Explainability and Hate Speech: Structured Explanations Make Social Media Moderators Faster

Agostina Calabrese[1*] **Leonardo Neves**[2] **Neil Shah**[2] **Maarten W. Bos**[2] **Björn Ross**[1]
Mirella Lapata[1] **Francesco Barbieri**[2]
School of Informatics, University of Edinburgh[1] Snap Inc.[2]
a.calabrese@ed.ac.uk

Abstract

Content moderators play a key role in keeping the conversation on social media healthy. While the high volume of content they need to judge represents a bottleneck to the moderation pipeline, no studies have explored how models could support them to make faster decisions. There is, by now, a vast body of research into detecting hate speech, sometimes explicitly motivated by a desire to help improve content moderation, but published research using real content moderators is scarce. In this work we investigate the effect of explanations on the speed of real-world moderators. Our experiments show that while generic explanations do not affect their speed and are often ignored, structured explanations lower moderators' decision making time by 7.4%.

1 Introduction

Social media provide a platform for free expression but users may abuse it and post content in violation of terms, like misinformation or hate speech. To fight these behaviours and enforce integrity on the platform, social media companies define policies that describe what content is allowed. Posts are then monitored through automatic systems that look for policy violations. While content that has been flagged by the system with high confidence is immediately removed, all other violations, including the ones reported by users, are *moderated* by trained *human* reviewers. These moderators are also responsible for reviewing user appeals and deciding when content has been flagged incorrectly. Therefore, a big challenge with enforcing integrity is the high volume of content that needs to pass the moderators' judgment (Halevy et al., 2022).

Previous work has claimed that moderators can be supported with explanations of why posts violate the policy (Calabrese et al., 2022; Nguyen et al., 2023). But while there have been studies showing the importance of explanations for users (Haimson et al., 2021; Brunk et al., 2019), the benefits of explanations for moderators have not been studied. Can explanations help moderators judge a post faster? And how much room for improvement is there? While social media share safety reports with statistics about the number and types of detected violations[1], data relative to moderator performance is not publicly available. Explanations might have a larger impact on the performance of crowdworkers who have only recently been trained on a policy, but smaller effects would be expected on the speed of moderators who know the policy by heart.

In this paper we conduct a study with *professional* moderators from an online social platform to answer the following research questions:

1. Do explanations make moderators faster?
2. Does the type of explanations matter?
3. Do moderators want explanations?

While online social platforms deal with several integrity issues, academic research has focused on a few specific ones. Hate speech is one of the most studied issues, and (English) hate speech is also the focus of our study. Our experiments show that despite their already impressive performance, structured explanations (that highlight which parts of a post are harmful and why) can make *experienced* moderators faster by 1.34s/post without any loss in accuracy. Considering that they spend an average of 18.14s/post, that is a time reduction of 7.4%, which is a meaningful improvement considering the scale at which online social platforms operate. Generic (pre-defined) explanations on the other hand have no impact.

An online survey further revealed that moderators strongly prefer structured explanations (84%).

[*] This work was done while the author was an intern at Snap Inc.

[1] e.g., https://about.fb.com/news/2023/05/metas-q1-2023-security-reports

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 398–408

In the case of generic explanations, most moderators admit to only looking at them when in doubt (80%) or ignoring them completely (12%).

2 Related Work

While some researchers have looked at hate speech[2] as a subjective matter (Davani et al., 2022; Basile et al., 2021), this paradigm is not suitable for the use case of content moderation, where a single decision has to be made for each post (Röttger et al., 2022). In this work we follow a prescriptive paradigm, and assume the existence of a ground truth that is determined by a policy.

Explainability is a key open problem for Natural Language Processing research on hate speech (Mishra et al., 2019; Mathew et al., 2021). Well documented model failures (Sap et al., 2019; Calabrese et al., 2021), together with EU regulations on algorithmic transparency (Brunk et al., 2019), call for the design of more transparent algorithms. However, the benefits of explainability on the moderators have been understudied. Wang et al. (2023) analysed the effect of explanations on annotators, observing that wrong explanations might dangerously convince the annotators to change their mind about whether a post contains hate speech. However, the experiment was run with crowdworkers and Abercrombie et al. (2023) has found that is not uncommon for non-professional moderators to change their opinion about the toxicity of a post over time, even when no additional information is provided. To the best of our knowledge, we are the first to explore how explanations can affect moderation speed of professional moderators although the need to support them with their unmanageable workload is well-documented[3].

