Delivering Fairness in Human Resources AI: Mutual Information to the Rescue

Léo Hemamou
iCIMS,
15 rue de Bucarest,
75008 Paris, France.
l.hemamou@gmail.com

William Coleman
iCIMS, Dogpatch Labs,
CHQ Building, Custom House Quay,
Dublin 1, D01 Y6H7, Ireland.
william.coleman@icims.com

Abstract

Automatic language processing is used frequently in the Human Resources (HR) sector for automated candidate sourcing and evaluation of resumes. These models often use pre-trained language models where it is difficult to know if possible biases exist. Recently, Mutual Information (MI) methods have demonstrated notable performance in obtaining representations agnostic to sensitive variables such as gender or ethnicity. However, accessing these variables can sometimes be challenging, and their use is prohibited in some jurisdictions. These factors can make detecting and mitigating biases challenging. In this context, we propose to minimize the MI between a candidate's name and a latent representation of their CV or short biography. This method may mitigate bias from sensitive variables without requiring the collection of these variables. We evaluate this methodology by first projecting the name representation into a smaller space to prevent potential MI minimization problems in high dimensions.

1 Introduction

There are numerous examples of Artificial Intelligence (AI) systems which fail to mitigate bias contained within datasets used to train models (Mehrabi et al., 2021; Crawford, 2021; Peña et al., 2020; Buolamwini and Gebru, 2018; Holstein et al., 2019). Bias can be introduced via human labelling or via data extracted from existing human processes which replicates societal biases (Barocas and Selbst, 2016). Left unchecked, machine learning models will reflect directly the data used to train them or possibly even exacerbate the effect of biased data. This is of particular concern in high-risk domains such as Human Resources (HR), where models can be used to assess candidates based on data provided in a Curriculum Vitae (CV) (Sánchez-Monedero et al., 2020a).

Large pre-trained language models (LLMs) have been the source of impressive performance gains in recent times on tasks such as question answering (Yan et al., 2021), common-sense reasoning (Wei et al., 2022), computer coding (Xu et al., 2022) and other domains. However, their capabilities are characterized poorly, requiring a greater understanding of their function to ameliorate potential harms (Srivastava et al., 2022). Fine-tuning LLMs on downstream tasks has become the gold standard for approaching many natural language processing (NLP) tasks (Ruder, 2021). However, the nature of this workflow means that practitioners who fine-tune such models on downstream tasks have little visibility of the data used to train the original model purely because of the volume of data involved. This lack of visibility can be problematic given that these models are trained on huge volumes of text data which may contain hidden biases (Crawford, 2021).

Mutual Information (MI) is a method for measuring the dependence between two features. It is the reduction in uncertainty for one random variable caused by knowledge of another. Cover and Thomas (1991) define it for two random variables X and Y, having a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$, as the relative entropy between the joint distribution and the product distribution:

$$MI(X;Y) = \mathbb{E}_{p(x,y)}\left[log\frac{p(X,Y)}{p(X)p(Y)}\right] \quad (1)$$

Here, $\mathbb{E}_{p(x,y)}$ is the expected value over the distribution p. MI is never negative, and values greater than zero indicate some degree of dependence between the variables (Kinney and Atwal, 2014). It has been extensively explored in domains such as statistics, robotics and bioinformatics (Cheng et al., 2020) in addition to machine learning (Pichler et al., 2022; Cheng et al., 2020; Chen et al., 2016; Alemi et al., 2017; Hjelm et al., 2019; Belghazi et al.,

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 867–882
November 20–23, 2022. ©2022 Association for Computational Linguistics

2021). In machine learning, it can be used to measure the amount of sensitive information, such as gender or ethnicity, contained in a CV in a hiring process. Using MI as a loss function's regularizer, the dependence between variables can be minimized (Cheng et al., 2020), thus disentangling sensitive and non-sensitive information in representations used to train models. In this work, we will refer to the process of applying MI to a representation to minimize the sensitive information held within it as 'disentanglement'.

We note that it is sometimes challenging to collect data on such sensitive variables due to privacy concerns and that it is even illegal in some jurisdictions (Lieberman, 2001). To overcome this problem, we propose using candidate names as a proxy for sensitive variables by reducing the MI between name and CV/BIOS embeddings.

We investigate three approaches to MI estimation which have already seen attention in the literature within a HR context: Info-NCE (van den Oord et al., 2019), CLUB (Cheng et al., 2020) and KNIFE (Pichler et al., 2022). Furthermore, we present low-dimensional versions of these algorithms, which are of interest given MI difficult to estimate in high dimensions (Kraskov et al., 2004; McAllester and Stratos, 2020; Pichler et al., 2020). We present results on experiments carried out on two datasets relevant to HR applications: FairCVTest (Peña et al., 2020; Morales et al., 2020), consisting of synthetic CV data and BIOS (De-Arteaga et al., 2019), a collection of freely available online short biographies in English.

Our contributions are as follows:

- We evaluate MI methods for disentangling sensitive information from unstructured data (i.e. image or text) in an HR application.
- We successfully remove sensitive information without accessing and retraining the pretrained backbone models and without requiring the collection of sensitive information, a critical point given that collecting such information is prohibited in some jurisdictions.
- Our proposed methodology simultaneously removes multiple biases (in the examples detailed, gender and ethnicity information).
- We show experimentally that this disentanglement leads to fairer models.

2 Related Work

This work motivates an investigation of MI estimators by highlighting the requirement for fairness procedures within AI-augmented systems for HR applications, such as hiring processes. We build on the MI estimators proposed by van den Oord et al. (2019) Info Noise Contrastive Estimation (InfoNCE), Cheng et al. (2020) Contrastive Log-ratio Upper Bound (CLUB) and Pichler et al. (2022) Kernelized-Neural Differential Entropy Estimation (KNIFE). To our knowledge, our work is the first wholly focused on the HR domain to investigate the potential use of MI in hiring processes. We note that Kamimura (2019) utilizes an HR dataset in his work, but it cannot be said to be focused entirely on the HR domain as his subject is validating a simplified method for calculating MI, which he demonstrates on HR, crab species and wholesale datasets.

2.1 Fairness via Privacy

HR processes are known to be sub-optimal as they are not free from bias introduced by practitioners (Sánchez-Monedero et al., 2020b). There is substantial literature on gender bias in the domain; for example, (Bertrand and Duflo, 2016; Bertrand and Mullainathan, 2003; Ginther and Kahn, 2004; Sarsons, 2017a,b). AI systems have the potential to address such problems, ensuring they do not themselves introduce or amplify bias must be prioritized (Köchling and Wehner, 2020; Giang, 2018; Wachter-Boettcher, 2017).

