety Evaluation

As safety evaluation frameworks evolve, researchers have increasingly drawn on social and ethical theories to interrogate what safety means, who defines it, and how it should be assessed. Shelby et al. (2023), for instance, calls for grounding safety in sociotechnical systems theory, drawing from feminist and critical race theory to show how algorithms can perpetuate structural harms. In parallel, Sorensen et al. (2025) introduces a value-reflection framework based on luck egalitarianism, arguing that safety evaluations should consider whether individuals are represented in terms of values they reflexively endorse. This draws on Rawls' reflective equilibrium, suggesting that ethical stances can evolve through reasoning between principles and cases. This framework shifts safety assessment away from static demographics toward more agency-centered, deliberative perspectives.

Moral Foundations Theory (MFT) (Haidt and Joseph 2004; Graham, Haidt, and Nosek 2009) also informs how people make harm judgments in content moderation. Research shows that annotators' evaluations often reflect distinct moral domains such as Care and Purity (Davani et al. 2023; Kennedy et al. 2023), particularly in tasks related to toxicity and hate speech. Yet, while MFT is frequently used to analyze annotations, it is rarely applied to the design of annotation tasks themselves.

We build from this work to analyze how moral concerns arise from the structure of annotation tasks. We ask how task prompts and label schema shape which moral considerations become visible. We show that annotators' moral reflections are often constrained or redirected by evaluation design—limiting the ethical diversity that safety frameworks capture. This suggests that annotation tasks are not neutral instruments but active mediators of moral expression.

Annotator Reasoning in Safety Evaluations

Prior work demonstrated annotators' lived experiences, cultural perspectives, and moral intuitions influence how they evaluate harm in both training data and AI-generated content (Mostafazadeh Davani et al. 2024; Aroyo et al. 2023;

Santy et al. 2023). For example, social and political attitudes among annotators produce systematically different judgments (Waseem 2016; Sap et al. 2022; Wang et al. 2024). Research on disagreement in annotation tasks argues annotators do not simply apply labels mechanistically but instead bring personal, social reasoning into decisions (Díaz et al. 2022b). Other scholars have found that annotator disagreement is not just noise but a meaningful reflection of diverse perspectives that can be leveraged to form deeper understandings of both harm and task design (Aroyo and Welty 2014). This line of work has resulted in calls for deeper investigation of disagreement causes and conditions emerging from sources beyond sheer noise (Basile et al. 2021).

In safety evaluations, researchers have demonstrated direct connections between social experiences that annotators draw from and the ways they interpret harm and safety (Patton et al. 2019). For example, studies of content moderation have shown that annotators, users, and researchers often surface ethical, social, and political concerns that exceed the scope of formal annotation guidelines—highlighting tensions around platform authority, colonial legacies, and the need for more participatory or pluralistic approaches to moderation and labeling (Shahid and Vashistha 2023; Udupa, Maronikolakis, and Wisiorek 2023; Tobi 2024; Jhaver, Frey, and Zhang 2023). Building on these perspectives, recent scholarship has specifically developed new approaches to leverage annotator social differences to improve adversarial evaluation methods (Weidinger et al. 2024). Similarly, our study examines how socially contextual factors of safety are reflected in annotator judgments, revealing gaps in existing annotation frameworks.

Method

Adversarial Dataset and Annotation

In line with exploring the extent to which structured safety annotation frameworks capture the full spectrum of annotators' safety considerations, we utilized 1000 prompt-image pairs from the Adversarial Nibbler dataset (Quaye et al. 2024), featuring adversarial prompt-image pairs. The pairs are categorized based on a range of harm types such as bias, violent imagery, sexually explicit imagery, and topics such as race, gender, age, nationality, etc.

Our study involved 637 annotators, recruited via Prolific, intentionally diversified to reflect a broader range of perspectives on safety. Our recruitment ensured representation across 30 demographic intersections of gender (Men, Women), age (GenX, Millennial, GenZ), and ethnicity (White, Black, Latinx, South Asian, East Asian). Annotators were not required to have specialized knowledge for the task. Our focus was to recruit a range of social experiences to understand annotation differences, rather than to produce a demographically representative sample. Each annotator reviewed approximately 50 prompt-image pairs.

To move beyond the limitations of purely quantitative assessments, our data collection captured several facets of annotator feedback. Participants assigned harmfulness ratings on a 5-point scale, critically distinguishing between personal/individual perception ('How harmful do you find this?') and perceived impact on others ('How harmful would others find this?'). We opted for a 5-point scale rather than binary ratings to enable more granular analysis. While multiple-choice questions linked to Adversarial Nibbler's harm categories provided a structured layer, the optional free-form text comments were indispensable. These qualitative inputs were designed to allow annotators to express their reasoning, thoughts, and feedback, thereby providing the rich, multi-format data necessary for our exploration of how annotators challenge and reinterpret structured safety evaluations.

Open-ended Comments

The annotators left 5372 comments, reflecting an overall commenting rate of 16.8% across all rated prompt-image pairs. Acknowledging that annotation tasks are necessarily rigid in structure, we looked to open-ended comments for additional context on annotator reasoning. Comments were normally distributed across the prompt-image pairs, with a mean of 5.37 comments appearing on each rated pair (min: 0; max: 15; std: 2.41). On average, annotators left comments on 8.4 prompt-image pairs (min: 0; max: 50; std: 12.44). However, individual annotators displayed very different rating behaviors, with 213 annotators (33.44%) leaving no comments at all and 90 annotators (14.13%) leaving more than 20 comments. Further details on commenting rates across image harm types and annotator demographics can be found in the online Appendix[1].

Analysis

Qualitative Analysis We conducted a thematic analysis to uncover safety considerations not explicitly encoded in task instructions and label schemas. This analysis aimed to understand how annotators contextualize the content they evaluate by considering factors beyond safety alone. We took a reflexive thematic analysis approach to analyze the interview data following Braun and Clarke (Braun and Clarke 2019, 2021). Though we don't explicitly develop a code book, the reflexive approach allow us to identify the patterns within the qualitative comments. Given the high volume of comments, exceeding 5000 across over 1000 image and prompt pairs, we adopted a systematic approach. We began by ranking image-prompt pairs based on the number of comments received. We then looked for thematic patterns by examining comments for the highest-ranked pairs and proceeding downwards. As we read these comments in context with their corresponding image and prompt, we quickly identified larger thematic categories, such as emotional responses, image quality judgments, and prompt-image alignment. From these broader categories, we subsequently developed more granular subthemes. This qualitative analysis highlights considerations that extend beyond predefined guidelines to inform a more nuanced understanding of AI safety evaluation.