3 Explainable Abuse Detection

We hypothesise that different types of explanations might lead to different results. Mishra et al. (2019) argue that explanations should at least indicate 1) the intent of the user, 2) the words that constitute abuse, and 3) who is the target. From a computational perspective, the cheapest way to achieve this goal is to define the task as multiple multi-class classification problems (Kirk et al., 2023; Saeidi et al., 2021; Vidgen et al., 2021b; Ousidhoum et al., 2019), where models choose between some predefined target groups (e.g., women, lgbt+) and types of abuse (e.g., threats, derogation). While the explanations provided by these approaches are limited to properties 1 and 3, some approaches have expanded the paradigm to also include rationales (i.e., spans of text from the post that suggest why a post is hateful) and satisfy property 2 (Vidgen et al., 2021a; Mathew et al., 2021). When dealing with implicit hate, where evidence cannot always be found in the exact words of a post, rationales have been replaced with free-text implied statements (ElSherief et al., 2021; Sap et al., 2020). Calabrese et al. (2022) introduce a more structured approach to explainability, where target, intent, and type of abuse are all indicated by means of *tagged* spans from the post. The popularity of prompt-based approaches has led to the generation of free-text explanations (Wang et al., 2023), with no guarantee that any of the above properties are satisfied.

4 Experimental Design

In this study we analyse the effect explanations have on the speed of professional moderators from an online social platform with millions of users. We use the term "generic" to describe explanations that can be obtained from a multi-class classification model. For instance, for the post *"immigrants are parasites"*[4], a generic explanation could be *"Content targeting a person or group of people on the basis of their protected characteristic(s) with dehumanising speech in the form of comparisons, generalisations or unqualified behavioural statements to or about insects"*[5]. This pre-defined explanation illustrates why the post violates the policy without reference to specific post content. "Structured" explanations are instead specific to the post, and indicate why a post violates the policy by highlighting relevant spans and specifying how they relate to the policy. In the framework introduced in Calabrese et al. (2022), the example above would be associated with a parse tree where *"immigrants"* is tagged as target and protected characteristic, and *"are parasites"* as dehumanising comparison. Our hypothesis is that structured explanations will help

[2]"Abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender or sexual orientation" (Warner and Hirschberg, 2012).

[3]e.g., https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/?sh=524ab91354d0 and https://www.wired.co.uk/article/facebook-content-moderators-ireland

[4]Example taken from Calabrese et al. (2022).

[5]https://transparency.fb.com/en-gb/policies/community-standards/hate-speech

Figure 1: Annotation interface for setting 2 (post+label), where moderators are shown a post and a description of the rule it is deemed to violate. We intentionally chose a generic policy paragraph for this example as we are not allowed to share the content of the internal policies.

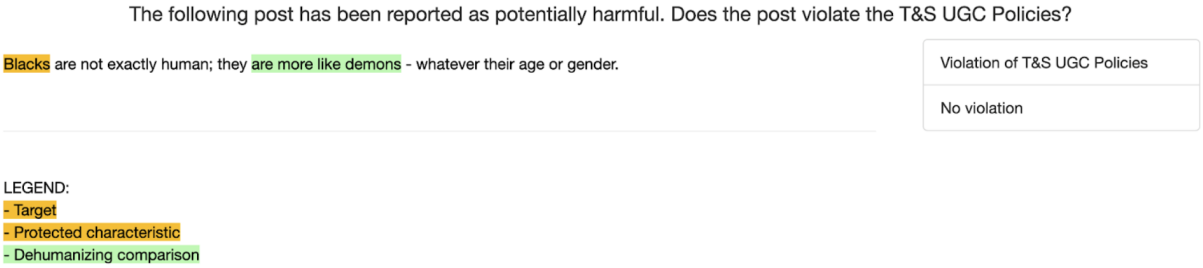

Figure 2: Annotation interface for setting 3 (post+tags), where moderators are shown the post with tagged spans as in Calabrese et al. (2022).

moderators judge posts faster, while generic explanations will not impact their speed. To verify our hypothesis, we asked 25 moderators to judge posts in three settings where they were shown: 1) only the post (**post-only**); 2) the post and the policy rule being violated (**post+policy**, which we refer to as generic explanations, Figure 1) (Kirk et al., 2023); 3) the post with tagged spans as in Calabrese et al. (2022) (**post+tags**, that is, structured explanations, Figure 2).

4.1 Data

For our experiment we used the PLEAD dataset (Calabrese et al., 2022). PLEAD contains 3,535 hateful and not-hateful posts annotated with the user intent (e.g., dehumanisation) and explanations in the form of parse trees. We include more details about PLEAD in Appendix A.1.

While there exist models that can generate structured explanations, the best model available in the literature achieved a production F1-score of 52.96% (Calabrese et al., 2022). We argue that using generated explanations in our study would bias the results. If the model gives wrong explanations half the time, then that prevents us from measuring how useful correct explanations are, or what "type" of explanations is most useful. In light of this, we used gold explanations from the PLEAD dataset.

