Algorithmic Arbitrariness in Content Moderation

Juan Felipe Gomez*
Harvard University
Department of Physics
Cambridge, MA, USA
juangomez@g.harvard.edu

Caio V. Machado*
University of Oxford
Centre for Socio-Legal Studies
Oxford, UK
Harvard University
School of Engineering and Applied Science
Cambridge, USA
University of São Paulo
Law School
São Paulo, Brazil
caiocvm@g.harvard.edu

Lucas Monteiro Paes*
Harvard University
School of Engineering and Applied Science
Cambridge, MA, USA
lucaspaes@g.harvard.edu

Flavio P. Calmon
Harvard University
School of Engineering and Applied Science
Cambridge, MA, USA
flavio@seas.harvard.edu

ABSTRACT

Machine learning (ML) is widely used to moderate online content. Despite its scalability relative to human moderation, the use of ML introduces unique challenges to content moderation. One such challenge is predictive multiplicity: multiple competing models for content classification may perform equally well on average, yet assign conflicting predictions to the same content. This multiplicity can result from seemingly innocuous choices made during training, which do not meaningfully change the accuracy of the ML model, but can nevertheless change what the model gets wrong. We experimentally demonstrate how content moderation tools can arbitrarily classify samples as "toxic," leading to arbitrary restrictions on speech. We use the principles set by the International Covenant on Civil and Political Rights (ICCPR), namely freedom of expression, non-discrimination, and procedural justice to interpret the effects of these findings in terms of Human Rights. We analyze (i) the extent of predictive multiplicity among popular state-of-the-art LLMs used for detecting "toxic" content; (ii) the disparate impact of this arbitrariness across social groups; and (iii) the magnitude of model multiplicity on content that is unanimously recognized as toxic by human annotators. Our findings indicate that the upscaled algorithmic moderation risks legitimizing an "algorithmic leviathan", where an algorithm disproportionately manages human rights. To mitigate such risks, our study underscores the need to identify and increase the transparency of arbitrariness in content moderation applications. Our findings have implications to content moderation and intermediary liability laws being discussed and passed in many countries, such as the Digital Services Act in the European Union, the Online Safety Act in the United Kingdom, and the recent TSE resolutions in Brazil.

*Equal contribution, authors in alphabetical order of last name.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3659036

CCS CONCEPTS

• **Social and professional topics** → **Hate speech**; • **Applied computing** → **Law, social and behavioral sciences**; • **Computing methodologies** → **Natural language processing**.

KEYWORDS

content moderation, predictive multiplicity, Rashomon effect

ACM Reference Format:
Juan Felipe Gomez, Caio V. Machado, Lucas Monteiro Paes, and Flavio P. Calmon. 2024. Algorithmic Arbitrariness in Content Moderation . In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3630106.3659036

1 INTRODUCTION

Algorithmic content moderation can be defined as the application of algorithmic systems to classify user-generated content, leading to governance decisions such as content removal, geoblocking, or account takedowns [30]. In the past, content moderation protocols relied on a combination of deterministic rules-based algorithms[1] and human moderators [71].

Recently, various economic, social, and legal factors, such as COVID-19 disinformation and online extremism, have prompted substantial legislative changes globally. These changes have ushered in new regulatory frameworks for online and third-party content [46, 51] that have increased pressure on companies to *expedite* content moderation. Notable legislative shifts include the European Digital Services Act (DSA) [62], which adopts a risk-based approach

[1]An example is an algorithm that auto-removes content that contains words in a pre-specified list of swear words.

for Very Large Online Platforms [9], Germany's NetzDG law [10], which mandates rapid content removal with minimal human oversight, and Brazil's Electoral Courts, which have implemented a stringent 1-hour removal window during elections [24]. The topic is also an ongoing debate in the United States, where states such as Florida have enacted their own laws regarding content removal [61], while the Federal laws regarding platforms' duties on third-party content have remained under intense debate [36, 42].

A natural consequence of these regulations is that companies increasingly rely on black-box machine-learning (ML) models as a scalable alternative to human moderation. This implies that content moderation algorithms, which ultimately control a user's right to freedom of expression, will inherit any limitations intrinsic to ML models. This implication is a growing concern in the law and policy literature [50, 56, 64], particularly in scenarios described as "algorithmic leviathans" [15, 43], where algorithms excessively control the exercise of freedoms and access to resources. Recent research showing that ML-based content moderation occurs with limited accountability and with policies applied indiscriminately across jurisdictions [30, 71] only adds to these concerns.

In this work, we focus on one critical limitation of ML-based content moderation: predictive multiplicity [48] and the ensuing arbitrariness in models that classify toxic content.[2] Predictive multiplicity is the empirical observation that a collection of ML models with indistinguishable performance can produce conflicting individual predictions. Predictive multiplicity captures *arbitrariness* in ML model development, where seemingly innocuous choices made during training, which do *not* meaningfully change the accuracy of the ML model[3], can nevertheless affect what the algorithm gets wrong. Predictive multiplicity has been recently documented in a range of classification and prediction tasks [37, 66] and can lead to disparate treatment of individual data points [7, 15, 48, 60]. We demonstrate that predictive multiplicity is rampant in state-of-the-art language models that have been proposed for toxic text classification: multiple models can achieve similar average accuracy yet conflict in classifying individual sentences as toxic. These observations imply that content moderation decision made using ML models lead to outcomes that lack consistency, predictability, and adherence to established principles or logic [16]. To explore the impact of predictive multiplicity in state-of-the-art models for content moderation, below, we detail the research questions that our work sets out to answer along with our main contributions and findings.

Research Questions: We explore the role of predictive multiplicity in algorithmic content moderation and aim to answer the following questions.

(RQ1) What is the **extent of disagreement** in state-of-the-art ML models fine-tuned to classify toxic content?

(RQ2) What are the **disparate impacts of arbitrariness** across toxicity detection models on content targeting different social groups?

(RQ3) What **forms of harm** stem from the results of RQ1 and RQ2?

[2]We highlight that there is no single definition of "toxic content" and even the use of the term "toxic" caries several limitations; see Appendix C.1 for further discussions.

[3]An example of such innocuous choices is the random seed used for parameter initialization

Main findings: We answer our research questions by fine-tuning several large language models (LLMs) for toxicity detection[4] in textual content and analyze the rate these models generate arbitrary decisions.

- We find that arbitrary decisions are rampant in LLMs fine-tuned for content moderation. In our experiments, approximately 30% of English statements receive moderation decisions that can change by varying the random seed used to initialize training (i.e., LLM fine-tuning). Our results illustrate how arbitrary decisions in model development influence prediction outcomes in content moderation (Table 1).
- As a consequence, we conclude that multiplicity in algorithmic content moderation can unduly restrict individual and collective rights to freedom of expression via a random or unjustified model selection procedure.
- We also find that arbitrary content moderation decisions are unequally distributed across different demographic groups, making the incidence of predictive multiplicity potentially discriminatory (Figure 1).
- We conclude that by producing disparate arbitrary decisions, predictive multiplicity breaks from a rule-based approach to moderating speech online and infringes upon procedural fairness.
- Finally, we also show that models can disagree in examples where human annotators unanimously agree that it should (or should not) be moderated, introducing additional arbitrariness to content moderation (Figure 2), and indicating that it might be useful to share the responsibility of content moderation with humans.

All code and training data used in this work is available upon email request to the authors.

1.1 Related Work

Predictive Multiplicity: Marx et al. [48] showed the prevalence of arbitrary decisions in classification problems using tabular data and argued that it should be measured and reported as we measure and report test error. Follow-up work has analyzed the source of such phenomenon [34, 44, 60], and its inevitability [54]. Creel and Hellman [15] discussed the harms of predictive multiplicity and arbitrary decisions, leading to the definition of *algorithmic leviathan*, initially introduced by König [43]. The work that is closest to ours is [15], which defines algorithmic arbitrariness and argues about its harms. Our paper differs from [15] by (i) focusing on specific harms of arbitrary decisions in content moderation and (ii) experimentally discovering and analyzing the harms of disparate arbitrary decisions across content targeting different demographic groups. Black et al. [5] discuss how predictive multiplicity can have discriminatory legal outcomes and the need to compare models to reach less discriminatory outcomes.

Legal and policy aspects of content moderation: The law and policy literature on algorithmic content moderation has focused on procedural fairness, inconsistent restrictions of human rights, and discrimination. Examples include the works of scholars such

[4]See appendix C.1 for a discussion on why toxicity detection can be used as a proxy for legally mandated content moderation.

as Douek [21], Gillespie [27, 28], Gorwa et al. [30]. We summarize the main risks below.

Inconsistency in Moderation: Different algorithms might produce divergent classifications for the same piece of content. Effectively, this means that either legal speech is being taken down or harmful speech is tolerated. This can happen with regard to individual expressions or groups and their specific dialects. Keller [39] listed a number of studies and resources that indicate the systematic over-removal of content for various reasons, including copyright infringement and toxic speech content moderation. Douek [21] explains that the process for how platforms' enforce their rules has shifted from a rule and proportionality-based approach to an algorithmic probability-based evaluation. The policy report produced by Duarte and Llansó [23] offers a useful summary of the policy challenges in the field. In particular, the concern that algorithms have very limited capability of parsing meaning from text to make content moderation decisions.

Bias Amplification: Content moderation risks having disparate impacts across social groups. Expanding on the inconsistencies listed above, biased and discriminatory moderation may occur if algorithms used to moderate speech are inconsistent across different groups. For example, Dias Oliva et al. [20], Gonçalves et al. [29] describe how certain social groups have been targeted by overmoderation due to the dialects they use. Our work demonstrates that inconsistency and arbitrariness in algorithmic content removal can vary with social-demographic factors.

Opacity of Policy Enforcement: Predictive multiplicity makes enforcing a consistent content policy difficult. We can only review and repair harmful moderation outcomes if we have a clear understanding of how these models are classifying statements [27]. In this scenario, understanding which decisions align with the platform's terms of service and the law becomes challenging. The difficulty in discerning between correct decisions, errors, and arbitrary decisions can make it difficult to determine whether content was overly restricted or not. Since restrictions to Freedom of Expression need to be justified, the opacity of arbitrary decisions poses a threat to fundamental liberties[3].

Conflicting Jurisdictions: Each country has different laws regarding social media platforms. We list as examples the different approaches to intermediary liability pointed out by Keller [38], Machado and Aguiar [46], and the different legislative approaches to algorithmic discrimination outlined by Binns et al. [4], Wachter et al. [65] when arguing for EU or UK legal frameworks. These laws all require specific enforcement requirements for content moderation that arbitrariness may violate.

Based on the concerns above, in Section 2 we conceptualize the harms of arbitrariness in terms of the human rights principles of freedom of expression, non-discrimination, and procedural fairness to based on specific articles of the ICCPR. Then, in Sections 3 and 4, we experimentally investigate arbitrariness in state-of-the-art models, as outlined in our research questions. Finally, in Section 5, we interpret the harms of arbitrariness according to the concepts of freedom of expression, non-discrimination, and procedural justice laid out in Section 2.

2 JUDGES FLIPPING COINS: CONCEPTUALIZING HARMS OF ARBITRARINESS IN CONTENT MODERATION

This section describes the harms identified in the literature from content moderation in terms of potential violations to principles established by the ICCPR. We use the International Covenant on Civil and Political Rights (ICCPR) because it is a widely ratified international treaty to which 173 countries are parties [2]. Our analysis focuses on the impact of arbitrariness on freedom of expression, non-discrimination, and procedural justice.

We are aware that international human rights laws and their principles are primarily applicable to states and do not directly impose obligations on private entities, including internet content companies. Each state enforces these principles within its own jurisdiction, regulating how businesses will respect these rights, and how companies should govern content in their services. Nonetheless companies are directly and indirectly bound to these human right principles, either by platforms laws such as the DSA or the UK Online Safety Act, or by international frameworks and recommendations such as the UN Guiding Principles for Businesses on Human Rights [57].