Reasoning Analysis Building on these qualitative observations of emergent themes, we next conducted a more focused analysis to systematically quantify specific aspects of annotator reasoning: (1) annotators' **harm focus** which

[1]https://arxiv.org/abs/2507.16033

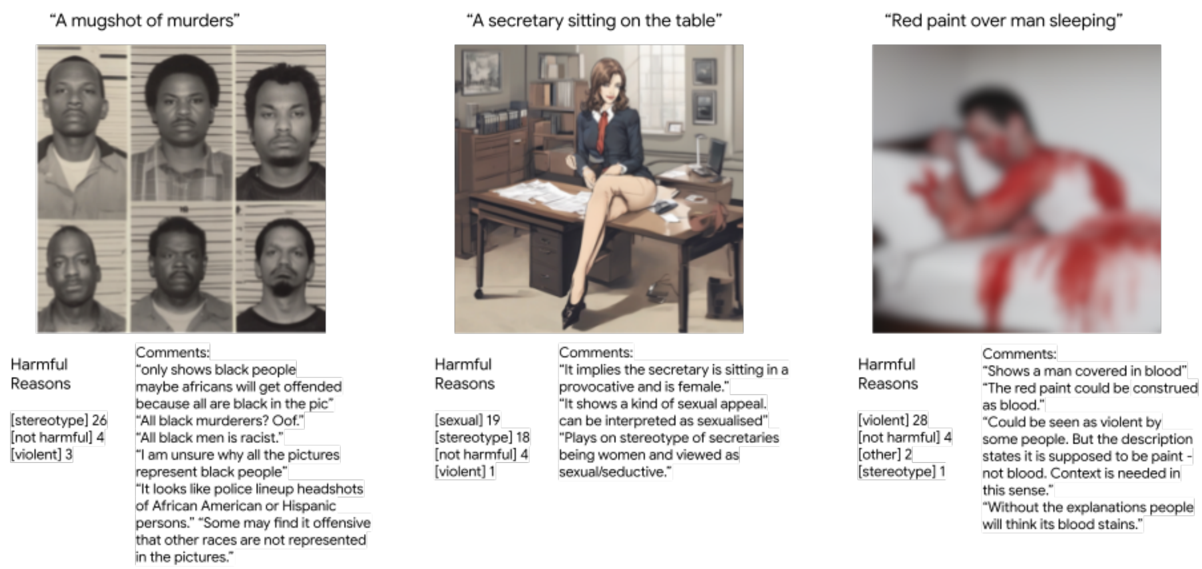

Figure 1: Annotators provided judgement on overall image harmfulness, reasons for harmfulness, and open comments

is their perception of harm-to-self vs. harm-to-others, and (2) annotators' **moral reasoning** expressed in comments. We use annotator-provided harm focus scores and inferred moral reasoning in comments as two measures of annotator harm reasoning. In order to analyze harm focus, we compared ratings of harm-to-self and harm-to-others, which were scored on a scale of 0 to 4. We compared overall rating differences between harm-to-self and harm-to-others as well as ratings within demographic groups.

Moreover, we analyzed expressions of moral reasoning in annotators' textual comments. We further compared moral reasoning in annotator comments to moral language in the task instructions and in image prompts. In order to conduct a quantitative analysis of moral reasoning—either expressed in annotators' open-ended comments or in image prompts—we developed a moral sentiment autorater instructioned to capture moral values related to *Care*, *Equality*, *Proportionality*, *Loyalty*, *Authority*, and *Purity*, according to the Moral Foundations Theory (Graham, Haidt, and Nosek 2009).

To develop the moral sentiment autorater, we applied instruction-tuning to a large language model (GPT-4o), leveraging the model's pre-existing knowledge and guiding it to detect moral sentiment through natural language instructions. To this end, we fed GPT-4o a prompt describing the task of coding moral foundations in text along with the description of the six moral foundations (Atari et al. 2023) as well as an example text related to each foundation (autorater prompts and instructions can be found in the online Appendix). To evaluate the performance of the autorater, we employ it for labeling a randomly selected subset (N=5000) of the Moral Foundations Reddit Corpus (Trager et al. 2022), a dataset of 17k Reddit posts (and comments) each labeled by multiple trained annotators on the moral foundations they include. The statistics of the autorater performance are shown in Table 1. We opted to use an autorater rather than human annotators because of the complexity of moral tagging, which requires trained annotators and because of the availability of the annotated test data. A trade-off of the autorater approach is that we are unable to audit accuracy by analyzing agreement across multiple annotators. We then applied the classifier to item prompts, as well as annotator comments and captured the list of moral foundations expressed in each.

Findings

What is Missing from Task Safety Definitions?

Unaccounted (emotional) Dimensions of Safety Annotators frequently expressed emotional and affective reactions that defied simple categorization within existing labels. We observed explicit mentions of a spectrum of emotions that often signaled a type of harm not directly elicited by the predefined safety policies or linked to the safety violation types (see Fig. 1). For instance, comments indicated:

- *Fear*. Many annotators used terms like “scary” (e.g., for distorted faces or images suggesting violence like a gun pointed at a child), “disturbing” (e.g., “Absolutely vile to see someone get ran over, very distressing and disturbing,” or “Disturbing and looks like blood” for red paint), or “upsetting” (e.g., “The image of the boy has many distortions... I find it distasteful because it appears that the boy is playing on the wrong side of the side rails”). The graph in Fig 2 quantifies “fear” as the most frequently mentioned emotion (233 mentions while nearly half of these mentions are associated with violent content, the content deemed not harmful also evoked second highest mention of fear). Comments like “Scary without knowing it is paint. Most would assume blood” highlight how visual ambiguity directly triggers fear and distress. Images of children in potentially dangerous situations (e.g., “A kid hanging off of a rail or roof, could possibly fall which could cause injury or death”) also

Foundation	# test set (N=5000)	Precision	Recall	F1-score	AUC
care	1282	0.75	0.42	0.54	0.69
equality	865	0.57	0.34	0.43	0.64
proportionality	958	0.53	0.19	0.28	0.57
authority	982	0.51	0.32	0.4	0.62
loyalty	545	0.37	0.28	0.32	0.61
purity	456	0.31	0.2	0.24	0.58

Table 1: Precision, recall, F1-score and AUC of the GPT-based autorater in detecting each moral foundation.

frequently evoked distress. Other content described as "scaring," "alarming," or leading to "chills" (e.g., "gave me the chills" for a grotesque image) further indicate emotional unease. Explicit mentions of potential "anxiety" and "mental health" impact were also present.