Since moderators would normally check posts that are "at risk", we reproduced their usual task by mostly sampling hateful posts. However, to keep the experiment realistic, we simulated some model errors: in each of the three settings we included posts that do not violate the policy (10%); posts that violate the policy but are shown together with wrong explanations (10%); the remaining posts are hateful (80%) and associated with the explanations from the dataset. While the simulated model accuracy is high, with 80% correct explanations and 90% correct predictions, we feared that trivial errors would still push the moderators towards ignoring the explanations (Dietvorst et al., 2015). To mitigate this issue, we first used heuristics to generate better explanations and then manually reviewed and edited the modified explanations (Appendix A.3). We sampled a batch of 100 posts for a pilot study and three batches of 800 posts for the final experiment, one for each setting. The distribution of the intents in each setting is the same as in PLEAD.

4.2 Method

We recruited 25 moderators from Snapchat, an online social platform with millions of users. All moderators had experience reviewing posts with

abusive language (as the platform policies are wider and contain many more phenomena) and posts that only contain text (as most moderators at the platform usually deal with multimodal content). We recognise that different levels of moderators experience might lead to different results. None of our moderators were new hires. Furthermore, we used mixed-effects models to analyse our results as a way to take into account different levels of experience and therefore "baseline" speed.

We asked moderators to annotate 2,400 posts, 800 for each setting, thus preventing moderators from encountering the same post twice and bias speed measurements. The order in which the settings were shown to moderators was randomised. Some moderators received setting 1 first, others received setting 2 first, etc. Each setting was shown as the first setting roughly the same number of times (respectively 8, 8 and 9). Each block of 800 posts was used for each setting a third of the time. This means that the observed results do not depend on the specific posts that occur in a block, because all blocks were used for all the settings. Posts within the same setting were also randomised, and shown to moderators in batches of 20 examples, one per page, on an internal annotation platform.

Moderators did not undertake any training for this task. We asked them to judge whether a post violated the policy, underlining not to judge whether the explanation was correct. We also informed them that annotation times were being recorded. Finally, we provided moderators with one example for each scenario, to illustrate what the annotation interface would look like. We ran a pilot study with one moderator to assess the clarity of the interface and the soundness of our mapping of PLEAD annotations onto internal policy rules (Appendix A.2). Details of the pilot can be found in Appendix A.4.

4.3 Evaluation Metrics

The annotation platform allowed us to record the timestamps at which posts were shown to moderators and when they moved to the next post, so for each post we stored the number of seconds it took to express a judgment. We also report moderator accuracy but do not expect an improvement from showing explanations, since these are professional moderators with a high degree of accuracy. Note also the limitation in accuracy measurements as this involves comparing the decisions of professional moderators – who are regarded by online social platforms to be the ground truth – against crowdsourced annotations.

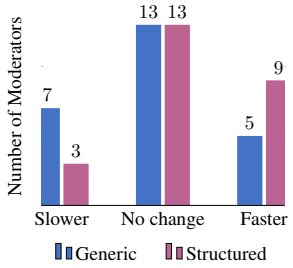

Figure 3: Effect of generic and structured explanations on the speed of each moderator (*No change*: $|z| < 2$).

5 Do Explanations Help Moderators?

Before analysing speed, we discarded the first 20 instances (0.025%) from each setting. We did this to provide a buffer to the moderators to adapt to a new setting and corresponding interface. Additionally we discarded for each moderator all data points with annotation time more than three standard deviations away from the moderator mean[6]. When moderators were prompted only with the post, the fastest and slowest moderators achieved a mean annotation speed of, respectively, 6.58s/post and 45.03s/post. To study the effect of generic and structured explanations on annotation time (*time*) while taking into account individual differences we fitted two linear mixed effects models to the data from *post-only* and *post+policy* or *post+tags*, respectively. We defined the two models as follows:

$$\text{time} \sim \text{length} + (1|\text{moderator})$$

$$\text{time} \sim \text{setting} + \text{length} + (1|\text{moderator})$$

where *length* is the length of the post, *setting* indicates whether the moderator was provided an explanation or not, and $(1|\text{moderator})$ accounts for individual differences of the moderators. We tested whether the explanations have a significant effect by testing whether the difference between the likelihood of these two models is significant using ANOVA. We found that in setting *post+policy* explanations did not affect the annotation time: the estimated effect is 0.02 ± 0.32 s, and is not significant ($\chi^2(1) = 0.005$, $p = .94$). When using structured explanations (*post+tags*) the estimated effect is -1.34 ± 0.32 s and is highly significant ($\chi^2(1) = 17.808$, $p < .001$), showing that moderators are faster with appropriate explanations.

[6]The number of outliers was comparable across settings.

