ML
MACHINE LEARNING
DEPARTMENT

Carnegie Mellon University
Heinz College
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Auton Lab

Machine Learning in High-Stakes Settings: Risks and Opportunities

Maria De-Arteaga

Dissertation

In partial fulfillment of the requirements for the degree of

Doctor of Philosophy in

Machine Learning and Public Policy

Maria De-Arteaga

Committee: Artur Dubrawski (co-Chair)
Alexandra Chouldechova (co-Chair)
Roni Rosenfeld
Adam Tauman Kalai

May 13, 2020

To my parents,

Acknowledgements

More than a thesis, a PhD is a journey. I am very grateful for the people it brought into my life, and for those who supported me along the way.

First, I am deeply grateful to my advisors, Artur and Alex, who have guided me and encouraged me when I have taken on new directions, and who have been the best example of what it means to be a caring leader.

To my internship mentors, Adam, Jennifer, Christian, Hanna, and Henning. And especially to Sumit, who signed up to be my mentor for three months and got stuck with me, and who helped me rediscover a part of me that I had lost along the way.

To the Auton Lab, especially to my collaborators Peter, Kyle, Jieshi, and Vincent. And to Predrag for his continuous support.

To Al Blumstein, for teaching me to always ask “so what?”.

To Will and Petar, without whom the beginning of the PhD would have been much harder than it was, if that is even possible. To Ellie, for always helping me keep perspective of what matters, and for exploring any and all of the coffee shops with me. To Zhe, who first showed me how to fall in love with Pittsburgh and who is one of the best gifts this city gave me. To Gabo and Valeria, for the climbing adventures. To Sina, Matt, Rahul, Yasmine, Mari, Tim, Micol, Willie, Momin, Ed, Vero, Eva, Dan, and Lauren, who made my time in Pittsburgh unforgettable. And to Apteka, the most heart warming of places.

To my family, who made everything possible. To my dad, for his unending encouragement. To my mom, who taught me never to be afraid to take up space. To Camilo, who continuously teaches me about kindness.

To Ben, for your support, your love, and for making this journey infinitely more fun. CMU will always be the place that brought us together, and the first of many adventures.

Abstract

Machine learning (ML) is increasingly being used to support decision-making in critical settings, where predictions have potentially grave implications over human lives. Examples include healthcare, hiring, child welfare, and the criminal justice system. In this thesis, I study the risks and opportunities of machine learning in high-stakes settings. In the first chapter I focus on opportunities of ML to support experts' decisions when dealing with high-resolution multivariate data, a type of data that is particularly hard for humans to interpret. I propose methodology to discover latent complex multivariate correlation structures and illustrate its use in two different domains: (1) identification of radioactive threats in nuclear physics, and (2) prediction of neurological recovery of comatose patients in healthcare. In the second chapter, focused on algorithmic fairness, I demonstrate how societal biases encoded in historical data may be reproduced and amplified by ML models, and introduce a new algorithm to mitigate biases without assuming access to protected attributes. Finally, in the third chapter I characterize challenges that arise from the limitations of available labels in decision support contexts–such as the selective labels problem and omitted payoff bias–and propose methodology to estimate and leverage human consistency to improve algorithmic recommendations and human-machine complementarity.

Contents

Introduction

In recent years, the use of machine learning to assist experts in high-stakes decision making has increased sharply. In this thesis, I study the risks and opportunities of machine learning in such high-stakes contexts.

Chapter 1 focuses on learning from high-resolution multivariate data, a type of data that experts often encounter and that is particularly hard for humans to parse. For example, physicians routinely make use of time-series collected via bedside monitoring to inform medical decisions. I propose novel methodology to learn from this type of data and show how it can be of use in two high-stakes settings: nuclear physics and healthcare. The proposed methodology, termed Canonical Autocorrelation Analysis, discovers multiple-to-multiple correlations within a set of features. Moreover, I also propose Canonical Autocorrelation Embeddings, a method for embedding sets of data points onto a space in which they are characterized in terms of their latent complex correlation structures, and where a proposed distance metric enables the comparison of such structures. This methodology is particularly fitting to tasks where each individual or object of study has a batch of data points associated to it, as in for instance patients for whom several vital signs or other health related parameters are recorded over time.

The discovered correlations can be used for anomaly detection, as in the case of radiation threat detection. In this domain, the proposed methodology enables the characterization of patterns of correlations between subsets of bins of gamma-ray spectra known to represent benign background radiation. Once such characterization is obtained, it is then possible to flag spectral measurements that do not follow the same patterns of correlations as anomalies potentially reflective of the presence of radiation threats. The resulting spectral anomaly detection technique performs substantially better than an unsupervised alternative prevalent in the domain, while providing valuable additional insights for threat analysis.

In addition, Canonical Autocorrelation Embeddings can also be used for classification. In this thesis, I apply the proposed methodology to characterize patterns of brain activity of comatose survivors of cardiac arrest, aiming to predict whether they would have a positive neurological recovery. Clinicians routinely face the ethically and emotionally charged decision of whether to continue life support for such patients or not. Both scenarios have potentially grave implications on patients and their close ones, so regardless of whether they believe they have enough information, clinicians are often forced to make a prediction. The empirical results show that we can identify with high confidence a substantial number of patients who are likely to have a good neurological outcome. Providing this information to support clinical decisions could motivate the continuation of life-sustaining therapies for patients whose data suggest it to be the right choice.

The positive results shown in Chapter 1 highlight the opportunities that machine learning presents to assist expert decision-making. However, there are several risks associated to the use of machine learning for decision support. In Chapter 2, I focus on how standard ML methods may reproduce and amplify societal biases, and propose a new algorithm to mitigate biases without assuming access to protected attributes.

