Personalization up to a Point: Why Personalized Content Moderation Needs Boundaries, and How We Can Enforce Them

Emanuele Moscato[1], Tiancheng Hu[2], Matthias Orlikowski[3], Paul Röttger[1], Debora Nozza[1]
[1]Bocconi University, Italy
[2]University of Cambridge, UK
[3]Bielefeld University, Germany

Abstract

Personalized content moderation can protect users from harm while facilitating free expression by tailoring moderation decisions to individual preferences rather than enforcing universal rules. However, content moderation that is fully personalized to individual preferences, *no matter what these preferences are*, may lead to even the most hazardous types of content being propagated on social media. In this paper, we explore this risk using hate speech as a case study. Certain types of hate speech are illegal in many countries. We show that, while fully personalized hate speech detection models increase overall user welfare (as measured by user-level classification performance), they also make predictions that violate such legal hate speech boundaries, especially when tailored to users who tolerate highly hateful content. To address this problem, we enforce legal boundaries in personalized hate speech detection by overriding predictions from personalized models with those from a boundary classifier. This approach significantly reduces legal violations while minimally affecting overall user welfare. Our findings highlight both the promise and the risks of personalized moderation, and offer a practical solution to balance user preferences with legal and ethical obligations.

Content warning: This paper contains obfuscated examples of hate speech.

1 Introduction

Content moderation on social media has to strike a careful balance between protecting users from harm while enabling freedom of expression. This is difficult because users disagree on what content should be moderated (Kumar et al., 2021; Jhaver et al., 2023; Pradel et al., 2024), meaning that any *platform-wide* moderation decision will necessarily go against the preferences of some users. *Personalized* content moderation has the potential to resolve this dilemma by enabling every user to set their own limits on what content is shown to them. Consequently, a large body of work has sought to develop classification models that reflect and predict the perspectives of individual users – or annotators – rather than single majority labels (see Frenda et al., 2024, for an overview).

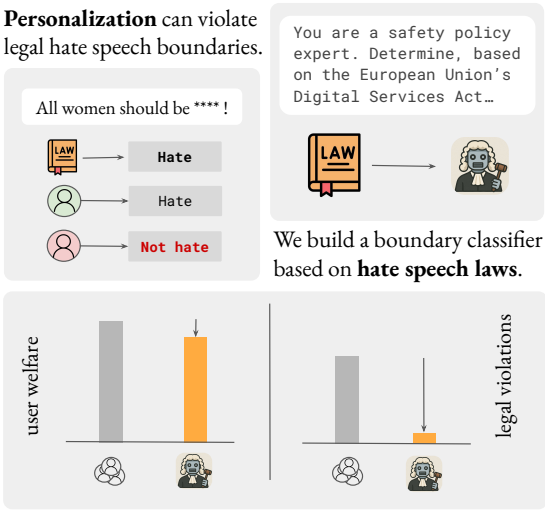


Figure 1: Personalized content moderation can violate legal hate speech boundaries. We address this issue by limiting personalization.

In this paper, we highlight one particular risk of personalized content moderation, which arises when classification models are tailored to the preferences of individual users, *no matter what these preferences are*. While content preferences are subjective, some kinds of content violate clear *boundaries*, such as laws set by regulators. The danger is that fully personalized moderation could allow even such content to go unmoderated (Figure 1).

As a case study, we focus on hate speech, which is among the most prominent types of content subject to moderation on social media. For hate speech, there are laws such as the European Union's Digital Services Act (DSA) that define certain types

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 34004–34018

of hate speech to be *illegal* (§2.3). Holocaust denial, for example, is illegal under the DSA and can lead to prison sentences for offending users in over a dozen EU Member States as well as large fines for platforms (Bakowski, 2021). Hate speech detection models that are fully personalized even to the preferences of extreme users (e.g., neo-Nazis) would plausibly classify such content as harmless, enabling its spread on social media among extreme users. This motivates our first research question:

> **RQ1**: To what extent do hate speech detection models personalized to the preferences of individual users violate independent legal boundaries of hate speech?

To answer this question, we train a personalized hate speech detection model on a large annotator-level dataset (§2.1), and then test this model on another dataset that matches the DSA's legal definition of hate speech (§2.3). In this setting, we show that **personalized hate speech detection models make predictions that violate legal boundaries of hate speech**, especially when personalized to extreme users who tolerate highly hateful content (§2.1). This finding suggests a need for setting boundaries to personalization, which motivates our second research question:

> **RQ2**: How can we enable personalized content moderation while also enforcing boundaries based on legal definitions?

To answer this question, we explore the use of a *boundary classifier* (§2.4), which overrides predictions from personalized models based on an enforcement threshold that can be calibrated. We show that **combining personalized models with a boundary classifier has minimal impacts on overall user welfare (as measured by user-level classification performance) while substantially reducing the rate at which personalized models violate legal boundaries of hate speech** (§3.2).

Overall, we make the following contributions:

1. We provide the first empirical evidence for fully personalized hate speech detection models violating legal boundaries of hate speech.

2. To address this risk, we introduce a simple yet effective method that is model-agnostic and adaptable to different legal standards.

3. We quantify trade-offs between personalization and enforcement of legal standards, and show that our method enables consistent enforcement with minimal impacts on overall user welfare.

All code and data is available at https://github.com/MilaNLProc/personal-hate-bounds.

2 Experimental Setup

2.1 Annotator-Level Hate Speech Dataset

Personalized hate speech detection models require datasets with hate speech ratings that can be attributed to individual annotators[1] for training and evaluation. Further, each annotator should provide enough ratings to enable meaningful personalization, and the dataset overall should contain reasonably many instances to allow for an effective classifier to be trained. For our experiments, we use the **DTC** dataset by Kumar et al. (2021), as it fully meets our requirements. The full dataset contains 107,620 English-language texts from social media annotated by 17,280 annotators in total for toxicity levels. Each text has exactly 5 annotations and each annotator labeled 31 texts on average. We drop non-unique (text, annotator) pairs, i.e., annotators that labeled the same text multiple times.

