Safety from Bigoted AI: A Framework for Online Debiasing

Jeffrey Cheng, Hirsh Guha

Abstract

Deployed models often stream high-impact outputs and actively train on new observed data. Many fairness algorithms provide extremely strong guarantees about statistical parity on protected groups in the batchwise setting or in the limit of training data. We are therefore motivated to extend these results to the online setting where we can impose safety conditions on the model's behavior throughout the trajectory of its deployment. We propose the ϵ-tolerant online debiaser as a rejection-sampling framework with a strong theoretical guarantee of statistical parity; we conclude with a demonstration of its effectiveness on minimizing bias from recidivism predictions on the ProPublica dataset.

1 Introduction

Suppose that one lives in a city where an AI sentences criminals. It would be of little comfort to understand that this AI's behavior is asymptotically fair or that some Chernoff bound will be small by the year 2080; one would vastly prefer a guarantee that the AI is fair **in the current moment**.

We are therefore interested in imposing a safety criterion over a model that holds true in the online setting; namely, that the algorithm makes unbiased decisions throughout the duration of training, not just in the batchwise setting or in the limit of training data.

Due to the highly fragmented nature of fairness frameworks from the Fairness, Accountability, and Transparency (FAT*) community, we will present our work in an unorthodox order. We will begin with the recidivism dataset and some exploratory observations on the bias of naive models. We will then motivate a selective review of relevant fairness frameworks and theoretical guarantees. Finally, we will present the ϵ-tolerant debiaser, its safety and liveness guarantees, and its emperical behavior on the recidivism dataset.

1.1 The Recidivism Dataset

We will use the publicly available ProPublica Compas Analysis dataset on criminal history, jail and prison times, and demographics for defendants in Broward county, Florida. The learning problem is to predict the probability of recidivism (committing another crime) within two years.

There is a clear a priori concern for models trained on this dataset to be racially biased; Propublica's work in open-sourcing this dataset revealed that COMPAS, a privately developed black-box algorithm used by court systems in multiple states [9], was not only unable to outperform linear regression [16] but also disproportionately erred on African-Americans. Black defendants who do not recidivate were almost twice as likely to be classified by COMPAS as a higher risk compared to white defendants (45 percent vs. 23 percent) [1].

For details on the bias of the COMPAS algorithm, our munging methodology and dataset schema, see Appendices 7.1, 7.2, and 7.3, respectively.