Bias mitigation in the HR domain has recently seen attention in the literature. Two main directions are being taken, namely "fairness through awareness" (Dwork et al., 2012; Kusner et al., 2017) and "fairness through unawareness" (Kusner et al., 2017; Grgic-Hlacˇa et al., 2016). In "fairness through awareness," researchers seek to make models more equitable by considering the sensitive variable. However, this approach may sometimes be inapplicable when affirmative action is prohibited by law (e.g., in France, the United Kingdom or Germany) (Lieberman, 2001). In "fairness through unawareness," researchers try to remove all information related to sensitive variables from the models. These approaches are closely related to privacy protection methods where network designers try to protect their system from attackers trying to extract personal information from latent representations, for example, through adversarial training (Jaiswal and Mower Provost, 2020a; Hemamou et al., 2021; Morales et al., 2020).

Lately, new methods using MI have emerged and have shown very good performance in disentangling representations (van den Oord et al., 2019; Cheng et al., 2020; Belghazi et al., 2021; Pichler et al., 2022). By minimizing the MI between the candidate name embedding and the latent representation of automatic models, we propose to evaluate these methods in the context of HR to obtain fairer models.

2.2 InfoNCE

In the approach outlined in (van den Oord et al., 2019), the authors utilize an encoder and autoregressive model to jointly optimize a loss based on Noise-Contrastive Estimation (NCE), which they term InfoNCE, to estimate a lower-bound for MI. Positive pairs, two representations from the same instance, are contrasted with negative pairs that contain a representation drawn from two different instances (which is, therefore, incorrect). We refer to the original work (van den Oord et al., 2019) for full technical details. One drawback of this method is that if there is some factor in a negative pair which has a positive association with the prediction task, this can mask the negative association we hope to capture in the negative pair. Additionally, this method may prove intractable if there is an extreme dimension mismatch between the two representations.

2.3 CLUB

Cheng et al. (2020) present an upper-bound MI estimator based on the difference of conditional probabilities between positive and negative sample pairs leveraging contrastive learning. Consider two random variables X and Y between which we want to measure the MI. The authors attempt to find a function that maps the mean and standard deviation for each dimension of Y for X. If these variables are related, the error will be much smaller than the estimated error observed in negative samples. However, the possibility of multiple dimensions in Y that are irrelevant to X is potentially problematic, as is the assumption of gaussian distributions for mean and standard deviation values.

2.4 KNIFE

Pichler et al. (2022) estimates differential entropy and conditional differential entropy to compute MI. Empirically, they show that KNIFE can adapt to distributions substantially different from the gaussian kernel shape contrary to the CLUB estimator. They validate this on text and image data. While reporting encouraging results, however, the architecture takes a long time to train, is complex and requires large data volumes to ensure it performs well. Similarly to CLUB, the potential high dimensionality of Y can be problematic and obscure the signal of the dimension of interest for MI estimation.

3 Methodology

Our method aims to minimize the MI between a latent representation of an individual's input (either BIOS or CV embedding) and a word embedding of their name. Our method comprises two steps. Firstly, we project the name embedding into a rich low-dimensional space to solve the curse of dimensionality problem for the MI estimators. Secondly, we minimize the MI between the representation of an individual's input and the latter disentangled representation. This method allows us to find possible sensitive, latent variables influencing the two views of the data (e.g. candidate gender influences the name of the candidate and the embedding of their CV) and to simultaneously mitigate the biases coming from these sensitive variables. We emphasize that this sensitive information is not used in the classifier but only in the disentanglement procedure. Therefore, it is unnecessary to collect the sensitive variables after deploying the classifier.

To generate name representations, we follow the approach of Romanov et al. (2019) who use FastText (Bojanowski et al., 2017) embeddings for this purpose. We note that this does not address the issue of Out of Vocabulary (OOV) names - a critical point in any real-world implementation of this method which would require robust testing for edge cases using different embedding schemas. We focus on comparing the different MI estimators; thus, we leave experimentation with name representations to subsequent experiments.

3.1 Formulation

We define x_i to be a data point of an individual (e.g. resume embedding or biography embedding), y_i to be its corresponding label (e.g. resume score or a job occupation), t_i to be the name embedding of the individual and s_i a private label (ethnicity, gender) used only for the disentanglement procedure. In our experiments, we decompose the primary task into two components, namely an encoder f_ϕ and a regression or classification head f_c:

$$z_i = f_\phi(x_i)$$
$$\hat{y}_i = f_c(z_i)$$

Here, z_i is a meaningful latent representation of x_i relative to the regression/classification task and $\hat{y}_i$ is the predicted score or probability label for x_i. Our method aims to minimize MI between z_i and t_i while maximizing performance on predicting $\hat{y}_i$.

3.2 Dimension reduction of target space via maximization of MI

The estimation and minimization of MI are challenging problems, especially between two high-dimensional continuous feature spaces. To address these issues, we propose to refine the continuous target space of the name embedding t_i by reducing it to a lower dimensional space. Thus, we propose to maximize a lower bound of MI via Noise Contrastive Estimation (NCE) based on a modified version of InfoNCE:

$$\mathrm{I}_{\text{NCE}} := \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{sim(\tilde{x}_i,\tilde{t}_i)})}{\frac{1}{N}\sum_{j=1}^{N}e^{sim(\tilde{x}_i,\tilde{t}_j)}}\right] \quad (2)$$

with

$$\tilde{x}_i = f_\psi(x_i)$$
$$\tilde{t}_i = g_\psi(t_i)$$

Here, f_ψ and g_ψ are two neural networks projecting in a lower dimensional continuous feature space, and sim calculates cosine similarity between two vectors. Finally, the expectation is over N samples $\{(x_i, t_i)\}_{i=1}^{N}$ drawn from the joint distribution $p(x, t)$. We expect to learn a useful encoder g_ψ projecting t_i into a rich lower dimension by maximising this lower bound.

3.3 Disentanglement via Minimization of MI

Once we obtain the rich low-dimensional representation $\tilde{t}_i$, we freeze the encoder g_ψ and we optimize the following loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}}(y_i, \hat{y}_i) + \lambda \cdot \mathrm{MI}(z_i, \tilde{t}_i) \quad (3)$$

In this case, MI refers to the value of MI computed by one of the MI estimators (InfoNCE, CLUB or KNIFE), and λ is a scaling factor to parameterize the degree of influence of MI for an experiment.

