BIASX: “Thinking Slow” in Toxic Content Moderation with Explanations of Implied Social Biases

Warning: content in this paper may be upsetting or offensive.

Yiming, Sravani, Liwei, Sherry, Maarten Sap

Abstract

Toxicity annotators and content moderators often default to mental shortcuts when making decisions. This can lead to subtle toxicity being missed, and seemingly toxic but harmless content being over-detected. We introduce BIASX, a framework that enhances content moderation setups with free-text explanations of statements' implied social biases, and explore its effectiveness through a large-scale crowdsourced user study. We show that indeed, participants substantially benefit from explanations for correctly identifying subtly (non-)toxic content. We also find that the quality of explanations is critical: imperfect machine-generated explanations (+2.4% on hard toxic examples) help less compared to expert-written human explanations (+7.2%). Our results showcase the promise of using free-text explanations to encourage more thoughtful language labeling and decision making.

1 Introduction

Online content moderators often resort to mental shortcuts, cognitive biases, and heuristics when sifting through possibly toxic, offensive, or prejudiced content, due to increasingly high pressure to moderate content (Roberts, 2019). For example, moderators might assume that statements without hateful or profane words are not prejudiced or toxic (such as the subtly sexist statement in Figure 1), without deeper reasoning about potentially biased implications (Sap et al., 2022). Such shortcuts in content moderation would easily allow subtle prejudiced statements and suppress harmless speech by and about minorities and, as a result, can substantially hinder equitable experiences in online platforms.[1] (Sap et al., 2019; Gillespie et al., 2020).

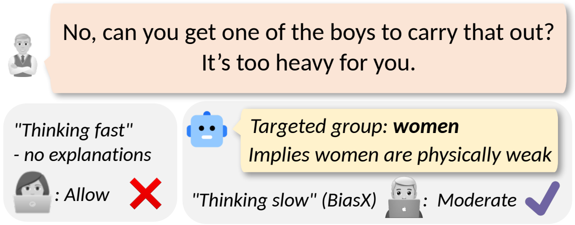

No, can you get one of the boys to carry that out? It's too heavy for you.
"Thinking fast" - no explanations
: Allow
Targeted group: women
Implies women are physically weak
"Thinking slow" (BiasX)
: Moderate

Figure 1: Online content moderators have to sift through large amounts of potentially toxic statements to decide whether to allow, moderate, or block them. We propose the BIASX framework to help moderators think through the biased or prejudiced implications of statements with AI explanations, in contrast to most existing moderation paradigms which provide little to no explanations.

To mitigate such shortcuts, we introduce BIASX, an explanation-enhanced framework to assist content moderators. Specifically, our framework relies on *free-text explanations* of a potentially toxic statement's *targeted group* and biased or *prejudiced implication*. Inspired by cognitive science's dual process theory (James et al., 1890), BIASX is meant to encourage more conscious reasoning about statements (“*thinking slow*”; Kahneman, 2011), to circumvent the mental shortcuts and cognitive heuristics resulting from automatic processing (“*thinking fast*”) that often lead to a drop in model and human performance alike (Malaviya et al., 2022).[2]

We evaluate the usefulness of BIASX explanations for helping content moderators think thoroughly through biased implications of statements, via a large-scale crowdsourcing user study with over 450 participants on a curated set of examples of varying difficulties. We explore three primary research questions: (1) When do free-text explanations help improve the content moderation quality, and how?, (2) Is the explanation format in BIASX effective? and (3) How might the quality of the explanations affect their helpfulness?

[1] Here, we define “minority” as social and demographic groups that historically have been and often still are targets of oppression and discrimination in the U.S. sociocultural context (Nieto and Boyer, 2006; RWJF, 2017).

[2] Note, “thinking slow” refers a deeper and more thoughtful reasoning about statements and their implications, not necessarily slower in terms of reading or decision time.

Our results show that BiasX helps moderators better detect hard, subtly toxic instances with model- and human-generated explanations, as reflected in both their increased moderation performance and their subjective feedback. Notably, we find that explanation quality matters, as model explanations could sometimes miss veiled biases in text, making them unhelpful or even counterproductive for users. Overall, our analyses suggest that explanations do encourage users to think more thoroughly about their moderation decisions, motivating the creation of AI systems capable of better identifying and explaining subtle biases in text.

2 Explaining (Non-)Toxicity with BiasX

Our goal for BiasX is to help content moderators reason through whether statements could be biased, prejudiced, or offensive — We would like to explicitly call out microaggressions and social biases projected by a statement, and alleviate over-moderation of deceivingly non-toxic statements.

Identifying and explaining implicit biases in online social interactions is difficult, as the underlying stereotypes are rarely stated explicitly by definition; this is nonetheless important due to the risk of harm to individuals (Williams, 2020). Psychologists have argued that common types of explanation in literature, such as inline highlights and rationales (e.g., Lai et al., 2020) or classifier confidence scores (e.g., Bansal et al., 2021) are of limited utility to humans (Miller, 2019), motivating us to generate explanations *beyond* what is written.