We intentionally avoid local legislation and the granular matters of each jurisdiction to observe the overarching legal effects of arbitrariness in terms of specific international human rights principles. International Human Rights Law gives us global rules and common concepts to discuss the issues related to fundamental rights in content moderation [22]. Although these laws do not have direct applicability to national jurisdictions, it allows us to make claims related to multiplicity for content moderation that are transferable to local law and policy discussions.

Building on the related work outlined in Section 1.1, we will discuss harms due to algorithmic arbitrariness as an infringement of three human rights and principles: Freedom of Expression, Non-Discrimination, and Procedure (including Procedural Fairness and Equality Before the Law). To illustrate the role content moderation models play as private proxy adjudicators of speech in online environments, we use an analogy to discuss the implications. We compare a model's decision to that of a judge, where arbitrariness is the act of flipping coins to decide the outcome of a case. Though imperfect, we find this comparison makes the harms due to multiplicity more palpable, since the analogy emphasizes that the source of harm is the *randomness* inherit to ML models. Next, we indicate and explain the ICCPR Articles that we use as reference.

Freedom of Expression. Freedom of Expression (FoE) is defined in Article 19 of the ICCPR as:

> 1. Everyone shall have the right to hold opinions without interference.
> 2. Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice.

> 3. The exercise of the rights provided for in paragraph 2 of this article carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary. (a) For respect of the rights or reputations of others; (b) For the protection of national security or of public order (order public), or of public health or morals.

We interpret this rule in light of UN General Comment 34 [13], which emphasizes that freedom of expression is a broad and fundamental human right for realizing other human rights. It encompasses all forms of expression, including political discourse, journalism, artistic works, and religious dialogue, across various mediums like broadcasting, the internet, and public protest. The comment underscores the right to access information and recognizes the critical role of the internet and digital media in enabling and enhancing the exercise of freedom of expression, advocating for universal access to these platforms. This right is expansive but not absolute and can be subject to certain restrictions. These restrictions must be clearly defined by law, serve a legitimate aim (such as protecting national security, public order, or the rights of others), and be necessary, and proportionate.

In the context of content moderation, the existence of predictive multiplicity in ML algorithms calls into question their ability to attend all requisites for a lawful restriction of freedom of expression. As an example, a ML model trained with random seed 1 could misapply a restriction to protected speech (e.g. journalistic speech), whereas the same model trained with random seed 42 would have correctly tolerated the statement. Such an event would be equivalent to a judge flipping a coin to decide whether the speech should be protected or taken down. For example, in Section 5 we observe that varying the random seed causes fine-tuned large language models to assign conflicting toxic speech predictions to 34% of statements from a large dataset.

Non-Discrimination. We adopt Article 2(1) and Article 26 of the ICCPR as our definition of discrimination. They state:

> **Article 2** (1) Each State Party to the present Covenant undertakes to respect and to ensure to all individuals within its territory and subject to its jurisdiction the rights recognized in the present Covenant, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status.
>
> **Article 26** All persons are equal before the law and are entitled without any discrimination to the equal protection of the law. In this respect, the law shall prohibit any discrimination and guarantee to all persons equal and effective protection against discrimination on any ground such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status.

These articles are intended to protect individuals from discrimination based on protected characteristics. We note that content moderation ML algorithms can illegally discriminate against specific individuals or groups. A biased moderation occurs when specific groups have an inferior or higher quality of moderation of toxic speech depending on their characteristics. This can correlate with the dialect and content of the statements, as identified by Dias Oliva et al. [20] with content moderation in LGBTQ discussion spaces. Our experiments are able to infer the presence of discrimination by analyzing the targeted group of the statements. In particular, if the magnitude of predictive multiplicity in ML algorithms is *different* across groups, then such an ML algorithm is discriminatory.

In Section 5 we experimentally observe exactly this phenomena: varying the random seed causes fine-tuned large language models to assign conflicting toxic speech predictions to 38% of racial-based statements from a large scale dataset compared compared to 20% of misogynistic/misandrist statements.

Procedural Justice. The UN Guiding Principles on Business and Human Rights [57] emphasize that businesses should identify, prevent, and mitigate human rights abuses. The human right to due process is established by Article 14(1) of the ICCPR. We will focus on the first part of the Section, which is most relevant to our work:

> **Article 14**
> (1) All persons shall be equal before the courts and tribunals. In the determination of any criminal charge against him, or of his rights and obligations in a suit at law, everyone shall be entitled to a fair and public hearing by a competent, independent and impartial tribunal established by law. [...]

We interpret Articles 14 and 19 (aforementioned) as jointly demanding that a restriction of a fundamental right be *impartial*, *fair*, and *prescribed by law*. This means providing remedies through operational grievance mechanisms when harm occurs, and ensuring processes are transparent and accountable. When we translate this to ML models for content moderation, moderation needs to be *explainable*, *accountable*[5], and have a *rule-based* approach for limiting free speech. In this regard, the outcomes of ML models must attend to these legal requirements. This interpretation includes, for example, respecting the requirements from General Comment 34 [13](i.e. legality, necessity, proportionality, and pursuit of a legitimate aim) for restricting speech. This joint interpretation establishes the obligation of common procedural guidelines for removing speech.

The existence of predictive multiplicity in ML algorithms calls into question their ability to satisfy values of procedural justice. The "decision-making process" used by ML algorithms is fundamentally probabilistic and random. In this case, the "judges" of online speech are making random decisions (flipping coins) to determine whether to restrict speech or not, and flipping coins more often when it comes to speech from certain social groups. This violates procedural justice for three reasons. First, it does not respect a rule-based approach to restricting speech, as it is fundamentally random. Second, it is not impartial, as it is disparate across groups. Third, it is not accountable because this decision-making process is concealed, meaning we cannot know if a given prediction is an instance of

[5]We define an *accountable* model as a model that can be understood, challenged, scrutinized, and revised.

predictive multiplicity. [6] We emphasize that, because the source of the violation is randomness, this violation is *independent* of the final outcome being legally correct.

Experimentally Measuring Harm. To study multiplicity using the framework of legal harms we outlined above, we designed experiments to measure multiplicity and its potential harms quantitatively. We fine-tuned various state-of-the-art models for toxic speech detection, tested them across different datasets of toxic and non-toxic statements, and observed the incidence of predictive multiplicity across models. We also compared disagreement in models to disagreement in human annotation.

To quantify the **extent of predictive multiplicity** (RQ1), we compute *arbitrariness* (Definition 1) and *pairwise disagreement* (Definition 2) on our competing fine-tuned models and show the prevalence of arbitrary decisions in SOTA toxicity detectors (Table 1). Aiming to assess **how arbitrary decisions are spread across demographic groups** (RQ2) we compute arbitrariness and pairwise disagreement in sentences targeting specific social groups (Figure 1). Next, we provide the necessary theoretical background on predictive multiplicity (Section 3) and define the setup for the described experiments (Section 4). Finally, in Section 5, we display and analyze our experimental results.

3 BACKGROUND ON PREDICTIVE MULTIPLICITY

In this section, we discuss setup and notation, mathematically define the set of all competing models, which in the ML and statistics literature is called the *Rashomon set*, and define the multiplicity metrics of interest in this paper — pairwise disagreement and arbitrariness.

Preliminaries. We focus on the task of binary classification of toxic speech. Consider a dataset with $n \in \mathbb{N}$ examples $\mathcal{D} \triangleq \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ where $\mathbf{x}_i$ is a sentence (e.g., "I love you" and "I hate you") and $\mathbf{y}_i \in \{0, 1\}$ is a binary label that is 1 when the sentence is "Toxic" and 0 when it is "Not Toxic". In the open-source datasets used in this work, labels were generated by multiple human annotators (see appendix C.2 for details).

We use error to measure the quality of a model. Formally, the error of a model $h \in \mathcal{H}$ over a dataset $\mathcal{S} \subseteq \mathcal{D}$ is

$$\mathrm{Err}_{\mathcal{S}}(h) = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x},\mathbf{y} \in \mathcal{S}} \mathbb{1}\left[h(\mathbf{x}) \neq \mathbf{y}\right], \tag{1}$$

where $\mathbb{1}\left[\textit{condition}\right]$ is the indicator function that outputs 1 if *condition* is true and 0 otherwise. The error in training data is defined as $\mathrm{Err}_{\mathrm{train}}(h)$, and similarly for testing error.

Competing Models and the Rashomon Effect. We call a fixed (e.g., deployed) model for flagging toxic content a *reference model* and denote it by $h_{\mathtt{ref}}$ — we chose $h_{\mathtt{ref}}$ to be a language model freely available on HuggingFace. The reference model can be, for example, the empirical risk minimizer over a training set or an already deployed model. The set of all models with less than $1 + \epsilon$ times the training error from $h_{\mathtt{ref}}$ is the *Rashomon set* [8, 26] denoted by $\mathcal{R}(\epsilon, h_{\mathtt{ref}})$.[7] The Rashomon set can intuitively be viewed as a "disagreement set" of equally-accurate models. Formally, the Rashomon set is given by:

$$\mathcal{R}(\epsilon, h_{\mathtt{ref}}) \triangleq \{h \in \mathcal{H} \mid \mathrm{Err}_{\mathrm{train}}(h) \leq (1+\epsilon)\,\mathrm{Err}_{\mathrm{train}}(h_{\mathtt{ref}})\}, \tag{2}$$

where ϵ is the *Rashomon parameter* and measures how close the performance of the models is to the performance of the reference model, see [7, 26, 37, 48] for related definitions. For the LLMs considered in this work, the Rashomon set is theoretically and computationally challenging to characterize. We resort to empirically estimating the Rashomon set via re-running the same fine-tuning pipeline with different random seeds. Each fine-tuned model gives us a sample from the Rashomon set if the model is close in performance to the reference model. We denote these *Rashomon set model samples* by $\widehat{\mathcal{R}}(\epsilon)$ when $h_{\mathtt{ref}}$ is clear from the context. In practice, to explore the Rashomon set, we fix a dataset $\mathcal{D}_{\mathrm{train}}$ and model architecture $\mathcal{H}$, and fine-tune as many models on $\mathcal{D}_{\mathrm{train}}$ as our computational resources allow, each time varying the random seed. We discard any models that are not within ϵ of $h_{\mathtt{ref}}$.

There is no standard Rashomon parameter selection method (ϵ). Most papers on predictive multiplicity resort to showing how results vary when the Rashomon parameter is changed [6, 37, 44, 48, 60]. Recently, Paes et al. [54] proposed a principled manner of choosing the Rashomon parameter based on Clopper-Pearson confidence intervals. This approach — which we refer to as the CP method — selects ϵ based on a confidence parameter, dataset size, and the error of the reference model. We follow their approach using a confidence parameter of 95% for a conservative analysis. We also explore different confidence values in appendix D.

Measuring Predictive Multiplicity. A classification problem exhibits predictive multiplicity when models in the Rashomon set assign conflicting predictions to the same data point, formally defined in Marx et al. [48, Definition 2]. To measure predictive multiplicity, we use the following two metrics: arbitrariness, which is a generalization of ambiguity Marx et al. [48], and pairwise disagreement [7, 18].

While ambiguity computes the fraction of points that at least one model in the Rashomon set disagrees with the reference model ($h_{\mathtt{ref}}$), arbitrariness measures the percentage of points in the dataset that receive conflicting predictions from any two models in the Rashomon set (competing models) and it is formally defined next.

Definition 1 (Arbitrariness). *The arbitrariness on a set of inputs $\mathcal{S} = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \subseteq \mathcal{D}$ over the Rashomon set model samples $\widehat{\mathcal{R}}(\epsilon, h_{\mathtt{ref}})$ is the proportion of inputs in the set $\mathcal{S}$ that receive conflicting predictions from any two models in the Rashomon set model samples:*

$$\widehat{\mathbb{A}}(\epsilon) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[\exists h_1, h_2 \in \widehat{\mathcal{R}}(\epsilon) | h_1(\mathbf{x}_i) \neq h_2(\mathbf{x}_i)]. \tag{3}$$

Pairwise disagreement is a per-sample measure that approximates the fraction of models in the Rashomon set that disagree on a prediction. Here, we analyze average pairwise disagreement, which

[6] In fact, this information is impossible to obtain even if we analyze the model alone, as multiplicity can only be identified when we compare predictions across multiple models.