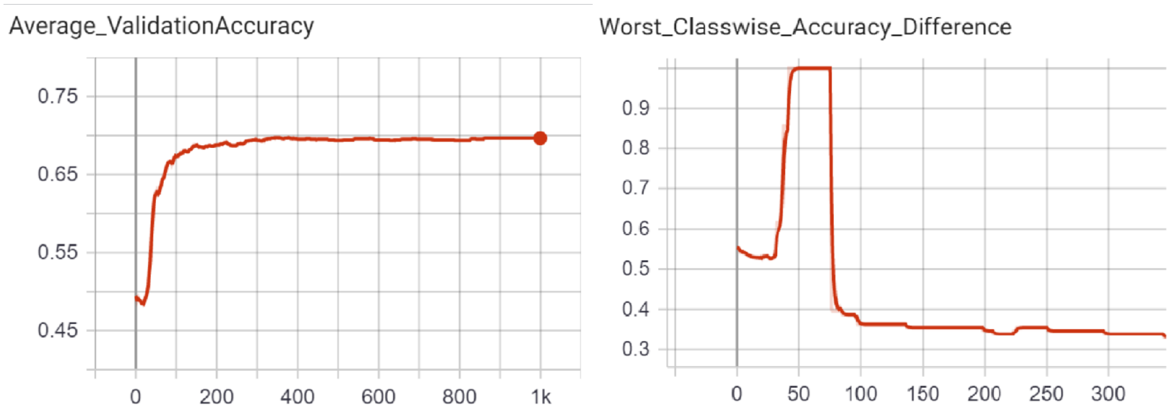


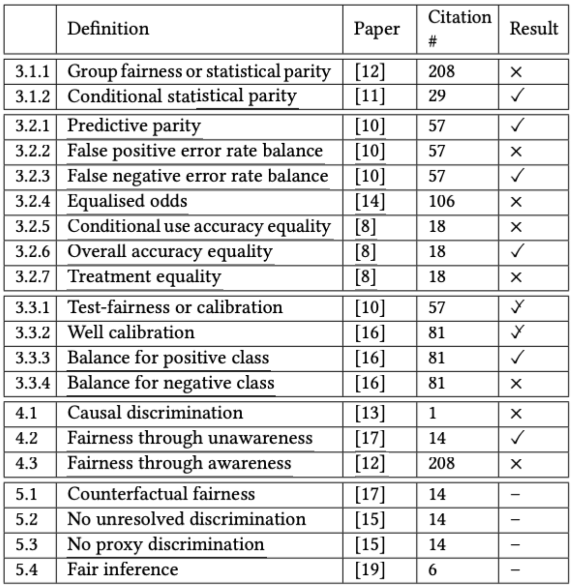
Figure 1: *A visualization of the distinction between low error and fairness. The x-axis represents the epoch number; validation metrics were measured at the end of each epoch. The y-axis on the right represents the absolute difference in accuracy between the highest accuracy over any rate and the lowest accuracy over any rate.*

We trained a shallow neural network on the recidivism dataset without any notions of fairness. As depicted on the right of Figure 1, though we immediately outperformed COMPAS' 65% accuracy[1], we observe on the left of Figure 1 that our model rocketed to extraordinary levels of bias during its early training. In fact, the worst classwise accuracy difference jumps to a perfect 1.0 between epochs 30 and 80, which indicates that our model had 100% error on one race and 0% error on another race for 50 consecutive epochs! And yet in the corresponding validation accuracy curve, we see only a well-behaved, smooth increase in accuracy. We attribute this behavior to two factors:

- Certain subsets of the data will have naturally lower error rates (e.g. more predictive covariates, lower real-life measurement error).
- Certain subsets of the data are faster to converge (e.g. curve-fitting over a very small, homogeneous distribution occurs much faster under stochastic gradient descent than the same process over a large, diverse one).

We are therefore motivated to avoid spikes of unfairness levels throughout training; we would prefer that our models improve on each protected class of individuals at a similar, steady rate.

1.2 Fairness Definitions



	Definition	Paper	Citation #	Result
3.1.1	Group fairness or statistical parity	[12]	208	×
3.1.2	Conditional statistical parity	[11]	29	✓
3.2.1	Predictive parity	[10]	57	✓
3.2.2	False positive error rate balance	[10]	57	×
3.2.3	False negative error rate balance	[10]	57	✓
3.2.4	Equalised odds	[14]	106	×
3.2.5	Conditional use accuracy equality	[8]	18	×
3.2.6	Overall accuracy equality	[8]	18	✓
3.2.7	Treatment equality	[8]	18	×
3.3.1	Test-fairness or calibration	[10]	57	✓×
3.3.2	Well calibration	[16]	81	✓×
3.3.3	Balance for positive class	[16]	81	✓
3.3.4	Balance for negative class	[16]	81	×
4.1	Causal discrimination	[13]	1	×
4.2	Fairness through unawareness	[17]	14	✓
4.3	Fairness through awareness	[12]	208	×
5.1	Counterfactual fairness	[17]	14	–
5.2	No unresolved discrimination	[15]	14	–
5.3	No proxy discrimination	[15]	14	–
5.4	Fair inference	[19]	6	–

Figure 2: *Verma et al's non-comprehensive list of broad fairness definitions. A check mark on the right denotes that a positive result exists on the German Credit Dataset [14].*

Dozens of fairness definitions have been proposed because the common-sense formalization is situational [14]. We will focus on statistical measures since those definitions align with the observed behavior in Figure 1.

Fairness over statistical measures is broadly defined by a requirement that some summary statistic must be statistically indistinguishable between two distributions of data. We further divide this set into two categories: outcome parity and error parity.