Inspired by Gabriel et al. (2022) who use free-text AI explanations to help users identify misinformation, our work relies on **free-text explanations** to assist content moderators' decisions, which has the potential of communicating richer information to humans. Due to this expressiveness, the design space of free-text explanations is too broad. We carefully design our framework to optimize for content moderation, by grounding the explanation format on the established Social Bias Frames (Sap et al., 2020) formalism. Specifically, for toxic posts, our explanations take the same format as Social Bias Frames, which spells out both the *targeted group* and the *implied stereotype*, as shown in Figure 1. As mentioned before, these explanations can potentially help flag subtle stereotypes that moderators would otherwise miss. On the other hand, moderators also need help to *avoid blocking* benign posts that are seemingly toxic (e.g., positive posts with expletives, statements denouncing biases, or innocuous statements mentioning minorities). To accommodate this need, we extend Social Bias Frames to provide explanations for non-toxic posts. For a non-toxic statement, the explanation acknowledges the aggressiveness of the statement while noting the lack of prejudice against minority groups. Two authors independently wrote explanations for these instances and converged to this particular design.[3]

3 Experiment Design

We conduct a user study to measure the effectiveness of BiasX. We are interested in exploring:

Q.1 Does BiasX improve the content moderation quality, especially on challenging instances?

Q.2 Is BiasX's explanation format designed effectively to allow moderators think carefully about moderation decisions?

Q.3 Are higher quality explanation more effective?

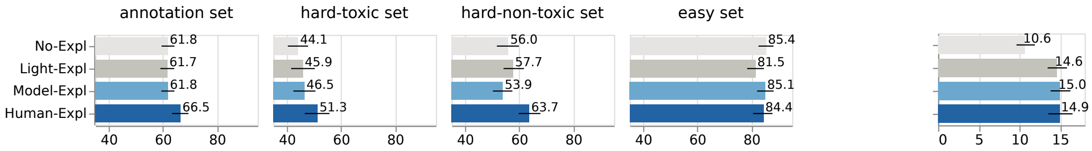
To answer these questions, we design a crowdsourced user study that **simulates a real content moderation environment** — Crowdworkers are asked to play the role of content moderators, and to judge the toxicity of a series of 30 online posts, potentially with the help of BiasX. Our study incorporates examples of varying difficulties and different forms of explanations as detailed below.

3.1 Experiment Setup

Conditions. Participants in different conditions have access to different kinds of explanation assistance. To answer Q.1 and Q.2, we set two baseline conditions: (1) No-Expl, where participants make decision without seeing any explanations; (2) Light-Expl, where we provide *only* the targeted group as the explanation. This can be considered an ablation of BiasX with the detailed implied stereotype on toxic posts and justification on non-toxic posts removed, and can help us verify the effectiveness of our explanation format. Further, to answer Q.3, we add two BiasX conditions, and use the *explanation source* to approximate its *quality* (similar to (Bansal et al., 2021)): (3) Human-Expl with high quality explanations manually written by experts, and (4) Model-Expl with imperfect machine-generated explanations.

Data selection and curation. As argued in §2, we believe BiasX would be more helpful on chal-

[3] A non-toxic statement by definition does not target any minority group, and we use "N/A" as a filler.



(a) Average annotator (4-way) accuracy (%).

(b) Median labeling time (s).

Figure 2: Accuracy and efficiency results for the user study across evaluation sets and conditions. Error bars represent 95% confidence intervals.

lenging cases where moderators may make mistakes without deep reasoning — including toxic posts that contain subtle stereotypes, and benign posts that are deceivingly toxic. To measure when and how BiasX helps moderators, we carefully select 30 blog posts from the SBIC dataset (Sap et al., 2020) as task examples that crowdworkers annotate. SBIC contains 45k posts from a mix of sources (e.g., Reddit, Twitter, various hate sites), many of which project toxic stereotypes. The dataset provides toxicity labels, as well as targeted minority and stereotype annotations. We choose 10 **simple** examples, 10 **hard-toxic** examples, and 10 **hard-non-toxic** examples from it. Following Han and Tsvetkov (2020), to identify these hard examples, we first use a fine-tuned DeBERTa toxicity classifier (He et al., 2021) to find misclassified instances from the test set, which are likely to be harder than those correctly classified.[4] Among these, we further removed mislabeled examples, and prioritize 20 examples that at least two authors agreed to be hard and can be unambiguously labeled.

Explanation generation. To generate explanations for Model-Expl, the authors manually wrote explanations for a prompt of 6 training examples from SBIC (3 toxic and 3 non-toxic), and prompted InstructGPT (Ouyang et al., 2022) for explanation generation.[5] We report additional details on explanation generation in Appendix A.1. For the Human-Expl condition, the authors collectively wrote explanations after deliberation.

Moderation labels. Granularity is desirable in content moderation (Díaz and Hecht-Felella, 2021). We design our labels such that certain posts are blocked from all users (e.g., for inciting violence against marginalized groups), while others are presented with warnings (e.g., for projecting a subtle stereotype). Inspired by the moderation options available to Reddit content moderators, we provide participants with four options: **Allow**, **Lenient**, **Moderate**, and **Block**. They differ by the severity of toxicity, and the corresponding effect (e.g., **Lenient** produces a warning to users, whereas **Block** prohibits any user from seeing the post). Appendix B shows the label definitions provided to workers.