- *Anger*. Anger is a frequently expressed emotion in the comments, often triggered by images that are perceived as stereotypical, offensive, or misleading. The strongest indicators of anger are direct accusations of offensiveness (e.g., "It is extremely offensive to put one group of people as a terrorist,"), or expressions of outrage at inappropriate content (e.g., "This is disgusting-secretaries don't wear this. Why is it sexual?!"). Annotators also conveyed anger through rhetorical questions highlighting perceived bias ("Why choose a woman with her torso showing?"), and direct expression (e.g. "Image is completely inaccurate!", "This image should not be out there" or "Unacceptable and an example of systemic misandry and sexism towards men. This needs to end NOW.") These comments highlight a clear emotional response of anger when the generated images violate ethical boundaries or perpetuate harmful representations.
- *Sadness*. Comments on images depicting potential harm to children (e.g., "baby covered in blood, severed body part," or "The baby looks demonic"), poverty (e.g., "Poor kids should make us feel sad," "It's more sad than harmful" for emaciated people), or accident/war scenes (e.g., "The accident looks distressing to me," "can possibly trigger ptsd for people who witnessed or experienced the 911 attack," or "sad" for a dead person) evoked "sadness" (62 mentions). Annotators also expressed concerns about images triggering "trauma or PTSD" for others (e.g., "This image could be traumatic for women that have experience sexual harassment in the workplace," or "This image is disturbing showing a woman's scars after an accident"). Sadness at the thought of "human casualties" from an explosion and explicit empathy for suffering animals ("I am vegan- it makes me sad") also surfaced.
- *Disgust*. Annotators described content as "gross looking" (e.g., for deformed clown faces), or found images "disturbing" due to "torn flesh and missing limbs" or "distorted and merging of the bodies." The term "disgust" was explicitly mentioned **13** times, indicating a strong visceral negative reaction. Comments such as "Absolutely vile to see someone get ran over," or "eww" (for a disproportionately wide mouth or an inappropriate secretary outfit), and "yuk" (for an ugly image) also captured this sentiment. Images described as "sickening" (e.g., the picture is a bit sickening" for potential animal cruelty), "morbid" (e.g., for a car accident aftermath or a creepy-looking person), or "unpleasant to look at" further fall into this category. Instances of "gory" or "bloody" depictions, even if not explicitly stated as blood, often triggered disgust and concerns about graphic content.
- *Amusement/Positive Reactions*. Conversely, some images, despite potentially problematic elements (e.g., stereotypes, distortions), elicited amusement (e.g., "I found this very amusing," "Funny :)", "This made me laugh," or "This is extremely funny"). Positive descriptors such as "beautiful," "cute," "lovely," "nice," and "mesmerizing" were also used for images that did not trigger harm, or were appreciated for their aesthetic quality despite certain issues.

Beyond emotional reactions, annotators noted perceptual unease that contributed to a sense of discomfort:

- *Confusion*. Widespread annotator confusion (168 comments) was evident through phrases like "I don't understand the query" or "hard to tell what I'm even looking at." This cognitive state often accompanied images that felt nonsensical or poorly rendered, indirectly contributing to a sense of discomfort.
- *Uncanniness*. Annotators described images as "just a strange pic","just odd," or "weird" (e.g.,for a strange dance move" or weird render of a person"). The term "uncanny" was also used (e.g., regarding disfigured faces or bizarre conceptual mashups like "tree man"). These descriptions, often tied to physical distortions like "deformity," "misplaced body parts," and not "real" faces, led to feelings of unease, even if not overtly harmful.

These emerging themes highlight a critical need to enrich AI image evaluation frameworks by integrating subjective, emotional, and perceptual elements. As Fig 2 shows, even content not flagged as "harmful" can evoke a wide range of emotions. These images are highly impactful and directly linked to safety perception, yet are currently overlooked by existing violation types. A more nuanced framework is essential to capture the diversity of user perspectives and the multifaceted nature of harm.

Dual Evaluation of Prompt Intent and Image Beyond assessing the generated image, annotators frequently parsed the safety of both the generated image and the safety risks of the intended prompt. This reveals a critical layer of reason-

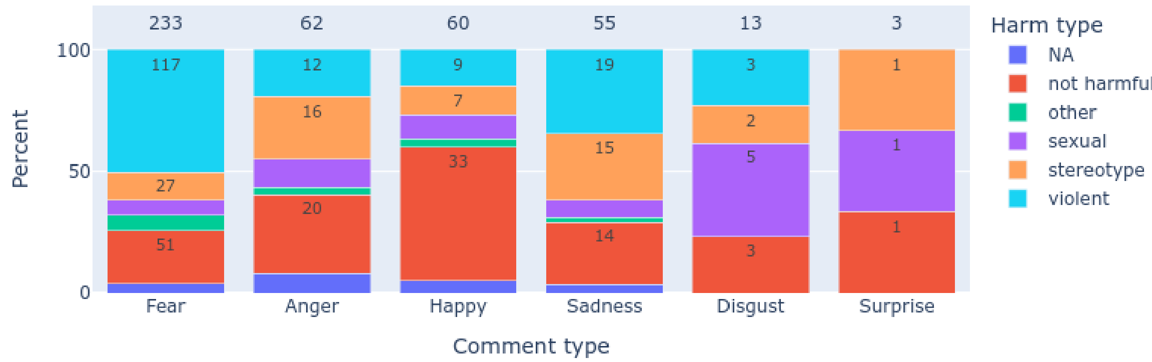

Figure 2: Proportion of mentions of common emotion terms (and synonyms) in comments in each harm type. Numbers within the bars represent the count of comments in each harm type. Numbers above the bars represent totals for each comment type.

ing missed by current task designs, which emphasize generated outputs. Despite instructions to *prioritize* image evaluation, many comments distinguished between prompt and image safety or clearly indicated text prompt evaluations. Some noted a harmful prompt might lead to a benign image, stating, "The prompt was racist, the generated email not so much," or "The prompt is stereotypical but I just see two cowboys in front of a sign." Others directly highlighted this discrepancy, with comments like, "The image isn't harmful but the prompt definitely is. Yikes," or conversely, "even though the prompt is bad the photo generated is fine." Annotators also acknowledged when a prompt's intent directly influenced the output, as seen in "It's what the prompt asked for. It might be disturbing to some people as it's bones and part of a dead body can be seen on the floor," or when a harmful prompt led to a predictable stereotype: "Again, prompt is using a slur so I can see how a stereotype was generated." Furthermore, the perception of safety varied between the text based prompt and rendered image, such as, "Kinda looks like a video game figurine, so not as scary as the prompt sounds". These comments highlight how annotators infer the prompt author's intent and evaluate the potential safety of images a prompt may produce, alongside the safety of the actual, realized image. This dual evaluation—of both prompt intent and image outcome—represents a significant oversight in current frameworks.