1.2.1 Outcome parity

In classification-based *statistical parity* (also known as group fairness, equal acceptance rate [17], and benchmarking), we require that the probability of being assigned to the positive class must be equal across values of the protected class. This condition is levied on the full joint distribution; in the case of recidivism, a model fails this fairness criterion if the aggregate predicted rate of recidivism in African-Americans and Caucasians are not statistically equal. Simoiu et al note that this assumption will fail whenever the true rates actually differ in the marginal distributions along the protected classes and name this condition *infra-marginality* [13]. As such, Dwork et al apply this framework only in cases where there is potential inherent benefit to outcome equality (e.g. affirmative action in diversifying college admissions) [5].

Conditional statistical parity subverts the problem of infra-marginality by making a more qualified criterion that the model's probability of assigning the positive class is identical across protected class values conditional on some set of legitimate covariates; e.g. in the recidivism, models may need to predict that African-Americans and Caucasians recidivate at the same rate conditional on income (in fact, the same recidivism dataset motivated Corbett-Davies et al towards this definition) [4]. The authors find that enforcing conditional statistical parity is more tractable than outcome parity. They also find that enforcing this condition increases the false negative rate and risks community safety by releasing dangerous individuals, and they name this phenomenon the "cost of fairness."

1.2.2 Error Parity

Error parity is a much simpler idea: some measure of validation error must be the same across classes. Predictive equality requires that the false positive rates be statistically equal [4] [3]; equal opportunity requires that the false negative rates be statistically equal [7]. Requiring both of these conditions is known as equalized odds or disparate mistreatment [10]. We note that since these definitions take the distributions of data labels into account, it's clear that there exist high-performing models that satisfy these definitions (whether these models are easily learnable is a separate question).

1.3 Fairness Guarantees

The FAT* community has wonderful guarantees on fair behavior under many settings and definitions of fairness. We wish to highlight two kinds of guarantees that motivate our work: data-driven and online guarantees.

1.3.1 Data-driven guarantees

For example, Zhang et al define a risk difference as the difference in positive classification rate between two distributions; they impose fairness as risk difference constraint on a convex relaxation of the non-convex model optimization problem:

$$\begin{aligned} &\min_{h \in \mathcal{H}} \mathbb{L}_\phi(h) \\ &\text{subject to } \mathbb{RD}_\kappa(h) \leq c_1, \\ &\qquad\qquad \mathbb{RD}_\delta(h) \leq c_2 \end{aligned}$$

where $\mathbb{L}_\phi$ represents a convex surrogate for the 0-1 empirical loss function, $h \in \mathcal{H}$ is a continuous model, and κ, δ are convex, concave surrogates for the risk difference function (respectively). The authors prove that this relaxation of this problem is itself a convex optimization problem, then solve the relaxation with Disciplined Convex-Concave Programming. [15] However, note that because this methodology uses convex optimization, it's not obvious how to apply this methodology to the online setting.

1.3.2 Online guarantees

Bechavod et al find a metric-free online fairness framework that can force a model to respect any form of fairness criterion. The input to this framework is an auditor that can observe a model's decision over two examples and decide whether fairness is violated; the form of fairness is implicit in the auditor's outputs. This a no-regret algorithm, which subverts the "cost of fairness" [2].

A similar finding from Gillen et al guarantees online fairness by learning an unknown similarity metric from weak feedback; the intuition is that this mimics "regulator who knows unfairness when he sees it but nevertheless cannot enunciate a quantitative fairness metric over individuals" [6].

We note with great intentionality that all data-driven fairness guarantees at the time of writing are either asymptotic or batchwise, and all online guarantees are auditor-driven (i.e. the fairness criterion is not evaluated from data, but rather from an oracle that outputs fairness violations).

1.4 Research Question

With the language and machinery of the FAT* body of literature at hand, we formulate our goal on the recidivism dataset: to create a data-driven debiasing algorithm that will learn a model in the online setting that needs neither a data-subverting oracle nor the complete data distribution. We require the following two conditions:

1. **Safety**: the model will never enter a state of disparate mistreatment; the difference in accuracy rate in predicting recidivism between ethnicities must never exceed a pre-specified threshold.
2. **Liveness**: the model must never be frozen into inaction by its safety criterion. A very strong fairness condition (such as group fairness) could conceivably be too strong as to prevent learning or improvement.