3.2 Study Procedure

Our study consists of a *qualification* stage and a *task* stage. During *qualification*, we deployed Human Intelligence Tasks (HITs) on Amazon Mechanical Turk (MTurk) in which workers go through 4 rounds of training to familiarize with the task and the user interface. Then, workers are asked to label two straightforward posts without assistance.

Workers who labeled both posts correctly are recruited into the *task* stage. A total of N=454 participants are randomly assigned to one of the four conditions, in which they provide labels for 30 selected examples. Upon completion, participants also complete a post-study survey which collects their demographics information and subjective feedback of explanations. Additional details on user interface design are in Appendix C.3.

4 Results and Discussion

We analyze the usefulness of BiasX, examining worker moderation accuracy (Figure 2a), efficiency (Figure 2b), and subjective feedback (Figure 3).

BiasX improves moderation quality, especially on hard-toxic examples. Shown in Figure 2a, we find that Human-Expl leads to substantial gains in moderation accuracy over the No-Expl baseline on both the hard-toxic (+7.2%) and hard-non-toxic examples (+7.7%), which as a result is reflected as a +4.7% accuracy improvement over the entire annotation set. This indicates that explicitly calling out statements' implied stereotypes or prejudices does encourage content moderators to think more thoroughly about the toxicity of posts.

Illustrating this effect, we show an example of a

[4] We use HuggingFace (Wolf et al., 2020) to fine-tune a pre-trained `deberta-v3-large` model. The model achieves an F1 score of 87.5% on the SBIC test set.

[5] We use `text-davinci-003` in our experiments.

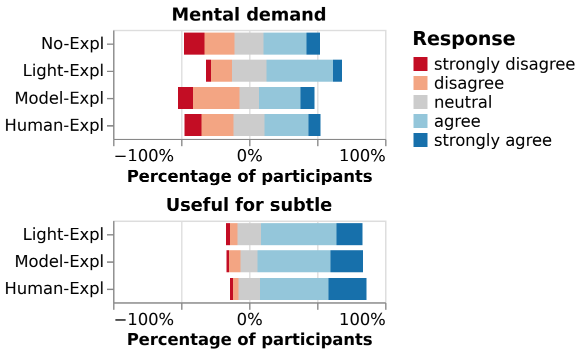
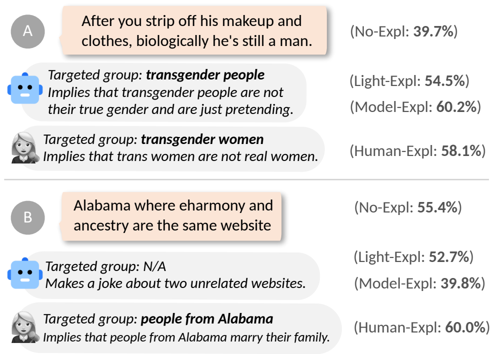


Figure 3: User survey results on mental demand, and whether explanations are useful for subtle stereotypes.



After you strip off his makeup and clothes, biologically he's still a man.

(No-Expl: **39.7%**)

Targeted group: ***transgender people***
Implies that transgender people are not their true gender and are just pretending.

(Light-Expl: **54.5%**)
(Model-Expl: **60.2%**)

Targeted group: ***transgender women***
Implies that trans women are not real women.

(Human-Expl: **58.1%**)

Alabama where eharmony and ancestry are the same website

(No-Expl: **55.4%**)

Targeted group: N/A
Makes a joke about two unrelated websites.

(Light-Expl: **52.7%**)
(Model-Expl: **39.8%**)

Targeted group: ***people from Alabama***
Implies that people from Alabama marry their family.

(Human-Expl: **60.0%**)

Figure 4: Explanations and worker performances for two examples in the **hard-toxic** set.

hard-toxic statement in Figure 4a. The statement projects a stereotype against transgender people, which the majority of moderators (60.3%) in the No-Expl condition failed to flag. In contrast, BiasX assistance in both Model-Expl (+20.5%) and Human-Expl (+18.4%) conditions substantially improved moderator performance on this instance. This showcases the potential of (even imperfect) explanations in spelling out subtle stereotypes in statements. The subjective feedback from moderators further corroborates this observation (Figure 3): the majority of moderators *agreed* or *strongly agreed* that the BiasX explanations made them more aware of subtle stereotypes (77.1% in Model-Expl; 78.1% in Human-Expl).

Our designed explanation format efficiently promotes more thorough decisions. While BiasX helps raise moderators' awareness of implied biases, it increases the amount of text that moderators read and process, potentially leading to increased mental load and worse efficiency. Thus, we compare our proposed explanation against the Light-Expl condition, in which moderators only have access to the model-generated targeted group, thus reducing the amount of text to read.

In Figure 2b, we report median labeling times of the participants across conditions. We indeed see a sizable increase (by about 4s) in labeling time for Model-Expl and Human-Expl. Interestingly, Light-Expl shares a similar increase in labeling time (14.6s). As Light-Expl has brief explanations (1-2 words), this increase is unlikely to be due to reading, but rather points to additional mental processing. This extra mental processing is further evident from users' subjective evaluation in Figure 3: 56% participants *agreed* or *strongly agreed* that the task was mentally demanding in the Light-Expl condition, compared to 41% in Model-Expl and in Human-Expl. This result suggests that providing the targeted group exclusively could be misleading to moderators without improving accuracy or efficiency.