4 Evaluation

4.1 Datasets

We assess the formulation proposed in Section 3 using two datasets: FairCVtest (Peña et al., 2020; Morales et al., 2020) and the BIOS dataset (De-Arteaga et al., 2019). These datasets are available publicly under standard licenses, and their usage in this work is consistent with their intended usage in a research context. Below, we offer basic descriptions and identify only where our approach differs from the original authors. We refer to the original papers for other implementation details.

4.1.1 FairCVtest

The FairCVtest dataset[1] consists of 24,000 synthetic CVs which contain both structured data in tabular format which present data about job proficiency and unstructured data such as face images and text (short biographies and experience profiles, for example). Gender and racially biased scores are applied consciously to each candidate (Peña et al., 2020). For this work, we use the same data splits as the authors and randomly select 10% of the training split as a validation set. We generate a name for each entry based on the gender and ethnicity specified using the same method used by (Romanov et al., 2019). FastText (Bojanowski et al., 2017) embeddings are used to represent candidate names in our algorithms (resp t_i).

4.1.2 BIOS

The BIOS dataset[2] consists of approximately 400,000 short biographies of individuals from twenty-eight different occupations where the classification task is to predict the individuals' occupation from the biography. Due to the dataset size, the authors provide code to generate the raw data. However, as the version of common-crawl used to generate the dataset is a more recent version than that used by the authors of the original paper (De-Arteaga et al., 2019) our understanding is that we cannot assert that the dataset used here is the same as theirs, though we expect that it is very similar. Extraction of each individual's name is possible because of the biography selection method used. We have augmented this dataset by inferring the individual's ethnicity using a dedicated neural network called *RaceBERT* (Parasurama, 2021). FastText

[1] https://github.com/BiDAlab/FairCVtest

[2] https://github.com/Microsoft/biosbias

embeddings represent the names, and the ethnicity variable is cast as binary (White/Non-White) to address the class imbalance. The biography embedding is generated from the last hidden state *CLS* token from a pre-trained *distilROBERTa* model.

4.2 Evaluation Metrics

We evaluate our experiences along three dimensions. The first dimension is performance: we ask whether our methods degrade performance on utility tasks. The second dimension is the private task: we evaluate the amount of information left to retrieve sensitive variables from the model. The last dimension consists of a fairness metric: we evaluate the possible bias in the trained models' scores between the different groups.

4.2.1 Performance Metrics

To evaluate the main task for the FairCVtest dataset, we use mean absolute error (MAE) as the label is a candidate score. In the case of the BIOS dataset, we use the balanced True Positive Rate (TPR) due to the uneven class distribution of the occupation target labels. Balanced TPR is the average of TPR for each job position.

4.2.2 Privacy Metrics

We train two diagnostic classifiers, XGBoost and Logistic Regression, to recover the sensitive variables of gender and ethnicity from the latent representation of the network. We use the Area Under the Curve (AUC-ROC) of these classifiers as only one of the categories (the ethnicity category for the BIOS dataset) is somewhat imbalanced (see Figure 10). Also, we care equally about the performance for all categories, which mitigates against use of the Precision-Recall AUC (AUC-PR), which is generally the appropriate metric for imbalanced classes (Saito and Rehmsmeier, 2015). We report the AUC-PR scores for both classes of the BIOS ethnicity category in Table 12 and Figure 4. If the performance of these models is good, it means the representation still contains sensitive information. This is a method widely used in the fairness and privacy literature (Jaiswal and Mower Provost, 2020b; Xie et al., 2017; Hemamou et al., 2021).

4.2.3 Fairness Metrics

We leverage two metrics to monitor fairness. In the case of the FairCVtest dataset, we report Kullback Leibler (KL) Divergence, a similarity measure for probability distributions. For the BIOS dataset, we follow the approach of Romanov et al. (2019), who compute a TPR ethnicity and gender gap defined as the differences in the TPRs between ethnicities and genders for each occupation. They define the gender TPR gap for an occupation c as:

$$Gap_{g,c} = TPR_{g,c} - TPR_{\sim g,c} \tag{4}$$

Here, g and $\sim g$ are binary genders, replaced with binary ethnicity values for the ethnicity metric. We also implement the same Root Mean Square (RMS) TPR gap metric as used by (Romanov et al., 2019), as it allows us to report a single score to quantify bias to provide ease of comparison. We square the gap values as we wish to mitigate more significant biases. This metric is formulated as follows in the case of gender:

$$Gap_g^{RMS} = \sqrt{\frac{1}{|C|}\sum_{c \in C} Gap_{g,c}^2} \tag{5}$$

We report the maximum TPR gap to facilitate worst-case analyses as per Romanov et al. (2019).

5 Experiments

We first examine the dimension reduction of target space to understand its utility. We then present the main results of our evaluation before discussing limitations and future works.