We used a z-test to compare individual performances across the settings (Figure 3). When shown generic explanations 52% of the moderators registered no significant change in speed (w.r.t. setting 1), 28% had a significant loss in performance, and only 20% improved. With structured explanations instead, 36% of the moderators had a significant improvement, 52% of the moderators registered no significant change, and 12% performed worse than without explanations[7]. We examined whether the different impact that explanations had on moderators was due to the experimental design by testing for correlations between said impact and the order in which the settings were shown to the moderators. With structured explanations, *all* moderators who registered a loss in performance were shown this setting first and the Pearson correlation between the impact (represented as -1 for loss, 0 for no change, and 1 for improvement) and the round in which setting 3 was shown is .66 ($p < .001$). However, the same trend was not observed for generic explanations. Moderators who registered a loss in performance were shown *post+policy* as either first or *last*, and the correlation score is .41 ($p = .04$) (Appendix B). We hypothesise that the posts from PLEAD might have been very different in language and topics from the ones moderators usually review, and therefore annotations in the first batch required moderators some extra adjustment time (regardless of the setting). However, the different trends observed for *post+policy* and *post+tags* demonstrate that the improvement recorded with structured explanations is not only related to the experimental design. Moreover, *post+tags* is the setting that was shown as first 1 time more than the other settings (9 instead of 8), and 2 of the corresponding 9 moderators still registered a significant improvement.

We did not observe any correlation between the impact of explanations and the specific sample of 800 posts that was selected for each setting (-.06 for setting 2 and .09 for setting 3) (Appendix C).

Finally, we looked at accuracy to ensure that faster annotation did not come at the price of more mistakes. In *post-only*, the highest and lowest recorded accuracy scores were 92.13% and 73.13%. We compared the accuracy of moderators across scenarios with a z-test between the accuracy of all moderators in setting 1 and 2 or 3. For both generic and structured explanations we did not observe a significant change ($z < 2$), not even when measuring accuracy only on not-hateful posts or hateful posts with wrong explanations (Appendix D).

[7]One of these three moderators declared in the follow-on survey to have ignored the explanations.

6 Do Moderators Want Explanations?

After the experiment was over, we asked the 25 moderators to complete a brief survey. A strong preference was expressed for the setting with structured explanations (84%), while 8% had no preference and 8% preferred generic explanations (Appendix E). When prompted with generic explanations, only 8% of the moderators consistently took them into account, while 80% only looked at the explanations when in doubt and 12% ignored them. The picture changes for structured explanations, where 60% of the moderators used them consistently, 32% looked at them when in doubt, and 8% ignored them. 48% of the moderators declared that the posts shown in this study were different from the ones they usually moderate. They differed in the use of abbreviations, slang and jargon, but also in topics, as the policy covers many phenomena and hate speech is not the most frequent. This supports our hypothesis that moderators required some extra adjustment time in the first setting.

7 Conclusions

In this work we investigated the impact of explainable NLP models on the decision speed of social media moderators. Our experiments showed that explanations make moderators faster, but only when presented in the appropriate format. Generic explanations have no impact on decision time and are likely to be ignored, while structured explanations made moderators faster by 1.34 s/instance. A follow-on survey further revealed that moderators prefer structured explanations over generic or none. These results were obtained simulating a model accuracy of 80%, with 10% of the posts misclassified as policy violations, and 10% correctly classified but associated with wrong explanations. Such accuracy is beyond the capabilities of available models, and yet resulted in criticism from the moderators who spotted the inaccuracies. We hope this study can encourage researchers to improve abuse detection models that produce structured explanations.

8 Limitations

In this work we focused on hate speech, but there may be other content forbidden by a platform's

terms that this work did not test. We focused on textual content and limited the study to English posts. These choices were merely driven by the lack of explainable multimodal and multilingual datasets for the task of integrity, or hate speech detection. Restricting the scope to English hate speech allowed us to compare the effects of different types of explanations on the same posts. We hope that the results reported in this study can promote the collection of structured explanations for new and existing multimodal or multilingual datasets.

9 Ethical Considerations

All the annotations in this study were produced by content moderators regularly employed at an online social platform. Although the posts they were asked to judge came from a public dataset and are different in style from the ones they usually review, dealing with hate speech is part of their role and they have been trained for handling such content. No user data from said platform was used in this study, and all annotations of the public posts have been released in anonymised format[8] to protect the identity of the moderators. We did not collect personal information about the moderators to protect their privacy, as 1) we are analyzing hate speech in a prescriptive paradigm that assumes the existence of a single ground truth and therefore it makes it less relevant to consider the demographics of individual annotators; 2) it would require asking platform employees for their protected characteristics.