Our experiments require building a training and a test split out of the full dataset, with specific constraints (e.g., no text should appear in both splits) that force us to subset the data, restricting both the number of unique texts and the number of annotators considered (details in Appendix A). In the end, we use 93,153 texts labeled by 15,563 annotators in total, with on average 5 annotations per text and 24 labeled texts per annotator. The percentage of hateful samples is 47.3% in the training split and 47.0% in the test split, and the train-test split at the instance level has 19.7% of the samples in the test set. When training the personalized model, we restrict to a subset of the annotators in the splits to guarantee we have enough training samples for each annotator. This procedure does not alter the set of texts in the splits (details in Appendix A).

Originally, annotators were asked to rate the toxicity of each instance on a 5-point scale from 0 (not

[1]We use the terms *annotator* and *user* interchangeably. Our research questions aim at content moderation in actual use. However, we do not have access to live user data. Instead, we use a hate speech dataset that provides annotations by individual annotators. We treat these annotations as a snapshot of each individual's perspective and preferences, just as users would provide for personalization.

at all toxic) to 4 (extremely toxic). We binarize this scale, considering all scores >0 to be toxic. Hate speech is generally considered a subset of toxic content (Poletto et al., 2021), meaning that content rated as hateful should also be rated as toxic. Therefore, the binarized version of DTC is compatible with our experimental setup, which is concerned with the lower boundaries to hate speech.

Extreme Annotators For the purpose of our analyses, we identify a subset of *extreme annotators*. Conceptually, we define extreme annotators as those annotators who are much more lenient in their ratings than most annotators, rating content as non-hateful even when most other annotators consider it hateful. For DTC, we consider an annotator to be extreme if, for more than 50% of the instances they rated across both the training and test split, they rated the instances as "non-hateful" despite the most common label for these instances across all annotators being "hateful". This results in 155 extreme annotators, i.e., around 4% of all annotators in our dataset.

2.2 Hate Speech Detection Models

For training personalized hate speech detection models, we use the SepHeads architecture introduced by Mostafazadeh Davani et al. (2022) and Heinisch et al. (2023). The SepHeads architecture contains a separate classification head (linear layer) for each annotator on top of a pre-trained text encoder model. During training, each annotator head is trained only on annotations by a specific annotator, referenced by an unique identifier. The text encoder receives updates from all annotations. During inference, the architecture produces a prediction for a specific annotator. We train a SepHeads model (SEP) on the DTC annotator-level dataset.

As a baseline that reflects standard, i.e., non-personalized hate speech detection, we also train a model on the instance-level majority-vote labels (MAJ) on DTC. The architecture used in this case is analogous to SepHeads, but with a single classification head, with no concept of annotator.

Functionally, given a text, SEP predicts one label per annotator, allowing for a simulation of the predictions from all possible annotators, while MAJ predicts a single label. SEP can only predict for annotators it has been trained on. Therefore, we make sure that SEP is trained and tested on labels from the same annotators on the DTC dataset. For further implementation details, see Appendix B.

2.3 Boundary Dataset: HateCheck

In order to measure whether personalized hate speech detection models violate independent legal standards of hate speech (RQ1), we need to 1) select a legal standard, and then 2) identify a dataset that reflects this standard for model evaluation.

1) We focus on the European Union's Digital Services Act (DSA), which sets a legal standard for hate speech in all 27 EU member states, affecting a combined population of around 450 million people. For defining illegal hate speech, the DSA refers to the EU's Council Framework Decision 2008/913/JHA, Article 1(1), which makes punishable by law content that "publicly incit[es] to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, color, religion, descent or national or ethnic origin". We choose the DSA because it is one of the clearest and most widely applied legal standards for hate speech. The US, for example, lacks equivalent regulation. However, our approach to boundaries for personalization in this paper is compatible with any legal standard or otherwise defined external set of rules (e.g., set by platform providers). Related work has similarly explored connecting legal definitions and hate speech detection (Khurana et al., 2022; Luo et al., 2023; Korre et al., 2025), situating our study within a broader effort to operationalize legal frameworks for moderation.

2) To our knowledge, there is no hate speech dataset annotated directly for whether content violates the DSA or not. As an approximation, we use the HateCheck test suite created by Röttger et al. (2021). HateCheck was originally designed as a set of functional tests for hate speech detection models, comprising manually constructed, human-validated examples of clear-cut hate and non-hate. Since our goal is to evaluate the enforcement of legal boundaries, where there must be no room for subjectivity or annotator disagreement, HateCheck's clear-cut labels make it a suitable dataset choice. More specifically, HateCheck contains 2,563 instances of very explicit hate (e.g., "I am going to shoot the next Muslim I see.") across categories such as "direct threat" and "explicit dehumanisation", which meet the criteria for illegal hate speech under DSA as outlined above. By construction, all hate in HateCheck is targeted at sociodemographic groups protected under DSA. HateCheck also contains 1,165 instances of contrasting non-hate (e.g., "I am going to shoot you", which is threatening

but does not reference any group characteristics), making it ideal for testing how well models can distinguish between hate as defined by the DSA and clear non-hate.

2.4 Boundary Classifier

To enable the enforcement of legal boundaries in personalized hate speech detection (RQ2), we develop a *boundary classifier* that detects DSA-violating hate speech. Specifically, we create a detailed classification prompt based on the DSA's legal definition of hate speech, as outlined above, along with targeted clarifications for edge cases such as counter-speech and reclaimed slurs scenarios where standard models often misclassify content. In addition to a binary assessment, the prompt also asks for a verbalized confidence score on a 0-100 scale.[2] In our later analyses, we vary an *enforcement threshold* based on this confidence score, to calibrate the extent to which predictions from personalized models are overridden by predictions from the boundary classifier (e.g., only when the boundary classifier is $\geq$X% confident that a piece of content is DSA-violating hate speech).

For this paper, we base our boundary classifier on GPT-4o-mini. We validate its performance on HateCheck, our boundary dataset, where it achieves a macro F1 score of 0.936 and a hateful-class recall of 0.996. High recall is especially important for our study, where failing to flag a boundary violation is more critical than false positives. These results indicate strong adherence to the defined boundaries and provide empirical support for the reliability of our boundary classifier.

3 Experiments

3.1 RQ1: Fully Personalized Models

Our first research question (see §1) concerns the extent to which personalized hate speech detection models violate independent legal definitions of hate speech, in terms of the predictions they make.