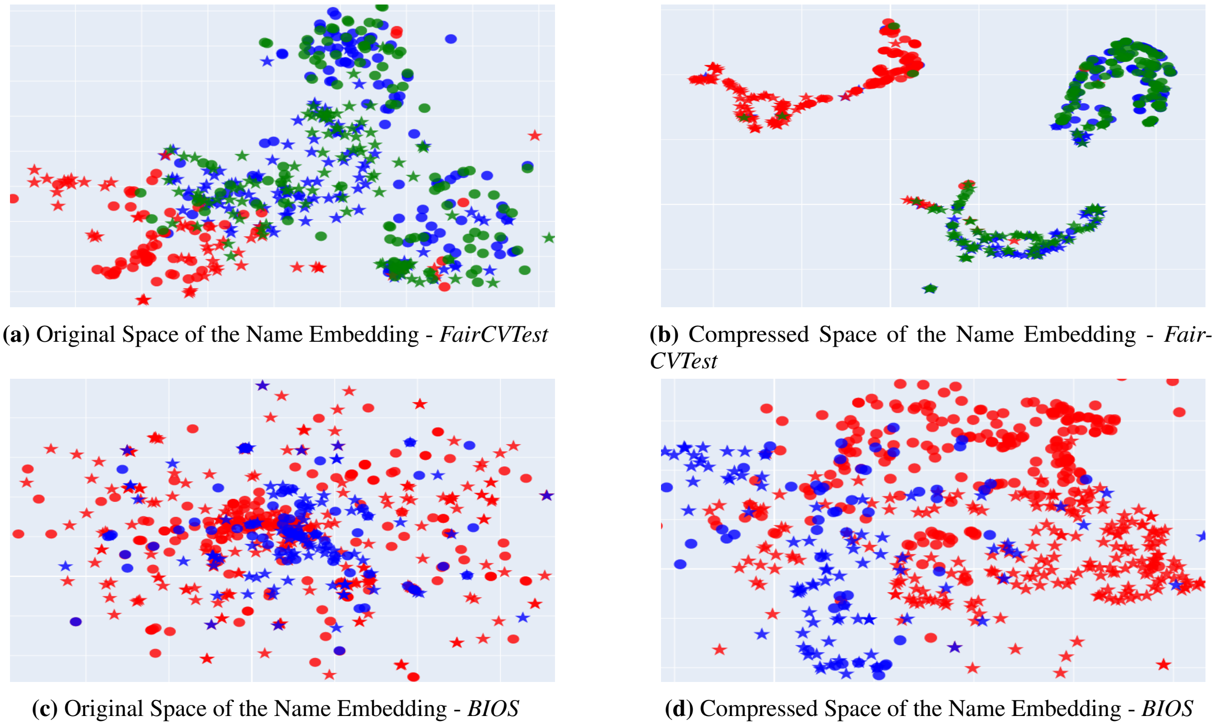
5.1 Low Dimensional Word Embeddings of Names as Proxies

In order to visualize and understand the usefulness of our proposed dimensionality reduction, we present in Figure 1 a 2-D UMAP[3] projection of the original space of the name embedding $\{t_i\}_{i=1}^N$ and the compressed space of the name embedding $\{\tilde{t_i}\}_{i=1}^N$.

Regarding gender (star vs circle in Figure 1), the separation is unclear in the original space projection for both datasets. In the projection of the compressed space, the separation is more apparent for both datasets.

Regarding ethnicity (i.e. color in Figure 1), the separation between groups is unclear in the original space projection for the FairCVTest dataset and even worse on the BIOS dataset. In the projection of the compressed space, the separation is better for both datasets. However, this separation is less pronounced on the BIOS dataset, possibly due to the non-synthetic nature of this dataset, which

[3] https://umap-learn.readthedocs.io/en/latest/



(a) Original Space of the Name Embedding - *FairCVTest*

(b) Compressed Space of the Name Embedding - *FairCVTest*

(c) Original Space of the Name Embedding - *BIOS*

(d) Compressed Space of the Name Embedding - *BIOS*

Figure 1: UMAP projections of proxy space on the FairCVTest and BIOS dataset. Color refers to ethnicity (red, blue and green are white, Asian and African American). The ethnicity class is reduced to binary "White" (in red) and "Non-White" (in blue) categories for the BIOS dataset. The symbol refers to gender, a circle for males and a star for females.

could indicate that the data reflects other dimensions such as socio-economic class, religion or age. The presence of multiple potential dimensions of bias in non-synthetic data is a factor that is ripe for further investigation. Finally, there is no clear separation between the African American (i.e. green) and Asian (i.e. blue) groups in the original and compressed space for the FairCVTest dataset.

This result shows that names encode sensitive variables such as ethnicity or gender. In addition, this demonstrates that it is possible through MI methods to obtain a lower dimensional representation better suited as a proxy for sensitive variables.

5.2 FairCVTest

Setup. The main task of this dataset is automatic CV scoring. From the original CV score, two biased labels are designed where additive biases depending on the sensitive classes are added. Without loss of generality, we treat this problem as a multi-task problem where we try to predict these two labels simultaneously.

Results. Figure 2 gathers results on the FairCVTest dataset. The red dotted line represents a vanilla model trained without MI minimization (case $\lambda = 0$). The green dashed line represents an oracle model trained with the input completely agnostic of gender and ethnicity. Note that biased models naturally perform better in the main tasks, as the label is biased towards sensitive categories. Thus, the oracle provides us with the information on the maximum performance on the main tasks without using any sensitive information. On gender and ethnicity, we can observe that InfoNCE-LD and Knife-LD perform better than the other MI estimators reaching nearly perfect privacy for the gender task while preserving performance on the primary task close to that of the oracle. However, a limit (AUC $\approx$ 0.7) seems to appear for ethnicity, which is in agreement with the observations of the section 5.1 regarding a lack of separation between the "Black" and "Asian" groups. Concerning the fairness metrics, the MI estimators' use of the low dimensional target space seems to perform better, especially for low lambda values (e.g. 0.1 or 1). With lambda greater than or equal to 10, KNIFE-LD and InfoNCE-LD reach near-perfect fair predictions with a KL divergence nearly equal to 0, showing the capability of MI minimization to reduce the potential bias of the classifier. Finally, we can note that the CLUB estimator does not improve with respect to the use of small dimensions for the target space.

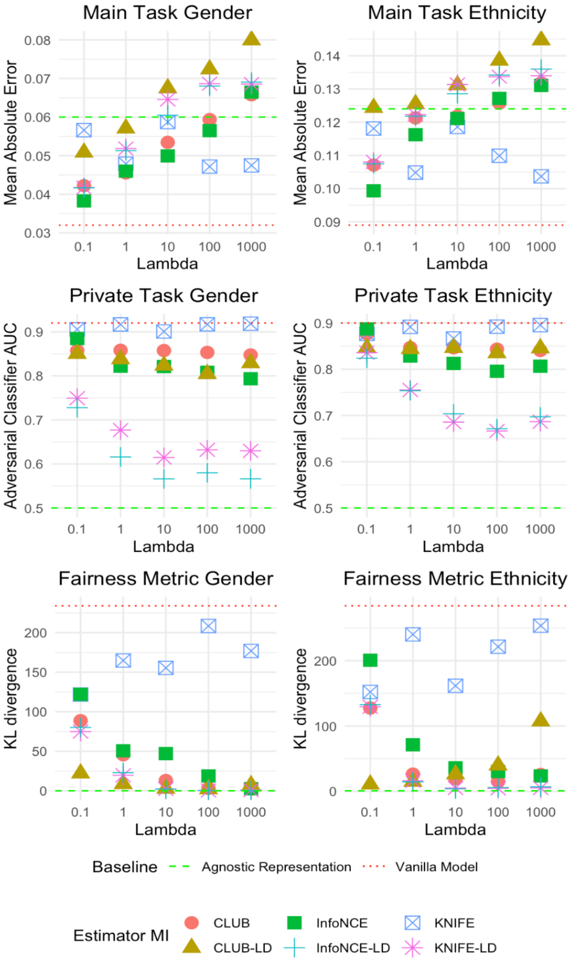


Figure 2: FairCVTest - Results on the MI training on the Resume Scoring Task, Private Task and Fairness Metrics depending on the lambda value and the MI estimator. The appendage of *-LD* indicates the application of the MI estimator and minimization on the compressed representation of the name embedding.