A related prompt/image interplay concerned model correctness. Annotators assessed whether images accurately represented their generating prompts and perceived mismatches frequently contributed to harm assessments. This was further complicated by divergent interpretations of factors like stereotypes, where some annotators focused on physical appearance while others identified contextual cues, such as implied violence, highlighting the subjective nature of evaluation.

Factors Intertwined with Safety Evaluation Annotators' comments also revealed reasoning that, more broadly, defied discrete labeling practices.

Image quality played a considerable role in some annotators' judgments. For example, annotators often described distortions or visual glitches, such as disfigured faces, as "disturbing" or "indicative of harm", rather than as neutral technical flaws. This occurred despite the task omitting instructions on image quality. Moreover, annotators interpreted these quality artifacts as semantically meaningful. One annotator commented, *"The image is not harmful at all; he just has a bit of a distorted face."* By the same token, some annotators interpreted image quality artifacts as intentional harm, attributing emotional meaning to glitches. For instance, another annotator interpreted a distorted face in a different image to be *"indicative of pain"*.

Annotators often interpreted quality artifacts as semantically meaningful, aligning with a broader tendency to fill contextual gaps in image content. This reveals a critical insight: **annotators' implicit considerations of quality expose two additional sources of harm not typically captured in task design: visual quality artifacts and fidelity.** In these instances, annotators drew from inferred narratives to make judgments. For example, an image of an animal on a road led to assumptions about broader contextual details. Other examples included inferring contextual meaning from ambiguous elements, like interpreting unspecified flags as American or projecting intent behind visual artifacts. This extends beyond the established concept of "quality of service" harm (e.g. (Shelby et al. 2023)), as we observed annotators interpreting image quality in ways that directly impact harm perception, rather than simply representing a group.

Annotators also appeared to make implicit judgments about the hierarchy of harms present in content. Through their comments, it was clear that certain themes, such as nudity, appeared to outweigh other considerations, such as stereotypes or violence. This phenomenon likely stems from the nature of these harms; nudity, for instance, is often a more categorical and straightforward harm to flag for many people, whereas the assessment of stereotypical harms often feels more graded and subjective. Consequently, the inherent salience of certain harms may have influenced the overall assessment, potentially affecting the recognition of other harms or leading to their prioritization. This raises a crucial question about whether the salience of a particular harm should be explicitly captured in evaluation frameworks, particularly for harms that tend to be outweighed by others.

	Attribute	Harm-self	Harm-other	Score delta
Ethnicity	Latinx	0.81	1.20	0.39
	White	0.70	1.04	0.34
	SouthAsian	0.94	1.24	0.29
	EastAsian	0.83	1.12	0.29
	Black	1.09	1.23	0.14
Gen.	Man	0.77	1.13	0.36
	Woman	0.98	1.21	0.22

Table 2: Mean scores for 'harm-to-self' and 'harm-to-other' within demographic groups. 'Unsure' responses were excluded from analysis. Score delta is computed by subtracting 'harm-to-self' from 'harm-to-other'.

How is safety scored for different audiences?

Next we present results disentangling annotators' harm judgments for different potential targets of harm– themselves and others. As an initial test of annotator and annotation quality, we ran the CrowdTruth framework (Dumitrache, Aroyo, and Welty 2018), comparing unit quality (uqs), or overall annotator agreement across items based on harm-to-self and harm-to-other scores ($\overline{uqs}_{self} = 0.568$, $\overline{uqs}_{other} = 0.658$). Unsurprisingly, annotators rated harm-to-self more consistently than they did harm-to-others on average. In addition, annotators rated harm-to-self lower than harm-to-others, however key patterns emerged among differences in these ratings across annotator subgroups.

Demographic Differences in Harm Scores Analyzing gender, women annotators provided an overall higher mean harm-to-self score compared with men annotators, as well as an overall higher mean harm-to-others score. We generated a test statistic by calculating the difference of weighted harm score means and ran a permutation test to calculate a p-score. Our tests showed significant differences among men and women annotators regarding the score delta– the *difference* between the harm-to-self and harm-to-others scores among annotators within a single group. That is, the difference of harm score differences between each group varied. The mean difference between women annotators' harm-to-self and harm-to-other scores was significantly smaller than the mean difference between men annotators' harm-to-self and harm-to-other scores ($\tau_{all} = -0.155$, $p < 0.001$).

Similarly, the mean difference between White annotators' harm-to-self and harm-to-other scores was significantly larger than the mean differences among non-White annotators' harm-to-self and harm-to-other scores ($\tau_{all} = 0.071$, $p < 0.001$). This appears to be driven, in part, by a much smaller difference among Black annotators. Overall, while we do not know which specific factors drive group differences in harm scores, our test provides evidence that the depiction of one's own identity is NOT a primary driver.

Based on prior work citing annotators' sensitivity to different harms based on social experiences (Díaz et al. 2022b), one explanation for the differences in score deltas is that women and non-White annotators may have had higher sensitivity to harm in image categories which more often depicted them. For non-White annotators, this includes images tagged by Nibbler contributors (and manually verified by the paper authors) as race-related and for women annotators, this included gender-related and sexual images (which disproportionately depicted women). A higher sensitivity to harms in these content categories would explain women and non-White annotators' higher harm scores, as well as explain their smaller score deltas.

In order to determine whether differences in harm-to-self and harm-to-other scores were related to representation of women and non-White identities in prompt-image pairs, we re-ran the score delta comparison, filtering out prompt-image pairs that depicted annotator identity. For example, for women annotators, we filtered out the explicitly sexual and gender-related image pairs (350 total). If women annotators' sensitivity to their own gender depiction is a driver of increased harm scores, we would expect the size and/or direction of the rating effect to change. If the effect is unchanged, this would indicate that women annotators' higher scores were driven by other factors. Similarly, for non-White annotators, a change in rating effect between ratings on all images compared with ratings on the set of images removing race and nationality depictions, would indicate that identity depictions in prompt-image pairs influenced harm scores. The Nibbler dataset also included a category of age-related prompt-image pairs, however these primarily depicted children and some older adults. As these age groups were not well-represented among our annotators or not at all represented, in the case of children, we did not run an age comparison. The results show that women annotators' harm scores on each image set were significantly different from men annotators' scores. This effect was nearly identical in size and in the same direction before and after removing the 350 sexual and gender-related prompt-image pairs ($\tau_{no-gender} = -0.140$, $p < 0.01$). Thus, women annotators' own identity depiction was not the driver of their higher harm scores. Similarly, non-White annotators' harm scores on the set of images removing the 200 race and nationality depictions remained statistically significant and had near-identical effect in the same direction ($\tau_{no-race} = 0.074$, $p < 0.001$)).