To answer this question, we evaluate the performance of the personalized hate speech detection models (SEP) against the instance-level majority models (MAJ) trained on DTC, testing on the DTC evaluation set and our legal boundary dataset, HateCheck. Table 1 reports both macro F1 and Recall$^+$ for the positive class. We measure Recall$^+$ because it specifically captures how well the model identifies hate speech, which is the primary focus of our analysis and directly relevant for assessing compliance with legal definitions. We show results by evaluation dataset (DTC test split and the boundary dataset HateCheck), model type (MAJ or SEP), and by annotator group (non-extreme or extreme annotators) when available. HateCheck cannot be used for evaluation at the annotator level for the Majority Vote models, as these models do not produce user-specific predictions and HateCheck itself includes only a single gold-standard label rather than multiple annotator judgments. We also compare results to two baselines: a Random classifier and the Boundary classifier described in §2.4 without considering the enforcement thresholds, i.e., accepting all predictions from the boundary classifier.

Personalization Evaluation We find that the personalized model SEP clearly outperforms the majority model MAJ on the DTC dataset. This is expected, since DTC is an annotator-level dataset, and personalized models are designed to better reflect individual annotator preferences.

Breaking down performance by annotator group, we observe a drop in performance for extreme annotators from both types of models, suggesting that classification is harder for annotators deviating from the majority. A notable exception arises in the case of the majority model, which achieves a substantially higher positive recall for extreme annotators compared to non-extreme annotators. This difference is due to the tendency of the majority model MAJ to predict more instances as hateful, aligning with the annotator consensus, which typically contrasts with the views of extreme annotators, which are a minority. As a result, MAJ captures nearly all truly hateful cases for the extreme group, at the cost of a low precision for the non-hateful class (0.480).

By contrast, the personalized model SEP tends to predict non-hateful labels more often for extreme annotators, where we observe high recall for the non-hateful class (0.888) and low precision for the hateful class (0.514). Notably, SEP always predicts the non-hateful class for all samples annotated by 62% of extreme annotators. This indicates that, while the personalized model better reflects the annotators' overall labeling patterns, in the context of hate speech detection, it is inadvertently reinforcing the biases of more extreme annotators.

Both SEP and MAJ consistently outperform the zero-shot boundary classifier, confirming the value of learning from the annotated data.

[2]For the full classification prompt, see Appendix C.

Eval Dataset	Model	F1 (All)	F1 (Non-Extreme)	F1 (Extreme)	Recall$^+$ (All)	Recall$^+$ (Non-Extreme)	Recall$^+$ (Extreme)
DTC	MAJ	0.690	0.693	0.544	0.678	0.676	**0.866**
	SEP	**0.746**	**0.746**	**0.673**	**0.735**	**0.738**	0.441
	Boundary	0.579	0.578	0.634	0.290	0.289	0.465
HateCheck	MAJ	0.613	–	–	0.941	–	–
	SEP	0.644	0.648	0.535	0.842	0.866	0.418
	Boundary	**0.936**	–	–	**0.996**	–	–

Table 1: Macro F1 and Recall$^+$ scores for MAJ and SEP evaluated on the DTC test split and HateCheck. Where possible, results are broken down by annotator group. For SEP on HateCheck, the annotator group breakdown is based on the learned annotators. Best performance highlighted in **bold**. Random baselines achieve F1 scores of 0.499 (DTC) and 0.482 (HateCheck). The boundary classifier does not apply any enforcement threshold.

Legal Boundary Violations When considering HateCheck, which serves as an external benchmark aligned with a legal definition of hate speech, the boundary classifier achieves the best performance, as it has been explicitly designed to adhere to this legal standard. Notably, the personalized model SEP consistently outperforms the majority model MAJ, which seems to suggest benefits of aligning predictions with individual annotator preferences even for HateCheck. However, a closer look reveals important nuances. When breaking down results by annotator group, the F1 score for extreme annotators under SEP is not only lower than the model's overall average but also lower than that of the majority model MAJ. This discrepancy becomes even more pronounced when examining recall for the hateful class: SEP has substantially lower recall than MAJ. Furthermore, recall for extreme annotators drops significantly under personalization, reinforcing the previous finding that personalized models, while better aligned with annotator preferences, lead to systematic under-enforcement of legal standards for more lenient annotators.

When examining the false negatives produced by both models, we observe a mix of expected and concerning patterns, particularly when analyzing the functional test categories (i.e., group-level classifications of expressions) provided by HateCheck. MAJ produces most of its false negatives due to spelling variations (53%) and derogation expressed in more implicit forms (18%). Similarly, the SEP's false negatives are largely driven by spelling variations, which make up 38% of its errors. However, the next most common source is explicit derogation, accounting for 11%. More problematically, the SEP also generates false negatives in categories not present in the majority model's errors, including threatening language, pronoun reference, and profanity usage. This suggests that for some extreme annotators, even clearly hateful content involving threats may not be labeled as hate speech, highlighting the risk of personalization reinforcing overly permissive interpretations of hate. Regarding false positives, both models predominantly struggle with counter speech, a category that is notoriously challenging to classify due to its overlap in form with actual hate speech despite its opposing intent (Chung et al., 2019; Gligoric et al., 2024).

3.2 RQ2: Models With Boundaries

Our second research question (see §1) concerns the development of personalized hate speech detection models that respect legal boundaries. For this purpose, we make use of the boundary classifier (§2.4) to override predictions from the personalized model (SEP), whenever the boundary classifier predicts something to be hateful but the personalized model does not. The boundary classifier gives a 0-100 confidence score alongside each binary prediction. We use these confidence scores as an enforcement threshold to decide when to accept the boundary classifier's prediction to override the personalized model, enforcing our legal boundaries. Accepting more overrides (lower thresholds) leads to less personalization in the final combined model.

Below, we first analyze results at the annotator level to assess user welfare, and then at the instance level, to contextualise overall model performance and identify common sources of error.