5.3 BIOS

Setup. This primary task here is to classify job positions based on the candidates' short biography. As in previous work, due to the strong class imbalance problem, we use a weighted cross-entropy loss as $\mathcal{L}_{\text{task}}$ with weights set to the values proposed by (Cui et al., 2019).

Results. Figure 3 presents results on the BIOS dataset. First of all, we can see that the representation of a LLM (Baseline Vanilla Model) indeed contains sensitive information and implies biases during training.

Concerning the main task, we can see that performance deteriorates for larger lambda values. Thus, for lambda = 5, a significant decrease is observed for the InfoNCE-LD estimator. For lambda = 10, this performance degradation is visible for all estimators except InfoNCE.

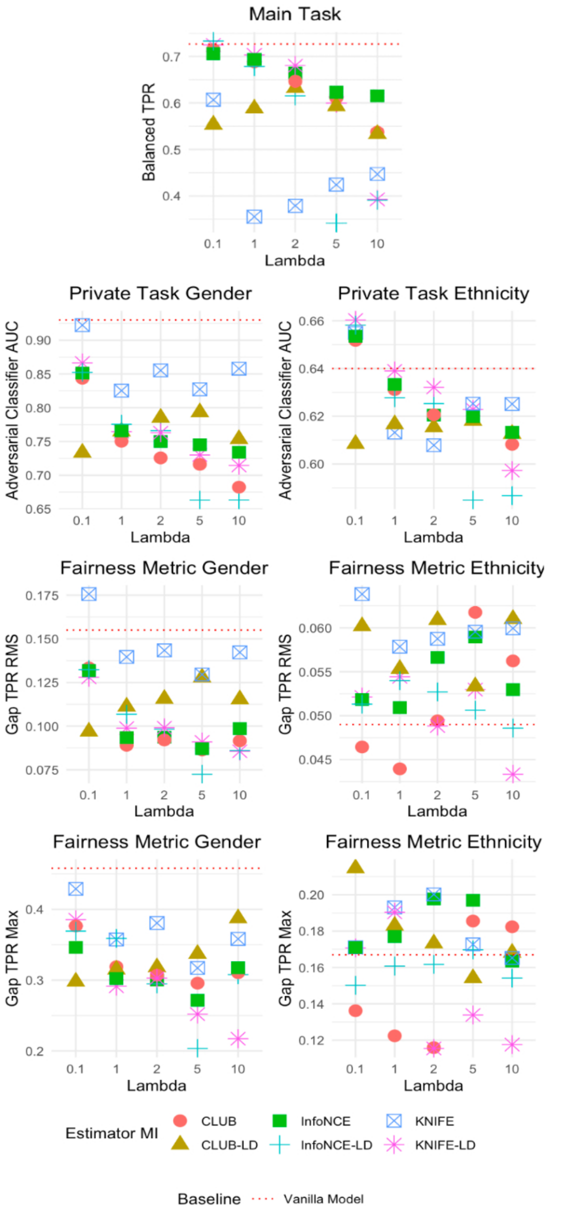


Figure 3: BIOS - Results of the MI training on the BIOS Job Classification Task, Private Task and Fairness Metrics depending on the lambda value and the MI estimator. The appendage of *-LD* indicates the application of the MI estimator and minimization on the compressed representation of the name embedding.

Considering the private tasks, we can see that a larger lambda reduces the capability of an adversarial classifier to retrieve sensitive attributes, particularly for the estimators CLUB, InfoNCE, InfoNCE-LD and KNIFE-LD.

Examining the Fairness Metrics, we can see that our method reduces the RMS error and the maximum TPR gender gap. Thus, when lambda is equal to 1 or 2, the RMS TPR Gap goes from 0.15 to 0.1, and the maximum TPR Gap goes from 0.50 to 0.3 for four estimators, namely: InfoNCE, InfoNCE-

LD, CLUB and KNIFE-LD. This improvement is not visible regarding the maximum and RMS TPR ethnicity gap, possibly because the original model is not specifically biased towards ethnicity and our method has no salient effect.

Regarding the beneficial effect of a compressed name representation, we can see that this is necessary for the KNIFE estimator. Concerning the CLUB estimator, using such a space seems to degrade the performance. The significant difference between the low dimensional space distribution and that of a gaussian distribution could explain this poorer result. Finally, contrary to the experiment on the FairCVTest dataset, no significant difference is visible for the InfoNCE estimator.

6 Conclusion and Discussion

In this paper, we propose the use of MI minimization in the context of HR to obtain fairer automatic models. In contrast to previous work that explicitly uses variables to be removed, we use a candidate name representation as a proxy. We show experimentally on two datasets that MI methods help obtain better-anonymized representations and fairer models while conserving task performance. Moreover, we show that the dimension reduction of candidate name word embeddings allows us to overcome some problems related to estimating MI in high dimensions. Overall, this work is the first to evaluate the use of MI in such an application context by considering the real-world limitations of sensitive data collection. Finally, we hope this work will attract research interest in this challenging and vital task.

6.1 Limitations

While the MI methods explored in this work are successful in mitigating biases, they are not successful in removing sensitive elements of representations entirely. Also, to simplify our analysis, we have been reductive in our treatment of some categories: simplifying the BIOS ethnicity category to white and non-white categories, for example. We justify this by pointing out we use this binary categorisation in the evaluation step only and that this approach follows established methods (Romanov et al., 2019). Neither have we controlled for factors such as religion, socio-economic status, age, or others, though we would note that an advantage of our method is that it uses MI minimization between two continuous representations. By doing so, we overcome the problem of categorizing or discretizing the name representation. Investigations of bias mitigation on categories such as religion, age and others, are suitable topics for future research requiring the annotation of datasets with these attributes to investigate if the results reported here are replicated for other categories.