Acknowledgements

We would like to thank Maryna Diakonova and the 25 Snapchat moderators who participated in our study. This work was supported in part by Huawei and the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics. Lapata gratefully acknowledges the support of the UK Engineering and Physical Sciences Research Council (grant EP/W002876/1) and the European Research Council (award 681760).

[8] https://github.com/Ago3/structured_explanations_make_moderators_faster

References

Gavin Abercrombie, Dirk Hovy, and Vinodkumar Prabhakaran. 2023. Temporal and second language influence on intra-annotator agreement and stability in hate speech labelling. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 96–103.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pages 15–21. Association for Computational Linguistics.

Jens Brunk, Jana Mattern, and Dennis M. Riehle. 2019. Effect of transparency and trust on acceptance of automatic online comment moderation systems. In *21st IEEE Conference on Business Informatics, CBI 2019, Moscow, Russia, July 15-17, 2019, Volume 1 - Research Papers*, pages 429–435. IEEE.

Agostina Calabrese, Michele Bevilacqua, Björn Ross, Rocco Tripodi, and Roberto Navigli. 2021. AAA: fair evaluation for abuse detection systems wanted. In *WebSci '21: 13th ACM Web Science Conference 2021, Virtual Event, United Kingdom, June 21-25, 2021*, pages 243–252. ACM.

Agostina Calabrese, Björn Ross, and Mirella Lapata. 2022. Explainable abuse detection as intent classification and slot filling. *Trans. Assoc. Comput. Linguistics*, 10:1440–1454.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Trans. Assoc. Comput. Linguistics*, 10:92–110.

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 345–363. Association for Computational Linguistics.

Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proc. ACM Hum. Comput. Interact.*, 5(CSCW2):466:1–466:35.

Alon Y. Halevy, Cristian Canton-Ferrer, Hao Ma, Umut Ozertem, Patrick Pantel, Marzieh Saeidi, Fabrizio Silvestri, and Ves Stoyanov. 2022. Preserving integrity in online social networks. *Commun. ACM*, 65(2):92–98.

Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. Semeval-2023 task 10: Explainable detection of online sexism. In *Proceedings of the The 17th International Workshop on Semantic Evaluation, SemEval@ACL 2023, Toronto, Canada, 13-14 July 2023*, pages 2193–2210. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14867–14875. AAAI Press.

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Tackling online abuse: A survey of automated abuse detection methods. *CoRR*, abs/1908.06024.

Tin Nguyen, Jiannan Xu, Aayushi Roy, Hal Daumé III, and Marine Carpuat. 2023. Towards conceptualization of "fair explanation": Disparate impacts of anti-asian hate speech explanations on content moderators. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9696–9717. Association for Computational Linguistics.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4674–4683. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B. 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, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 175–190. Association for Computational Linguistics.

Marzieh Saeidi, Majid Yazdani, and Andreas Vlachos. 2021. Cross-policy compliance detection via question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8622–8632. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1668–1678. Association for Computational Linguistics.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5477–5490. Association for Computational Linguistics.

Bertie Vidgen, Dong Nguyen, Helen Z. Margetts, Patrícia G. C. Rossini, and Rebekah Tromble. 2021a. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2289–2303. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021b. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1667–1682. Association for Computational Linguistics.

Han Wang, Ming Shan Hee, Md. Rabiul Awal, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2023. Evaluating GPT-3 generated explanations for hateful content moderation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 6255–6263. ijcai.org.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.

A Experimental Design

A.1 PLEAD

PLEAD is an extension of the LFTW dataset (Vidgen et al., 2021b) where the hateful and not-hateful posts have been enriched with span-level annotations for the task of intent classification and slot filling. Slots represent properties like “target” and

"protected characteristic", while intents are policy rules or guidelines (e.g., "dehumanisation"). PLEAD contains 3,535 posts, 25% of which are not-hateful, while the remaining posts correspond to the intents of dehumanisation (25%), threatening (17%), derogation (28%) and support of hate crimes (5%).

A.2 Policy Adaptation

PLEAD was annotated using the codebook for hate speech annotations designed by the Alan Turing Institute (Vidgen et al., 2021b), and although everything that is labelled as hate speech in PLEAD also violates social media policies[9], the converse does not apply. Specifically, threats and harassment are not allowed by social media even when targeted at groups that are not protected. Therefore we manually reviewed all the not-hateful posts containing threats or derogatory expressions in the parse tree and labelled as policy violations all the posts in which such expressions are targeted at people. For the second setting, where posts are shown together with a description of the violated rule, we adapted the wording in the explanations to match the internal policy the moderators are familiar with.