User Welfare We aim to measure "user welfare", which we define as the extent to which a classifier makes predictions that match this user's preferences, as measured by the labels (hate / not hate) provided by this user. If the classifier was used for content moderation, and it made predictions that

Eval Dataset	Enforcement Threshold	% Annotators with overrides	% Annotators Welfare ↑	% Annotators Welfare ↓	Median F1 Increase	Median F1 Decrease
DTC	100	0.00	0.00	0.00	–	–
	95	6.04	3.30	2.74	0.12	-0.06
	90	8.32	4.44	3.88	0.11	-0.06
	85	33.02	17.34	15.57	0.10	-0.07
HateCheck	100	88.62	88.62	0.00	0.00	–
	90	91.47	91.36	0.1	0.04	0.00
	95	91.43	91.33	0.10	0.03	0.00
	85	91.54	91.43	0.10	0.04	0.00

Table 2: Annotator-level statistics for the personalized SepHeads model with legal boundary enforcement via the boundary classifier, evaluated on the DTC test split and HateCheck.

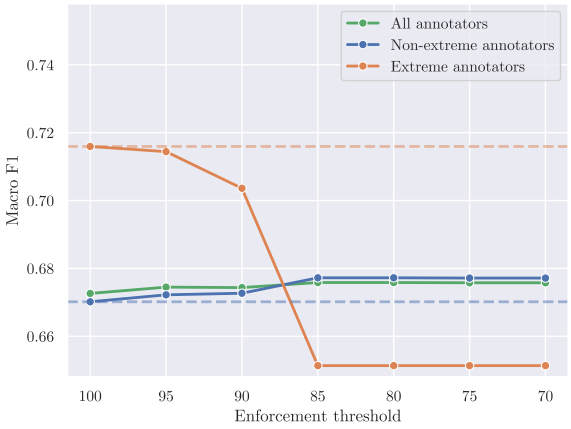


Figure 2: Average annotator-level macro F1 for the personalized SEP model with legal boundary enforcement, evaluated on the **DTC** evaluation dataset. Results are shown for extreme, non-extreme and all annotators across legal boundary enforcing levels. Dashed lines represent baseline SEP performance without boundary overrides for extreme and non-extreme annotators.

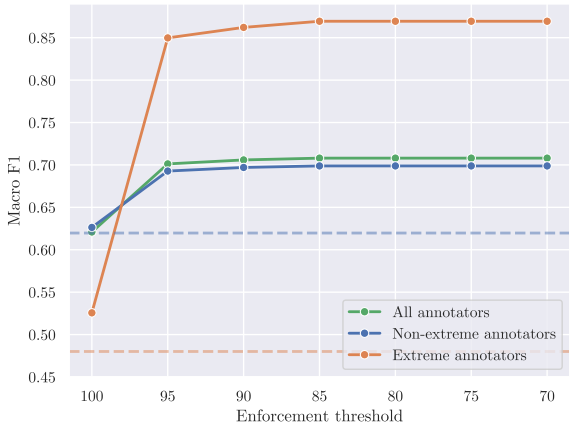


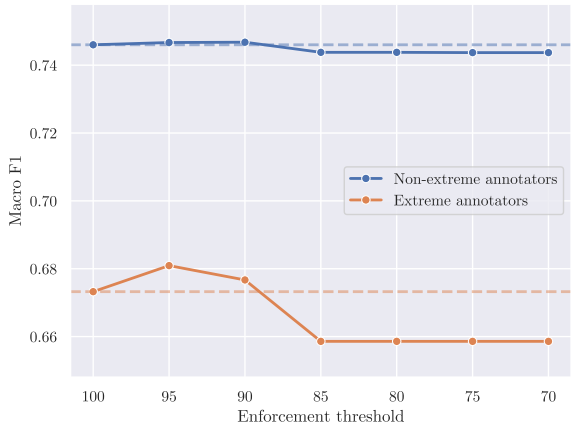
Figure 3: Average annotator-level macro F1 for the personalized SEP model with legal boundary enforcement, evaluated on the **HateCheck** dataset. Results are shown for extreme, non-extreme and all annotators across legal boundary enforcing levels. Dashed lines represent baseline SEP performance without boundary overrides for extreme and non-extreme annotators.

do not match a user's labels, the user would either have been exposed to content they consider hateful, or deprived of content they considered not hateful.

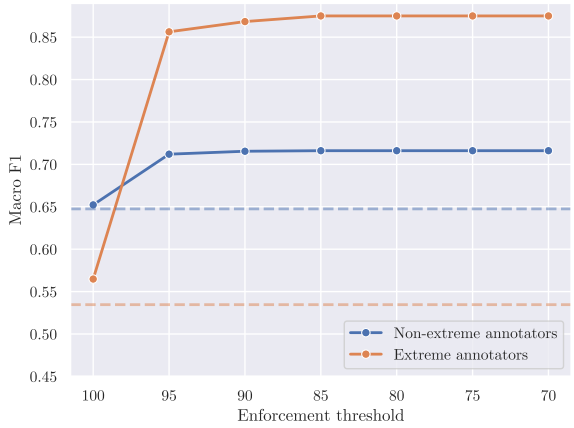
Table 2 reports the effect of the boundary classifier on DTC when overriding predictions from the personalized model (SEP). We compute metrics at annotator level and evaluate the percentage of annotators that are better off and worse off due to the overrides, as measured by F1 scores with the overridden predictions compared to predictions from SEP. We define the annotators with overrides as the ones with at least one prediction overridden by the boundary model. As the enforcement thresholds lowers, we observe a notable percentage of the annotators with overridden predictions, with performance getting better for roughly half of them and worse for the other half.

Notably, the median increase in F1 score for annotators who benefit from the overrides is larger than the median decrease for those whose performance declines. This suggests that, even though the number of annotators affected positively or negatively is similar, the typical performance gains tend to outweigh the typical losses, pointing to a modest net benefit from applying the boundary classifier. This is also reflected in the overall average annotator-level macro F1 (Figure 2), where we observe that **the enforcement of legal boundaries does not harm user welfare, on aggregate**, with overall performance varying by less than 1% across enforcement thresholds.

Figure 2 also breaks down the macro F1 between extreme and non-extreme annotators, as well as providing the comparison with the personalized model



(a) DTC.



(b) HateCheck.