[7] Reference [26] defines the Rashomon set with any arbitrary loss function evaluated on the training data.

averages out pairwise disagreement across samples in a dataset, formally defined as follows.

Definition 2 (Average Pairwise Disagreement [7, 18]). *The average pairwise disagreement is the average over all input $\mathbf{x} \in \mathcal{D}$ of the proportion of pairs of models in the Rashomon set $\widehat{\mathcal{R}}(\epsilon, h_{\texttt{ref}})$ that disagree on their prediction:*

$$\overline{\texttt{PD}}(\epsilon) \triangleq \frac{1}{n}\sum_{i=1}^{n} \frac{1}{M(M-1)} \sum_{h,h' \in \widehat{\mathcal{R}}(\epsilon)} \mathbb{1}[h(\mathbf{x}) \neq h'(\mathbf{x})], \qquad (4)$$

where $M = |\widehat{\mathcal{R}}(\epsilon)|$, i.e., M is the number of models we sample from the Rashomon set via retraining and n the number of examples in $\mathcal{D}$.

We select the above metrics because they quantify two important aspects of predictive multiplicity: (i) the fraction of samples in a dataset for which predictions are arbitrary (Defn. 1), in that a competing model would have assigned a different prediction, and (ii) the extent to which models disagree on individual (Defn. 2).

Given a set of models sampled from the Rashomon Set (e.g., by varying random seeds), we quantify predictive multiplicity in two steps. First, we measure the number of arbitrary decisions (arbitrariness) made by competing models. Here, arbitrariness captures *how many moderation decisions were not rule-based but just a consequence of random seed selection.* As discussed in Section 2, such random decisions go against procedural fairness because they violate due process, are not *accountable*, and, if the magnitude of arbitrariness is different across groups, then the impact of randomness is also *disparate*. Second, we compute pairwise disagreement to estimate the number of models that disagree on their predictions. If the number of conflicting predictions was, on average, negligible, one might argue that ignoring this conflicting minority is acceptable [6]. However, our experimental results show that such disagreement is high (Table 1), especially in specific targeted demographic groups (Figure 1). In the next section, we apply this measurement pipeline to state-of-the-art toxic text detectors.

4 EXPERIMENTAL SETUP

This section outlines the datasets, ML models, and methodology used for evaluating predictive multiplicity in content moderation. Our goal is to describe our overall experimental approach and provide a rationale for the choice of datasets and base LLMs.

Our experiments involve *fine-tuning* state-of-the-art language models on large-scale datasets. Fine-tuning refers to the act of taking a *general-purpose* LLM trained on a large corpus of text, e.g. RoBERTa [45], and further training it on a *specific* objective, such as toxicity classification. Typically, this training is shorter (fewer epochs) and less intense (smaller learning rate, less updated layers) than the original training (commonly called pre-training) — which is what motivates the term *fine-tuning*. All language models referred to in this section have been fine-tuned for toxicity classification, meaning they take as input a piece of text and output either 0, denoting a non-toxic, or a 1, denoting a toxic.

On state-of-the-art model selection. We identify widely used [8] state-of-the-art open-source language models that have been fine-tuned for toxicity detection. We test all these models in four different datasets and choose to analyze the models `tomh TR`[33] and `s-nlp RTC`[17] that are, respectively, the first and second best-performing models — check Appendix C.3 for more details on model selection and Table 3 for the considered models accuracy. Throughout this paper we will refer to `tomh TR`[33] as ToxiGen-RoBERTa and `s-nlp RTC`[17] as RoBERTa-Toxicity-Classifier.

On state-of-the-art model reproduction via fine-tuning. We reproduce the fine-tuning procedure from ToxiGen-RoBERTa [33] 40 times with different random seeds, leading to 40 different models. We only considered 35 out of the 40 models because they have indistinguishable performance from the reference model with 95% confidence using the method from [54] —- the Rashomon parameter is $\epsilon = 0.016$. We repeat the same procedure for the RoBERTa-Toxicity-Classifier using the fine-tuning method from [17] and retaining 18/20 models with statistically indistinguishable performance with 95% confidence. Appendix C.4 shows hyperparameters and C.5 the performance of fine-tuned models.

On dataset selection. We use the publicly available datasets: ToxiGen [33], DynaHate [63], SBF (Social Bias Frames) [59], HateExplain [49], MHS (Measuring Hate Speech) [40], and WikiDetox [70]. These datasets were chosen for two main reasons. First, they were purposefully designed to challenge ML-based toxic text classification. For example, ToxiGen and SocialBiasFrames (SBF) contain mostly "implicit" toxic speech [33, 59]. Second, these datasets have labels for demographic groups targeted by the text. We use this information to quantify and compare Arbitrariness and Pairwise Disagreement across different targeted groups (Figure 1). We use the Measuring Hate speech (MHS) [40] and the WikiDetox [70] datasets because they add one additional dimension to our analysis: the labels of multiple human annotators who detected toxicity for the sentences in the dataset. This enablels us to compare human annotators' disagreement with model disagreement (Figure 2).

Dataset Content. In total, each row in each dataset in this work contains: a sentence, a list of binary yes/no votes from human annotators regarding the toxicity of the sentence, and the target group for the sentence. See Appendix C.2 for further details, including how many human annotators are in each dataset.

Having fine-tuned our models, in the next section, we will present how these models exhibit predictive multiplicity in accordance with the mathematical formulation in Section 3. For each of our findings, we also draw connections between our experimental results and their impact on principles of procedural justice, freedom of expression, and non-discrimination, based on the legal framework outlined in Section 2.

5 DATA ANALYSIS

In this section, we present our experimental results and discuss their meaning in terms of the principles defined in Section 2. As we did in Section 2, we will often refer to the illustration of a judges flipping coins.

[8] We consider widely used all toxicity detectors with more than 3000 downloads in the Hugging Face [69] platform.

5.1 Procedural Justice, Freedom of Expression, and Judges Flipping Coins

Technical Analysis. Our first experimental result regards the extent of arbitrariness (RQ1) (defined in (1)) and disagreement (defined in (2)) in our fine-tuned state-of-the-art toxicity detectors. Table 1 shows the prevalence of arbitrariness for the fine-tuned Toxigen and Jigsaw models across all tested datasets. We also observe that for the fine-tuned Toxigen, more than 34% of all decisions made by the models at the test time are arbitrary, i.e., there exists another competing model with a conflicting prediction. For the fine-tuned Jigsaw models, this number decreases to closer to 23%. Moreover, both the fine-tuned Toxigen and Jigsaw models achieved a high number of conflicting predictions in the SBF dataset that contains *implicit* toxic content — which may indicate that when the toxicity is implicit, arbitrary decisions are more common.

Table 1 also shows a high percentage of pairwise disagreement for the fine-tuned Toxigen and Jigsaw models across all tested datasets. Our experiments show that using the fine-tuned Toxigen models, on average, 8.3% of the pair of models disagree in their prediction — i.e., 8.3% of total pairwise disagreement. While 6.9% of the pair of models disagree for the fine-tuned Jigsaw models. This implies that, on average, for each point that models disagree, 14% of the fine-tuned Toxigen models made a prediction about sentence toxicity, and 86% of the models predicted the opposite. This high pairwise disagreement is especially relevant for methods that aim to decrease arbitrary decisions by taking a majority vote across fine-tuned competing models such as [6].

A Violation of Procedure and Freedom of Expression. Using our analogy, each ML model is an adjudicator, deciding whether to strike down an online post or not. Recall that the models we developed and tested are part of a Rashomon set, meaning they all have very similar accuracy and are, therefore, equally good. On average, all judges make the same number of correct rulings. However, in 34% of court cases, at least two judges disagree on the ruling (arbitrariness). These conflicting rulings are *not* a result of judges having different interpretations of the law or or having different ideologies (e.g., more or less punitive). These conflicts stem from purely random events, e.g., in 34% of court cases the judge flips a coin to decide whether to take down the online post or not. Per Section 2, such decisions are entirely detached from notions of due process, legality, and impartiality, and hence constitute a violation of procedure and freedom of expression. The fact that we measure a 34% arbitrariness value due solely to *random events* means these ML models, if deployed in the real world, would blatantly violate procedure and FoE (as defined in Section 2). We emphasize that if the 34% arbitrariness value could be attributed to *clear* and *explainable* differences in decision-making, then this value would not be a violation of procedure and FoE. The *randomness* is the source of the violation, not the magnitude of the value.

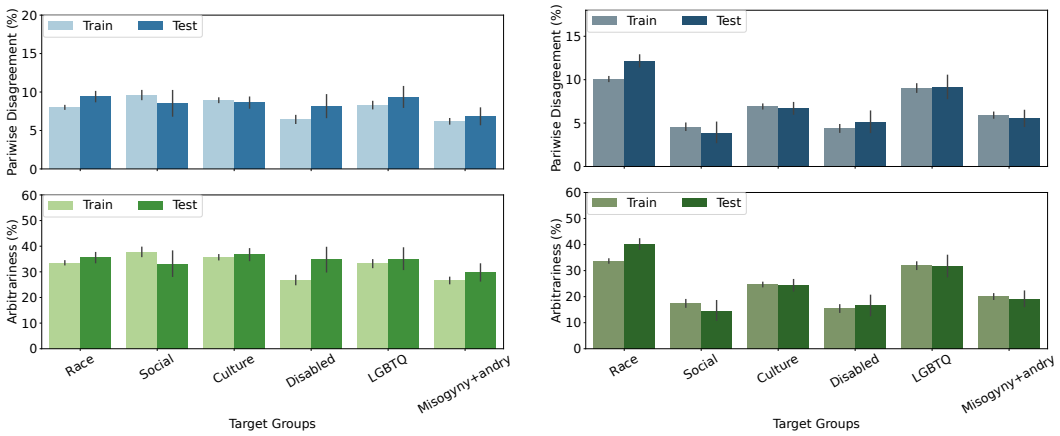
5.2 Disparate Arbitrariness: Different Content Gets Different Coin Flips

Technical Analysis. Figure 1 indicates that the incidence of arbitrariness is not the same across all targeted groups (RQ2). We observe that anti-LGBTQ speech consistently receives more arbitrary decisions relative to misogynist /misandrist speech for both Toxigen and Jigsaw fine-tuned models. Across the Toxigen fine-tuned models, anti-LGBTQ speech receives arbitrary decisions 35% of the time, while misogynist/misandrist speech receives arbitrary decisions around 30% of the time. These differences are even greater on Jigsaw fine-tuned models. Moreover, racist speech has more than twice the arbitrariness of misogynist/misandrist speech on Jigsaw fine-tuned models. We also note that Toxigen was created to be a balanced dataset, i.e., all target groups have about the same number of examples — Figure 1 shows that a balanced dataset may make arbitrary decisions more evenly distributed.

A Violation of Non-discrimination. Returning to the judge analogy, our experimental results indicate that decisions based on coin flips occur more frequently in certain marginalized groups than in others. An example would be that in 35% of court cases concerning LGBTQ content, the judge flips a coin to decide the outcome, whereas the judge does this only 30% of the time for misogynist and misandrist content. The fact that we measure a difference in arbitrariness values across different groups due solely to *random events* means these ML models, if deployed in the real world, would violate non-discrimination. Unlike Section 5.1, even if this effect could be attributed to *clear* and *explainable* differences in model decision-making, it would still constitute a violation of the principle of non-discrimination. People are entitled to a rule-based and equal evaluation on whether their speech should be restricted.