A domain in which the use of ML is increasingly popular–and in which unfair practices can lead to particularly negative consequences–is that of online recruiting and automated hiring. Through a large-scale study of gender bias in occupation classification, this thesis studies the potential allocation harms that can result from the use of predictive models in automated recruiting. A new dataset of over 400,000 online biographies written in third person was collected and made publicly available. This dataset was then used to study the biases present in algorithms trained to predict a person's occupation from their online biography. Several algorithms and semantic representations were explored, ranging from a bag-of-words representation as input for a logistic regression, to a word embedding as input for a deep recurrent neural network. The empirical results show that differences in true positive rates between genders are correlated with existing gender imbalances in occupations, even when explicit gender indicators such as gender pronouns were "scrubbed" from the text. Moreover, this work provides a theoretical demonstration that whenever there is a positive correlation between differences in true positive rates across groups and previous group imbalances, imbalances will be compounded. That means that if a group is underrepresented in a certain occupation, it will be further underrepresented amongst the individuals who are correctly predicted to be in that occupation. This effect can be related to compounding injustices—an existing notion of indirect discrimination in the political philosophy literature that holds that it is a general moral duty to refrain from taking actions that would harm people when those actions are informed by, and would compound, prior injustices suffered by those people [1]. This work has important practical relevance at a time when automated recruiting is becoming widespread and companies are grappling with the consequences of these technologies[1].

Moreover, even before a supervised learning task is defined (such as predicting a person's occupation), the choice of how to represent the data may itself lead to representational harms. Bias in word embeddings has received considerable attention in the past years, but until recently its study depended on querying for specific biases, such as determined re-

[1] Amazon scrapped a secret AI recruitment tool that showed bias against women: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

lationships or pre-defined protected groups. The second part of Chapter 2 proposes an algorithm to automatically enumerate biases in word embeddings. The algorithm is highly unsupervised–it does not even require the sensitive features to be pre-specified. This is desirable because: (a) many forms of discrimination–such as racial discrimination–are linked to social constructs that may vary depending on the context, rather than to categories with fixed definitions; and (b) it makes it easier to identify biases against intersectional groups, which depend on combinations of sensitive features. The utility of the approach is demonstrated on publicly available word embeddings, and the output is evaluated using crowdsourcing. Through its application, a large number of offensive associations related to sensitive features such as race, religion and gender were found on widely used embeddings, including a supposedly "debiased" embedding. A crowd-sourcing evaluation shows that this associations aligned with societal stereotypes.

Characterizing the risks of using ML for decision support may help (i) inform policy, and (ii) frame novel research to tackle these challenges. The last section of Chapter 2 proposes one of the first methodologies to reduce bias in predictive models without requiring access to protected attributes. The underlying intuition of the proposed method is to "fight bias with bias", leveraging the biases discovered in word embeddings to mitigate compounding imbalance effects in classification. The biases found in Section 2.2 are used to mitigate the bias characterized in Section 2.1. Specifically, the method penalizes correlations between the predicted probability of an individual's true class in a classification task and a word embedding of their name. The results demonstrate that this strategy significantly reduces gaps in true positive rates across race and gender groups, thereby mitigating the compounding imbalances effect observed in our earlier work. By design, name information is only present during training, which means that no private information is required during deployment, and that gains extend to individuals for whom protected attributes may be poorly proxied by their name.

Finally, even if data does not encode societal biases, the sensitive task of developing decision support tools is complicated by several factors, many of which stem from the nature of the labels available for training predictive models. Chapter 3 focuses on the limits of learning from observed outcomes to train decision-support systems, and proposes methodology to overcome this. There are two central challenges this chapter focuses on: omitted payoff bias and the selective labels problem. First, it is often the case that experts care about constructs that are not well captured in the available labels. This leads to omitted payoff bias, where there is a disconnect between the prediction loss function and the true payoff function. Second, these tools are often constructed and validated on data that is the result of historical human decisions. In such settings we commonly observe labels only for certain decisions–a phenomenon known as the selective labels problem [2]–, and the decisions themselves may have affected the observed outcomes. These issues limit the validity and utility of predictive models learned from the data using standard methods.

To overcome these challenges, Chapter 3 introduces methodology to leverage information contained in the historical human decisions, a rich but messy source of information.

Drawing inspiration from the literature on crowd-sourcing and wisdom of the crowds, the aim is to tackle some of the limits of learning from observed outcomes alone by also learning from consistency amongst experts. However, while in crowd-sourcing the same instance is assessed by multiple people, in the settings considered here each instance observed in the historical data is assessed by a single expert, such as a physician or a judge. This Chapter proposes an influence-function-based method to estimate human consistency in this setting. Under the assumption that human consistency is indicative of correctness, this human knowledge can then be incorporated into a model trained to predict observed labels through *label amalgamation*, an approach introduced in this Chapter. Through semi-synthetic experiments, it is shown how the proposed approach successfully incorporates human knowledge in different decision-making scenarios. Empirical experiments conducted on data from a child abuse hotline setting indicate that the proposed methodology successfully incorporates human knowledge, increasing recall for cases whose risk is not well captured in the available labels. Finally, for domains where it cannot be assumed that consistency is indicative of correctness, this Chapter introduces an influence-driven second opinion recommender algorithm, which identifies which expert is most likely to provide an alternative opinion for a given case.

Chapter 1

Learning from multivariate high-resolution data

Chapter partially based on:
M. De-Arteaga, J. Chen, P. Huggins, J. Elmer, G. Clermont, A. Dubrawski, Predicting Neurological Recovery with Canonical Autocorrelation Embeddings, *PLoS ONE*, 2019.