Figure 4: Personalized model performance comparison of extreme and non-extreme annotators with boundary classifier intervention on DTC and HateCheck at different levels of legal boundary enforcing. Dashed lines indicate baseline performance of the SepHeads model without boundary overrides for extreme and non-extreme annotators.

with no legal boundary enforcement. The results show that performance for non-extreme annotators remains largely stable (no statistically significant difference), consistent with the overall trend, whereas performance for extreme annotators decreases marginally (independent-sample t-test with $p = 0.0831$, significant at the 0.1 level), suggesting that overridden predictions are less aligned with annotator-level labels. This reflects exactly what we would want from the enforcing of legal boundaries: non-extreme annotators' performance is not hurt (or even increases), while extreme annotators are heavily penalized, reflecting the fact that **the legal boundary is naturally enforced on the part of the annotator base for which it is most relevant**. Note that this is not reflected on the overall score, as the non-extreme annotators make up for around 96% of all annotators and are thus main drivers when aggregating. The percentage of extreme annotators with overrides is consistently above that of the non-extreme ones, with a 42% against 31% difference at enforcement threshold 85, providing further evidence that extreme annotators are more affected by the boundary model.

Table 2 also reports the effect of the boundary classifier on the predictions of SEP using the boundary dataset HateCheck as evaluation. In this case, we used the predictions of SEP on the HateCheck instances, which yields one prediction per learned annotator. These annotator-level predictions are then compared against the instance-level labels provided by HateCheck in order to compute F1 scores and assess the impact of overriding personalization. Compared to DTC, the percentage of annotators with overrides is much higher for HateCheck, being close to 90% at enforcement threshold 100. This means that, even if we only accept boundary classifications when the model claims 100% confidence, 90% of the annotators have at least one prediction overridden. Moreover, enforcing the boundary classifier on the personalized model yields a strongly positive effect, with around 90% of annotators showing improved performance and almost none experiencing a decline.

In Figure 3, we assess the effect of the boundary classifier on the annotator-level performance on the overall annotator set (with and without override). The results show an increase from thresholds 100 to 95, with the annotator-averaged macro F1 jumping by 8 percentage points (statistically significant, independent-sample t-test with $p < 0.001$), followed by a plateau. This increase is even more visible in the annotator groups breakdown, showing that the performance for both extreme and non-extreme annotators against HateCheck's gold labels benefits from enforcing the legal boundary, especially in the extreme annotators case, whose macro F1 increases by more than 30 percent points. As for the DTC evaluation set, the percentage of extreme annotators with overrides from the boundary model is consistently higher than that of non-extreme annotators. However, both groups are broadly affected: at an enforcement threshold of 95, 91% of non-extreme annotators and nearly all extreme annotators have at least one overridden prediction.

However, the extent of overrides or the absolute performance gain from overriding predictions is not the key takeaway here: we already know that, in our

specific evaluation setting, we could get the best performance on HateCheck by always accepting the boundary classifier's prediction. Basically, the more we override, the better our performance. But, more importantly, the performance gains show that we can steer personalized models away from ignoring illegal cases of hate, in particular for extreme annotators. Thus, enforcing legal boundaries on the boundary dataset HateCheck further demonstrates that personalized models can not only uphold legal standards without compromising user welfare, but also significantly reduce legal violations.

Models with Boundaries Evaluation Figure 4 shows the personalized model macro F1 scores evaluated on DTC (Figure 4a) and on HateCheck (Figure 4b), comparing extreme and non-extreme annotators at different boundary classifier enforcement thresholds, this time **at the instance level** instead of at the annotator level. This can be used to evaluate how enforcing legal boundaries affects the overall predictive performance of the model, irrespective of the specific annotator.

Aggregating at the instance level brings a completely different evaluation of performance w.r.t. the annotator-level case for DTC (Figure 4a), with a similar trend but higher F1 scores for non-extreme annotators and lower ones for the extreme ones. This is due to the fact that the evaluation of the instance-level performance depends on the total number of overridden instances between the two groups of annotators, thus better reflecting a measure of the overall performance of the classifier.

For HateCheck, Figure 4b shows that the macro F1 behavior at the instance level closely mirrors that at the annotator level, primarily because each annotator is predicted for the same set of instances.

These results validate the design of the personalized model with the enforcement of legal boundaries, demonstrating that, at the instance level, the model preserves the benefits of personalization for non-extreme annotators while substantially improving legal compliance for extreme ones.

4 Related Work

Prior work has highlighted that **perceptions of toxic content and preferences for moderation vary across people** from different backgrounds (Talat, 2016; Binns et al., 2017; Larimore et al., 2021; Jiang et al., 2021; Sap et al., 2022; Goyal et al., 2022; Hettiachchi et al., 2023; Rastogi et al., 2024; Mishra et al., 2025). Importantly, differences were found to extent to the individual level and vary within demographic groups (Salminen et al., 2018; Mostafazadeh Davani et al., 2024). This variation is often seen manifested as low inter-annotator agreement for annotations of toxicity and hate speech (Vidgen and Derczynski, 2020). In this situation, when aggregating to a single gold label for each piece of annotated content, minority views might be overwritten - potentially (but not exclusively, as we show) views of groups who are affected the most (Prabhakaran et al., 2021).

Poor agreement and fairness issues motivated **annotator-level modeling** and personalization in research on detecting toxic or hateful content (Mostafazadeh Davani et al., 2022; Orlikowski et al., 2023; Fleisig et al., 2023; Weerasooriya et al., 2023; Hu and Collier, 2024; Mokhberian et al., 2024; Jaggi et al., 2024; Anand et al., 2024). "Jury Learning", for example, uses a recommender architecture to learn individuals' toxicity perceptions to predict the distribution of views in defined subpopulations (Gordon et al., 2022). Some works effectively personalize classifiers using only background information such as demographics (Tahaei and Bergler, 2024), while others argue that attitudinal information (Jiang et al., 2024; Hu and Collier, 2025) or personal values (Sorensen et al., 2025; Hu and Collier, 2025) lead to stronger personalization. However, individual-level performance seems to be highest when learning from examples of each individual using unique identifiers (Orlikowski et al., 2025), as we do in our experiments.