With these qualitative goals in mind, we will now notate and specify our algorithm design. We will define and prove the safety and liveness conditions formally in Section 3.

1.5 Notation

Definition 1.1. Let the data distribution be a mapping $\mathcal{D} : \{(\vec{x}_i, y_i)\forall i\} \to \mathbb{R}$. We notate the **distributions over protected class values** $\mathcal{C} = \{c_1, c_2, ...\}$ as $\mathcal{D}_{c_1}, \mathcal{D}_{c_2}....$ In this paper, these will refer to the subset data distributions over Caucasians, African-Americans, etc.

Definition 1.2. A model f is **ϵ-biased** over a protected class $\mathcal{C}$ on loss function $\mathcal{L}$ if $(\max_{c_i \in C} \mathcal{L}(f, c_i)) - \left(\min_{c_j \in C} \mathcal{L}(f, c_j)\right) > \epsilon$.

Definition 1.3. An **ϵ-tolerant online debiaser** is a learning system that shields an online model such that it is never ϵ-biased.

Definition 1.4. Let $\mathcal{F}$ be a family of models. Let $\varepsilon^{Bayes}_{\mathcal{D}_i,\mathcal{L},\mathcal{F}} = \min_{f \in \mathcal{F}} \mathcal{L}_{\mathcal{D}_i}(f)$ be the **Bayes error** for loss function $\mathcal{L}$ over data distribution $\mathcal{D}_i$ on the model family $\mathcal{F}$. When $\mathcal{F}$ is not specified, assume the Bayes error is inclusive of all models.

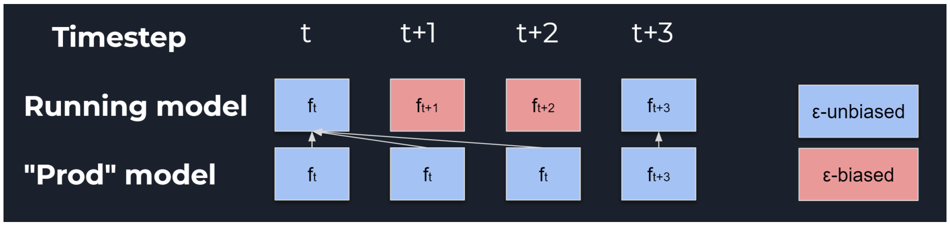


Figure 3: *If the running model is ϵ-unbiased at time t and is biased for 2 timesteps thereafter, the production model will point to the parametrization at timestep t for 3 timesteps.*

2 Algorithm

Algorithm 1 ϵ-tolerant Debiaser

```
procedure DEBIASER(Minibatch stream $D$, validation set $V$)
    Initialize running model as neural network $N$
    Initialize loss function $L$, optimizer $A$.
    Set production model $R \leftarrow N$
    for minibatch $(x, y)$ in stream $D$ do
        Compute gradient update $g \leftarrow L(N(x), y)$.
        Execute an optimization step $N \leftarrow A(N, g)$.
        Calculate maximum classwise loss difference of $N$ on $V$.
        if $N$ is not $\epsilon$-biased then
            $R \leftarrow N$
    return $N$
```

We will first establish some intuition, then formalize the ϵ-tolerant debiaser algorithm.

We wish to obtain models within the distribution of high-performing, ϵ-unbiased models; however, it is extremely difficult to sample from this distribution. We turn to computational Bayesian methods for inspiration, where we sample from arbitrarily complex non-standard distributions via *rejection sampling*. For example, the Metropolis-Hastings algorithm uses a simple proposal distribution $g(x' \mid x_t)$ that is easy to sample from (e.g. a Gaussian centered at the previous sample) and a rejection sampling routine where the proposal x' is accepted with probability $\alpha = \frac{f(x')}{f(x_t)}$, where f is the density of the desired distribution. [8]

Our approach will be to use an increasingly high-performing proposal distribution of models by offering snapshots of a neural network throughout its training process and accepting/rejecting the model parametrization at each snapshot based on whether it is ϵ-biased. As a matter of implementation, we will train a neural network in the online setting, and the "samples" we are optimizing (referred to as the "production" model) will be pointers to the most recent parametrization of this neural network that is not ϵ-biased. We depict an example in Figure 3 and the formal algorithm in Figure 1. Note that in this implementation, the production model is allowed to be ϵ-biased at initialization but never again after the first reparametrization.