A.3 Error Simulation

To simulate model errors we tweaked some of the parse trees from PLEAD. Not-hateful posts are labelled as such when they lack at least one tag in the parse tree to violate the policy (e.g., they do not contain a reference to a protected group) or when a span of text tagged as negative stance is present (e.g. they quote a hateful expression only to disagree with it). For the 10% of the posts that we sampled among the not-hateful ones, we either hallucinated new tagged spans, or deleted a negative stance tag. To prevent the moderators from associating obviously inaccurate explanations with the not-hateful class, we also simulated mistakes in the explanations of 10% of the hateful posts. For these instances we dropped one tagged span from the parse tree, and hallucinated a new one to keep a policy violation. We first used heuristics to generate better explanations by only selecting noun phrases when hallucinating tags like *target* and verb phrases for, e.g., *threat*. We then manually reviewed and edited the modified explanations. Examples of wrong explanations are shown in Table 1.

A.4 Pilot Study

We ran a pilot study with one of the moderators to assess the clarity of the interface and the soundness of our mapping of PLEAD annotations onto internal policy rules. We intentionally decided against asking more of the moderators to take the pilot, to avoid learning effects that could affect the final results. The pilot moderator was shown the same 100 posts in each setting, and achieved an accuracy of 93% in all of them. This suggests that the interface did not confuse the moderator into judging the coherence of the explanations instead of the posts themselves, and that the mapping between the policies was accurate. Since the posts were the same, it is not meaningful to compare the speed across the settings. The moderator started from setting 3 (posts + structured explanations) and took on average 13.11 seconds per instance. The re-annotation of the same posts in the following settings was faster, as expected.

B Effect of Settings Order

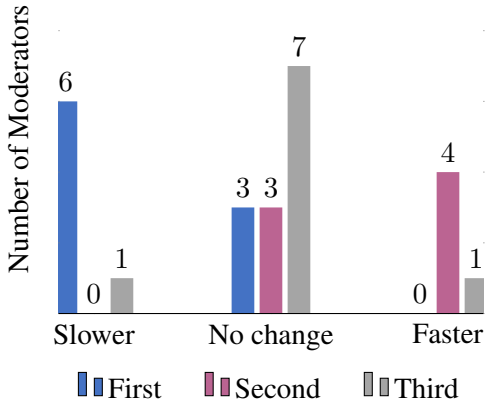

Figure 5: Effect of generic explanations on the speed of individual moderators, grouped depending on which round they were shown this setting (*No change*: $|z| < 2$).

We tested for correlations between the impact that explanations had on moderators speed and the order in which the settings were shown to the moderators. Figure 6 shows that with structured explanations, *all* moderators who registered a loss in performance were shown this setting first. However, the same trend was not observed for generic explanations, where moderators who registered a

[9] e.g., https://transparency.fb.com/policies/community-standards/ or https://values.snap.com/en-GB/privacy/transparency/community-guidelines

Original Explanation	Wrong Explanation
My agenda is buying a new car and driving everywhere I can to see the world with my trans boyfriend apart from [ProtectedCharacteristic *muslim*] countries, they are full of [Target, DehumanisingComparison *terrorists*]. And we just bought one yesterday so be careful! Agenda in motion!	My agenda is buying a new car and driving everywhere I can to see the world with [Target, ProtectedCharacteristic *my trans boyfriend*] apart from muslim countries, they are full of [DehumanisingComparison *terrorists*]. And we just bought one yesterday so be careful! Agenda in motion!
[Target, ProtectedCharacteristic *Women*] like me [DehumanisingComparison *are scum*]? [NegativeStance *Shut up*]	[Target, ProtectedCharacteristic *Women*] like me [DehumanisingComparison *are scum*]? Shut up

Table 1: Example of hateful post (first row) and not-hateful post (second row) with the original (left column) and generated wrong (right column) explanation.

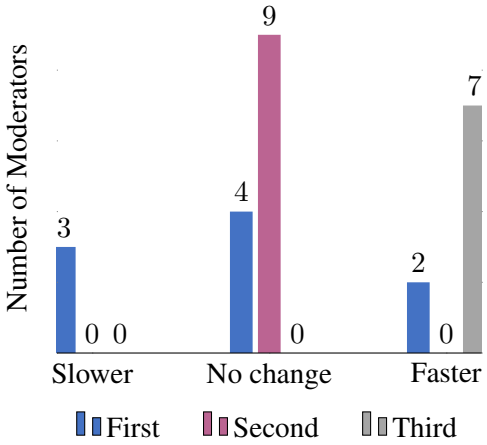

Figure 6: Effect of structured explanations on the speed of individual moderators, grouped depending on which round they were shown this setting (*No change*: $|z| < 2$).