Plepi et al. (2022) discuss annotator modeling in relation to established settings of **personalization**, highlighting the shared goal of modeling identifiable individuals. Although some works discuss the issue of extreme annotators who are insensitive to hateful content (Sachdeva et al., 2022), the problem of limiting personalization went largely unexplored (see, e.g., Jhaver et al. 2023). As exceptions, boundaries for personalization are discussed by Kirk et al. (2024) and River Dong et al. (2025), albeit in the context of Large Language Model alignment. Kirk et al. (2024) propose a hierarchy of bounds where the lowest tier is formed by regulatory boundaries, followed by organization-specific bounds, while River Dong et al. (2025) demonstrate that personalization can introduce significant safety misalignments. These discussions, however, have so far remained mostly conceptual or evaluative.

Our work advances the study of personalization in the context of hate speech moderation and, to the

best of our knowledge, is the first to introduce the concept of enforcing legal boundaries within personalized moderation models, showing that fully unconstrained personalization can lead to the propagation of harmful or even illegal content, particularly among extreme users who tolerate or endorse such speech. Our contribution is both conceptual and practical: we propose a simple yet effective framework for enforcing legal boundaries that is model-agnostic and adaptable to different regulatory standards.

5 Conclusion

Personalized content moderation offers a promising way to protect users from harm while respecting individual preferences. However, as we have demonstrated, fully personalized moderation models, tailored without constraints, can inadvertently allow highly harmful and even illegal hate speech to spread, particularly among extreme users who tolerate or endorse such content.

In this work, we addressed this critical risk by enforcing legal boundaries based on legal definitions on personalized hate speech detection models. Our findings show that enforcing these legal boundaries drastically reduces violations while maintaining a high degree of model performance and without compromising user welfare. This demonstrates that it is both feasible and beneficial to integrate legal constraints into personalized moderation systems, offering a safer and more responsible path forward for content platforms.

Overall, this paper contributes empirical evidence of the risks inherent in unregulated personalized content moderation and offers a scalable solution that safeguards users and platforms by harmonising personalization with legal accountability.

Limitations

Our annotator-level analysis is based on a single dataset (DTC), which we selected specifically because it aligns closely with the goals of our study, namely, examining how personalized moderation interacts with legal boundaries of hate speech. This dataset allowed us to explore the core questions of our work in a focused and principled manner. However, we acknowledge that results may vary across different datasets, particularly those reflecting other languages, cultural norms, or moderation practices. While broader validation is an important direction for future work, our aim here is not to benchmark performance across datasets, but to foreground a key conceptual risk: the potential for personalization to conflict with legal constraints.

Because the dataset we used is broadly aligned with the legal boundary definitions we apply, we observe relatively minimal negative impact on user welfare when those constraints are enforced. However, in contexts where user preferences diverge more significantly from legal standards, such as on platforms with highly extremist user bases, strict enforcement may lead to more substantial trade-offs in user welfare. We discuss the dynamics of moderation frameworks and the individuals, platforms, or institutions responsible for their implementation in the Ethical Considerations section.

Our findings regarding legal boundary violations rely in part on our operational definition of extreme users. We adopted a threshold that we consider reasonable and interpretable, enabling a meaningful analysis of how personalized models behave for users with fringe preferences. Nonetheless, we recognize that the exact threshold value for when to consider a user to be extreme is, to some degree, arbitrary. A stricter threshold might have captured more ideologically extreme users, but it would yield too small a sample to generalize from. Conversely, a looser threshold would have resulted in a group more representative of the average user, potentially obscuring the effects we aimed to study. Our results should be understood in light of this methodological trade-off.

Ethical Considerations

A central ethical challenge in personalized content moderation is balancing individual user preferences with shared societal standards (Kirk et al., 2024). These standards may be defined by law, platform policy, or broader community norms. While our approach allows for moderation to be tailored through adjusting the boundary classifier prompt, it raises important questions about which boundaries should be respected, who defines them, and how strictly they should be enforced.

Legal definitions of hate speech vary significantly across jurisdictions, and platform policies differ in how permissive or restrictive they are. Some platforms may choose to enforce stricter content rules than what the law requires, reflecting internal values or the expectations of their user base. Our framework is designed to accommodate for this diversity. The boundary classifier can

be trained to reflect any agreed-upon threshold, whether legal, ethical, or policy-based. This makes the system both flexible and adaptable to a wide range of moderation regimes.

However, this flexibility introduces a meaningful ethical risk. While we maintain that legal standards should serve as the minimum threshold for moderation, there is a possibility that our approach could be used to weaken enforcement. This concern points to a deeper normative issue: the question of who has the authority to define and enforce the limits of acceptable speech.

We argue that any system of personalized moderation must be grounded in a clear and enforceable baseline that at least reflects legal requirements. Personalization should improve user experience within these boundaries, not undermine them. In this sense, our dual-layer framework is more than a technical mechanism. It also serves as a structure for promoting accountability by enabling flexible content moderation while maintaining a firm commitment to fundamental standards.

Acknowledgments

Debora Nozza's research is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101116095, PERSONAE). Tiancheng Hu is supported by Gates Cambridge Trust (grant OPP1144 from the Bill & Melinda Gates Foundation). Emanuele Moscato's research was funded by the European Union - NextGenerationEU, in the framework of the FAIR - Future Artificial Intelligence Research project (FAIR PE00000013 – CUP B43C22000800006). Matthias Orlikowski's research was funded by Volkswagen Foundation as part of the "Bots Building Bridges (3B)" project under the "Artificial Intelligence and the Society of the Future" programme. Paul Röttger was supported by a MUR FARE 2020 initiative under grant agreement Prot. R20YSMBZ8S (INDOMITA). Emanuele Moscato, Paul Röttger and Debora Nozza are members of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis. The authors thank the MilanLP group at Bocconi University for feedback on an earlier version of this paper.

References

Abhishek Anand, Negar Mokhberian, Prathyusha Kumar, Anweasha Saha, Zihao He, Ashwin Rao, Fred Morstatter, and Kristina Lerman. 2024. Don't blame the data, blame the model: Understanding noise and bias when learning from subjective annotations. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 102–113, St Julians, Malta. Association for Computational Linguistics.

Piotr Bakowski. 2021. Holocaust denial in criminal law: Legal frameworks in selected eu member states. *Think Tank*.

Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation. In *Social Informatics*, Lecture Notes in Computer Science, pages 405–415, Cham. Springer International Publishing.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, pages 1–28.