3 Theoretical Behavior

We formalize the safety condition outlined in the research goal as the following theorem.

Theorem 3.1. *The production model of an ϵ-tolerant online debiaser is always ϵ-unbiased.*

We further claim this theorem is true by the definition of the production model.

Subject to assumptions on our choice of tolerance ϵ and the expressiveness of the learner family, we claim (in informal language) that the ϵ-tolerant debiaser will never put its production model into a perpetual frozen state. Before proving this statement, we prove a brief lemma about the nature of Bayes errors on partitions of distributions.

Lemma 3.2. *Suppose over sample space $\Omega \subset X \times Y$ with measure λ, we induce a finite partition $\Omega = \Omega_1 \cup \Omega_2 \cup ...\Omega_n$. Further suppose that for each element of the partition Ω_i we assign a distribution F_i with zero support in $\Omega - \Omega_i$, and we define the distribution of the aggregate sample space Ω to be $F(\omega) = \sum_i \frac{\lambda(\Omega_i)}{\lambda(\Omega)} F_i(\omega)$.*

Then any model $M : X \to Y$ that achieves $\varepsilon^{Bayes}_{F,\mathcal{L}}$ must also achieve $\varepsilon^{Bayes}_{F_i,\mathcal{L}}$ $\forall$ F_i.

Proof. Suppose for the sake of contradiction that model M achieves $\varepsilon^{Bayes}_{F,\mathcal{L}}$ but fails to achieve $\varepsilon^{Bayes}_{F_j,\mathcal{L}}$ for some j. We begin by expanding the Bayes error as an integral with respect to the sample space Ω.

$$\varepsilon^{Bayes}_{F,\mathcal{L},M} = \int_{\omega=(x,y)\in\Omega} F(\omega)\cdot\mathcal{L}(M(x),y)d\omega$$

We note that by construction, every term in the expansion $F(\omega) = \sum_i \frac{\lambda(\Omega_i)}{\lambda(\Omega)} F_i(\omega)$ evaluates to zero support except for one term because the set $\{\Omega_i \forall i\}$ is a partition. We then collect over the partitions in the Bayes error expansion.

$$\varepsilon^{Bayes}_{F,\mathcal{L},M} = \int_\Omega \left(\sum_i \frac{\lambda(\Omega_i)}{\lambda(\Omega)} F_i(\omega)\right)\cdot\mathcal{L}(M(x),y)d\omega$$

(By definition of F)

$$= \int_\Omega \left(\sum_i a_i \cdot F_i(\omega)\right)\cdot\mathcal{L}(M(x),y)d\omega$$

(Define $a_i = \frac{\lambda(\Omega_i)}{\lambda(\Omega)}$)

$$= \int_\Omega \left(\sum_i a_i \cdot F_i(\omega)\mathbb{I}(\omega\in\Omega_i)\right) \\ \cdot\mathcal{L}(M(x),y)d\omega$$

(Each ω is only in one Ω_i.)

$$= \sum_i \left(a_i \int_{\Omega_i} F_i(\omega)\cdot\mathcal{L}(M(x),y)d\omega\right)$$

(Addition rule)

$$= \sum_i \left(a_i \cdot \varepsilon^{Bayes}_{F_i,\mathcal{L}}\right)$$

Now, we consider the specific loss on model $M : \sum_i \left(a_i \cdot \varepsilon^{M}_{F_i,\mathcal{L}}\right)$. We know that for all $i \neq j$, model M cannot perform better than $\varepsilon^{Bayes}_{F_i,\mathcal{L}}$. Therefore, for each of those $n-1$ terms, the loss on model M is at least as large as the corresponding term in the Bayes error expansion by the definition of Bayes error.