Introduction

In many domains experts are routinely tasked with interpreting high-resolution multivariate data as part of their decision-making process. Examples include physicians who track multiple vital signs of patients in intensive care units, and security experts who monitor gamma-ray spectra in order to identify potential radioactive threats. This Chapter presents Canonical Autocorrelation Analysis (CAA), a method for automated discovery of multiple-to-multiple correlation structures within a set of features. Through the introduction of a distance metric between CAA correlation structures it is possible to obtain a feature space embedding–termed Canonical Autocorrelation Embeddings (CAE)–where each individual/object is represented by the set of its multivariate correlation structures. This methodology is particularly fitting to tasks where each individual or object of study has a batch of data points associated to it, e.g., patients for whom several vital signs or other physiological measures are recorded over time. Using the proposed feature space embedding, traditional machine learning algorithms that rely on distance metrics, such as nonparametric clustering and k-nearest neighbors (k-nn), can be applied straightforwardly to leverage similarities or dissimilarities of correlation structures characteristic to individuals.

The first part of this Chapter introduces CAA and its application for anomaly detection, and illustrates its use in the context of radiation threat detection in nuclear physics. The second part introduces CAE and its use for supervised learning, demonstrating its use for predicting neruological recovery of comatose survivors of cardiac arrest. The relevance of the two application domains explored in this Chapter is explained below.

Nuclear physics Ever since the invention of nuclear weapons, radiation threat detection has been a security priority around the globe. Even though the total number of weapons has declined since the Cold War, a continued investment in nuclear arsenal has increased the destruction capacity of existing warheads, thus the threats that characterized the Cold War are still a main concern for the international community [3]. Furthermore, a vast number of such weapons are tactical nuclear weapons, characterized by their incapability to inflict strategically decisive damage to the military or economy of the target, a trait that has kept them out of current nuclear arms control arrangements. Many of them are kept under dubious security standards, as is the case of many that are stored in remote, hard-to-defend locations. These small and portable warheads, although incapable of devastating a country's economy in a single blow, would cause large-scale harms if used. Robbery of stolen fissile material that can be used to build radiological devices, commonly known as "dirty bombs", is also a concern for governments [4]. This problem reached its peak after the collapse of the Soviet Union, when the risk of people who are unaware of the dangers of radioactive material getting hold of it was illustrated by the case of a man who died of radiation sickness after storing material stolen from a nuclear waste facility in a kitchen cabinet [5]. Such risks are still present; in 2015 international alerts were triggered after a container full of medical isotopes was stolen in Mexico, where two years earlier thieves accidentally got hold of a container with radioactive material used in medical equipment [6]. Even though the destruction capacity of such devices cannot be compared to that of a nuclear warhead [4], they could expose thousands of people to dangerous levels of radiation [7].

Thus, effectively monitoring borders to prevent smuggling of radioactive threats, as well as monitoring the interior for signatures of possible threats, are crucial needs for many countries. This, however, is not an easy task. Radioactive materials are typically shielded, and the shielding can be engineered to make detection harder. In addition, faint sources of potentially harmful radiation can be hard to detect in scenes where intensity and spectral characteristics of benign background radiation vary significantly, as is the case in human-made environments. An additional challenge comes from the fact that different types of threats follow different spectral patterns, and even if templates of some common threat types are available, relying on supervised analysis of field data is risky. Supervised detectors may fail to detect threats that were not present in the training data, or which were shielded in a particularly unexpected fashion. Therefore, efforts have been made to develop unsupervised methods that successfully detect threats without relying on source templates.

Applying CAA for radiation threat detection enables us to characterize harmless radiation with a structure of correlations spanning sets of energy bins. Once this characterization is established, it can be used for spectral anomaly detection, as threats can be expected to deviate from the correlations characterizing harmless ambience. The experiments show that the ability of CAA to identify parsimonious subsets of features and later use it to model background radiation variability makes it more robust at threat detection than one of the most popular unsupervised methods used in the domain: a Principal Component Analysis (PCA) spectral anomaly detection method that considers all dimensions of spectra in linear combinations corresponding to subsequent principal components [8].

Healthcare In the healthcare domain, characterizing the current state of a patient through correlation structures can make it possible to leverage potentially under-appreciated forms of information, such as the interdependencies and interactions between different parts of the human body. The utility of CAE in this domain is demonstrated by focusing on the specific example of predicting neurological outcomes for comatose survivors of cardiac arrest based on electroencephalographic (EEG) measurements.

Cardiac arrest is the most common cause of death in high-income nations [9]. In the United States alone, over 350.000 people suffer a sudden out-of-hospital cardiac arrest each year [10]. Despite advances in care, only a minority of patients that survive to hospital admission after cardiac arrest are discharged alive, and even fewer enjoy a favorable neurological recovery [10, 11]. Among non-survivors, the most common proximate cause of death is withdrawal of life-sustaining therapy based on perceived poor neurological prognosis [11, 12]. This decision may be motivated by the rarity of favorable neurological recovery, the emotional difficulty for families faced with even a few days of critical care of a comatose loved one, and fear of survival with severe disability.

Unfortunately, accurate neurological prognostication after cardiac arrest is challenging, particularly in the first 3 to 5 days after initial resuscitation [13]. Modalities to facilitate early prognosis of recovery have been explored [14, 15, 16], in an attempt to augment medical knowledge and provide decision support systems to inform physicians as they continually reassess whether to continue or withdraw life-sustaining therapy. However, current methods are inadequate. Life-sustaining therapy is still often withdrawn before prognosis is certain as a result of “therapeutic nihilism” that may undermine otherwise effective post-arrest critical care that could have resulted in good recovery [17, 11, 18, 19]. At the same time, patients with brain injury that will ultimately be deemed irrecoverable are often still supported for days while providers gather prognostic data.