Kristina Gligoric, Myra Cheng, Lucia Zheng, Esin Durmus, and Dan Jurafsky. 2024. NLP systems that can't tell use from mention censor counterspeech, but teaching the distinction helps. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5942–5959, Mexico City, Mexico. Association for Computational Linguistics.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 1–19, New York, NY, USA. Association for Computing Machinery.

Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *Preprint*, arXiv:2006.03654.

Philipp Heinisch, Matthias Orlikowski, Julia Romberg, and Philipp Cimiano. 2023. Architectural sweet spots for modeling human label variation by the example of argument quality: It's best to relate perspectives! In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11138–11154, Singapore. Association for Computational Linguistics.

Danula Hettiachchi, Indigo Holcombe-James, Stephanie Livingstone, Anjalee de Silva, Matthew Lease, Flora D. Salim, and Mark Sanderson. 2023. How Crowd Worker Factors Influence Subjective Annotations: A Study of Tagging Misogynistic Hate Speech in Tweets. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 11:38–50.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

Tiancheng Hu and Nigel Collier. 2025. iNews: A multimodal dataset for modeling personalized affective responses to news. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25000–25040, Vienna, Austria. Association for Computational Linguistics.

Harbani Jaggi, Kashyap Coimbatore Murali, Eve Fleisig, and Erdem Biyik. 2024. Accurate and data-efficient toxicity prediction when annotators disagree. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21910–21917, Miami, Florida, USA. Association for Computational Linguistics.

Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X. Zhang. 2023. Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2):289:1–289:33.

Aiqi Jiang, Nikolas Vitsakis, Tanvi Dinkar, Gavin Abercrombie, and Ioannis Konstas. 2024. Re-examining sexism and misogyny classification with annotator attitudes. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15103–15125, Miami, Florida, USA. Association for Computational Linguistics.

Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R. Brubaker. 2021. Understanding international perceptions of the severity of harmful content online. *PLOS ONE*, 16(8):e0256762. Publisher: Public Library of Science.

Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens. 2022. Hate speech criteria: A modular approach to task-specific hate speech definitions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 176–191, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6(4):383–392. Publisher: Nature Publishing Group.

Katerina Korre, Arianna Muti, Federico Ruggeri, and Alberto Barrón-Cedeño. 2025. Untangling hate speech definitions: A semantic componential analysis across cultures and domains. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3184–3198, Albuquerque, New Mexico. Association for Computational Linguistics.

Deepak Kumar, {Patrick Gage} Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Proceedings of the 17th Symposium on Usable Privacy and Security, SOUPS 2021*, Proceedings of the 17th Symposium on Usable Privacy and Security, SOUPS 2021, pages 299–317. USENIX Association.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Chu Luo, Rohan Bhambhoria, Samuel Dahan, and Xiaodan Zhu. 2023. Legally enforceable hate speech detection for public forums. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10948–10963, Singapore. Association for Computational Linguistics.

Pushkar Mishra, Charvi Rastogi, Stephen R. Pfohl, Alicia Parrish, Roma Patel, Mark Diaz, Ding Wang, Michela Paganini, Vinodkumar Prabhakaran, Lora Aroyo, and Verena Rieser. 2025. Nuanced safety for generative ai: How demographics shape responsiveness to severity. *Preprint*, arXiv:2503.05609.

Negar Mokhberian, Myrl Marmarelis, Frederic Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman.

2024. Capturing perspectives of crowdsourced annotators in subjective learning tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7337–7349, Mexico City, Mexico. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Diaz, Dylan K Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18511–18526, Miami, Florida, USA. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2092–2111, Vienna, Austria. Association for Computational Linguistics.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1017–1029, Toronto, Canada. Association for Computational Linguistics.

Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. Unifying data perspectivism and personalization: An application to social norms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7391–7402, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Franziska Pradel, Jan Zilinsky, Spyros Kosmidis, and Yannis Theocharis. 2024. Toxic speech and limited demand for content moderation on social media. *American Political Science Review*, 118(4):1895–1912.

Charvi Rastogi, Tian Huey Teh, Pushkar Mishra, Roma Patel, Zoe Ashwood, Aida Mostafazadeh Davani, Mark Diaz, Michela Paganini, Alicia Parrish, Ding Wang, Vinodkumar Prabhakaran, Lora Aroyo, and Verena Rieser. 2024. Insights on disagreement patterns in multimodal safety perception across diverse rater groups. *Preprint*, arXiv:2410.17032.

Yijiang River Dong, Tiancheng Hu, Yinhong Liu, Ahmet Üstün, and Nigel Collier. 2025. When personalization meets reality: A multi-faceted analysis of personalized preference learning. *arXiv e-prints*, pages arXiv–2502.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Joni Salminen, Fabio Veronesi, Hind Almerekhi, Soon-Gvo Jung, and Bernard J. Jansen. 2018. Online Hate Interpretation Varies by Country, But More by Individual: A Statistical Analysis Using Crowdsourced Ratings. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 88–94.

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 *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Taylor Sorensen, Pushkar Mishra, Roma Patel, Michael Henry Tessler, Michiel Bakker, Georgina Evans, Iason Gabriel, Noah Goodman, and Verena Rieser. 2025. Value Profiles for Encoding Human Variation. *arXiv preprint*. ArXiv:2503.15484 [cs].

Narjes Tahaei and Sabine Bergler. 2024. Analysis of annotator demographics in sexism detection. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 376–383,

Bangkok, Thailand. Association for Computational Linguistics.

Zeerak Talat. 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*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300. Publisher: Public Library of Science.

Tharindu Cyril Weerasooriya, Sarah Luger, Saloni Poddar, Ashiqur KhudaBukhsh, and Christopher Homan. 2023. Subjective crowd disagreements for subjective data: Uncovering meaningful CrowdOpinion with population-level learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 950–966, Toronto, Canada. Association for Computational Linguistics.

A Dataset Construction

The construction of the training and test splits from the full DTC dataset requires careful attention to two key criteria: (i) ensuring that no data contamination occurs between the splits (each text only appears either in the training or in the test split) and (ii) maintaining a consistent set of annotators across both splits. To meet these requirements, an initial data cleaning phase is conducted. During this phase, all duplicate (text, annotator) pairs, which result from cases in which an annotator labeled the same text multiple times, are removed to prevent ambiguity. In addition, all samples associated with annotators who have labeled fewer than 18 texts in total are excluded from the dataset.