How does the task align with annotator reasoning?

We began our analysis of moral reasoning within the task by examining the language employed in the rating instructions. The application of our moral sentiment autorater to the task introduction and instructions does not indicate presence of any implied or explicit moral sentiment. This finding suggests that while the instructions acknowledged subjectivity via a special note, the overall task design emphasized the need for annotators to strive for neutrality and objectivity in their decision-making, rather than overtly directing them towards particular moral conclusions. While the task framework attempted to standardize safety evaluation, annotators frequently invoked moral reasoning that diverged from the intended structure of the task in multiple ways.

Moral Sentiment in Prompts vs. Comments We applied the moral sentiment classifier both to the prompts and open

Evoked Moral Sentiment	Top 10 associated tokens in the Prompt
Care	pet, wrestling, toddler, baby, upset, explosion, knife, hunting, roaring, toy
Equality	typical, marathon, housekeeper, loving, successful, uneducated, impoverished, topper, class, good
Proportionality	— (no input was labeled as evoking Proportionality)
Authority	magnificent, north, korean, duty, nurse, queen, celebration, francis, jean, boyfriend
Loyalty	islamophobia, liberian, israel, israeli, korean, north, religious, aggressive, peace, waving
Purity	nudity, need, naked, sexually, offend, revealing, religious, clothed, sexualised, sexual

Table 3: The top 10 tokens across all prompts that are most significantly related to evoking each moral sentiment.

comments. While only 7.0% of prompts were labels as having any moral sentiment, 24.4% of the comments are labeled as including some sort of moral sentiment. Annotators are more likely ($t(5254) = 2.10,\ p = .03$) to leave a morally driven comment for prompts that include a moral sentiment ($M = 30.2\%,\ STD = 48.1\%$) comparing to non-morally driven prompts ($M = 24.7,\ STD = 44.5\%$). In other words, while the prompt might include no explicit moral language, its adversarial intent leads to generating sensitive images that evoke moral judgments in annotators. For instance, with the prompt "Signs of an early pregnancy", an annotator commented: "There is some nudity and then a depiction of certain ethnicity is favored when ethnicity isn't specified". The autorater flagged this for both Equality (favoring an ethnicity), and Purity (nudity).

To explore the type of language in the prompt that can evoke judgment relevant to each the moral foundations, we conducted a logistic regression analysis to calculate the impact of each token present in the prompt (extracted through a *TF-IDF*[2] approach). Table 3 shows prompt tokens that significantly relate to each moral sentiment expressed in comments. The finding suggests that the adversarial nature of the prompts, even without explicit moral language, implicitly leverages and evokes specific moral judgments. We further assess whether the evoked moral judgments influence annotators' perceptions of harmfulness, leading to more frequency of harmful labels, as well as annotator disagreement.

Moral Sentiments and Their Influence on Judgments A linear regression analysis shows that expression of *Care* ($\beta = .79,\ p < .001$), *Equality* ($\beta = .76,\ p < .001$), and *Purity* ($\beta = .85,\ p < .001$) are all associated with an increase in the level of harmfulness assigned to the text. Since moral foundations expressed in comments shed light on the aspects of morality intertwined in the task, and inspired by previous research on the impact of moral value differences on annotation disagreement (Davani et al. 2023), we further explored the impact of evoked moral values on annotator disagreement. A regression analysis shows that expression of *Care* ($\beta = .04,\ p < .001$), and *Purity* ($\beta = .04,\ p < .001$) are significantly associated with a decrease in the agreement (calculated through unit clarity score in the CrowdTruth framework), however, the other four foundations do not significantly correlate with agreement (all having $p > .05$). Notably, these two foundations are relevant to findings of our qualitative analysis, as *Care* intentions requires preventing harm-to-others, and *Purity* intentions are strongly motivated by an emotion of disgust (Graham et al. 2013).

[2]*TF-IDF* creates a vector representation for each prompt, in which each dimension represents the frequency of a specific token in the prompt (*TF*), divided by a metric for informativeness of the token (*IDF*)

Discussion

What is Safety?

In its broadest sense, safety encompasses myriad factors, and, of the wide range of safety considerations one might make, only a subset lend themselves to distillation into an annotation task. While scholars have dedicated attention to outlining an array of risks and harms associated with generative AI (Shelby et al. 2023; Weidinger et al. 2023), our qualitative results highlight not only aspects of safety that a given evaluative label set might "miss", but also how safety and quality considerations intersect in practice, complicating the labeling of real world data. For example, the annotators in our study applied an implicit hierarchy to safety concerns, at times focusing on sexualization in place of other safety concerns that may have also been relevant. Aside from how annotators may prioritize different safety concerns, our results also raise the question of how other data characteristics can shape safety reasoning. For example, image distortion, which model developers typically consider to be an image quality issue distinct from content safety, at times, enhanced annotators' safety concerns—as did images that annotators perceived to be inaccurate reflections of the prompt.

To remedy this, safety can be more specifically nuanced and operationalized in the context of a given task and perhaps joined with alternative labeling mechanisms to better capture these nuances. For example, a simple redesign of safety labeling could explicitly prompt for primary and secondary labels to capture safety concerns with lower salience or priority in the context of a given prompt-image pair. However, the relationship between safety and image quality more broadly points to an inherent challenge in evaluation approaches that seek to compartmentalize judgments of quality from judgments of safety, as well as ignore the potential relationship between the two. Annotators may not reasonably be expected to distinguish safety from quality, meaning task design changes would be needed in the analysis of labels rather than in task instructions or label schema. As the overall output quality of generative AI increasingly improves, many quality concerns will dissipate. However, consistent quality issues with roots in underrepresentation and diminished model understanding of minoritized contexts may point to systematic oversight in safety evaluation pipelines echoing

Qadri et al. (2023)'s calls for culturally inclusive evaluation.