The methods explored here vary in complexity, and their computational intensity is another non-trivial factor. Implementation requires an understanding of the influence of hyperparameters and an ability to enter into a computationally intensive grid search which may be infeasible for companies without dedicated machine learning resources. Also, we note that these methods rely on recognizing the existence of certain biases. They are not a protection against bias that is unknown or unacknowledged.

6.2 Risks

We have outlined a series of experiments that address bias mitigation in a laboratory setting. Real-world implementation must build on these methods and address some of the simplifications introduced to facilitate ease of analysis. While we have demonstrated some success in bias mitigation in the foregoing, we cannot presume these methods can remove all bias. We have used name embeddings as proxies for sensitive information, but names may not be a foolproof method to reflect social attributes. People can change their names or manifest different characteristics from others with similar names. We, therefore, argue that the results presented here are promising but not a complete solution to a problem area that requires further investigation.

To counter this, while we achieve partial success, in this case, we would also caution against the risk of refusing to implement these methods because they are only a partial solution. Solely human-based hiring processes are biased (Mehrabi et al., 2021). The application of these methods can reduce these biases.

Acknowledgements

We would like to thank our colleagues in the Talent Cloud AI team for their encouragement and also iCIMS for their support.

References

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2017. Deep Variational Information

Bottleneck. In *ICLR*. arXiv.

Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review*, 104:671.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R. Devon Hjelm. 2021. MINE: Mutual Information Neural Estimation.

Marianne Bertrand and Esther Duflo. 2016. Field Experiments on Discrimination. Technical Report w22014, National Bureau of Economic Research, Cambridge, MA.

Marianne Bertrand and Sendhil Mullainathan. 2003. Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. 2020. CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information. *arXiv:2006.12013 [cs, stat]*.

Thomas M Cover and Joy A Thomas. 1991. *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley & Sons Inc., New York.

Kate. Crawford. 2021. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, New Haven and London.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, NY, USA. Association for Computing Machinery.

Vivian Giang. 2018. The Potential Hidden Bias In Automated Hiring Systems. https://www.fastcompany.com/40566971/the-potential-hidden-bias-in-automated-hiring-systems.

Donna K. Ginther and Shulamit Kahn. 2004. Women in Economics: Moving Up or Falling Off the Academic Career Ladder? *Journal of Economic Perspectives*, 18(3):193–214.

Nina Grgic-Hlac ˇa, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2016. The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making. *NIPS Symposium on Machine Learning and the Law*, 1(2):11.

Léo Hemamou, Arthur Guillon, Jean-Claude Martin, and Chloé Clavel. 2021. Don't judge me by my face: An indirect adversarial approach to remove sensitive information from multimodal neural representation in asynchronous job video interviews. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8.

R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudík, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Mimansa Jaiswal and Emily Mower Provost. 2020a. Privacy Enhanced Multimodal Neural Representations for Emotion Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7985–7993.

Mimansa Jaiswal and Emily Mower Provost. 2020b. Privacy Enhanced Multimodal Neural Representations for Emotion Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:7985–7993.

Ryotaro Kamimura. 2019. Supposed Maximum Mutual Information for Improving Generalization and Interpretation of Multi-Layered Neural Networks. *Journal of Artificial Intelligence and Soft Computing Research*, 9(2):123–147.

Justin B. Kinney and Gurinder S. Atwal. 2014. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359.

Alina Köchling and Marius Claus Wehner. 2020. Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, 13(3):795–848.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Physical Review E*, 69(6):066138.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Robert Lieberman. 2001. A Tale of Two Countries: The Politics of Color Blindness in France and the United States. *French Politics, Culture & Society*, 19(3):32–59.

David McAllester and Karl Stratos. 2020. Formal Limitations on the Measurement of Mutual Information. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 875–884. PMLR.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):115:1–115:35.

Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, and Ruben Tolosana. 2020. SensitiveNets: Learning Agnostic Representations with Application to Face Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP.

Prasanna Parasurama. 2021. raceBERT - A Transformer-based Model for Predicting Race and Ethnicity from Names. *ArXiv*.

Alejandro Peña, Ignacio Serna, Aythami Morales, and Julian Fierrez. 2020. Bias in Multimodal AI: Testbed for Fair Automatic Recruitment. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 129–137.

Georg Pichler, Pierre Colombo, Malik Boudiaf, Gunther Koliander, and Pablo Piantanida. 2022. KNIFE: Kernelized-Neural Differential Entropy Estimation.

Georg Pichler, P. Piantanida, and Günther Koliander. 2020. On the Estimation of Information Measures of Continuous Distributions. *ArXiv*.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Tauman Kalai. 2019. What's in a Name? Reducing Bias in Bios without Access to Protected Attributes. *arXiv:1904.05233 [cs, stat]*.

Sebastian Ruder. 2021. Recent Advances in Language Model Fine-tuning. https://ruder.io/recent-advances-lm-fine-tuning/.

Takaya Saito and Marc Rehmsmeier. 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE*, 10(3):e0118432.

Javier Sánchez-Monedero, Lina Dencik, and Lilian Edwards. 2020a. What does it mean to 'solve' the problem of discrimination in hiring? social, technical and legal perspectives from the UK on automated hiring systems. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 458–468, New York, NY, USA. Association for Computing Machinery.

Javier Sánchez-Monedero, Lina Dencik, and Lilian Edwards. 2020b. What does it mean to 'solve' the problem of discrimination in hiring? social, technical and legal perspectives from the UK on automated hiring systems. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 458–468, New York, NY, USA. Association for Computing Machinery.

Heather Sarsons. 2017a. Gender Differences in Recognition for Group Work. *Journal of Political Economy*, 129(1).

Heather Sarsons. 2017b. Interpreting Signals in the Labor Market: Evidence from Medical Referrals. *Job Market Paper*, pages 141–45.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748 [cs, stat]*.

Sara Wachter-Boettcher. 2017. Why You Can't Trust AI to Make Unbiased Hiring Decisions. https://time.com/4993431/ai-recruiting-tools-do-not-eliminate-bias/.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. 2017. Controllable Invariance through Adversarial Feature Learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Frank F. Xu, Uri Alon, Graham Neubig, and Vincent Josua Hellendoorn. 2022. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, MAPS 2022, pages 1–10, New York, NY, USA. Association for Computing Machinery.