Then, we proceed with the following filtering steps:

1. We randomly choose 20% of the texts to be in the test split (the rest of the texts will be in the training one).

2. We drop the annotators that appear only in one split.

3. We drop the annotators that have less than 12 samples in the training dataset.

4. We repeat the previous points 100 times and keep the version of the splits with the most total samples.

This procedure allows for cleaner and more reliable splits, though it may influence the class distribution and the proportion of samples assigned to each split. Nevertheless, we verify a posteriori that the class distribution remains consistent between the training and test sets (approximately 47%), and that the resulting split maintains a desirable balance, with roughly 80% of instances in the training set and 20% in the test set.

The SEP model is trained on a restricted subset of the DTC training split. This subset includes all samples from the 2,730 annotators who contributed at least 35 annotations to the training set, as well as the samples annotated by the 155 extreme ones. In the end SEP is trained $152,528$ samples and tested on $35,447$. The original training and testing split are in fact built so that the texts are the same with and without this subsampling: this way MAJ is trained and tested on datasets containing the same texts SEP is trained upon, allowing for a comparison like-for-like of the two models. Moreover, this allows the majority-vote labels to be computed on the full annotators population

B Model Training

Following the architectures introduced by Heinisch et al. (2023), our non-personalized hate speech detection model, MAJ, adopts the *Majority* architecture, while our personalized, annotator-specific model, SEP, is based on the *SepHeads* architecture. For both models, we use the `microsoft/deberta-v3-base`[3] (He et al., 2021) pre-trained model as text encoder, which is then fine-tuned along with the classification head(s) (see below). The annotator heads (or single classification head, in the MAJ case) share the same architecture as the default DeBERTa classifier from Hugging Face (a pooling linear layer mapping sequences to latent vectors of dimension 768 followed by a linear readout outputting class probabilities, with each layer having its own dropout). The MAJ model contains $184,423,682$ parameters in total.

[3] https://huggingface.co/microsoft/deberta-v3-base

In the SEP model, the classification head to use is selected during the forward pass using the annotator's ID. Training batches are created randomly to avoid batches of samples all from the same annotator, which might bias the weight updates of the text encoder. The total number of parameters for SEP, with $2,883$ classification heads, is $188,856,198$.

Training is performed using the Hugging Face API, with the default AdamW optimizer (Loshchilov and Hutter, 2019) and linear learning rate decay. MAJ is trained with a starting learning rate of 10^{-6} and a training batch size of 64, while SEP is trained with a starting learning rate of $5 \cdot 10^{-6}$ and a training batch size of 16. We report the results for a single run. Both models are trained for 10 epochs on two NVIDIA RTX A5000 GPUs. Training times amount to ~ 3.5 hours for MAJ and ~ 24 hours for SEP.

C Legal Boundary Classifier

The boundary classifier consists of the o4-mini-2025-04-16 model invoked via the OpenAI API using a structured prompting approach, with all data collected between April 17th and May 10th 2025. We extracted scores by parsing the JSON responses returned by the model in structured output mode. The temperature parameter was set to 0 to ensure deterministic and consistent outputs.

We use the following prompt to implement the boundary classifier:

> You are a safety policy expert. Your task is to determine whether a given TEXT violates the safety policy below. Your decision must strictly follow the specific language used in the safety policy.
> <SAFETY_POLICY>
> This safety policy is based on: - "Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law" - Article 8 of the "Directive (EU) 2024/1385 of the European Parliament and of the Council of 14 May 2024 on combating violence against women and domestic violence" - The inclusion of additional protected groups (e.g. based on sexual orientation) in Member States' laws.
>
> The following types of content violate this safety policy: 1) Content that publicly incites to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent, national or ethnic origin, gender, sex, sexual orientation, age, or disability; 2) Content that publicly condones, denies, or grossly trivializes genocide, crimes against humanity and war crimes as defined in Articles 6, 7 and 8 of the Statute of the International Criminal Court, directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent, national or ethnic origin, gender, sex, sexual orientation, age, or disability, when the conduct is carried out in a manner likely to incite to violence or hatred against such a group or a member of such a group;
> </SAFETY_POLICY>
>
> Note that texts which clearly oppose hatred or discrimination, including educational content, journalistic reporting, or counter-arguments that quote problematic language specifically to refute it, do *not* violate the safety policy.
>
> Note also that texts which use reclaimed slurs in clearly innocuous contexts do not violate the safety policy.
>
> Does the TEXT below violate the safety policy? Your answer must start with "yes" or "no". After your answer, provide a score between 0-100 that reflects your confidence in your assessment. Do not write anything else. Here is the text for you to classify.
> TEXT:

We conducted limited internal ablations of the boundary prompt to assess its sensitivity to different formulations. In particular, we tested a variant where the model is not provided with a safety policy or article references but is simply instructed to recall the DSA regulations, using the simplified safety policy *"The European Union's Digital Services Act (DSA) and its definition of illegal hate speech."* The recall remains identical (0.996), but our original prompt achieves higher precision, resulting in a better macro F1 score: 0.936 (ours)

versus 0.927 (recall DSA only, without detailed legal clauses). We therefore rely on the original prompt formulation in this work, as it provides a more reliable balance between recall and precision.

Regarding the confidence scores, we adopt the method of (Tian et al., 2023), who show that verbalized confidence scores provide well-calibrated estimates. In our case, we assessed calibration on HateCheck, where we found consistently high confidence levels (minimum approximately 75) and already strong overall performance (hateful-class recall of 0.996). As a result, standard calibration diagnostics provided limited additional insight in this setting. To investigate further, we turned to the DTC dataset. We used the standard deviation of the gold-label toxicity scores across annotators as a proxy for sample difficulty, based on the assumption that lower variance reflects higher agreement and therefore easier instances. We then computed the Pearson correlation between this difficulty measure and the boundary classifier's confidence scores. We observed a strong negative correlation ($R = 0.81$), indicating that the classifier assigns lower confidence to samples with greater annotator disagreement. This supports the conclusion that the confidence scores are meaningfully calibrated and reflect item-level uncertainty.