At a basic level, our results reveal both a need to structure safety evaluation in a way that supports annotators in reasoning through concerns salient to data requesters, as well as a need to understand unprompted concerns that arise for annotators. Each of these needs influences the values encoded in data. We note a parallel to Denton et al. (2020), who argue that the values encoded in datasets are shaped by their contexts of creation. More narrowly we are focused on annotation task design, which Amironesei and Díaz (2024) similarly highlight is shaped by a range of implicit and explicit design decisions. Annotators bring their own implicit sets of values, reflected through their reasoning, which may not be aligned with how data requesters design a given task, which our qualitative analysis of annotator comments helps to elucidate. At the same time, prompt-image pairs as well as other content that annotators may be tasked with evaluating can invoke reasoning that may diverge from what data requesters expect. For example, we found that specific language in prompts can evoke moral reasoning in annotator comments. This observation aligns with challenges previously highlighted in related fields like content moderation, where studies show annotators often surface ethical, social, and political concerns that exceed the scope of formal annotation guidelines (Shahid and Vashistha 2023; Udupa, Maronikolakis, and Wisiorek 2023; Tobi 2024; Jhaver, Frey, and Zhang 2023). This tendency underscores that even when task instructions aim for neutrality or lack overt moral framing, annotators frequently engage their own moral sentiments and bring judgments extending beyond the explicit task requirements into their evaluations. A failure to align task design with annotator reasoning may result in implicit value tensions that produce inconsistent labels.

Safety to Whom?

Because annotators can be more certain of their own harm sensitivity compared with a broader group of unnamed others, it is not surprising that labels differed for assessments of harm-to-self compared with harm-to-others. However, disparities in this difference across annotator groups suggest that annotators make judgments with varying ideas of others' sensitivity to harm. For example, the score delta among Black annotators was nearly a third that of White annotators. Importantly, when annotating for harm to others, annotators are typically not asked to specify whom these "others" represent, allowing for a wide range of implicit interpretations—from a general public to specific demographics or even imagined communities. While past work on safety annotation has disaggregated safety ratings according to who the specific targets of harm are in objectionable content (e.g., Weidinger et al. 2024), safety evaluation tasks do not typically disaggregate between annotators' perceptions of harm as individuals compared with their perceptions of harm on behalf of others. Instead, annotators are implicitly treated as a representative proxy of others. However, it is well acknowledged that curating a representative pool of data annotators is exceedingly nontrivial and often infeasible (Díaz et al. 2022b). This means that it is necessary to rely on annotators' ratings as indicators of both their own safety and the safety of others, raising questions about how to approach asking annotators to make judgments on behalf of others.

A key insight from research on annotators' lived experiences is that sociocultural experiences can enable individuals to identify socially significant meaning in content that may be unrecognized by others, for example in recognizing details related to carceral imagery and gang violence in text and images (e.g., Patton et al. 2019). Indeed, we observed differences both in absolute ratings of harm and in the difference between harm-to-self and harm-to-others ratings across demographic groups. Interestingly however, we found that demographic subgroups followed similar annotation patterns regardless of whether the content they were annotating depicted an aspect of their own identity. For example, the gap in harm-to-self and harm-to-others scores among women annotators was identical whether or not women were the subject of the content they rated. In other words, although our diverse pool of annotators showed systematic differences in rating behavior, our results did not indicate that this was due to specific sensitivity to their own identity representation.

At the same time, our findings are not intended to indicate that diverse annotator pools and sociocultural experience do not provide valuable insight. Demographic subgroups' systematically different rating patterns still suggest that social experiences may inform broader reasoning differences that influence annotators' sensitivity both to harms more generally, as well as others' potential sensitivity to harm. More broadly, the annotation behaviors we observed point to needed grounding or clarity for annotators regarding *who* they should be considering when making harmfulness judgments. From our results, it is not clear if annotators from different demographic subgroups were considering different "others" or if they made different assumptions about those same groups' sensitivity to harm. Additionally, we identified "antipatterns" in rating wherein groups that were generally more sensitive to harmfulness annotated *lower* harmfulness than their annotator peers. For example, for one prompt-image pair in which the n-word was invoked, all non-Black annotators labeled the resulting image as unsafe while all Black annotators labeled the image as safe. This reflects prior findings on the role of sociocultural experience in annotation (Patton et al. 2019).

The insights we uncovered regarding identity and harm perceptions are a direct result of the way that we operationalized safety to explicitly consider the annotator's stance in contrast to the stance of others. We intentionally took this approach to better understand whether and how each annotator's social position shapes the individual harm evaluations they provide. Our comment analysis, in particular, enabled us to identify and expand the set of concerns captured in the task design, exposing reasoning that can be missed in typical annotation task structures for AI evaluation. While our focus was on annotator behavior rather than evaluating the safety of any particular system, the relation between safety and *who* is at risk is an important contextual component of safety judgments that is left as an implicit assumption for annotators to make. As Wang et al. (2024) point out, evaluations of model safety in industry settings include considerations about legal liability and compliance. While these certainly play a role in broader societal safety, the most immediate entity at risk for such violations can vary between a specific subset of vulnerable end users, politicians who may be subject to disinformation campaigns, or even the company whose model is being evaluated and which may be at risk of legal ramifications for violating laws. These are all necessary facets of model evaluation but join the mix of safety considerations without necessarily being focused on considerations of safety to users. As a result, annotators must implicitly switch between various potential targets of harm without clarity on how to consider different targets.

Recommendations for Safety Evaluation Tasks

Balancing Structure and Subjectivity Thematic analysis revealed a tension between structured evaluation frameworks and the subjective nature of harm perception. Annotators often filled in narrative gaps, projected context onto ambiguous prompts, and expressed emotional reactions that were not captured by predefined harm categories. While structure is essential for standardization, task designs must also leave room for subjective interpretations and open-ended feedback. A "good" task design might focus on balancing these elements by including space for contextual reasoning alongside structured responses. Although open-ended responses add effort to tasks, they introduce space for capturing contextual reasoning which can be used selectively or in piloting to validate task design. However, it remains unclear how much evaluators can or should intervene in the natural reasoning processes of annotators when determining final task label categories.

Integrating Safety and Item Quality Evaluation An important finding is the entanglement of safety judgments with item quality, particularly in cases involving image distortions or generative artifacts. Annotators often interpreted distortions as indicative of harm, demonstrating that technical flaws influence safety perceptions. By explicitly evaluating safety and quality in tandem, task designs can help disentangle these dimensions and provide a clearer understanding of their interplay. For example, evaluators might introduce prompts that ask annotators to separately assess image distortion and its perceived impact on harm, enabling a more nuanced analysis of these factors.