By our contradiction assumption, we assume that $\varepsilon^{M}_{F_j,\mathcal{L}} > \varepsilon^{Bayes}_{F_j,\mathcal{L}}$. Since each of the terms in the n-term expansion is larger, then the overall error on model M is larger than the Bayes error $\varepsilon^{Bayes}_{F,\mathcal{L},M}$. This concludes the proof by contradiction.

□

We now formalize our theorem as follows:

Theorem 3.3. *Suppose that a sigmoid-activation neural network f is trained by an ϵ-tolerant debiaser that is instantiated on data distribution $\mathcal{D}$, loss function $\mathcal{L}$, and protected classes $\mathcal{C}$. Suppose further that:*

1. *f is arbitrarily wide.*

2. *The parametrization at time $t-1$, f_{t-1}, is ϵ-unbiased.*
3. *The parametrization at time t, f_t, is ϵ-biased.*
4. $\epsilon > \left(\max_{i\in\mathcal{C}} \varepsilon^{Bayes}_{\mathcal{D}_i,\mathcal{L}}\right) - \left(\min_{i\in\mathcal{C}} \varepsilon^{Bayes}_{\mathcal{D}_i,\mathcal{L}}\right)$.

Then there exists some time $t^* > t$ for which f_{t*} is ϵ-unbiased.

Proof. Suppose for the sake of contradiction that there does not exist a timestep t^* that allows the production model to escape its pointer at time $t-1$. Intuitively, neural network f will then observe an infinite number of data points in the data stream D.

Since f is assumed to be arbitrarily wide and because the sigmoid activation is bounded, we can invoke the equivalence between the family of infinite-width neural networks and the Gaussian process prior, where a specific parametrization of an infinite-width neural network is equivalent to a sampled function from the Gaussian process. [11] Thus, any non-parametric properties of Gaussian processes must apply to f as well.

Gaussian processes are consistent estimators. [12] Therefore, f will converge to Bayes error $\varepsilon^{Bayes}_{\mathcal{D},\mathcal{L}}$.

Since the distributions D_i are a partition generated from the protected class values C, we apply the earlier lemma. We now have that f will converge to the Bayes error $\varepsilon^{Bayes}_{\mathcal{D}_\rangle,\mathcal{L}}$ for every distribution D_i.

Therefore, the maximum classwise loss error difference is:

$$\left(\max_{c_i\in C} \mathcal{L}(f, c_i)\right) - \left(\min_{c_j\in C} \mathcal{L}(f, c_j)\right)$$
$$= \left(\max_{i\in\mathcal{C}} \varepsilon^{Bayes}_{\mathcal{D}_i,\mathcal{L}}\right) - \left(\min_{i\in\mathcal{C}} \varepsilon^{Bayes}_{\mathcal{D}_i,\mathcal{L}}\right)$$

Since ϵ was constructed to be larger than the RHS of the above equation, it must be larger than the left-hand side. Since the convergence limit of the consistent estimator is strictly less than ϵ, there exists some time t^* where f achieve classwise accuracy difference strictly lower than ϵ. By definition, f_{t^*} is ϵ-unbiased. This concludes the proof by contradiction.

□

Note that generalizing this theorem from sufficiently large neural networks to any consistent estimator is trivial.

4 Empirical Results

4.1 Removal of Racial Bias

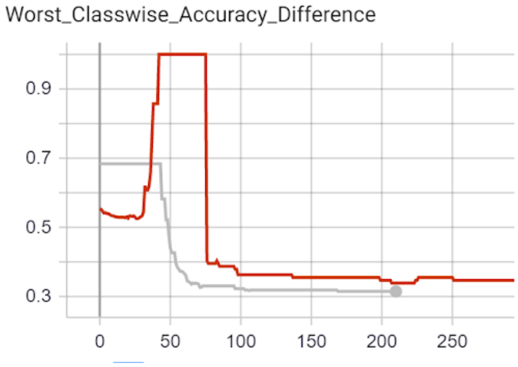


Figure 4: *The vanilla model from Figure 1 is shown in red, and the model with the debiaser running is shown in grey. Note the grey curve stays below its initialization bias.*