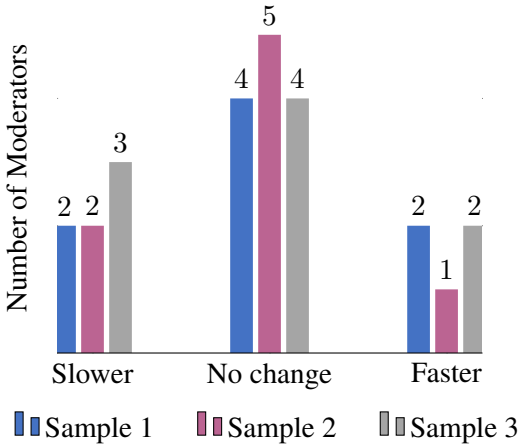

Figure 7: Effect of generic explanations on the speed of individual moderators, grouped depending on which sample of 800 posts was used for this setting (*No change*: $|z| < 2$).

loss in performance were shown *post+policy* as either first or *last* (Figure 5).

C Effect of Post Samples

We tested for correlations between the impact that explanations had on moderators speed and the specific sample of 800 posts that was selected for each setting. As Figure 7 and 8 show no clear pattern emerged, and the correlation between impact and sample was -.06 for *post+label* and .09 for *post+tags*.

D Accuracy

We compared the accuracy of moderators across scenarios with a z-test between the accuracy of all moderators in setting 1 (*post-only*) and 2 (*post+policy*) or 3 (*post+label*). For both generic and structured explanations we did not observe a significant change ($z < 2$, Figure 9), not even when measuring accuracy only on not-hateful posts (Figure 10) or hateful posts with wrong explanations (Figure 11).

E Moderators' Preference

Figure 12 summarises the moderators' preferences among the three settings. Only 8% of the moderators expressed a preference for generic explanations, and this is coherent to the level of engagement that this type of explanations registered (Figure 13). 84% of the moderators expressed a preference for the structured explanations, with only 8% who declared to have ignored the explanations during the annotation (Figure 14). The criticisms raised about these explanations concerned their accuracy and the need to sometimes still read the whole post to grasp the context in which the highlighted expressions were used. Overall moderators did not think the design of the structured explanations could be further improved to optimise their decision speed. They stressed the importance of using the explanations as a guide while still reading the posts for context, leaving no margin for improvement on this metric.

When asked what the most common reasons were for them to be unsure about how to judge a post during their regular job, they indicated slang, unknown words/symbols and the lack of cultural

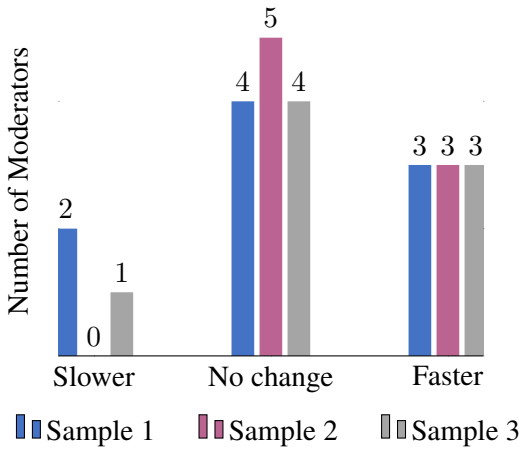

Figure 8: Effect of structured explanations on the speed of individual moderators, grouped depending on which sample of 800 posts was used for this setting (*No change*: $|z| < 2$).

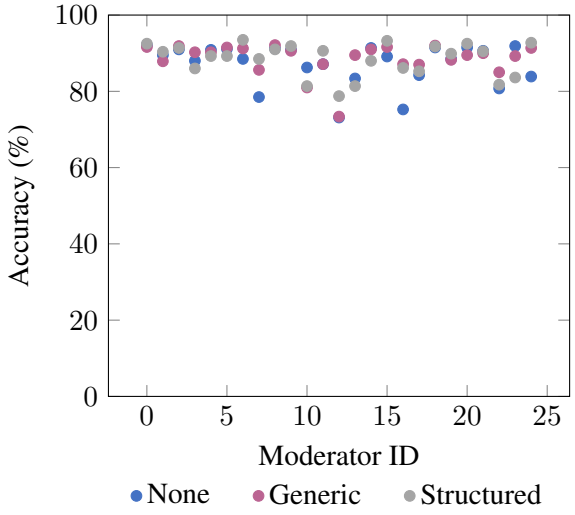

Figure 9: Accuracy score achieved by each moderator with no, generic or structured explanations on the 3 different samples of 800 posts.

context. Combining structured explanations with additional free-text explanations could be a way to support moderators when judging complex posts, improving their accuracy (but not speed).