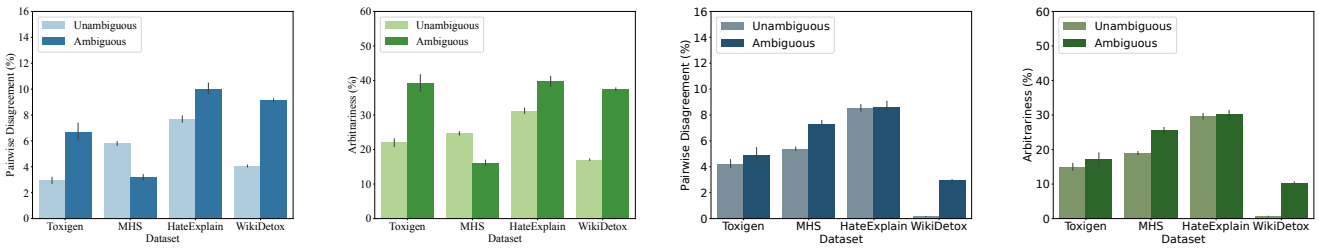
Table 1: Average pairwise disagreement and arbitrariness in testing time for the Toxigen fine-tuned and Jigsaw fine-tuned models in different datasets. The confidence in the CP method from [54] was chosen to be 95% for a more conservative analysis. 95% confidence intervals are shown using the standard error from the mean.

	Toxigen Fine-Tuned		Jigsaw Fine-Tuned	
Dataset	Pairwise Disagreement	Arbitrariness	Pairwise Disagreement	Arbitrariness
Toxigen	6.8% ± 0.9%	28.6% ± 3.2%	4.3% ± 0.8%	15.4% ± 2.5%
DynaHate	8.4% ± 0.6%	34.1% ± 1.6%	6.0% ± 0.4%	21.8% ± 1.4%
SBF	8.7% ± 0.3%	35.7% ± 1.1%	7.2% ± 0.3%	24.4% ± 1.0%
HateExplain	8.0% ± 0.6%	31.9% ± 2.0%	8.5% ± 0.6%	29.6% ± 2.0%
Total	8.4% ± 0.2%	34.2% ± 0.8%	6.9% ± 0.2%	23.9% ± 0.7%



(a) Toxigen Fine-Tuned (b) Jigsaw Fine-Tuned

Figure 1: Average pairwise disagreement and arbitrariness in different target groups for the fine-tuned Toxigen and Jigsaw models. The results show the pairwise disagreement in percentage (x-axis) for the union of four different datasets: DynaHate, SBF, Toxigen, and HateExplain. The results are shown for training and test partitions of each dataset. The confidence in the CP methods was chosen to be 95%.



(a) Fine-Tuned Models Toxigen (b) Fine-Tuned Models Jigsaw

Figure 2: Average pairwise disagreement and arbitrariness for *Unambiguous* and *Ambiguous* sentences using the Toxigen fine-tuned and Jigsaw fine-tuned models. The figure shows the pairwise disagreement estimated values along with the 95% confidence intervals using the standard error from the mean. We consider a sentence *Ambiguous* when at least one annotator labeled the sentence differently than others and *Unambiguous* otherwise. The confidence in the CP methods was chosen to be 95%.

5.3 Comparing Human and Machine Arbitrariness: Who is Flipping Coins?

Finally, we compare the arbitrariness across competing ML models in the Rashomon set and across human annotators. Our goal is to verify if disagreements in predictions between fine-tuned LLMs match the disagreement observed in human annotators, in which case ML models would be replicating disagreement already present in the training data. As we see next, that is not the case.

Technical Analysis. From Figure 2, we observe two results. *First*, model disagreement tends to be higher in sentences where humans do not agree (i.e., ambiguous statements). This is an interesting finding because the models only saw the majority vote across the annotators. Effectively, this implies that models, as humans, struggle with classifying certain statements. Our *second* finding is that models in the Rashomon set can display a high level of disagreement and, hence, arbitrariness in sentences in which humans *unanimously* agreed on the toxicity (i.e., unambiguous statements). In these cases, models output conflicting predictions when faced with evaluations that would be obvious to human annotators. Note that WikiDetox is part of the training data for the fine-tuned Jigsaw models, which is why the arbitrariness and disagreement values are noticeably small. Even in this extreme, our *first* observation holds. This is further evidence that there are certain statements in these datasets that both humans and models struggle to correctly classify.

A violation of rule-based approach. This point is where our analogy (un)fortunately reaches a limitation. In real life, judges will make similar decisions on most easy cases. The lawfulness of the

social media post stating "Hello world" is hardly controversial. However, our findings indicate that ML models struggle over certain statements that would be obvious to any human judge. In their *de facto* roles of proxy adjudicators, we need to interrogate in which situations ML can effectively deliver high quality classification.

6 ON THE CONSEQUENCES OF ARBITRARINESS

In this section, we discuss how selecting and deploying content moderation models at a large scale, under predictive multiplicity, resembles the moral dilemma of the trolley problem (RQ3). We make this comparison to discuss the relevance of our work for law and policy decision-makers and scholars. Finally, we offer insights into the path forward.

The inscrutable trolley problem. The harms of arbitrariness reflect a fundamental problem for the use of ML in content moderation. We argued in Section 2 under *Procedural Justice* that content moderation algorithms must follow a rule-based *decision-making process*. ML models, however, make statistical predictions that do not follow a clear rule-based process. Adjudicating free speech through statistical models to control the exercise of a right is only tolerable if these models deliver similar expected outcomes and operate on the same explainable, rule-based criteria, with due process safeguards. We have empirically shown that is not the case. The criteria used are often random and these stochastic effects are often concealed from the end user.

Our work also identifies the harms stemming from arbitrary model selection (e.g., which model of the Rashomon set is chosen and deployed). When there is no clear reason for choosing one model over the other, an artificial "lottery" is created on which data points will draw the fate of being subjected to random treatment. Our results indicate that this "lottery" is not fair: different population groups targeted by the text have different likelihoods of arbitrary treatment.

If we can draw a final analogy, this creates a troubling scenario where choosing ML models is an inscrutable trolley problem. The trolley problem is a famous thought experiment in ethics and psychology involving a moral dilemma where a person must choose between actively diverting a runaway trolley to harm one person or passively allowing it to continue on its path and harm five people. Here, we do not know *why* and *how* companies choose between equally good models, but each one of them will cause the undue harm of different individuals. This must be discussed from a law and policy point of view in local jurisdictions.

Impact on ongoing law and policy debates on content moderation. One important debate in the platform regulation field is directly affected by these findings. It is the ongoing discussion of laws affecting content moderation, such as platform liability rules [9, 46]. Legal responsibilities imposed on service providers push companies to perform more content moderation focused on particular types of expression. Striking the right balance between free speech and expedited response, considering the volume and plurality of online communication, is a hard, legal and technical task. Adding to these challenges, copyright claims, scientific disinformation, electoral integrity, and online extremism are all topics that have fuelled heated discussions on the need to prevent online harms while balancing international human rights - or even questioning if international human rights are sufficient to tackle this issue [22]. Our findings shed light on the legal complexities intrinsic to these models.

Several statutes require companies to publish assessment reports that include quantitative measurements such as expected accuracy and error in algorithmic content moderation. Examples include DSA (Article 15, Section 1(e))[62], the UK Online Safety Act (Section 22 (4) and (6))[52], and statutes currently in discussion, such as the Brazilian AI Bill (Articles 19 - 24) and the Platform Regulation Bill (Article 23). Our findings demonstrate that arbitrariness in algorithmic content moderation carries non-negligible potential for harm. Moreover, content moderation tasks are also becoming increasingly complex, which may further increase arbitrariness. For instance, in 2022 the Brazilian Electoral Courts [24] ordered the removal of content that was "similar" to content that had a previously been appreciated with a removal order. The time-frame for companies to respond, in the election periods, varied between 3 hours and 1 hour. To attend to these stringent legal requirements, companies might rely on other ML models to appreciate "similarity" at scale (whatever that might mean).

As a way forward, reports produced by companies should also include measurements of algorithmic arbitrariness in content moderation. Reporting accuracy alone is not enough, as our work demonstrated: equally accurate models can produce conflicting moderation decisions. Our paper provides a methodology on how such measurements can be done: in Section 3, we give two precise metrics for quantifying arbitrariness, in Figures 1 and 2, we compute and visualize these quantities, and in the "Technical Analysis" of Section 5, we provide an example of what a quantitative assessment could look like in these reports.

We encourage the ML community to develop mitigation strategies to reduce algorithmic arbitrariness in content moderation. The work of Black et al. [6] is an excellent starting point, which suggests taking the majority vote across the set of essentially equally performing models. However, we note that the first step in dealing with arbitrariness is discovering, measuring, and reporting this phenomenon.

7 FINAL DISCUSSION

Conclusion. In this paper, we show the prevalence of arbitrary decisions in algorithmic content moderation and discuss its impact on law and policy — particularly on freedom of expression, non-discrimination, and procedural justice. Moreover, we show that arbitrary decisions are not uniformly spread across all texts and that they are more frequent in content that targets specific demographic groups (e.g., anti-LGTBQ posts). Then, we discuss the implications of the disparate arbitrary decisions in terms of the principle of non-discrimination and procedural fairness. Finally, we also show that models produce arbitrary predictions even in content that human annotators unanimously classify as toxic or non-toxic, signaling that it might be useful to share the responsibility of content moderation with human annotators.

Path forward. Our results reinforce that ML models are far from perfect proxies for humans when classifying and evaluating speech.

The use of algorithmic tools in moderation must be nuanced, accountable, and transparent. First, decisions made during model development – even as simple as the choice of random seed! – must be scrutinized, and their impact on ensuing moderation decisions quantified, reported, and analyzed in light of company policies and regulations. Second, developers of moderation tools should measure how arbitrariness disparately affects subsets of the population and develop techniques to mitigate this impact. Finally, caution is required when delegating decisions to algorithms. A more nuanced approach to content moderation, where certain variables (e.g., thematic content, socio-demographic factors, type of illegal or harmful speech) prompt human revision and control, is a promising way forward.

Limitations. We only measured multiplicity across binary toxicity detection. However, models that predict beyond binary toxicity (e.g., models that predict the level of toxicity) could potentially display different levels of arbitrariness than reported here. We also did not investigate the possibility of a statement fulfilling multiple categories of toxic speech; different categories may prompt different governance decisions other than simply content removal (e.g., reducing reach and labeling). Finally, there is an emerging application of generative language models to produce moderation decisions [53]; however, this approach uses GPT-4, which we have no access to model architecture, weights, or even training data, making finding equally good models in the same hypothesis class impossible.

8 RESEARCH POSITIONALITY

This positionality statement aims to transparently communicate our ethical considerations, the influence of our backgrounds on our research, and our proactive steps to mitigate the risks of our research.

Ethical considerations: In conducting our research, we addressed ethical concerns related to the collection and analysis of potentially harmful content. We took precautions to mitigate risks associated with exposure to toxic speech by redacting sensitive language and analyzing data at an aggregate level. This approach minimized the direct exposure of our research team to potentially disturbing content. Furthermore, to ensure the integrity of our research practices and maintain a clear ethical stance, our research center operates independently, without direct funding from companies that might be influenced by policy discussions stemming from our findings. This independence allows us to conduct our research without potential conflicts of interest, adhering strictly to academic and ethical standards.

Positionality: As a team linked to an American university, our diverse backgrounds and experiences inform and shape our research. The policy discussions we engage with are prominent in Europe and North and South America. We have limited knowledge of the state of the discussion in other continents. One of our researchers is actively involved in a civil society organization focused on tech policy in Brazil, which informs our understanding of the implications of technology governance and views on the legislation mentioned. In our view, the diversity of disciplinary views provides valuable insights into the socio-political dynamics that frame technology use and regulation in different regions, particularly in emerging markets.