Our key finding is that the ϵ-tolerant online debiaser does remove the characteristic early-training spike in bias. In addition, the debiaser seems to make entire classwise accuracy difference curve well-behaved. As depicted in Figure 4, not only does the debiased grey curve avoid the characteristic 1.0-accuracy difference spike between epochs 30 and 80, it very nearly decreases monotonically after leaving a 40-epoch freeze at initialization.

4.2 "Cost of Fairness"

Since the imposition of our ϵ-bias constraint does not help fulfill an objective (it, in fact, actively hampers the training objective), we do not expect the ϵ-tolerant debiaser to improve overall accuracy performance.

Recall that Corbett-Davies et al found a significant tradeoff between fairness and accuracy; a recidivism model with conditional statistical parity with respect to race introduces risk by letting more potentially dangerous criminals into the community, thus creating a "cost of fairness" [4]. We similarly calculate our own cost of fairness.

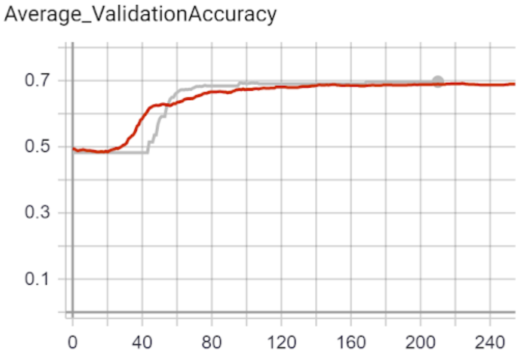


Figure 5: *Both the vanilla and the debiased models converge to the same accuracy around 70%.*

We observe in Figure 5 that although the debiased production model is initially slower to train, it catches up fairly quickly. In fact, due to sample variation between these two trajectories, the debiased model actually converges slightly faster. We therefore conclude that the cost of fairness is negligible.

4.3 Liveness in Practice

The liveness condition predicts that the production model will never be frozen indefinitely, but does not preclude intolerably long freezes.

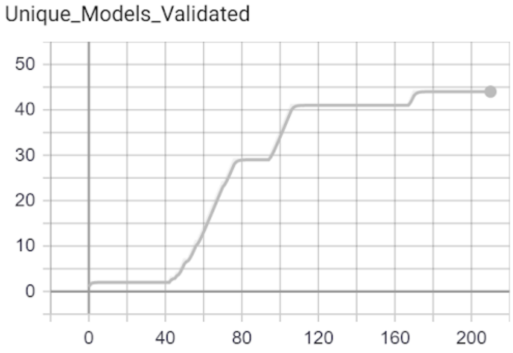


Figure 6: *The number of unique production models throughout training.*

Furthermore, the liveness property only applies in the infinite-width limit. However, we see our model never stays permanently frozen in Figure 6, and the longest stretch where no new models are being validated is 40 epochs.

5 Conclusion

We reasoned about safety-critical applications of online models with high social-impact outputs and implemented the ϵ-tolerant online debiaser, a shielding method with successful empirical reduction of bias on the recidivism dataset as well as theoretical guarantees in both safety and liveness. Surprisingly, the Corbett-Davies cost of fairness is negligible. We hope that this finding is particularly useful to practitioners who must deploy live models with active learning because this subverts the efficient frontier between iteration speed and safety.

We propose as next steps an augmentation of the debiasing algorithm that can manipulate gradient flow to "push" the running model into good parametrizations without violating the safety and liveness guarantees. This presents challenges because doing so introduces non-stationarity into the data distribution.

6 References

[1] Julia Angwin, Jeff Larson, and Lauren Kirchner. How we analyzed the compas recidivism algorithm, May 2016.

[2] Yahav Bechavod, Christopher Jung, and Zhiwei Steven Wu. Metric-free individual fairness in online learning, 2020.