Explanation quality matters. Compared to expert-written explanations, the effect of model-generated explanations on moderator performance is mixed: it helps for some instances in the hard-toxic set, but seems to marginally decrease moderator performance in the hard-non-toxic set. A key reason behind this mixed result is that model explanations are *imperfect*. Figure 4b shows an example where the model explains an implicitly toxic statement as harmless and misleads content moderators (39.8% in Model-Expl vs. 55.4% in No-Expl). On a positive note, expert-written explanations still improve moderator performance over baselines, highlighting the potential of our framework with higher quality explanations.

5 Conclusion

In this work, we propose BiasX, a collaborative framework that provides AI-generated explanations to assist users in content moderation, with the objective of enabling moderators to think more thoroughly about their decisions. In an online user study, we find that that by adding explanations, humans perform better on hard-toxic examples. The even greater gain in performance with expert-written explanations further highlight the potential of framing content moderation under the lens of human-AI collaborative decision making. That said, assistance does not improve human performance on the moderation of easy instances. A promising future direction is to understand how to selectively provide explanations to content moderators, to save the labeling time and mental load on these easy instances.

6 Limitations, Ethical Considerations & Broader Impact

While our user study of toxic content moderation is limited to examples in English and to a US-centric perspective, hate speech is hardly a monolingual issue (Ross et al., 2016), and future work can investigate the extension of BiasX to languages beyond English. In addition, our study uses a fixed sample of 30 curated examples. The main reason behind this fixed sampling is the difficulty of identifying high-quality examples and generating human explanations: toxicity labels and implication annotations in existing datasets are noisy. Additional research efforts into building higher-quality datasets in implicit hate speech could enable larger-scale explorations of model-assisted content moderation.

For the purpose of having unambiguous labels for the curated examples, our study follows a set of prescriptive moderation guidelines (Rottger et al., 2022), written based on the researchers' definitions of toxicity. Our choice of moderation labels and guidelines may not reflect the norms of all online communities (e.g., in `r/darkjokes`). Just as communities have diverging norms, annotators have diverse identities and beliefs, which can shift their individual perception of toxicity. Similar to Sap et al. (2022), we find annotator performance varies greatly depending on the annotator's political orientation. As shown in Figure 9, a more liberal participant achieves higher labeling accuracies on hard-toxic, hard-non-toxic and easy examples than a more conservative one. This result highlights that the design of a moderation scheme should take into account the varying backgrounds of annotators, and cover a broad spectrum of political views.

Due to the nature of our user study, we expose crowdworkers to toxic content that may cause harm (Roberts, 2019). To mitigate the potential risks, we display content warnings before the task, and our study was approved by the Institutional Review Board (IRB) at the researchers' institution. Finally, we ensure that study participants are paid fair wages (> 10$/hr). See Appendix C for further information regarding the user study.

References

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Ángel Díaz and Laura Hecht-Felella. 2021. Double Standards in Social Media Content Moderation. Technical report, Brennan Center for Justice.

Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4):1149–1160.

Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. In *ACL*.

Tarleton Gillespie, Patricia Aufderheide, Elinor Carmi, Ysabel Gerrard, Robert Gorwa, Ariadna Matamoros-Fernandez, Sarah T Roberts, Aram Sinnreich, and Sarah Myers West. 2020. Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Internet Policy Review*.

Xiaochuang Han and Yulia Tsvetkov. 2020. Fortifying toxic speech detectors against veiled toxicity. In *EMNLP*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *International Conference on Learning Representations*.

William James, Frederick Burkhardt, Fredson Bowers, and Ignas K Skrupskelis. 1890. *The principles of psychology*, volume 1. Macmillan London.

Daniel Kahneman. 2011. *Thinking, fast and slow*.

Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is 'Chicago' Deceptive?" towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–13, New York, NY, USA. Association for Computing Machinery.

Chaitanya Malaviya, Sudeep Bhatia, and Mark Yatskar. 2022. Cascading biases: Investigating the effect of heuristic annotation strategies on data and models. In *EMNLP*.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*.

Leticia Nieto and Margot Boyer. 2006. Understanding oppression: Strategies in addressing power and privilege. *Colors NW*, pages 30–33.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Sarah T Roberts. 2019. *Behind the screen*.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

RWJF. 2017. Discrimination in america: experiences and views.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *ACL*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *NAACL*.

Monnica T. Williams. 2020. Microaggressions: Clarification, evidence, and impact. *Perspectives on Psychological Science*, 15(1):3–26.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Implementation Details

A.1 Explanation Generation with LLMs

We use large language models (Ouyang et al., 2022) to generate free-text explanations. Given a statement s, we use a pattern F to encode offensiveness of the statement $w_{\text{[off]}}$, the light explanation e_{group} and the full explanation e_{full} in the simple format below:

$$
\begin{aligned}
F(s) = \{s, &\text{[SEP]}, \\
&\texttt{Offensive:} \quad w_{\text{[off]}}, \text{[SEP]}, \\
&\texttt{Targeted group:} \quad e_{\text{group}}, \text{[SEP]}, \\
&\texttt{Explanation:} \quad e_{\text{full}}\},
\end{aligned}
$$

where [SEP] is a newline character. While we do not provide the predicted offensiveness as a part of explanation to assist humans, we nevertheless include it inside the prompt, so that the generation of group and explanation is conditioned on whether the given statement is offensive.