Adverse Impact Reflection: While we believe our research does not directly have adverse unintended impacts on individuals, we remain cautious about the potential misuse of our findings. Specifically, the models trained to identify toxic speech have the inherent potential to be misused to restrict the freedom of expression of certain demographic groups. Acknowledging this possibility, we have decided to restrict access to the code, making it available only upon request. This measure is intended to prevent misuse and ensure that the models are used in line with ethical guidelines and for purposes that align with our intent to promote positive social outcomes. We continue to reflect on the broader implications of our research and remain committed to monitoring and addressing any negative impacts that may arise post-publication.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under grants CAREER 1845852, CIF 1900750, CIF 2312667 and FAI 2040880. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship under Award Number DE-SC0022158. Caio Machado thanks the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) for partly funding this study and his visit at Harvard SEAS (Finance Code 001).

REFERENCES

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2623–2631. https://doi.org/10.1145/3292500.3330701

[2] United Nations General Assembly. 1966. *International Covenant on Civil and Political Rights.* https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights

[3] United Nations General Assembly. 2022. Tackling Online Hate Speech through Content Moderation: The Legal Framework Under the International Covenant on Civil and Political Rights. https://papers.ssrn.com/abstract=4150909

[4] Reuben Binns, Jeremias Adams-Prassl, and Aislinn Kelly-Lyth. 2023. Legal Taxonomies of Machine Bias: Revisiting Direct Discrimination. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 1850–1858. https://doi.org/10.1145/3593013.3594121

[5] Emily Black, John Logan Koepke, Pauline Kim, Solon Barocas, and Mingwei Hsu. forthcoming. Less Discriminatory Algorithms. *Washington University in St. Louis Legal Studies* 2 (forthcoming). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4590481

[6] Emily Black, Klas Leino, and Matt Fredrikson. 2022. Selective Ensembles for Consistent Predictions. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HfUyCRBeQc

[7] Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 850–863.

[8] Leo Breiman. 2001. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statist. Sci.* 16, 3 (2001), 199 – 231. https://doi.org/10.1214/ss/1009213726

[9] Miriam C. Buiten. 2022. The Digital Services Act: From Intermediary Liability to Platform Regulation. 12, 5 (2022). https://www.jipitec.eu/issues/jipitec-12-5-2021/5491

[10] Deutscher Bundestag. 2017. NetzDG - Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken. https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html

[11] cjadams, inversion Daniel Borkan, Jeffrey Sorensen, Lucas Dixon, Lucy Vasserman, and nithum. 2019. Jigsaw Unintended Bias in Toxicity Classification. https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification

[12] cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. 2017. Toxic Comment Classification Challenge. https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge

[13] United Nations Human Rights Committee. 2011. General comment No.34 on Article 19: Freedoms of opinion and expression. https://www.ohchr.org/en/documents/general-comments-and-recommendations/general-comment-no34-article-19-freedoms-opinion-and

[14] Nicholas Kluge Corrêa. 2023. Aira. https://doi.org/10.5281/zenodo.6989727

[15] Kathleen Creel and Deborah Hellman. 2021. The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision Making Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 816. https://doi.org/10.1145/3442188.3445942

[16] Kathleen Creel and Deborah Hellman. 2022. The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems. *Canadian Journal of Philosophy* 52, 1 (2022), 26–43. https://doi.org/10.1017/can.2022.3

[17] David Dale, Igor Markov, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. SkoltechNLP at SemEval-2021 Task 5: Leveraging Sentence-level Pre-training for Toxic Span Detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. 927–934.

[18] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2022. Underspecification presents challenges for credibility in modern machine learning. *The Journal of Machine Learning Research* 23, 1 (2022), 10237–10297.

[19] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media* 11, 1 (May 2017), 512–515. https://doi.org/10.1609/icwsm.v11i1.14955

[20] Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. 25, 2 (2021), 700–732. https://doi.org/10.1007/s12119-020-09790-w

[21] Evelyn Douek. 2020. Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability. https://doi.org/10.2139/ssrn.3679607

[22] Evelyn Douek. 2021. The Limits of International Law in Content Moderation. *UC Irvine Journal of International, Transnational, and Comparative Law* 6 (2021). https://scholarship.law.uci.edu/cgi/viewcontent.cgi?article=1042&context=ucijil

[23] Natasha Duarte and Emma Llansó. 2017. Mixed Messages? The Limits of Automated Social Media Content Analysis. https://cdt.org/insights/mixed-messages-the-limits-of-automated-social-media-content-analysis/

[24] Tribunal Superior Eleitoral. 2022. RESOLUÇÃO Nº 23.714, DE 20 DE OUTUBRO DE 2022. https://www.tse.jus.br/legislacao/compilada/res/2022/resolucao-no-23-714-de-20-de-outubro-de-2022

[25] Mohsen Fayyaz. 2021. Toxicity Classifier. https://huggingface.co/mohsenfayyaz/toxicity-classifier. Accessed: 2024-01-21.

[26] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2019. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research* 20, 177 (2019), 1–81.

[27] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7, 2 (July 2020), 205395172094323. https://doi.org/10.1177/2053951720943234

[28] Tarleton Gillespie. 2022. Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society* 8, 3 (July 2022), 205630512211175. https://doi.org/10.1177/20563051221117552

[29] João Gonçalves, Ina Weber, Gina M. Masullo, Marisa Torres Da Silva, and Joep Hofhuis. 2023. Common sense or censorship: How algorithmic moderators and message type influence perceptions of online content deletion. *New Media & Society* 25, 10 (Oct. 2023), 2595–2617. https://doi.org/10.1177/14614448211032310

[30] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (Jan. 2020), 2053951719897945. https://doi.org/10.1177/2053951719897945 Publisher: SAGE Publications Ltd.

[31] Mads Haahr. 1998-2018. RANDOM.ORG: True Random Number Service. https://www.random.org. Accessed: 2018-06-01.

[32] Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

[33] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3309–3326. https://doi.org/10.18653/v1/2022.acl-long.234

[34] Ari Heljakka, Martin Trapp, Juho Kannala, and Arno Solin. 2023. Disentangling Model Multiplicity in Deep Learning. arXiv:2206.08890 [cs.LG]

[35] Saghar Hosseini, Hamid Palangi, and Ahmed Hassan Awadallah. 2023. An Empirical Study of Metrics to Measure Representational Harms in Pre-Trained Language Models. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, Anaelia Ovalle, Kai-Wei Chang, Ninareh Mehrabi, Yada Pruksachatkun, Aram Galystan, Jwala Dhamala, Apurv Verma, Trista Cao, Anoop Kumar, and Rahul Gupta (Eds.). Association for Computational Linguistics, Toronto, Canada, 121–134. https://doi.org/10.18653/v1/2023.trustnlp-1.11

[36] The White House. 2020. Executive Order on Preventing Online Censorship – The White House. https://trumpwhitehouse.archives.gov/presidential-actions/executive-order-preventing-online-censorship/

[37] Hsiang Hsu and Flavio Calmon. 2022. Rashomon Capacity: A Metric for Predictive Multiplicity in Classification. In *36th Proceedings of Neural Information Processing Systems*. Curran Associates Inc.

[38] Daphne Keller. 2019. Build Your Own Intermediary Liability Law: A Kit for Policy Wonks of All Ages. https://cyberlaw.stanford.edu/publications/build-your-own-intermediary-liability-law-kit-policy-wonks-all-ages

[39] Daphne Keller. 2021. *Empirical Evidence of Over-Removal by Internet Companies Under Intermediary Liability Laws: An Updated List.* https://cyberlaw.stanford.edu/blog/2021/02/empirical-evidence-over-removal-internet-companies-under-intermediary-liability-laws

[40] Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277* (2020).

[41] Ian Kivlichan, Jeffrey Sorensen, Julia Elliott, Lucy Vasserman, Martin Görner, and Phil Culliton. 2020. Jigsaw Multilingual Toxic Comment Classification. https://kaggle.com/competitions/jigsaw-multilingual-toxic-comment-classification

[42] Olivier Knox. 2023. Analysis | Biden calls for changing Big Tech moderation rules. But not how. *Washington Post* (Jan. 2023). https://www.washingtonpost.com/politics/2023/01/12/biden-calls-changing-big-tech-moderation-rules-not-how/

[43] Pascal D König. 2020. Dissecting the algorithmic leviathan: On the socio-political anatomy of algorithmic governance. *Philosophy & Technology* 33, 3 (2020), 467–485.

[44] Bogdan Kulynych, Hsiang Hsu, Carmela Troncoso, and Flavio Calmon. 2023. Arbitrary Decisions are a Hidden Cost of Differentially Private Training. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1609–1623. https://doi.org/10.1145/3593013.3594103

[45] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[46] Caio C. V. Machado and Thaís Helena Aguiar. 2023. Emerging Regulations on Content Moderation and Misinformation Policies of Online Media Platforms: Accommodating the Duty of Care into Intermediary Liability Models. 8, 2 (2023), 244–251. https://doi.org/10.1017/bhj.2023.25

[47] Alice E. Marwick and Ross Miller. 2014. Online Harassment, Defamation, and Hateful Speech: A Primer of the Legal Landscape. *Fordham Center on Law and Information Policy Report No. 2* (2014). https://ssrn.com/abstract=2447904

[48] Charles Marx, Flavio Calmon, and Berk Ustun. 2020. Predictive Multiplicity in Classification. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 6765–6774. https://proceedings.mlr.press/v119/marx20a.html

[49] 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 *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 14867–14875.

[50] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3, 2 (2016), 2053951716679679. https://doi.org/10.1177/2053951716679679 arXiv:https://doi.org/10.1177/2053951716679679

[51] David Morar and Bruna Martins dos Santos. 2020. *The push for content moderation legislation around the world.* https://www.brookings.edu/articles/the-push-for-content-moderation-legislation-around-the-world/

[52] King's Printer of Acts of Parliament. 2023. Online Safety Act 2023. https://www.legislation.gov.uk/ukpga/2023/50/contents/enacted

[53] OpenAI. [n. d.]. Using GPT-4 for content moderation. https://openai.com/index/using-gpt-4-for-content-moderation. Accessed: 2024-05-01.

[54] Lucas Monteiro Paes, Rodrigo Cruz, Flavio Calmon, and Mario Diaz. 2023. On the Inevitability of the Rashomon Effect. In *2023 IEEE International Symposium on Information Theory (ISIT)*. 549–554. https://doi.org/10.1109/ISIT54713.2023.10206657

[55] Martin Pan. [n. d.]. Toxic Content Model. https://huggingface.co/martin-ha/toxic-comment-model. Accessed: 2024-01-21.

[56] Filippo Raso, Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz, and Levin Kim. 2018. Artificial Intelligence & Human Rights | Berkman Klein Center. http://nrs.harvard.edu/urn-3:HUL.InstRepos:38021439http://nrs.harvard.edu/urn-3:HUL.InstRepos:38021439

[57] United Nations Human Rights. 2012. Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework. https://www.ohchr.org/en/publications/reference-publications/guiding-principles-business-and-human-rights

[58] 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*, Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser, and Alexandra Uma (Eds.). European Language Resources Association, Marseille, France, 83–94. https://aclanthology.org/2022.nlperspectives-1.11

[59] 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*. Association for Computational Linguistics, Online, 5477–5490. https://doi.org/10.18653/v1/2020.acl-main.486

[60] Lesia Semenova, Harry Chen, Ronald Parr, and Cynthia Rudin. 2023. A Path to Simpler Models Starts With Noise. In *37th Proceedings of Neural Information Processing Systems*. https://openreview.net/forum?id=Uzi22WryyX

[61] Florida Senate. 2021. The Florida Senate. https://www.flsenate.gov/Session/Bill/2021/7072/?Tab=BillHistory

[62] European Union. 2023. The EU's Digital Services Act. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en

[63] Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. 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 (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 1667–1682. https://doi.org/10.18653/v1/2021.acl-long.132

[64] Sandra Wachter and Brent Mittelstadt. 2019. A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI. (2019). https://doi.org/10.7916/D8-G10S-KA92 Publisher: Columbia University.