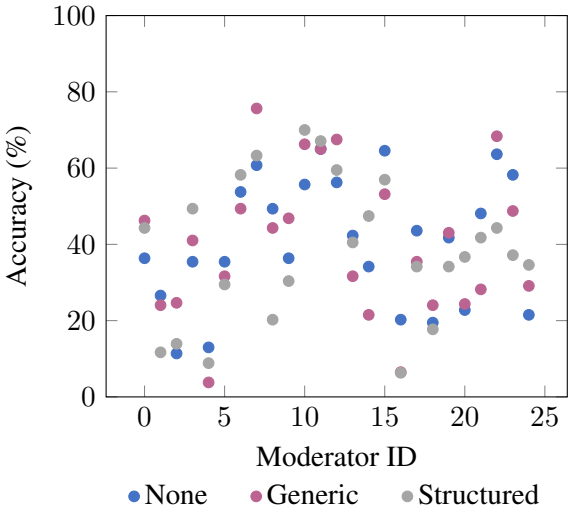

Figure 10: Accuracy score achieved by each moderator with no, generic or structured explanations on the 80 not-hateful instances of the 3 different samples.

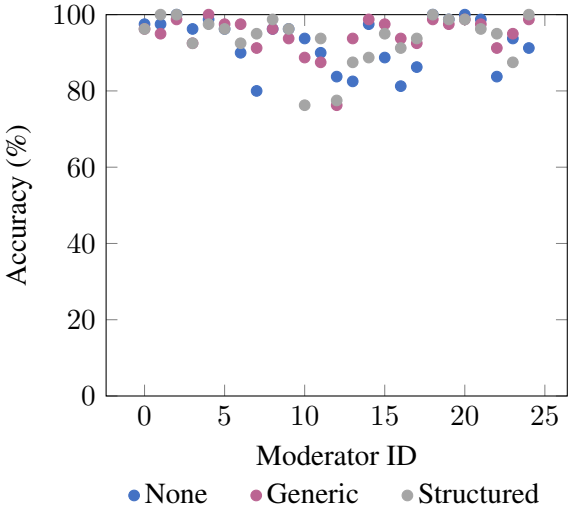

Figure 11: Accuracy score achieved by each moderator with no, generic or structured explanations on the 80 hateful instances of the 3 different samples that were shown with wrong explanations.

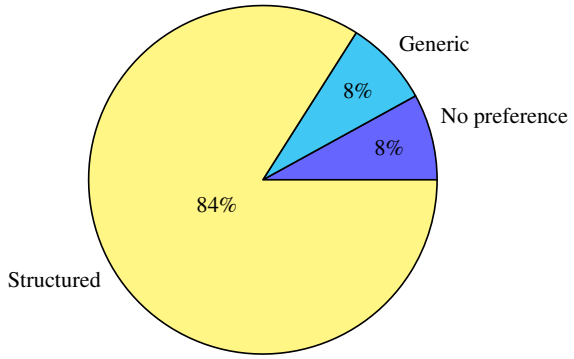

Figure 12: We asked the 25 moderators whether they preferred the setting with generic explanations, structured explanations, or had no preference. The great majority preferred the setting with structured explanations.

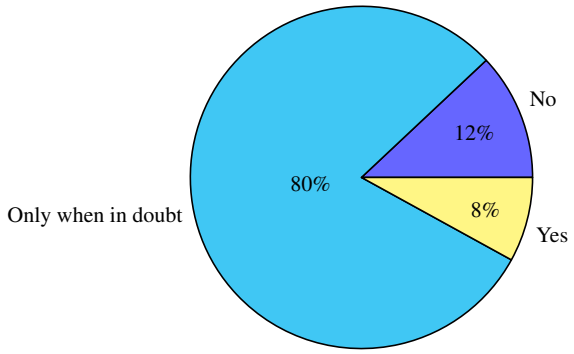

Figure 13: We asked the 25 moderators whether they used the generic explanations or ignored them. 80% of the moderators declared to have used the explanations only when in doubt, and a further 12% ignored the explanations.

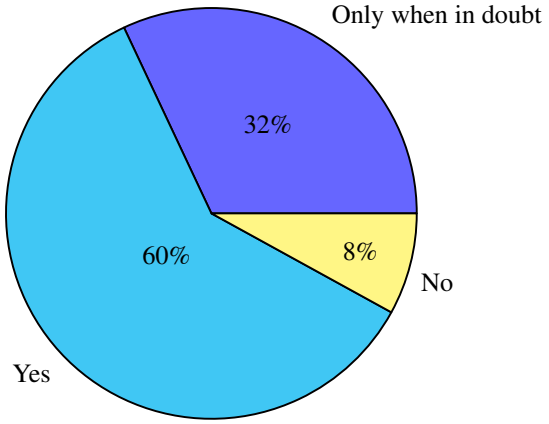

Figure 14: We asked the 25 moderators whether they used the structured explanations or ignored them. 60% of the moderators declared to have used the explanations consistently, and a further 32% relied on them when in doubt.