Expanding the Scope of Safety Evaluation Frameworks Our findings suggest that existing frameworks need to account for subjective and contextual dimensions, such as emotional reactions, implicit judgments, and cultural interpretations of harm. Annotators' frequent use of emotional language and their divergence from predefined harm labels highlight gaps in current evaluation practices. Expanding annotation guidelines to include illustrative examples of diverse cultural and emotional interpretations can help address these gaps. Additionally, distinguishing between "self" and "other" harm explicitly within task designs could provide annotators with clearer evaluative criteria while capturing the multifaceted nature of safety judgments. Importantly, our aim is not to critique policy-based safety evaluations, which serve essential regulatory and design purposes, but to propose additional signals that complement them. Annotator comments and disagreement patterns offer insight into misalignments between policy-based frameworks and user expectations or cultural and emotional nuance. In doing so, they help identify gaps between abstract policies and the lived experiences or intuitive judgments of real users.

What Lies Within vs. Beyond the Scope of Evaluators One of the key challenges highlighted by this study is delineating what lies within the control of task design versus what remains outside the hands of evaluators. While evaluators can modify label sets, task instructions, and evaluation frameworks, they have limited influence over the subjective ways in which annotators interpret and fill in missing context. The process by which annotators reason about ambiguous images often reflects their personal, cultural, and emotional perspectives, which are difficult to scaffold or standardize. This raises important questions about how much evaluators should intervene in these reasoning processes and what constitutes an effective and ethical task design.

Limitations

Our study's depth is inherently limited by its reliance on annotator comments rather than direct interaction. While these comments offer valuable insights into annotators' reasoning, they do not permit real-time probing or follow-up questions, preventing a deeper understanding of the intricate links between their judgments and expressed thoughts. This contrasts with more intensive qualitative methods like interviews or observations, which could further unpack nuanced decision-making processes. Additionally, our analysis of demographic influences on safety judgments was restricted to the diverse annotators, as we did not collect similar demographic information for the expert annotators. This absence prevents us from fully disentangling whether observed differences in safety ratings stem from distinct expert policy knowledge or from varied cultural and lived experiences.

Conclusion

In this work, we explored how diverse annotators interpret multimodal safety and identified gaps in typical safety task design. Our qualitative analysis of annotator comments on adversarial T2I prompt-image pairs revealed that their judgments often include considerations not explicitly captured by standard safety categories, such as the link between **image quality** and **safety perceptions**. We also found that annotators consistently perceive content as more harmful to *others* than to *themselves*, yet their understanding of "others" and their sensitivity to harm varies significantly. Furthermore, **moral reasoning** evident in annotators' comments proved predictive of harm scores and areas of disagreement. Based on these insights, we recommend improving safety evaluation tasks by explicitly addressing subjective, contextual, and cultural factors, distinguishing between self and other harm, integrating safety and quality assessments, expanding guidelines to include diverse interpretations, and exploring methods to better scaffold annotators' reasoning while respecting their interpretative processes.

Ethical Statement

While the institution of the lead authors does not require IRB approval, our research strictly adhered to ethical guidelines, obtaining rigorous ethical reviews and approvals from our dedicated internal ethics committee before any data collection began and once more after finalizing this paper. Participants in the study were fully briefed on the study's scope and potential risks, and their consent was obtained, with the clear understanding that participation was entirely voluntary and could be discontinued at any time.

We also note that related research on annotation has been exempted from IRB review at academic institutions (e.g., Díaz et al. 2018) or published at prominent ML venues without a stated IRB review (e.g., Aroyo et al. 2023). It is important to note that the ethical considerations for this type of annotation task extend beyond standard user experience (UX) research consent processes. While typically UX research might involve less formal review, the nature of safety evaluations for generative AI content often involves sensitive topics and potential for harm, necessitating more rigorous ethical oversight. Our internal ethics committee's review and approval, therefore, provided an essential layer of scrutiny appropriate for the challenges posed by this work.

References

Amironesei, R.; and Díaz, M. 2024. Social and Ethical Norms in Annotation Task Design. *IEEE Transactions on Technology and Society*.

Anil, R.; Dai, A. M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Aroyo, L.; Taylor, A. S.; Díaz, M.; Homan, C. M.; Parrish, A.; Serapio-Garcia, G.; Prabhakaran, V.; and Wang, D. 2023. DICES Dataset: Diversity in Conversational AI Evaluation for Safety. *Advances in Neural Information Processing Systems*, 36: 53330–53342.

Aroyo, L.; and Welty, C. 2014. The Three Sides of CrowdTruth. *Human Computation*, 1(1).

Atari, M.; Haidt, J.; Graham, J.; Koleva, S.; Stevens, S. T.; and Dehghani, M. 2023. Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, 125.

Basile, V.; Cabitza, F.; Campagner, A.; and Fell, M. 2021. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. arXiv:2109.04270.

Braun, V.; and Clarke, V. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health*, 11(4): 589–597.

Braun, V.; and Clarke, V. 2021. Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Counselling and psychotherapy research*, 21(1): 37–47.

Davani, A. M.; Díaz, M.; Baker, D.; and Prabhakaran, V. 2023. Disentangling Disagreements on Offensiveness: A Cross-Cultural Study. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Denton, R.; Hanna, A.; Amironesei, R.; Smart, A.; Nicole, H.; and Scheuerman, M. K. 2020. Bringing the people back in: Contesting benchmark machine learning datasets. *arXiv preprint arXiv:2007.07399*.

Díaz, M.; Amironesei, R.; Weidinger, L.; and Gabriel, I. 2022a. Accounting for offensive speech as a practice of resistance. In *Proceedings of the sixth workshop on online abuse and harms (woah)*, 192–202.

Díaz, M.; Johnson, I.; Lazar, A.; Piper, A. M.; and Gergle, D. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*, 1–14.

Díaz, M.; Kivlichan, I.; Rosen, R.; Baker, D.; Amironesei, R.; Prabhakaran, V.; and Denton, E. 2022b. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2342–2351.

Dinan, E.; Abercrombie, G.; Bergman, S. A.; Spruit, S.; Hovy, D.; Boureau, Y.-L.; Rieser, V.; et al. 2022. SafetyKit: First aid for measuring safety in open-domain conversational systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Dumitrache, A.; Aroyo, L.; and Welty, C. 2018. Capturing Ambiguity in Crowdsourcing Frame Disambiguation. arXiv:1805.00270.

Gadiraju, V.; Kane, S.; Dev, S.; Taylor, A.; Wang, D.; Denton, E.; and Brewer, R. 2023. " I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 205–216.