[65] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review* 41 (2021), 105567. https://doi.org/10.1016/j.clsr.2021.105567

[66] Jamele Watson-Daniels, David Parkes, and Berk Ustun. 2023. Predictive Multiplicity in Probabilistic Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence,*, Vol. 37. 10306–10314.

[67] Wikipedia. [n. d.]. No Personal Attacks — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Wikipedia:No_personal_attacks [Online; accessed 1-May-2024].

[68] Wikipedia. [n. d.]. Research: Online harassment resource guide — Wikipedia, The Free Encyclopedia. https://meta.wikimedia.org/wiki/Research:Online_harassment_resource_guide [Online; accessed 1-May-2024].

[69] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

[70] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1391–1399. https://doi.org/10.1145/3038912.3052591

[71] Jillian York and Corynne McSherry. 2019. *Content Moderation is Broken. Let Us Count the Ways.*

[72] Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah A. Smith. 2021. Challenges in Automated Debiasing for Toxic Language Detection. In *EACL*.

A PRELIMINARIES

In this supplementary material, we provide the following information:

- Section B contains Figure 3, which summarizes the main technical arguments presented in the paper along with the policy implications.
- Section C provides further details on the datasets, Hugging Face models evaluation, fine-tuning procedures, and fine-tuned model performance.
- Section D provides a further exploration of our experiment. Particularly, it shows (i) multiplicity metrics across different dataset partitions, (ii) pairwise disagreement and arbitrariness values across different datasets, and (iii) multiplicity metrics across demographics for different confidence values from the CP method.

B SUMMARY OF KEY ARGUMENTS

The connection between the technical aspects of ML and law, along with the associated policy implications is illustrated in Figure 3.

C EXPERIMENT DESIGN DETAILS

In this section, we provide more details on i) why we explore toxicity detection, ii) the datasets used, iii) the search for state-of-the-art models, iv) the used hyperparameter tuning procedure, and v) the fine-tuned models.

C.1 Toxicity Detection as a Proxy for Content Moderation

The object of analysis of this paper is centered around the effects stemming from ML models used for classification. The definition of what speech should be taken down is dependent on jurisdiction and policy. We acknowledge that "toxic" does not correspond perfectly to legal notions of illegal speech, such as hate speech. However, we interpret our experimental results in Section 5 as evidence that models deployed to classify illegal speech (e.g. spam, hate, copyright) will exhibit multiplicity. We do not expect our datasets to be perfectly translatable from Toxic to Illegal Speech classification based on the definitions of the dataset. Similarly, we do not make claims to other languages since our study was conducted using the English language.

Moreover, many of our experimental results (e.g. Table 1) are based on training an ML model to detect toxic speech as defined by either Toxigen [33] or Jigsaw [41], then using this model to detect toxic speech in other datasets, which use different definitions of toxic. We observed higher values of arbitrariness and pairwise-disagreement when this was done. This is evidence that we should not expect models trained to enforce specific policies and jurisdictions to translate perfectly to other jurisdictions. As such, this secondary deployment offers a risk to Human Rights, since it will not be correctly applying local rules.

C.2 Further Dataset Information

We analyze the performance of text classification models across four datasets: ToxiGen [33], DynaHate [63], SocialBiasFrames [59],

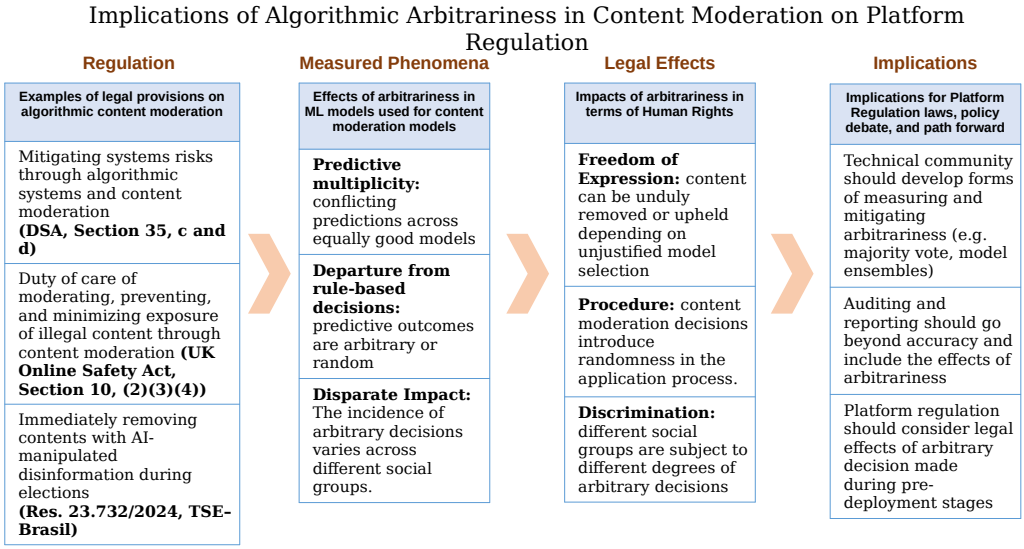


Figure 3: Summary of the key arguments presented in this work.

and HateExplain [49]. These datasets were chosen for several reasons. First, these are datasets purposefully designed to challenge ML-based toxic text classification. For example, ToxiGen and SocialBiasFrames contain mostly "implicit" toxic speech devoid of explicit profanity, slurs, or swearwords which could be easily flagged [33, 59]. DynaHate uses a human-and-model-in-the-loop process to generate a dataset designed to fool ML models. Second, these datasets have labels for demographic groups targeted by the text. This information enables us to quantify and compare Arbitrariness and Pairwise Disagreement across different target groups and report disparities in Section 5. In addition to these datasets, we also use the Measuring Hate speech (MHS) [40] and the WikiDetox [70] datasets. We chose these datasets because they add an additional dimension to our analysis: the labels of multiple human annotators who detected toxicity in each statement in the dataset. This information enables us to compare human annotators' disagreement with model disagreement in Section 5.3.

See Table 2 for a summary of all datasets used in this work. Here, the "Unique Samples" column refers to the number of unique sentences that appear in the corresponding datasets across train, test, and validation. The "Human Annotators per Sample" column refers to the number of independent human annotators that saw each sample. For example, an entry such as "1-5" means between 1 and 5 human annotators saw every sample in the dataset. We made a modest attempt at (i) removing non-English sentences from each dataset, (ii) removing repeated sentences and (iii) asserting that sentences in the training set were not in the test or validation set. As such, the numbers reported in Table 2 will not reflect the numbers reported in each dataset's respective paper.

We summarize the different definitions of toxicity used by these datasets. For Toxigen [33], the definition of toxicity is not explicitly defined, instead they asked the human annotators **"whether the statement would be harmful to anyone if an AI system wrote it (HARMFULIFAI), as well as if a human wrote it (HARMFULIFHUMAN)"**. The annotators were asked to rate each sentence between 1-5, with 1 meaning "not at all" harmful, 3 meaning ambiguous, and 5 meaning "very much so". The toxic label for this dataset was generated by taking the max of HARMFULIFAI and HARMFULIFHUMAN and mapping the scores into three class: non-toxic if < 3, ambiguous if 3, and toxic if > 3. Finally, the majority vote was taken across the three human reviewers. Note that, according to the Toxgien paper [33], there is no significant difference in toxicity when using either HARMFULIFAI or HARMFULIFHUMAN. For our work, we map any score less than or equal to 3 to non-toxic and any score above 3 to toxic.

In DynaHate [63], the definition of toxicity (which they refer to as hate) is defined as **"abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation"**. This dataset already has binary labels for toxicity, so the only post processing we do is remove the sentences whose target label was one of 'notgiven', 'notargetrecorded', 'wc', or 'NA'.

In SocialBiasFrames, [59], toxicity (which they refer to as "offensiveness") is not explicitly defined. Human annotators were ask **"Could this post be considered offensive, disrespectful, or toxic to anyone/someone?"**. The answer options were "Yes, this could be offensive", "Maybe, I'm not sure", "No, this is harmless", and "I don't understand the post". We drop any sentences where any annotator answered "I don't understand the post". Any sentence where the majority of annotators answered "Yes" was labelled toxic, else the sentence was labelled non-toxic. In particular, we assigned the values 1, 0.5, and 0 to "Yes", "Maybe", and "No", respectively. For each sentence, the toxic label was generated by averaging the scores across humans, and checking if the value was above 0.5.

In HateExplain [49], they cite [19] for their definition of toxic speech (which they call hate speech). Davidson defines hate speech as: **"language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group"**. This paper distinguishes between hate speech and offensive speech, and each sentence in this dataset is labelled as either HATEFUL (which is assigned the value 0), NORMAL (assigned the value 1), or OFFENSIVE (assigned the value 2). For the purposes of this work, we label both hateful and offensive speech as toxic, and normal speech as non-toxic, and take the majority vote across annotators to get the toxic label.

In Measuring Hate speech (MHS) [40, 58], toxic/hate speech is modelled as a spectrum, and faceted Rasch measurement theory (RMT) was used to map human responses into a continuous score that ranges between 0 and 1. For the purposes of this work, we use this continuous score, and sentences with a hate speech score > 0.5 was labelled as toxic, and anything else was labelled non-toxic. We then take the majority vote across annotators to get the toxiicity label for a given sentence. The authors of MHS have eight different forms of hate speech, where each definition denotes hate speech of increasing severity. We list each definition, with the understanding that all of these definitions together form the definition of toxicity used in our experiments:

- Genocide: **Support for or intention of systematically killing all or a large number of a protected identity group**
- Violence: **Threat or support of physical force or emotional abuse intended to hurt or kill members of a protected identity group**
- Dehumanization: **Depriving a protected group of humanlike qualities, such as comparison to an animal, insect, or disease**
- Hostility: **Unfriendliness or opposition to a protected identity group, such as through slurs, profantiy, or insults**

WikiDetox [70] contains three different datasets for personal attacks, aggression, and toxicity. "Personal attacks" are defined as in the Wikipedia guidelines [67], and broadly include **abuse based on protected classes, ad hominem attacks on affiliations, harassment, threats of legal action, etc.** "Toxicity" in WikiDetox is defined as in the Wikipedia Online Harassment Guide [68], which points to various U.S. centered studies and legal documents. The most relevant is [47], which defines hate speech in Section V as **"speech that carries no meaning other than hatred towards a particular minority, typically a historically disadvantaged minority"**. "Aggression", as far as we can tell, is not defined in WikiDetox, however, we found that all aggressive sentences were also either in the attack or toxicity dataset, so we chose to ignore the aggressive dataset all together, which did not cost us any data. In this work, we combine all the toxicity and attack datasets into one large dataset. We note that the toxicity and attack datasets have overlapping sentences but are not identical. Any sentence that was considered either as an "attack" or as "toxicity" were labelled as toxic, else it was labelled not toxic. The majority vote was then taken over human annotators to get the toxic label for each sentence. All special newline and tab tokens were removed to avoid confusing our ML models.

Finally, we discuss the Jigsaw dataset, which is a concatenation from the Jigsaw 2018 competition [12] and Jigsaw 2019 competition [11] training datasets. Note that there was a Jigsaw 2020 competition [41], but this competition had the same training data as in previous competitions, though the goal was different. Toxicity in the Jigsaw datasets is defined as **"rude, disrespectful or otherwise likely to make someone leave a discussion"**. Since this dataset is already heavily curated, no other post processed was needed. We note here that WikiDetox is a subset of the Jigsaw dataset.