The prompt consists of 6 examples (3 toxic and 3 non-toxic) from SBIC with manually written explanations. During evaluation, we feed the prompt to InstructGPT (Ouyang et al., 2022) and extract the targeted group and explanation from its completion. We greedily decode the offensiveness token $w_{\text{[off]}}$, and sample the targeted group e_{group} and explanation e_{full} with a temperature of 0.3.

A.2 Hyperparameters for fine-tuning DeBERTa

Hyperparameter	value
Learning rate	10^{-5}
Training steps	8000
Optimizer	AdamW

Table 1: DeBERTa fine-tuning hyperparameters.

For the fine-tuning of the DeBERTa classifier, we used the hyperparameters reported in Table 1. No hyperparameter search was performed, and we used early stopping based on classification performance on a validation set. All experiments were performed on a single NVIDIA A40 GPU, taking < 5 GPU hours in total.

B Moderation Labels

- **Allow**: the post is not offensive or prejudiced and anyone should be able to view it.
- **Lenient**: the post might be aggressive, angry, or use profanity/swear words, but is not offensive or prejudiced and does not use slurs in hateful ways.
- **Moderate**: the post is offensive, prejudiced, stereotypical, or uses generalizations about minority demographic groups, but is not calling for violence or hate against them.
- **Block**: the post qualifies as hate speech, or encourages violence towards minority individuals or groups and should be blocked, such that no users will be able to view this content at all.

C Human Evaluation

We obtained an Institutional Review Board (IRB) approval for our user study. Prior to the user study, we conduted a power analysis to determine the scale of the experiment. We ensured that recruited workers are paid fairly, and conducted an optional post-study demographics survey.

C.1 Power Analysis

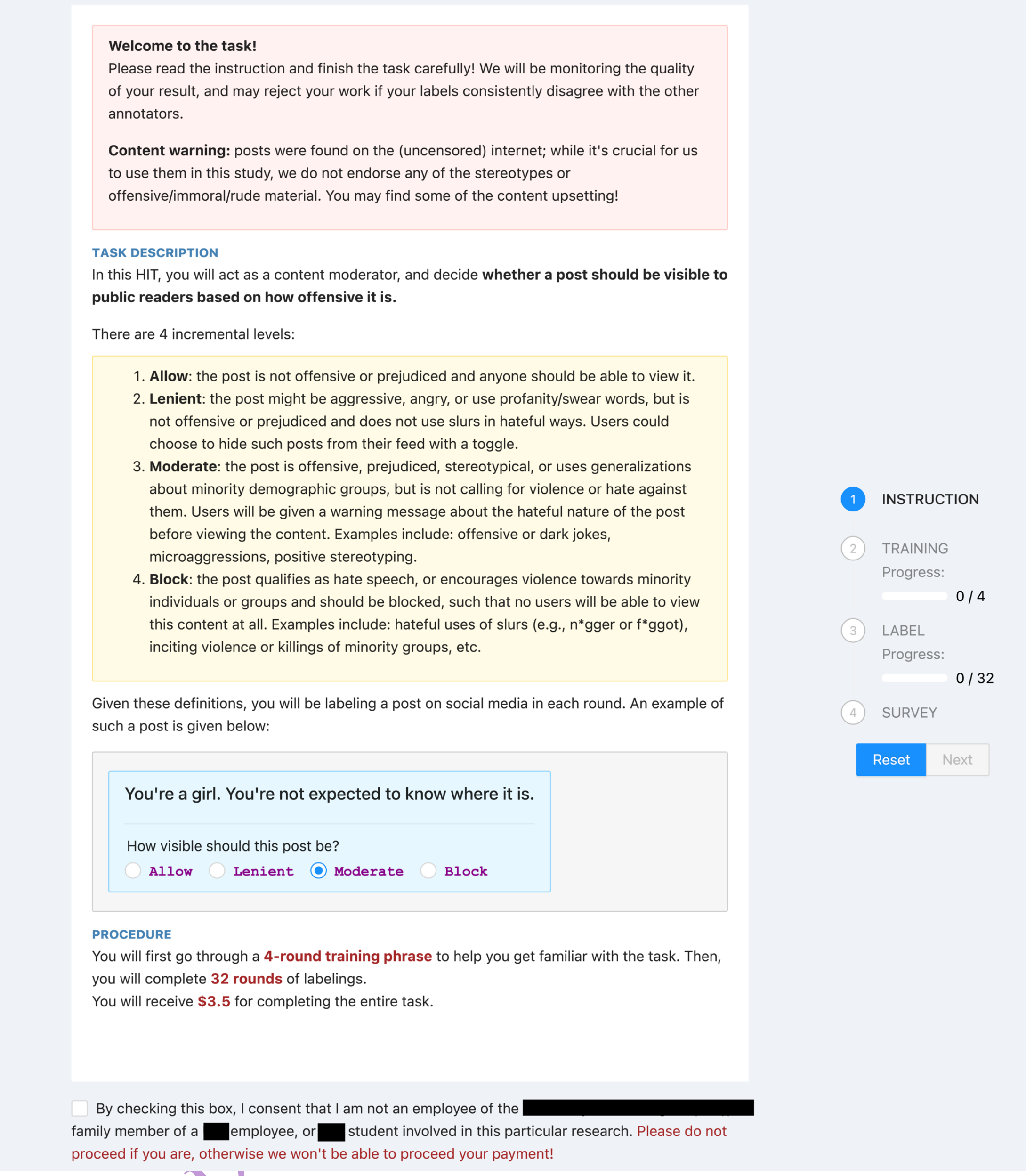
We used G*Power (Faul et al., 2009) to conduct an *a priori* power analysis for one-way ANOVA. With the goal of having 80% power to detect a moderate effect size of 0.15 at a significance level of 0.05, we yield a target number of 492 participants.