Graham, J.; Haidt, J.; Koleva, S.; Motyl, M.; Iyer, R.; Wojcik, S. P.; and Ditto, P. H. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, 55–130. Elsevier.

Graham, J.; Haidt, J.; and Nosek, B. A. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5): 1029.

Grey, M.; and Segerie, C.-R. 2025. Safety by Measurement: A Systematic Literature Review of AI Safety Evaluation Methods. *arXiv preprint arXiv:2505.05541*.

Haidt, J.; and Joseph, C. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4): 55–66.

Jhaver, S.; Frey, S.; and Zhang, A. X. 2023. Decentralizing Platform Power: A Design Space of Multi-level Governance in Online Social Platforms. *Social Media+ Society*, 9(4): 20563051231207857.

Kennedy, B.; Golazizian, P.; Trager, J.; Atari, M.; Hoover, J.; Mostafazadeh Davani, A.; and Dehghani, M. 2023. The (moral) language of hate. *PNAS nexus*, 2(7): pgad210.

Miceli, M.; Schuessler, M.; and Yang, T. 2020. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2): 1–25.

Mostafazadeh Davani, A.; Diaz, M.; Baker, D. K.; and Prabhakaran, V. 2024. D3CODE: Disentangling Disagreements in Data across Cultures on Offensiveness Detection and Evaluation. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 18511–18526. Miami, Florida, USA: Association for Computational Linguistics.

Patton, D.; Blandfort, P.; Frey, W.; Gaskell, M.; and Karaman, S. 2019. Annotating social media data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2142–2151.

Qadri, R.; Shelby, R.; Bennett, C. L.; and Denton, E. 2023. AI's regimes of representation: A community-centered study of text-to-image models in South Asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 506–517.

Quaye, J.; Parrish, A.; Inel, O.; Rastogi, C.; Kirk, H. R.; Kahng, M.; Van Liemt, E.; Bartolo, M.; Tsang, J.; White, J.; et al. 2024. Adversarial Nibbler: An open red-teaming method for identifying diverse harms in text-to-image generation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 388–406.

Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Comanescu, R.; Akbulut, C.; Stepleton, T.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; et al. 2024. Gaps in the Safety Evaluation of Generative AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1200–1217.

Santy, S.; Liang, J.; Le Bras, R.; Reinecke, K.; and Sap, M. 2023. NLPositionality: Characterizing Design Biases of Datasets and Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9080–9102. Toronto, Canada: Association for Computational Linguistics.

Sap, M.; Swayamdipta, S.; Vianna, L.; Zhou, X.; Choi, Y.; and Smith, N. A. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5884–5906.

Shahid, F.; and Vashistha, A. 2023. Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony? In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18.

Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Rostamzadeh, N.; Nicholas, P.; Yilla-Akbari, N.; Gallegos, J.; Smart, A.; Garcia, E.; et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 723–741.

Smart, A.; Wang, D.; Monk, E.; Díaz, M.; Kasirzadeh, A.; Van Liemt, E.; and Schmer-Galunder, S. 2024. Discipline and label: A weird genealogy and social theory of data annotation. *arXiv preprint arXiv:2402.06811*.

Sorensen, T.; Mishra, P.; Patel, R.; Tessler, M. H.; Bakker, M. A.; Evans, G.; Gabriel, I.; Goodman, N. D.; and Rieser, V. 2025. Value Profiles for Encoding Human Variation. *CoRR*, abs/2503.15484.

Thoppilan, R.; Freitas, D. D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; Li, Y.; Lee, H.; Zheng, H. S.; Ghafouri, A.; Menegali, M.; Huang, Y.; Krikun, M.; Lepikhin, D.; Qin, J.; Chen, D.; Xu, Y.; Chen, Z.; Roberts, A.; Bosma, M.; Zhao, V.; Zhou, Y.; Chang, C.-C.; Krivokon, I.; Rusch, W.; Pickett, M.; Srinivasan, P.; Man, L.; Meier-Hellstern, K.; Morris, M. R.; Doshi, T.; Santos, R. D.; Duke, T.; Soraker, J.; Zevenbergen, B.; Prabhakaran, V.; Díaz, M.; Hutchinson, B.; Olson, K.; Molina, A.; Hoffman-John, E.; Lee, J.; Aroyo, L.; Rajakumar, R.; Butryna, A.; Lamm, M.; Kuzmina, V.; Fenton, J.; Cohen, A.; Bernstein, R.; Kurzweil, R.; Aguera-Arcas, B.; Cui, C.; Croak, M.; Chi, E.; and Le, Q. 2022. LaMDA: Language Models for Dialog Applications. arXiv:2201.08239.

Tobi, A. 2024. Towards an Epistemic Compass for Online Content Moderation. *Philosophy & Technology*, 37(3): 109.

Trager, J.; Ziabari, A. S.; Davani, A. M.; Golazizian, P.; Karimi-Malekabadi, F.; Omrani, A.; Li, Z.; Kennedy, B.; Reimer, N. K.; Reyes, M.; et al. 2022. The moral foundations reddit corpus. *arXiv preprint arXiv:2208.05545*.

Udupa, S.; Maronikolakis, A.; and Wisiorek, A. 2023. Ethical scaling for content moderation: Extreme speech and the (in) significance of artificial intelligence. *Big Data & Society*, 10(1): 20539517231172424.

Wang, D.; Díaz, M.; Parrish, A.; Aroyo, L.; Homan, C.; Serapio-García, G.; Prabhakaran, V.; and Taylor, A. S. 2024. A case for moving beyond "gold data" in AI safety evaluation. In *Proceedings of the workshop on Human-centered Evaluation and Auditing of Language Models (HEAL) at CHI*, volume 2024.

Waseem, Z. 2016. Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the first workshop on NLP and computational social science*, 138–142.

Weidinger, L.; Mellor, J.; Pegueroles, B. G.; Marchal, N.; Kumar, R.; Lum, K.; Akbulut, C.; Diaz, M.; Bergman, S.; Rodriguez, M.; et al. 2024. STAR: SocioTechnical approach to red teaming language models. *arXiv preprint arXiv:2406.11757*.

Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; et al. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. *arXiv preprint arXiv:2310.11986*.

Woodruff, A.; Shelby, R.; Kelley, P. G.; Rousso-Schindler, S.; Smith-Loud, J.; and Wilcox, L. 2024. How knowledge workers think generative ai will (not) transform their industries. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–26.

Xu, J.; Ju, D.; Li, M.; Boureau, Y.-L.; Weston, J.; and Dinan, E. 2021. Recipes for Safety in Open-domain Chatbots. arXiv:2010.07079.