Improving care and making better decisions requires more predictive power and a better understanding of the brain early after cardiac arrest. Although many modalities may inform neurological prognostication in these patients [20], of particular interest is the rich EEG data that may be obtained. Qualitatively, some EEG patterns such as seizures suggest severe brain injury [21]. Quantitatively, patterns with strong correlations between channels or over time, such as burst suppression with identical bursts, are suggestive of nonsurvivable injury [22, 23]. Research indicates that there are EEG signals that can improve prediction accuracy [15, 24, 23]. Within EEG signals, as in many biological systems, entropy is a marker of information content [25]. By contrast, strong spatial or temporal correlations are an ominous predictor of severe brain injury [22, 23, 15]. Figure 1.1 shows an example of an EEG of a post-arrest patient with mild brain injury who goes on to enjoy a favorable recovery and an example of an EEG of a patient with severe brain injury, for which correlations across channels are very strong. However, in some cases these correlations may be subtle and complex, making them unapparent to physicians that qualitatively interpret an EEG recording. Motivated by this, our goal is to characterize patients in terms of their multivariate, non-linear structures of correlation and use the resulting featurization to predict their neurological outcome.

A proof of concept is presented to illustrate the potential utility of CAE by applying it to characterize electroencephalographic recordings from 80 comatose survivors of cardiac arrest, aiming to identify patients who will survive to hospital discharge with favorable functional recovery. The results show that with very low probability of making a Type 1 error, it is possible to identify 32.5% of patients who are likely to have a good neurological outcome, some of whom have otherwise unfavorable clinical characteristics. Importantly, some of these had 5% predicted chance of favorable recovery based on initial illness severity measures alone. Providing this information to support clinical decision-making could motivate the continuation of life-sustaining therapies for these patients.

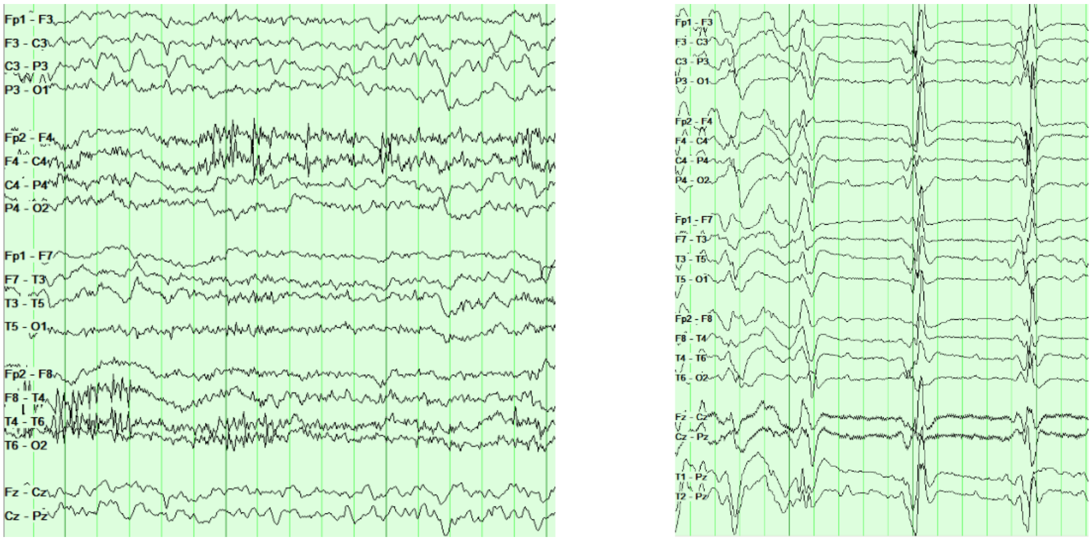


Figure 1.1: Left: EEG of a post-arrest patient who goes on to recover. Right: EEG of a patient with poor neurological prognosis. Strong correlations across channels suggest severe brain injury.

1.1 Related work

Canonical Correlation Analysis (CCA) is a statistical method first introduced by [26], useful for exploring relationships between two sets of variables. It is used in machine learning, with applications to medicine, biology and finance, e.g., [27, 28, 29, 30]. Sparse CCA, an ℓ_1 variant of CCA, was proposed by [30, 31]. This method adds constraints to guarantee sparse solutions, which limits the number of features being correlated. Given two matrices $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{n \times q}$, CCA aims to find linear combinations of their columns that maximize the correlation between them. Usually, X and Y are two disjoint matrix representations for one set of objects, so that each matrix is using a strictly different set of variables to describe them. Assuming X and Y have been standardized, the constrained optimization problem is shown in Eq. 1.1. When c_1 and c_2 are small, solutions will be sparse and thus only a few features are correlated.

$$\begin{aligned} & max_{u,v} u^T X^T Y v \\ ||u||_2^2 \leq 1, ||v||_2^2 \leq 1 \quad & ||u||_1 \leq c_1, ||v||_1 \leq c_2 \\ \text{for} \quad & 0 \leq c_1 \leq 1, 0 \leq c_2 \leq 1 \end{aligned} \tag{1.1}$$

The extension of Sparse CCA for discovery of multivariate correlations within a single set of features to study brain imaging has been previously explored in [27, 28]. Using the notion of autocorrelation, the authors attempt to find underlying components of functional magnetic resonance imaging (fMRI) and EEG, respectively, that have maximum autocorrelation. The types of data used in these works are ordered, both temporally and spatially. To find temporal autocorrelations, X is defined as the original data matrix and Y is constructed as a translated version of X, such that $Y_t = X_{t+1}$.

Canonical Autocorrelation Analysis (CAA), the methodology introduced in this Chapter, is a generalized approach to discovering multiple-to-multiple correlations within a set of features. Figure 1.2 illustrates the different use cases of Sparse CCA and CAA. The proposed formulation of CAA also allows for the user to select sets within which correlations are forbidden, which is useful when trivial correlations should be avoided.

Other methods for finding sparse representations of data comprised in a single matrix include the well-known Sparse Principal Component Analysis (Sparse PCA). While CAA resembles Sparse PCA in the sense that it finds sparse representations of data contained in one matrix, Sparse PCA maximizes retained *variance* of data in one-dimensional projections, while CAA finds two-dimensional projections where *correlation* across two subsets of features is maximized. CAA specifically seeks projections composed by pairs of strongly correlated linear combinations of features, enabling discovery of hidden characteristic correlations in data, which cannot be easily found with other methods such as Sparse PCA. Appendix A.2 explores the difference between the two methods in more detail and from a theoretical perspective.