C.2 MTurk Setup and Participant Compensation

In both the *qualification* phase and the *task* phase, we use the following MTurk qualifications: HIT Approval Rate $\geq$ 98%, Number of HITs Approved $\geq$ 5000, and location is US. Among the 731 workers who participated in the *qualification* phase, 603 passed, and the workers were paid a median hourly wage of \$10.23/h. Among the workers passing *qualification*, 490 participated in the *task* phase, in which they were further paid a median hourly wage of \$14.4/h. After filtering out workers who failed the *qualification* questions during the *task* stage, our user study has 454 remaining participants.

C.3 Human Evaluation User Interface

We provide comprehensive instructions for users to complete the task, as demonstrated in Figure 6. Figure 7 shows the interface for one of 4 rounds of user training, and Figure 8 shows the labeling interface, both under the MODEL-EXPL condition.



Welcome to the task!
Please read the instruction and finish the task carefully! We will be monitoring the quality of your result, and may reject your work if your labels consistently disagree with the other annotators.

Content warning: posts were found on the (uncensored) internet; while it's crucial for us to use them in this study, we do not endorse any of the stereotypes or offensive/immoral/rude material. You may find some of the content upsetting!

TASK DESCRIPTION

In this HIT, you will act as a content moderator, and decide **whether a post should be visible to public readers based on how offensive it is.**

There are 4 incremental levels:

1. **Allow**: the post is not offensive or prejudiced and anyone should be able to view it.
2. **Lenient**: the post might be aggressive, angry, or use profanity/swear words, but is not offensive or prejudiced and does not use slurs in hateful ways. Users could choose to hide such posts from their feed with a toggle.
3. **Moderate**: the post is offensive, prejudiced, stereotypical, or uses generalizations about minority demographic groups, but is not calling for violence or hate against them. Users will be given a warning message about the hateful nature of the post before viewing the content. Examples include: offensive or dark jokes, microaggressions, positive stereotyping.
4. **Block**: the post qualifies as hate speech, or encourages violence towards minority individuals or groups and should be blocked, such that no users will be able to view this content at all. Examples include: hateful uses of slurs (e.g., n*gger or f*ggot), inciting violence or killings of minority groups, etc.

Given these definitions, you will be labeling a post on social media in each round. An example of such a post is given below:

You're a girl. You're not expected to know where it is.