Yuanmeng Yan, Rumei Li, Sirui Wang, Hongzhi Zhang, Zan Daoguang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Large-Scale Relation Learning for Question Answering over Knowledge Bases with Pre-trained Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3653–3660, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Appendix

This Appendix provides additional experiment details, such as model parameters, results tables, additional plots and dataset details. The batch size used for all experiments was 128. The results reported are the average across three random seeds. NVIDIA Tesla K80 GPUs were used to carry out the training on a cloud computing platform, which provided the run metrics reported in Table 7. In total, 1,627.5 GPU hours were expended running these experiments.

A.1 Model Parameters for FairCVtest Dataset

Encoder Model f_ϕ - FairCVtest

Layer Type	Input Length	Output Length
Fully Connected	32	20
Activation	Hyperbolic Tangent	
Fully Connected	16	20
Activation	Hyperbolic Tangent	

Table 1: Dimensions and details for the encoder model in the FairCVtest dataset experiments.

Regression Head Model f_c - FairCVtest

Layer Type	Input Length	Output Length
Fully Connected	20	1
Activation	Sigmoid	

Table 2: Dimensions and details for the regression head in the FairCVtest dataset experiments.

Name Embedding Encoder g_Ψ - FairCVtest

Layer Type	Input Length	Output Length
Fully Connected	100	16
Activation	Hyperbolic Tangent	
Fully Connected	16	16
Activation	Hyperbolic Tangent	
Fully Connected	16	16
Activation	Hyperbolic Tangent	

Table 3: Dimensions and details for the name embedding encoder in the FairCVtest dataset experiments.

A.2 Model Parameters for BIOS Dataset

The input embedding for the BIOS encoding model is generated from the last hidden state CLS token of a pre-trained distilROBERTa model.

Encoder Model f_ϕ - BIOS

Layer Type	Input Length	Output Length
Fully Connected	768	50
Activation	Hyperbolic Tangent	
Fully Connected	50	50
Activation	Hyperbolic Tangent	

Table 4: Dimensions and details for the encoder model in the BIOS dataset experiments.

Label Model f_c - BIOS

Layer Type	Input Length	Output Length
Fully Connected	50	28

Table 5: Dimensions and details for the label model in the BIOS dataset experiments.

Name Embedding Encoder g_Ψ - BIOS

Layer Type	Input Length	Output Length
Fully Connected	100	12
Activation	Hyperbolic Tangent	
Fully Connected	12	12
Activation	Hyperbolic Tangent	
Fully Connected	12	12
Activation	Hyperbolic Tangent	

Table 6: Dimensions and details for the name embedding encoder in the BIOS dataset experiments.

A.3 Supplemental Figure for the BIOS Ethnicity Category - AUC-PR

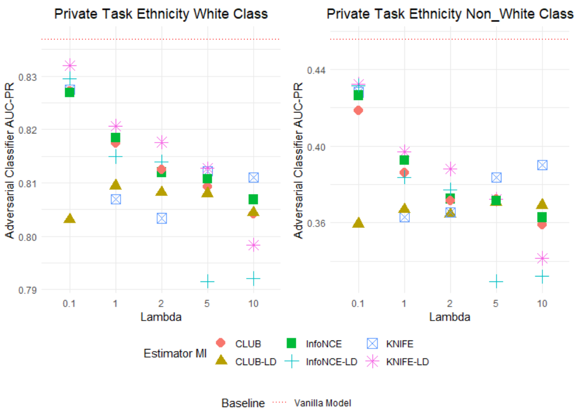


Figure 4: The ethnicity category for the BIOS dataset is imbalanced in a ratio of 3:1 for white versus non-white categories. For this reason, we generate the AUC-PR scores for this category as it is the appropriate metric for imbalanced data. We observe the same pattern here as with the AUC-ROC scores presented in Figure 3: the model becomes less accurate at predicting ethnicity as values of lambda increase, indicating the MI process is successful at removing ethnicity information from the representation.

A.4 GPU Training Hours per Mutual Information Estimator

Estimator	FairCVtest	BIOS	Total
KNIFE	390	300	690
CLUB	75	90	165
InfoNCE	75	90	165
KNIFE-LD	97.5	127.5	225
CLUB-LD	90	105	195
InfoNCE-LD	82.5	105	187.5
Total	**810**	**817.5**	**1627.5**

Table 7: GPU hours expended per MI estimator.

A.5 Dataset Characteristics

Dataset Split Sizes			
Dataset	Train	Validation	Test
FairCVtest	17,280	4,800	1,920
BIOS	247,010	38,571	94,435

Table 8: Details of splits used for each dataset.

	FairCVtest		
Label	**Train**	**Validation**	**Test**
	Ethnicity		
White	5765	1598	637
Asian	5695	1640	665
African-American	5820	1562	618
	Gender		
Male	8636	2400	964
Female	8644	2400	956
Total	**17280**	**4800**	**1920**

Table 9: Details of data splits used for the FairCVTest dataset including class sizes for the gender and ethnicity categories.

	BIOS		
Label	**Train**	**Validation**	**Test**
	Ethnicity		
White	183,048	28,660	70,035
Non-White	63,962	9,911	24,400
	Gender		
Male	113,414	17,731	43,559
Female	133,596	20,840	50,876
Total	**247,010**	**38,571**	**94,435**

Table 10: Details of data splits used for the BIOS dataset including class sizes for the gender and ethnicity categories.

A.6 Tables of Results

See following pages.