Extraction of informative projections has been tackled in the past [32, 33]. The work presented in this Chapter differs from the existing methodology in two primary ways. First,

$$\left[\underbrace{\begin{matrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \ldots & x_{n,p} \end{matrix}}_{X} \middle| \underbrace{\begin{matrix} y_{1,1} & \cdots & y_{1,q} \\ \vdots & \ddots & \vdots \\ y_{n,1} & \cdots & y_{n,q} \end{matrix}}_{Y}\right] \qquad \left[\underbrace{\begin{matrix} x_{1,1} & \cdots & \cdots & x_{1,m} \\ \vdots & \ddots & \ddots & \vdots \\ x_{n,1} & \cdots & \cdots & x_{n,m} \end{matrix}}_{X}\right]$$

Figure 1.2: Comparison between scenarios where Sparse CCA and CAA can be used. **(Left)** Sparse CCA set up: X and Y are two matrices where the rows correspond to the same items but the columns represent separate sets of variables. Sparse CCA finds sparse multiple-to-multiple linear correlations between subsets of the features in matrix X and subsets of features in matrix Y. **(Right)** CAA set up: In cases where there is no natural or intuitive division of the features into two sets, a possible division represented by the dotted line is no longer given. Instead, all features are part of one matrix X. CAA finds multiple-to-multiple correlations between subsets of features in this matrix.

each of CAA projection axes is defined by a linear combination of features, rather than a single feature, which helps discover complex structures if they exist. Secondly, rather than finding projections where classes are well-separated, the proposed methodology is unsupervised and it is aimed at characterizing objects or individuals that have a batch of data points associated to them, yielding an embedding where standard machine learning methodologies can be used with minor modifications. In that sense, the extracted projections are different both in their form and in their purpose.

The comparison of correlation structures and principal components has been explored in the literature for decades. Most prominently, [34] discusses comparison of principal components between groups. To do so, they propose a metric inspired by the concept of congruence coefficient [35], which is nothing but the cosine of the angle between the two p-dimensional vectors. Also related is [36], where a metric between covariance matrices is proposed. The notion of a distance metric between canonical autocorrelation structures differs from these in that CAA finds a factorization of the correlation matrix where each portion of the correlation matrix is expressed as the outer product of a pair of orthonormal vectors, which define a bi-dimensional space in which the projected data follows a linear correlation. Section 1.3.4 discusses the proposed metric.

1.2 Data

1.2.1 Nuclear physics

Radiation is often characterized using gamma-ray spectra, which are typically represented as vectors of photon counts registered by the sensor at subsequent discrete and disjoint intervals of energy. These vectors, called in the application domain the energy spectra, become data points for analysis. In this Chapter, 128 energy bins are used, thus each data point is a vector in $\mathbb{R}^{128}$.

There are two types of data used in this research: harmless background and threat-infected background.

- *Harmless background* Over a period of five consecutive days a truck drove around downtown Sacramento, California, with a double 4x16" NaI planar detector on its back. The data contains approximately 70,000 one-second observations collected enroute, that reflect background radiation as well as any nuisance sources.
- *Threat-injected background* Synthetic threat injections done at the Lawrence Livermore National Laboratory. Simulated data mimics 15 types of sources. For each source, a data set of 10,000 observations is created by embedding synthetic threat signatures in harmless background data.

Once the model is trained using this data, it can be used on data collected by mobile detectors of radiation threat.

1.2.2 Prediction of neurological recovery

The data used in this study is derived from 451 comatose survivors of cardiac arrest treated at a single academic medical center between 2010 and 2015 [15, 37]. For each patient, quantitative EEG (qEEG) summary measures at one-per-second resolution are available for continuous EEG recordings averaging about 36 hours per patient. These qEEG features were calculated using FDA-approved clinical software (Persyst(R) Version 12, Persyst Development Corp, Prescott AZ), using standard signal processing engines. The total number of qEEG features is 66 and include seizure probability, amplitude-integrated EEG for the left and right hemisphere, spike detections, and suppression ratio, among other features that doctors have identified as clinically useful. The raw EEG data were not available. The full list of features can be found in Appendix A.7.

The data makes it known whether a patient survived to hospital discharge. For those who lived, it is known whether they experienced a functionally favorable recovery as measured by one of two standard outcome scales: the Cerebral Performance Category and a modified Rankin Scale. For those who died, the proximate cause of death is available. Figure 1.3 shows this information in detail. As it is discussed in more detail in Section 1.4, the data used in our experiments corresponds to the 80 patients who survived hospital discharge and who were monitored for at least 36 hours, 40 of whom had a positive neurological recovery and 40 who did not.

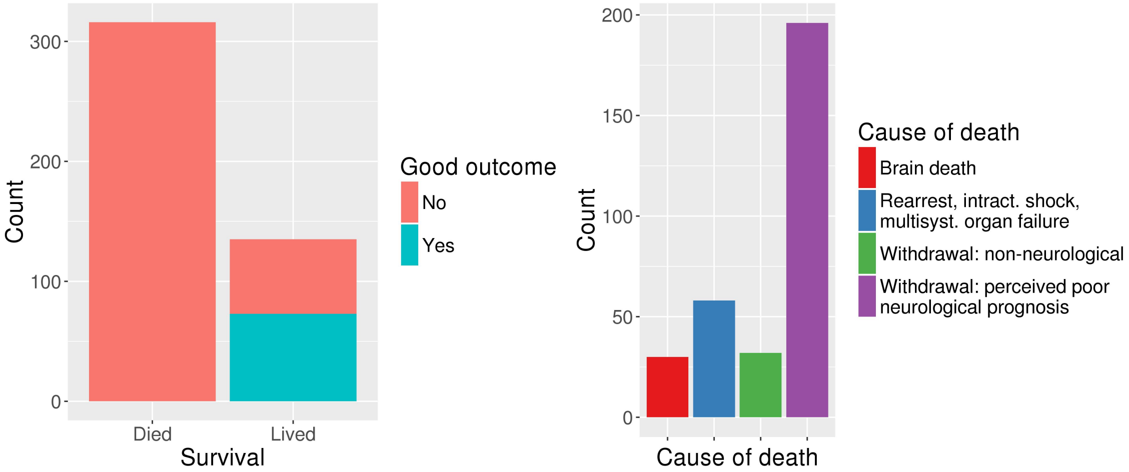


Figure 1.3: Patient labels. Survival and outcome (Left), and cause of death (Right).