[3] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, 2016.

[4] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. *CoRR*, abs/1701.08230, 2017.

[5] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. *CoRR*, abs/1104.3913, 2011.

[6] Stephen Gillen, Christopher Jung, Michael J. Kearns, and Aaron Roth. Online learning with an unknown fairness metric. *CoRR*, abs/1802.06936, 2018.

[7] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016.

[8] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[9] Keith Kirkpatrick. It's not the algorithm, it's the data. *Communications of the ACM*, 60:21–23, 01 2017.

[10] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness, 2018.

[11] RM Neal. *Bayesian learning for neural networks*. PhD thesis, University of Toronto, 1995.

[12] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

[13] Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. The problem of inframarginality in outcome tests for discrimination, 2017.

[14] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, page 1–7, New York, NY, USA, 2018. Association for Computing Machinery.

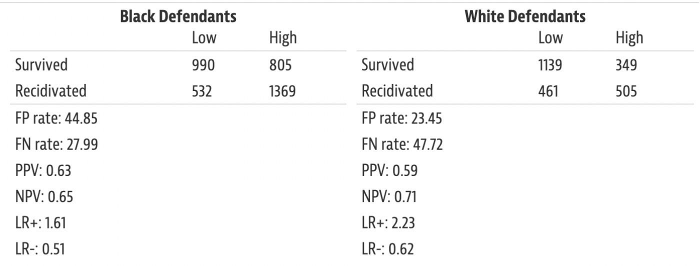
[15] Yongkai Wu, Lu Zhang, and Xintao Wu. Fairness-aware classification: Criterion, convexity, and bounds. *CoRR*, abs/1809.04737, 2018.

[16] Ed Yong. A popular algorithm is no better at predicting crimes than random people. *The Atlantic*, pages 55064–6, 2018.

[17] Indre Zliobaite. On the relation between accuracy and fairness in binary classification. *CoRR*, abs/1505.05723, 2015.

7 Appendices

7.1 The COMPAS Algorithm



Black Defendants	Low	High
Survived	990	805
Recidivated	532	1369

FP rate: 44.85
FN rate: 27.99
PPV: 0.63
NPV: 0.65
LR+: 1.61
LR-: 0.51

White Defendants	Low	High
Survived	1139	349
Recidivated	461	505

FP rate: 23.45
FN rate: 47.72
PPV: 0.59
NPV: 0.71
LR+: 2.23
LR-: 0.62

Figure 7

Figure 7 shows logistic analysis of the dataset broken down by race performed on Compas by ProPublica. Here we see Black defendants who do not recidivate were almost twice as likely to be classified as a higher risk compared to white defendants, as the false positive(FP) rate is almost double for Black defendants.

7.2 Munging the Recidivism Dataset

Not all of the rows are useful for analysis, particularly due to missing data and reasonable constraints on timeframe and offense.

- If the charge date of a defendants COMPAS scored crime was not within 30 days from when the person was arrested, we assume that because of data quality reasons, that we do not have the right offense.
- Ordinary traffic offenses – those with a c_charge_degree of 'O' – will not result in Jail time are removed.
- We filtered the underlying data from Broward county to include only those rows representing people who had either recidivated in two years, or had at least two years outside of a correctional facility.
- rows that were missing race were also filtered out to provide proper sectioning.

7.3 Dataset Schema & Sample Data

```
sex                         category
age                         int64
age_cat                     category
race                        category
```

```
decile_score                int64
priors_count                int64
days_b_screening_arrest     int64
c_charge_degree             category
is_recid                    int64
score_text                  category
two_year_recid              int64
```

sex	age	age_cat	race	decile_score	priors_count	c_charge_degree	is_recid	score_text
Male	69	Greater than 45	Other	1	0	F	0	Low
Male	34	25 - 45	Black	3	0	F	1	Low
Male	24	Less than 25	Black	4	4	F	1	Low
Male	44	25 - 45	Other	8	0	M	0	Low
Male	41	25 - 45	Caucasian	6	14	F	1	Medium

Table 1: *Munged & truncated Broward County defendant history data set*