		Main Task (MAE)		Private Task (AUC-ROC)		Fairness Metric (KL Divergence)	
Estimator	**lambda_c**	**Gender**	**Ethnicity**	**Gender**	**Ethnicity**	**Gender**	**Ethnicity**
CLUB	0.1	0.042	0.107	0.857	0.877	88.756	127.554
CLUB	1	0.046	0.121	0.858	0.847	45.654	26.377
CLUB	10	0.053	0.122	0.857	0.846	13.095	19.203
CLUB	100	0.059	0.126	0.853	0.843	3.166	15.003
CLUB	1000	0.066	0.132	0.847	0.841	2.359	25.881
CLUB-LD	0.1	0.047	0.118	0.847	0.846	76.712	59.562
CLUB-LD	1	0.053	0.124	0.851	0.837	46.472	18.305
CLUB-LD	10	0.06	0.129	0.844	0.832	31.273	28.124
CLUB-LD	100	0.067	0.133	0.833	0.82	9.331	25.276
CLUB-LD	1000	0.072	0.139	0.836	0.823	9.568	45.959
KNIFE	0.1	0.057	0.118	0.906	0.876	121.768	152.077
KNIFE	1	0.048	0.105	0.917	0.892	164.8	240.278
KNIFE	10	0.059	0.119	0.901	0.867	155.534	161.667
KNIFE	100	0.047	0.11	0.917	0.892	208.462	221.448
KNIFE	1000	0.047	0.104	0.918	0.896	176.905	253.58
KNIFE-LD	0.1	0.042	0.107	0.824	0.862	111.392	160.064
KNIFE-LD	1	0.052	0.119	0.759	0.797	42.881	72.188
KNIFE-LD	10	0.06	0.127	0.701	0.748	21.757	37.958
KNIFE-LD	100	0.068	0.132	0.682	0.755	7.38	21.052
KNIFE-LD	1000	0.068	0.132	0.705	0.752	9.526	28.477
InfoNCE	0.1	0.038	0.099	0.885	0.887	122.047	200.76
InfoNCE	1	0.046	0.116	0.822	0.829	50.614	71.302
InfoNCE	10	0.05	0.121	0.821	0.813	47.074	36.173
InfoNCE	100	0.056	0.127	0.808	0.795	18.851	30.402
InfoNCE	1000	0.066	0.131	0.793	0.807	2.744	23.718
InfoNCE-LD	0.1	0.04	0.104	0.805	0.849	103.715	169.672
InfoNCE-LD	1	0.048	0.118	0.722	0.792	54.501	49.611
InfoNCE-LD	10	0.057	0.124	0.68	0.75	22.705	26.756
InfoNCE-LD	100	0.063	0.13	0.686	0.74	9.577	14.96
InfoNCE-LD	1000	0.068	0.133	0.675	0.738	8.452	17.478

Table 11: Table of results that corresponds to Figure 2

		Main Task	Private Task (AUC-ROC)		Private Task Ethnicity (AUC-PR)		Fairness (Gap TPR RMS)		Fairness (Gap TPR Max)	
Estimator	**lambda_c**	**(Bal. TPR)**	**Gender**	**Ethnicity**	**Pos. Label=0**	**Pos. Label=1**	**Gender**	**Ethnicity**	**Gender**	**Ethnicity**
CLUB	0.1	0.715	0.844	0.652	0.827	0.418	0.133	0.046	0.377	0.136
CLUB	1	0.689	0.75	0.631	0.817	0.386	0.089	0.044	0.319	0.122
CLUB	2	0.647	0.726	0.621	0.813	0.372	0.092	0.049	0.307	0.116
CLUB	5	0.611	0.716	0.619	0.809	0.372	0.086	0.062	0.296	0.186
CLUB	10	0.537	0.682	0.608	0.804	0.359	0.092	0.056	0.31	0.182
CLUB-LD	0.1	0.553	0.733	0.608	0.803	0.359	0.097	0.06	0.298	0.215
CLUB-LD	1	0.588	0.765	0.617	0.809	0.367	0.111	0.055	0.315	0.183
CLUB-LD	2	0.632	0.785	0.615	0.808	0.365	0.116	0.061	0.318	0.173
CLUB-LD	5	0.593	0.793	0.618	0.808	0.371	0.128	0.053	0.337	0.154
CLUB-LD	10	0.533	0.754	0.613	0.804	0.369	0.115	0.061	0.387	0.168
KNIFE	0.1	0.607	0.922	0.655	0.827	0.428	0.176	0.064	0.429	0.171
KNIFE	1	0.355	0.825	0.613	0.807	0.363	0.14	0.058	0.357	0.193
KNIFE	2	0.379	0.855	0.608	0.803	0.365	0.143	0.059	0.381	0.2
KNIFE	5	0.424	0.827	0.625	0.812	0.384	0.13	0.06	0.317	0.173
KNIFE	10	0.447	0.858	0.625	0.811	0.39	0.142	0.06	0.358	0.165
KNIFE-LD	0.1	0.725	0.866	0.66	0.832	0.432	0.128	0.052	0.385	0.171
KNIFE-LD	1	0.704	0.765	0.639	0.821	0.397	0.099	0.054	0.291	0.19
KNIFE-LD	2	0.681	0.763	0.632	0.818	0.388	0.099	0.049	0.303	0.115
KNIFE-LD	5	0.599	0.73	0.623	0.813	0.372	0.091	0.053	0.252	0.134
KNIFE-LD	10	0.393	0.715	0.597	0.798	0.342	0.086	0.043	0.217	0.118
InfoNCE	0.1	0.706	0.851	0.654	0.827	0.426	0.132	0.052	0.346	0.171
InfoNCE	1	0.694	0.766	0.633	0.818	0.393	0.093	0.051	0.302	0.177
InfoNCE	2	0.664	0.75	0.62	0.812	0.373	0.094	0.057	0.3	0.198
InfoNCE	5	0.623	0.745	0.62	0.811	0.371	0.087	0.059	0.271	0.197
InfoNCE	10	0.615	0.734	0.613	0.807	0.363	0.099	0.053	0.317	0.163
InfoNCE-LD	0.1	0.733	0.852	0.658	0.83	0.432	0.132	0.051	0.369	0.15
InfoNCE-LD	1	0.678	0.775	0.628	0.815	0.384	0.107	0.054	0.359	0.161
InfoNCE-LD	2	0.615	0.766	0.625	0.814	0.377	0.098	0.053	0.295	0.162
InfoNCE-LD	5	0.342	0.663	0.585	0.792	0.33	0.072	0.051	0.203	0.17
InfoNCE-LD	10	0.391	0.663	0.587	0.792	0.332	0.086	0.049	0.308	0.154

Table 12: Table of results that corresponds to Figure 3

A.7 Parameters for the Mutual Information Estimator

Mutual Information Estimator		
Parameter	FairCVtest	BIOS
Low Dimensional Space	16	12
λ	[0.1, 1, 10, 100, 1000]	[0.1, 1, 2, 5, 10]
MI Learning Rate	0.01	
Context Learning Rate	0.01	
MI Layers	3	
Warm Up Epochs	15	
Main Training Epochs	15	
Validation Epochs	4	
Optimizer	Adam	

Table 13: Here we present the parameters used for MI estimation detailed per dataset. A single value indicates that this parameter remained unchanged between datasets. The estimators compared were InfoNCE, CLUB, and KNIFE, along with low-dimensional versions. Baseline comparisons with $\lambda = 0$ were made to demonstrate the effect of removing MI entirely.