1.3 Methodology

1.3.1 Canonical Autocorrelation Analysis

The goal of Canonical Autocorrelation Analysis (CAA) is to find multivariate sparse correlations within a single set of variables, yielding multiple bi-dimensional projections where each axis corresponds to a linear combination of a subset of the features and the projected data follows a linear trend. In the Sparse CCA framework, this could be understood as having identical matrices X and Y. Applying Sparse CCA when $X = Y$ results in solutions $u = v$, corresponding to Sparse PCA solutions for X [31]. This issue is tackled by introducing a penalty for overlapping feature support. The resulting optimization problem for CAA is shown in Eq. 1.2.

$$\begin{aligned} & max_{u,v} u^T X^T X v \\ ||u||_2^2 \leq 1, ||v||_2^2 \leq 1 \quad & ||u||_1 \leq c_1, ||v||_1 \leq c_2 \\ & \sum_{i=1}^{m} |u_i v_i| = 0 \\ \text{for} \quad & 0 \leq c_1 \leq 1, 0 \leq c_2 \leq 1 \end{aligned} \tag{1.2}$$

Understanding this as a new generalization of the PMD decomposition [31], the solution for CAA is analogous to that of other PMD-based approximations, although necessary adjustments have to be made to account for the additional constraint. Note that in the CAA optimization problem seen in Eq. 1.2, the equality constraint can be seen as a weighted L_1 penalty when either u or v are fixed. Replacing the equality constraint by an inequality constraint gives a biconvex problem, while resulting in the same solution. Therefore, it can be solved through alternate convex search [38], as shown in Algorithm 1.

Algorithm 1: CAA via alternate convex search

```
Initialize $v$ s.t. $||v||_2 = 1$;
repeat
    $u \leftarrow \arg\max_u u^T X^T X v$
    s.t. $||u||_2^2 \leq 1$, $||u||_1 \leq c_1$, $\sum_{i=1}^m |u_i||v_i| = 0$
    $v \leftarrow \arg\max_v u^T X^T X v$
    s.t. $||v||_2^2 \leq 1$, $||v||_1 \leq c_1$, $\sum_{i=1}^m |u_i||v_i| = 0$
until $u$, $v$ converge;
$d \leftarrow u^T X^T X v$;
```

At each iteration, the resulting convex problem can be solved through the Karush-Kuhn-Tucker (KKT) conditions. The pseudo-code for solving the convex problems at each iteration of the alternate convex search is provided in Algorithm 2, where it is solved for u without loss of generality. For a detailed derivation see Appendix A.1.

Algorithm 2: CAA alternate convex search iteration via KKT conditions

```
$\lambda_1 = \max_i \frac{|(X^T X v)_i|}{|v_i|}$;
if $||\frac{S_{\Phi(v\lambda_1,0)}(X^T X v)}{||S_{\Phi(v\lambda_1,0)}(X^T X v)||_2^2}||_1 \leq c_1$ then
    return $u = \frac{S_{\Phi(v\lambda_1,0)}(X^T X v)}{||S_{\Phi(v\lambda_1,0)}(X^T X v)||_2^2}$
else
    Binary search to find $\lambda_2$ s.t. $||\frac{S_{\Phi(v\lambda_1,\lambda_2)}(X^T X v)}{||S_{\Phi(v\lambda_1,\lambda_2)}(X^T X v)||_2^2}||_1 = c_1$;
    return $u = \frac{S_{\Phi(v\lambda_1,\lambda_2)}(X^T X v)}{||S_{\Phi(v\lambda_1,\lambda_2)}(X^T X v)||_2^2}$
```

To find multiple pairs of CAA canonical vectors, Algorithm 1 can be repeated iteratively, replacing $X^T X$ with a matrix from which the already found correlations are removed, as shown in Eq. 1.3.

$$X^T X - d(uv^T + vu^T) \tag{1.3}$$

1.3.2 CAA-based anomaly detection

How can the outcome of CAA be used once it has been applied to a matrix X? CAA produces several multiple-to-multiple linear correlation patterns. If the only goal is that of characterizing and understanding the data, one can analyze the coefficients in the canonical projections to understand which correlations are characteristic in a certain data set. Such projections can also be used as the basis of an anomaly detection method, introduced below.