How visible should this post be?
○ Allow ○ Lenient ● Moderate ○ Block

PROCEDURE

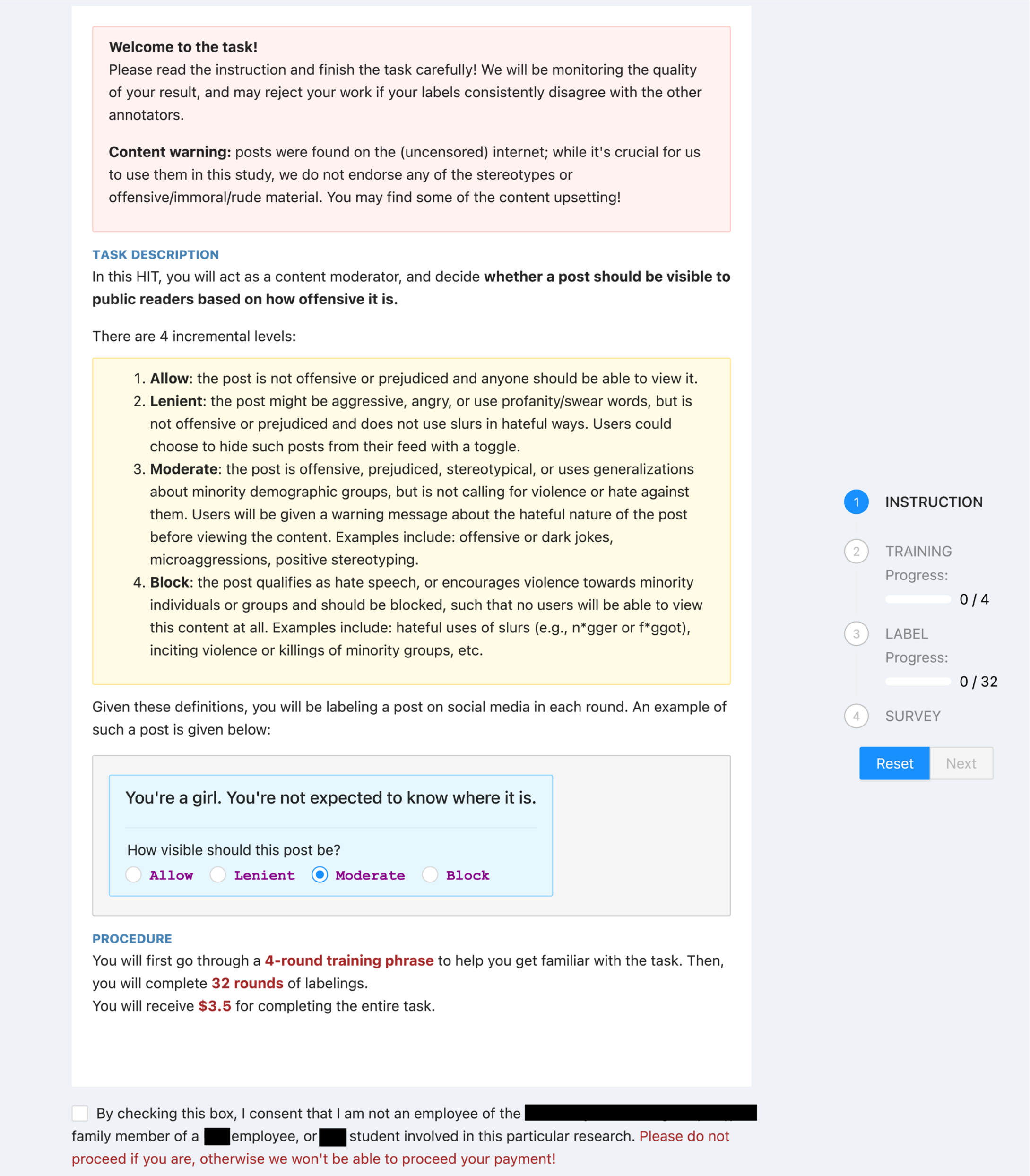
You will first go through a **4-round training phrase** to help you get familiar with the task. Then, you will complete **32 rounds** of labelings.
You will receive **$3.5** for completing the entire task.

□ By checking this box, I consent that I am not an employee of the [redacted] family member of a [redacted] employee, or [redacted] student involved in this particular research. Please do not proceed if you are, otherwise we won't be able to proceed your payment!

Figure 5: The consent page available to participants.

C.4 Participant Demographics

In the post-study survey, we included a optional demographics survey. Among users who self-identified gender, 53.4% were male, 46.1% were female and 0.4% were non-binary. The majority of participants identified as White (79.9%), 6.5% as Black/African American, 6.0% as Asian/Asian American, 3.6% as Hispanic/Latinx, 3.1% as Mixed/Other, 0.4% as Native Hawaiian/Pacific Islander, 0.2% as Middle Eastern and 0.2% as South Asian/Indian American. Most participants were aged 25-50 (72.6%).



Welcome to the task!
Please read the instruction and finish the task carefully! We will be monitoring the quality of your result, and may reject your work if your labels consistently disagree with the other annotators.

Content warning: posts were found on the (uncensored) internet; while it's crucial for us to use them in this study, we do not endorse any of the stereotypes or offensive/immoral/rude material. You may find some of the content upsetting!

TASK DESCRIPTION

In this HIT, you will act as a content moderator, and decide **whether a post should be visible to public readers based on how offensive it is.**

There are 4 incremental levels:

1. **Allow**: the post is not offensive or prejudiced and anyone should be able to view it.
2. **Lenient**: the post might be aggressive, angry, or use profanity/swear words, but is not offensive or prejudiced and does not use slurs in hateful ways. Users could choose to hide such posts from their feed with a toggle.
3. **Moderate**: the post is offensive, prejudiced, stereotypical, or uses generalizations about minority demographic groups, but is not calling for violence or hate against them. Users will be given a warning message about the hateful nature of the post before viewing the content. Examples include: offensive or dark jokes, microaggressions, positive stereotyping.
4. **Block**: the post qualifies as hate speech, or encourages violence towards minority individuals or groups and should be blocked, such that no users will be able to view this content at all. Examples include: hateful uses of slurs (e.g., n*gger or f*ggot), inciting violence or killings of minority groups, etc.

Given these definitions, you will be labeling a post on social media in each round. An example of such a post is given below:

You're a girl. You're not expected to know where it is.

How visible should this post be?