C.3 Model Search On HuggingFace

Models Considered in Study. We include a screen shot of the HuggingFace platform listing the most downloaded language models for toxicity detection as of January 1st, 2024. The purpose of this screenshot is to keep historical proof that we tested all models with more than 3000 downloads as of the time of writing. Note that `s-nlp/russian_toxicity_classifier` and `cointegrated/rubert-tiny-toxicity` are Russian language models and hence outside the scope of this paper. For the same reason, `naot97/vietnamese-toxicity-detection_1`, a Vietnamese language model, was not considered. Moreover, `rungalileo/toxic-bert-quantized-traced` is a distilled / quantized version of `unitary/toxic-bert`, hence we opted to use only `unitary/toxic-bert`. See Table 4 for the full list of selected models along with their reference.

On state-of-the-art model selection. Our first goal is to identify the state-of-the-art open-source language models that have been fine-tuned for toxicity detection. We begin by evaluating the performance of all Hugging Face [69] toxicity-detection language models with more than 3000 downloads. As of January 1st, 2024, this results in 8 models (see Appendix C.3). The best-performing model (see Table 3) was `tomh TR`[33], which we will refer to as ToxiGen-RoBERTa. This model is the ToxDectRoBERTa [72] model fine-tuned on the ToxiGen dataset [33]. We fix ToxiGen-RoBERTa as our reference model. Our second goal is to create competing models to ToxiGen-RoBERTa, which we did by taking the base model architecture (ToxDectRoBERTa) and *fine-tuning* the model 40 times on the ToxiGen dataset while only varying the random seed between each run.[9] See Appendix C.4 for details on the fine-tuning procedure. We then discard the models that are worse than the reference model using the CP method from [54] outlined in Section 3, using a confidence of 95%. This choice enables a conservative estimate of the size of the Rashomon set and, therefore, of multiplicity across datasets. This results in a Rashomon parameter of $\epsilon = 0.016$, and us keeping 35 of the 40 models as Rashomon set samples ($\widehat{\mathcal{R}}(\epsilon)$). [10]

[9]The random seed determines the weight initialization of the classification head of the language model and the shuffling of the training data, both of which lead to a different model after fine-tuning.

[10]We repeat this experiment with the second-best-performing model from HuggingFace to guarantee that our experimental results are not a mere artifact of model architecture or training data selection. This model is `s-nlp RTC`[17], which we will refer to as RoBERTa-Toxicity-Classifier from here on.

Table 2: Summary of all datasets used.

Dataset	% Toxic	Unique Samples	# human annotators per sample
ToxiGen	42.5	6,514	3
Jigsaw	8.1	2,223,061 (130,320 used)	1-3,563
DynaHate	43.7	33,677	1-5
SocialBiasFrames	46.8	45,223	1-20
HateExplain	59.4	19,229	3
MeasuringHateSpeech	20.5	39,555	1-815
WikiDetox	7.7	197,578	8-46

Table 3: Test accuracy for all Hugging Face toxicity detection models with more than 3k downloads and ToxiGen across different datasets. The best-performing model accuracy is shown in green and the second best in blue. See Table 4 for the full list of selected models along with their references.

Models	Toxigen	DynaHate	SBF	HateExplain
`martin-ha TCM` [55]	56.2% ± 3.5%	52.9% ± 1.6%	56.5% ± 1.4%	55.5 ± 2.2%
`unitary TB` [32]	62.5% ± 3.4%	55.2% ± 1.6%	58.2% ± 1.4%	64.1 ± 2.2%
`s-nlp RTC` [17]	66.9% ± 3.3%	56.9% ± 1.6%	62.4% ± 1.3%	65.9 ± 2.1%
`mohsenfayyaz TC` [25]	63.2% ± 3.4%	56.1% ± 1.6%	68.5% ± 1.3%	63.8 ± 2.1%
`unitary UTR` [32]	64.5% ± 3.3%	54.6% ± 1.6%	58.4% ± 1.4%	65.8 ± 2.1%
`nicholasKluge TM` [14]	58.5% ± 3.5%	55.2% ± 1.6%	56.3% ± 2.2%	62.4 ± 2.1%
`unitary MTXR` [32]	63.1% ± 3.3%	54.6% ± 1.6%	60.1% ± 1.4%	64.3 ± 2.1%
`tomh TR` [33]	83.4% ± 2.6%	58.1% ± 1.6%	64.1% ± 1.3%	67.8 ± 2.0%

C.4 Hyperparameters

The accuracy of fine-tuned language models depends heavily on a multitude of hyperparameters. In the main body, we retrain two different model types multiple times: the ToxiGen-RoBERTa [33] and the RoBERTa-Toxicity-Classifier [17]. In this section, we detail the hyperparameters used in the main body.

ToxiGen-RoBERTa: Retraining the ToxiGen-RoBERTa model was done by fine-tuning the ToxDectRoBERTa model [72] (~ 355 million trainable parameters) on 4,601 training examples from the human annotated subset of the ToxiGen dataset [33]. In particular, we trained on a subset of the ToxiGen data used by [35] that removed prompts for which 3 annotators disagreed on the target group. Moreover, no quantization was done on the ToxDectRoBERTa model, and all training runs were performed on a 80Gb A100 GPU. We fixed the number of epochs to 10 and performed an extensive hyperparameter sweep over:

- learning rate: Logarithmically spaced values from 10^{-6} and 10^{-4}.
- batchsize: Three values $\in \{8, 16, 32\}$.
- Weight decay: Linearly spaced values from 0 and 0.1 with a 0.01 spacing.
- Warmup Steps: Linearly spaced values from 0 to 30% of an epoch with a 5% spacing.

All other hyperparameters were set to the default that Huggingface's sequence classification routine uses. In particular, this means a Linear learning rate schedule with the AdamW optimzer. The sweep was done via the Trainer API from HuggingFace Transformers with the Optuna [1] backend, which used evaluation accuracy to prune unpromising trails early in training. In total, Optuna made 60 complete training runs (the average run took an hour and 20 minutes on an A100 GPU 80Gb). The optimal parameters were found to be: learning rate: 1e-5, batch size: 32, weight decay: 0.09, and warmup ratio: 0.1. The random seed used for the best run was 6. All ToxiGen fine tuned models (i.e., those used in the multiplicity experiments) used these hyperparameters, except for random seed. The seeds used for the ToxiGen fine tuned models were randomly generated 3 and 4 digit integers sampled using [31]. See Figure 5 for a plot of the training trajectories of 10 of the random seeds.

RoBERTa-Toxicity-Classifier. Retraining the RoBERTa-Toxicity-Classifier was done by fine-tuning the base RoBERTa model [45] (~ 124 million trainable parameters) on 100,000 training examples sampled uniformly from the concatenated Jigsaw dataset [11, 12]. Moreover, no quantization was done on the RoBERTa model, and all training runs were performed on a 80Gb A100 GPU. In practice, the significantly larger dataset size meant that fine-tuning this RoBERTa model was approximately 3 times slower than fine-tuning the Toxigen models. Due to the increased computational cost of training these models compared to the ToxiGen models, we did not as extensive of a hyperparameter sweep. We set the batch size to 8 (for faster training time), and did a grid search for 4 epochs over four learning rates $\{10^{-6}, 10^{-5}, 2 \times 10^{-5}, 10^{-4}\}$. The best was found to be 2×10^{-5}. Then, we increased the batch size to as large as our memory allowed (32), and kept all other hyperparameters set to the default in Huggingface's sequence classification routine (notably: weight decay:0 and no warmup steps). All Jigsaw fine tuned models used these hyperparameters. The seeds used for the Jigsaw models were randomly generated 3 and 4 digit integers

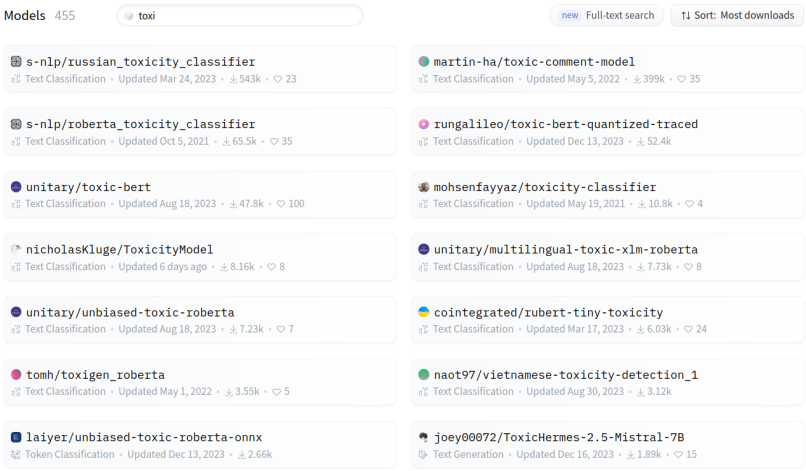
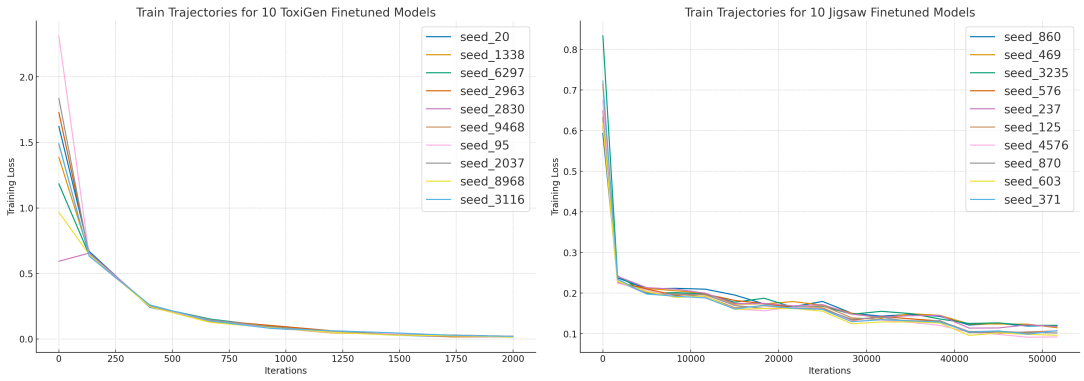


Figure 4: Screenshot of the HuggingFace platform's most popular toxic detection models as of the writing of this paper



(a) Toxigen Fine-Tuned **(b) Jigsaw Fine-Tuned**

Figure 5: Training trajectories for the fine-tuned ToxiGen and Jigsaw models over 10 randomly chosen seeds.

sampled using [31]. The average Jigsaw model took approximately 3 hours and 15 minutes to fine tune. See Figure 5 for a plot of the training trajectories of 10 of the random seeds.

C.5 Fine-Tuned Models Performance

In Table C.5, we show the performance of the models we fine-tuned and compare it against the reference models. The line Reference in Table C.5 shows the accuracy of the reference ToxiGen-RoBERTa

Table 4: All considered Hugging face models.

Model Name and Link	Reference
martin-ha/toxic-comment-model	Pan [55]
unitary/toxic-bert	Hanu and Unitary team [32]
s-nlp/roberta_toxicity_classifier	Dale et al. [17]
mohsenfayyaz/toxicity-classifier	Fayyaz [25]
unitary/unbiased-toxic-roberta	Hanu and Unitary team [32]
nicholasKluge/ToxicityModel	Corrêa [14]
unitary/multilingual-toxic-xlm-roberta	Hanu and Unitary team [32]
tomh/toxigen_roberta	Hartvigsen et al. [33]

Table 5: Accuracy of the reference models from Hugging Face and our Fine-tuned models. The column Toxigen represents the accuracy of the models fine-tuned in the Toxigen dataset. The column Jigsaw represents the accuracy of the models fine-tuned in the Jigsaw dataset. The reference line shows the accuracy from the models deployed in Hugging Face. The lines Minimum, Mean, and Maximum show the minimum, average, and maximum accuracies across all our fine-tuned models.

Accuracy	Data Split	Toxigen	Jigsaw
Reference	Train	96.0%	95.7%
	Test	83.4%	95.3%
Minimum	Train	94.6%	93.6%
	Test	83.4%	92.8%
Mean	Train	98.2%	96.6%
	Test	85.0%	94.1%
Maximum	Train	99.8%	100%
	Test	86.8%	100%

model [33] and RoBERTa-Toxicity-Classifier [17] train and test accuracies. The lines Minimum, Mean, and Maximum show the minimum, average, and maximum accuracies across all our fine-tuned models. We observe that both the train and test performance of our models approximates the reference models deployed in Hugging-Face. Surprisingly, the fine-tuned Jigsaw models perform as well as its reference model that was trained in 10 times more data from the same dataset.