CAA can be applied to a set ($X \in \mathbb{R}^{n\times m}$) of data points that are assumed to not be anomalous. The result will be $k \leq m$ pairs of canonical vectors, where the ith pair is refered to as $(u^{(i)}, v^{(i)})$, for $i = 1, ..., k$. Each of these vector pairs maps the data into a new bi-dimensional space, where the x-axis corresponds to ${u^{(i)}}^t X^t$ and the y-axis corresponds to $Xv^{(i)}$. The projection of the data onto the ith canonical space is defined as in Equation 1.4.

$$X_i^{proj} = ({u^{(i)}}^t X^t, Xv^{(i)}) \tag{1.4}$$

The top canonical projections yield representations in which the training data shows a strong diagonal tendency if the underlying correlations indeed exist.The resulting distribution can be characterized by fitting a parametric density model (e.g. bivariate Gaussian) or using a non-parametric density model (e.g. kernel density estimation),

$$X_i^{proj} \sim F_i(\boldsymbol{\theta}_i) \tag{1.5}$$

This yields a characterization of the non-anomalous data points by:

$$\begin{cases} \text{Canonical vectors } (u^{(i)}, v^{(i)}) & for \quad i = 1, ..., k \\ \text{Distributions } F_i \text{ parameterized by } \boldsymbol{\theta}_i & for \quad i = 1, ..., k \end{cases}$$

Given a new data point x, it can be projected onto the k canonical projections and a score of anomalousness can be computed for each projection using the likelihood (Equation 1.6).

$$s_i(x) = \mathbb{P}(x|\boldsymbol{\theta}_i) \tag{1.6}$$

Finally, a cumulative single score can be computed using an aggregation metric $M(\cdot)$, where the choice of this metric depends on the particular application (e.g., minimum or product), as shown in Equation 1.7.

$$S(x) = M_{\{i=1:k\}}(s_i(x)) \tag{1.7}$$

For the experiments in this Chapter, it is assumed that the training data follows a bivariate Gaussian in each of the canonical projections, i.e,

$$X_i^{proj} \sim \mathcal{N}(\mu_i, \Sigma_i). \tag{1.8}$$

Therefore, the data set can be characterized by Equation 1.9.

$$\begin{cases} (u^{(i)}, v^{(i)}) & for \quad i = 1, ..., k \quad k \leq n \\ (\mu_i, \Sigma_i) & for \quad i = 1, ..., k \end{cases} \tag{1.9}$$

Each i yields a characterization of the training data that involves multiple features. A new data point x can be simultaneously mapped onto all k canonical spaces, and given the assumption of a bivariate Gaussian, a Mahalanobis distance metric can be used as an equivalent to the likelihood. Therefore, our score $s_i(x)$ is given by

$$\begin{aligned} s_i(x) &= D_{M_i}(x_i^{proj}) \\ \text{where } x_i^{proj} &= ({u^{(i)}}^T x^T, xv^{(i)}) \\ D_{M_i}(x_i^{proj}) &= \sqrt{(x_i^{proj} - \mu_i)\Sigma_i^{-1}(x_i^{proj} - \mu_i)} \end{aligned} \tag{1.10}$$

Note $D_{M_i}(\cdot)$ is the Mahalanobis distance from x_i^{proj} to $N(\mu_i, \sigma_i)$, where x is the current observation.

If the new data point follows the same correlation patterns as the training data, all of the Mahalanobis distances computed for it should be small. It can be expected that a data point that is anomalous in the CAA sense would not match that behavior. It will likely fail to follow one or multiple of these characterizations, which will result in one or multiple large Mahalanobis distances. To marginalize the resulting distribution of scores into a total score $S(x)$, one conceivable option is maximization, as in Equation 1.11.

$$S(x) = \max_{i=1,...,k} D_{M_i}(x^{proj}) \tag{1.11}$$

Maximization is only one of many possible ways to aggregate scores from multiple CAA projections. This approach has proved to be effective in the threat detection application because it is typically sufficient for a gamma-ray spectrum measurement to substantially deviate from the expectation in only a few energy bins to warrant attention. However, in other applications alternative forms of $S(x)$ may be more relevant and effective.

The threat detection threshold is calibrated following [39], by assuming a particular rate of nuisance positives in training data (2-5%).

1.3.3 Non-linear and forbidden correlations

When looking for latent structures of correlation in data, it may often be useful to consider non-linear relationships. Stemming from Weierstrass approximation theorem [40], the most straightforward way of doing so is by extending the feature space with subsequent powers of the original features. This increases the potential power of expression of the resulting models, but using it directly with the current formulation of CAA would likely result in each feature being trivially correlated with its own exponential transformations. Similar useless effects could be expected if the available data contains features that are already known to be mutually correlated by design. For example, in the data considered in this project, several features correspond to basic statistics of the amplitude integrated EEG.

To overcome those limitations, the optimization problem is modified to extend the concept of disjoint support to sets of features. Assuming each feature x_i has a subset S_i of associated indices of other features that should not be included as correlates of x_i, the resulting optimization problem follows Eq. 1.12.

$$\begin{aligned} & max_{u,v} u^T X^T X v \\ ||u||_2^2 \leq 1, ||v||_2^2 \leq 1 \quad & ||u||_1 \leq c_1, ||v||_1 \leq c_2 \\ & \sum_{i=1}^{m} \sum_{j \in S_i} |u_i v_i| = 0 \\ \text{for} \quad & 0 \leq c_1 \leq 1, 0 \leq c_2 \leq 1 \end{aligned} \tag{1.12}$$

The new constraint for disjoint support can still be understood as a weighted-L_1 penalty at each iteration of the biconvex optimization algorithm. Hence, the problem can still be solved in the way presented in Section 1.3.1, with the only difference that the parameters of the soft-thresholding operator will change.

1.3.4 Canonical Autocorrelation Embeddings

CAA enables the discovery of bi-dimensional projections where the data closely follows a linear distribution. Each axis of these projections corresponds to a linear combination of the original features, and their respective coefficients are represented in a pair of vectors $u, v \in \mathbb{R}^m$. Each pair u, v constitutes a *CAA canonical space*, and each CAA model may consist of one or more such canonical spaces.