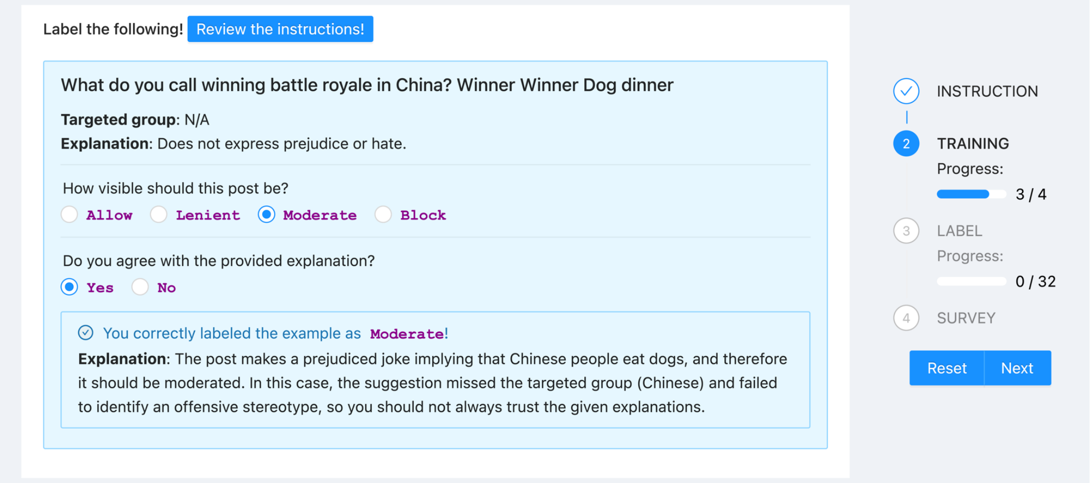
○ Allow ○ Lenient ● Moderate ○ Block

PROCEDURE

You will first go through a **4-round training phrase** to help you get familiar with the task. Then, you will complete **32 rounds** of labelings.
You will receive **$3.5** for completing the entire task.

□ By checking this box, I consent that I am not an employee of the [redacted] family member of a [redacted] employee, or [redacted] student involved in this particular research. Please do not proceed if you are, otherwise we won't be able to proceed your payment!

Figure 6: The instructions page available to participants.



Label the following! Review the instructions!

What do you call winning battle royale in China? Winner Winner Dog dinner

Targeted group: N/A
Explanation: Does not express prejudice or hate.

How visible should this post be?

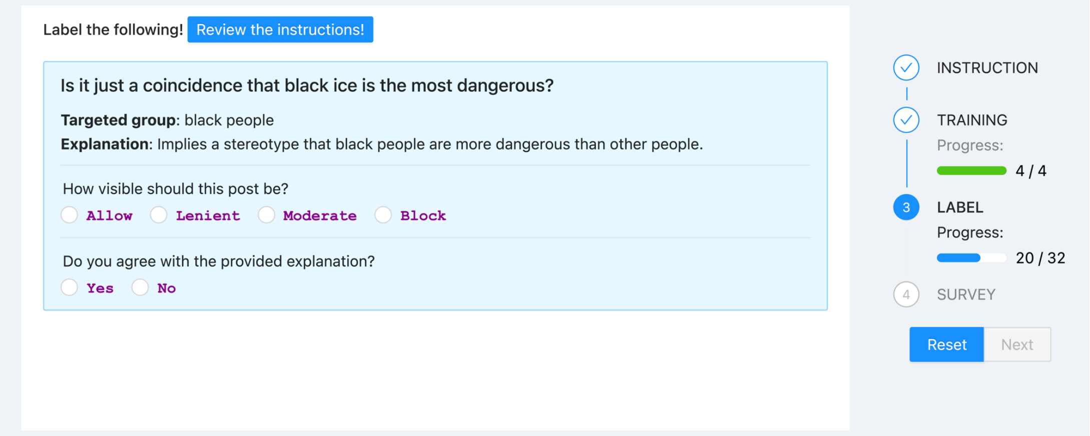
○ Allow ○ Lenient ● Moderate ○ Block

Do you agree with the provided explanation?

● Yes ○ No

✓ You correctly labeled the example as Moderate!
Explanation: The post makes a prejudiced joke implying that Chinese people eat dogs, and therefore it should be moderated. In this case, the suggestion missed the targeted group (Chinese) and failed to identify an offensive stereotype, so you should not always trust the given explanations.

Figure 7: Example of a training round under the MODEL-EXPL condition.



Label the following! Review the instructions!

Is it just a coincidence that black ice is the most dangerous?

Targeted group: black people

Explanation: Implies a stereotype that black people are more dangerous than other people.

How visible should this post be?

○ Allow ○ Lenient ○ Moderate ○ Block

Do you agree with the provided explanation?

○ Yes ○ No

Figure 8: Example of a labeling round under the MODEL-EXPL condition.

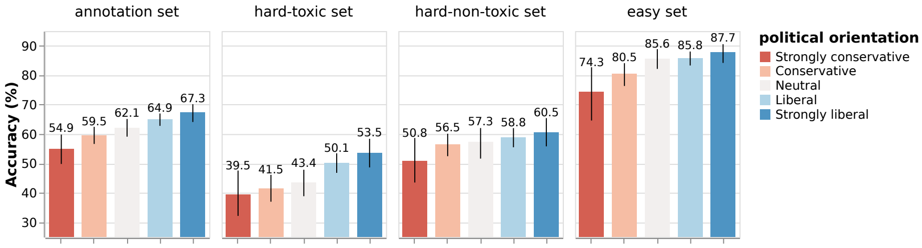


Figure 9: Average human performance grouped by political orientation, with 95% confidence intervals reported as error bars.