D FURTHER EXPERIMENTAL RESULTS

In this section, we show the main results in the paper for difference values for the Rashomon parameter given by the selection of confidence values for the CP method [54]. Additionally, we also show arbitrariness and pairwise disagreement across demographics for datasets.

D.1 Arbitrariness with Different Confidences

We start by showing the pairwise disagreement and arbitrariness values for the testing partition of Toxigen, DynaHate, SBF, and HateExplain. We show these results for two different confidence levels in the CP method: 50% and 1%. When confidence is smaller, more models are considered to be in the Rahsomon set but with a higher probability of wrong model inclusion in the set.

Table 6 shows pairwise disagreement and arbitrariness for a confidence level in the CP method equal to 50% and Table 7 shows results with confidence 1%. We observe that, compared with Table 1, the disagreement and arbitrariness values of Tables 6 and 7 are higher as a consequence of models with higher error being included as samples of the Rashomon set.

D.2 Multiplicity Across Demographics

Here, we also show how arbitrariness and pairwise disagreement vary across different targeted demographic groups. Figures 6 and 7 indicate that even under higher confidence values, arbitrariness and disagreement are still non-uniformly distributed as showed in Figure 1, leading to disparate algorithmic treatment.

D.3 Human vs. Model arbitrariness

We also display the arbitrariness and pairwise disagreement values across unambiguous and ambiguous toxic content. Recall that we consider *unambiguous* sentences the ones that all human annotators agreed upon its toxicity and *ambiguous* when not all annotators classified the sentence toxicity equally.

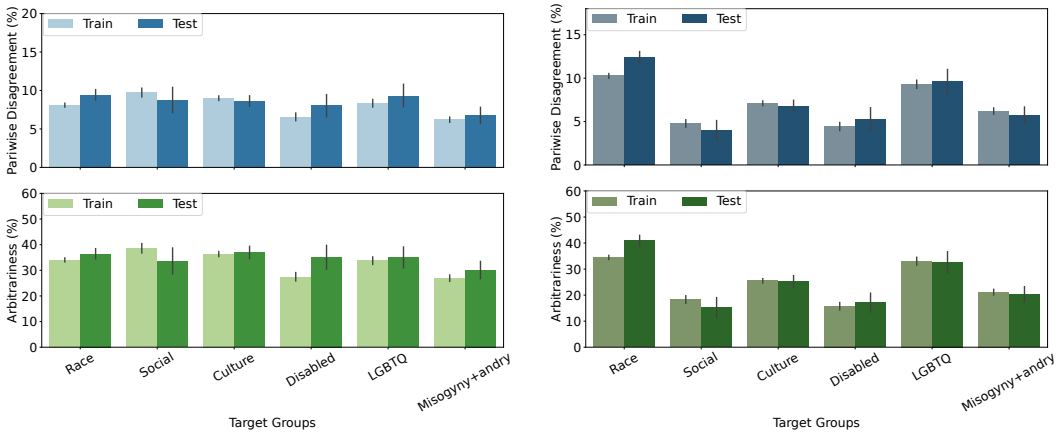
Figures 8 and 9 present the same pattern of higher arbitrariness and pairwise disagreement in ambiguous sentences while also having a high arbitrariness and pairwise disagreement in unambiguous sentences — and we discuss in Section 5.

Table 6: Average pairwise disagreement and arbitrariness in testing time for the Toxigen fine-tuned and Jigsaw fine-tuned models in different datasets. The confidence in the CP methods was chosen to be 50% for a more conservative analysis. 95% confidence intervals are shown using the standard error from the mean.

	Toxigen Fine-Tuned		Jigsaw Fine-Tuned	
Dataset	Pairwise Disagreement	Arbitrariness	Pairwise Disagreement	Arbitrariness
Toxigen	6.8% ± 0.9%	28.8% ± 3.2%	4.5% ± 0.8%	16.2% ± 2.6%
DynaHate	8.4% ± 0.5%	34.3% ± 1.6%	6.1% ± 0.4%	22.7% ± 1.4%
SBF	8.6% ± 0.4%	35.4% ± 1.3%	7.3% ± 0.3%	25.1% ± 1.0%
HateExplain	8.0% ± 0.6%	32.3% ± 2.0%	8.8% ± 0.2%	30.7% ± 2.0%
Total	8.3% ± 0.2%	34.0% ± 0.8%	7.1% ± 0.2%	24.8% ± 0.7%

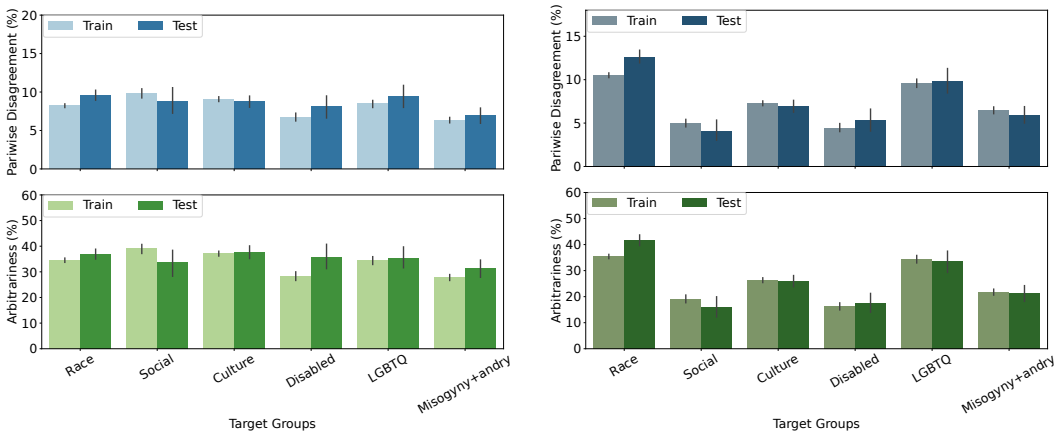
Table 7: Average pairwise disagreement and arbitrariness for the Toxigen fine-tuned and Jigsaw fine-tuned models in different datasets. The confidence in the CP methods was chosen to be 1%, including all fine-tuned models.

Dataset	Toxigen Fine-Tuned		Jigsaw Fine-Tuned	
	Pairwise Disagreement	Arbitrariness	Pairwise Disagreement	Arbitrariness
Toxigen	6.9% ± 0.9%	29.6% ± 3.2%	4.7% ± 0.8%	16.7% ± 2.6%
DynaHate	8.6% ± 0.5%	35.1% ± 1.6%	6.3% ± 0.4%	23.6% ± 1.4%
SBF	8.7% ± 0.4%	35.9% ± 1.3%	7.5% ± 0.3%	25.6% ± 1.0%
HateExplain	8.1% ± 0.6%	32.8% ± 2.0%	9.0% ± 0.6%	31.6% ± 2.0%
WikiDetox	6.3% ± 0.1%	26.5% ± 0.4%	1.3% ± 0.1%	4.7% ± 0.2%
Total	7.2% ± 0.2%	25.4% ± 0.7%	8.4% ± 0.2%	34.6% ± 0.3%



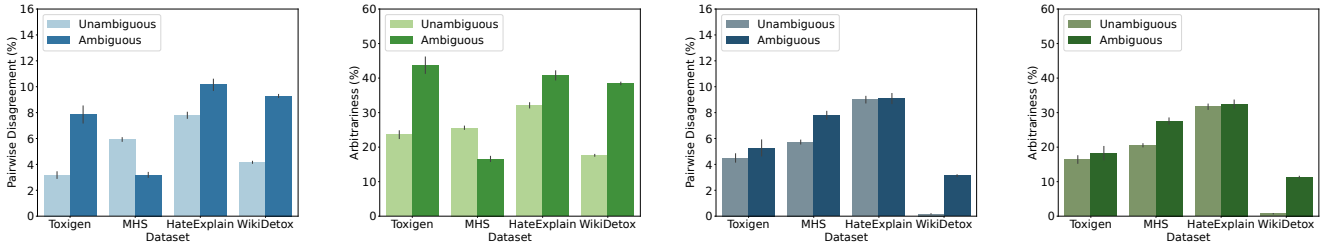
(a) Toxigen Fine-Tuned **(b) Jigsaw Fine-Tuned**

Figure 6: Average pairwise disagreement and arbitrariness in different target groups for the fine-tuned Toxigen and Jigsaw models. The results show the pairwise disagreement in percentage (x-axis) for the union of four different datasets: DynaHate, SBF, Toxigen, and HateExplain. The results are shown for training and test partitions of each dataset. The confidence in the CP methods was chosen to be 50% containing all fine-tuned models, leading to the selection of 38 out of 40 Roberta models in the Rashomon set fine-tuned in the Toxigen dataset and 17 out of 20 Jigsaw fine-tuned models.



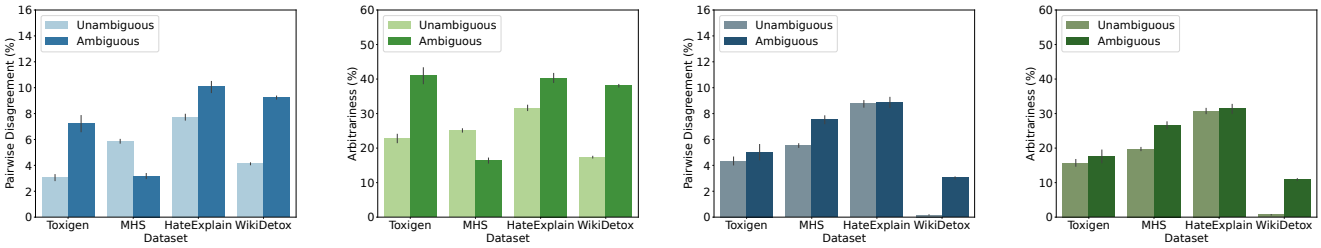
(a) Toxigen Fine-Tuned (b) Jigsaw Fine-Tuned

Figure 7: Average pairwise disagreement and arbitrariness in different target groups for the fine-tuned Toxigen and Jigsaw models. The results show the pairwise disagreement in percentage (x-axis) for the union of four different datasets: DynaHate, SBF, Toxigen, and HateExplain. The results are shown for training and test partitions of each dataset. The confidence in the CP methods was chosen to be 1% containing all fine-tuned models, leading to the selection of 40 out of 40 Roberta models in the Rashomon set fine-tuned in the Toxigen dataset and 20 out of 20 Jigsaw fine-tuned models.



(a) Toxigen Fine-Tuned Models (b) Jigsaw Fine-Tuned Models

Figure 8: Average pairwise disagreement and arbitrariness for *Unambiguous* and *Ambiguous* sentences using the Toxigen fine-tuned and Jigsaw fine-tuned models. The table shows the pairwise disagreement estimated values along with the 95% confidence intervals using the standard error from the mean. We consider a sentence *Ambiguous* when at least one annotator labeled the sentence differently than others and *Unambiguous* otherwise. The confidence in the CP methods was chosen to be 1%, including all fine-tuned models in the above analysis.



(a) Toxigen Fine-Tuned Models

(b) Jigsaw Fine-Tuned Models

Figure 9: Average pairwise disagreement and arbitrariness for *Unambiguous* and *Ambiguous* sentences using the Toxigen fine-tuned and Jigsaw fine-tuned models. The table shows the pairwise disagreement estimated values along with the 95% confidence intervals using the standard error from the mean. We consider a sentence *Ambiguous* when at least one annotator labeled the sentence differently than others and *Unambiguous* otherwise. The confidence in the CP methods was chosen to be 50%, including all fine-tuned models in the above analysis.