Since the correlations discovered by CAA are defined by pairs of vectors in $\mathbb{R}^m$, it is possible to measure the distance between two CAA canonical spaces in terms of Euler angles defining the rotation from one pair of axes to the other. Given that measuring the angle between two vectors is equivalent to measuring the arc between them, and that $||u^{(i)}||_2 = ||v^{(i)}||_2 = 1 \ \forall i$, the distance between two CAA canonical spaces C_1 and C_2 can be defined as shown in Eq. 1.13. Note that the minimum is simply used to find the best (the smallest angle) of two possible ways of aligning arbitrary C_1 and C_2.

$$d(C_1, C_2) = \min(||u_1 - u_2||_2 + ||v_1 - v_2||_2 \ , \ ||u_1 - v_2||_2 + ||v_1 - u_2||_2) \tag{1.13}$$

It is easy to show that this metric satisfies the necessary conditions for a well-defined distance (see Appendix A.3 for the proof). Moreover, if two CAA canonical spaces represent the same correlation structure, the vectors defining them must be equal. This stems from the fact that such correlation structure would take the form of a matrix $Co \in \mathbb{R}^{m \times m}$, therefore, Eq 1.14 can be seen as a system of linear equations with at most one solution.

$$Co = uv^T \tag{1.14}$$

Even though Eq. 1.13 provides a distance metric that captures desired characteristics, this is not the only nor necessarily the best such metric, and it is appropriate to continue exploring alternatives. Appendix A.4 contains a short discussion of why the "principal angles", one of the metrics most commonly used to measure distance between subspaces and which naturally comes to mind in this setting, is not well-suited for our current task.

1.3.5 Classification and K-Nearest Correlations

Having formulated a distance metric between pairs of CAA canonical spaces enables the use of a range of distance-based machine learning algorithms, such as k-means or hierarchical clustering or k-nn, to leverage similarities among correlation structures present in data. One complexity that arises while doing so is that each subset of data being compared may be represented by more than one CAA canonical space, and therefore more than one point in the embedding.

This setting can be incorporated into the k-nn framework by calculating the class probability for each correlation structure through the votes of their k nearest neighbors, and then aggregating over all correlations associated to an object using log-odds, as shown in Eq. 1.15, where $n_{p,i,j}$ denotes the class label of the jth neighbor of the ith correlation of patient p.

$$\begin{aligned} q_i &= \frac{\sum_{j=1}^{k} n_{p,i,j}}{k} \\ \hat{y}_p &= \log(\prod_{i=1}^{m_p} \frac{q_i}{1-q_i}) \end{aligned} \tag{1.15}$$

However, it is likely that some type of correlation structures will be common to both classes, while others are discriminative. To reduce noise and allow for those discriminative correlations to lead the decision, a threshold t is incorporated, so that log-odds are only calculated over those correlation structures with a class probability that is discriminative enough, as shown in Eq. 1.16. Incorporating this threshold also enhances interpretability of the comparisons, as it reduces the number of structures that are used for making a prediction, making it easier for practitioners to understand which correlations appear relevant for the task at hand. The parameters k, indicating the number of neighbors, and t can be tuned through cross-validation.

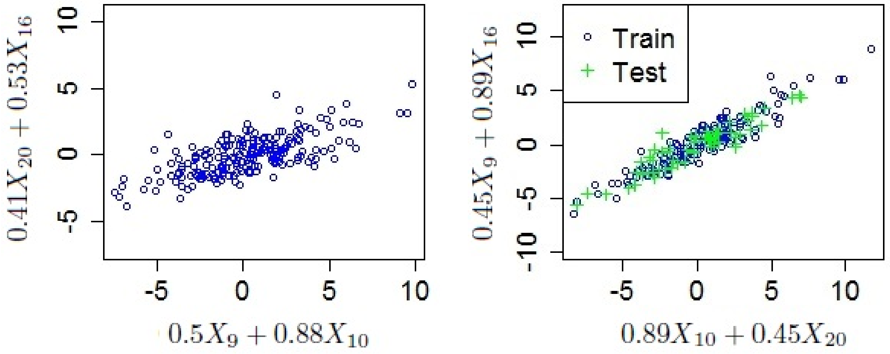
$$\hat{y}_p = \log(\prod_{i=1}^{m_p} \mathbb{I}_{(|q_i-0.5|>t)} \frac{q_i}{1-q_i}) \tag{1.16}$$

1.4 Results and Analysis

1.4.1 Nuclear physics

Synthetic data

The first experiment aims to illustrate how known correlations can be successfully retrieved by CAA.



(a) Synthetic correlation

(b) Retrieved correlation

Figure 1.4: Comparison between a synthetic correlation pattern and the correlation pattern retrieved by CAA. Equations have the form of $k_iX[,i] + k_jX[,j]$, where $X[,i]$, $X[,j]$ are the ith and jth columns of X and k_i, k_j are the linear combination coefficients.

A Gaussian bivariate distribution is generated with an assumed mean and covariance, and 200 data points are sampled from it. A matrix X of dimensions 200 × 20 is created such that there exist sparse vectors u, v, each with two non-zero components, for which (u^TX^T, Xv) correspond to the previously generated Gaussian. Next, 70% of data is used to train a CAA model and the rest is used for testing. Figure 1.4a contains a scatter plot of the data sampled from the Gaussian, where the axes indicate the linear combinations of columns of X that map the original data onto the Gaussian. Figure 1.4b shows the projection of both training and testing data onto the space determined by the first pair of canonical vectors retrieved by CAA, where the equations on the axes correspond to the correlation they establish. Note that the method is able to successfully identify the existing multiple-to-multiple linear correlation, even though the features are not grouped identically as in the original design.

Nuclear physics

The radiation data used in our experiments is featurized into 128 disjoint energy bins, and reflects photon counts obtained from gamma-ray spectrometer measurements. There are 20,000 records available for harmless background data, and 10,000 records for each of 15 types of threat-injected data, to simulate various radiological threats.

As it was previously explained, PCA-based spectral anomaly detection assumes that the top few principal components represent the expected envelope of background variation, and uses the residual after removing these top components as a spectral anomaly score. In the case of CAA, multiple-to-multiple combinations of energy bins that are well correlated provide a characterization model for background radiation. This model can be used as the basis of the anomaly detection method described in Section 1.3.2, which identifies threats when radiation spectra depart from the expected patterns of correlation.

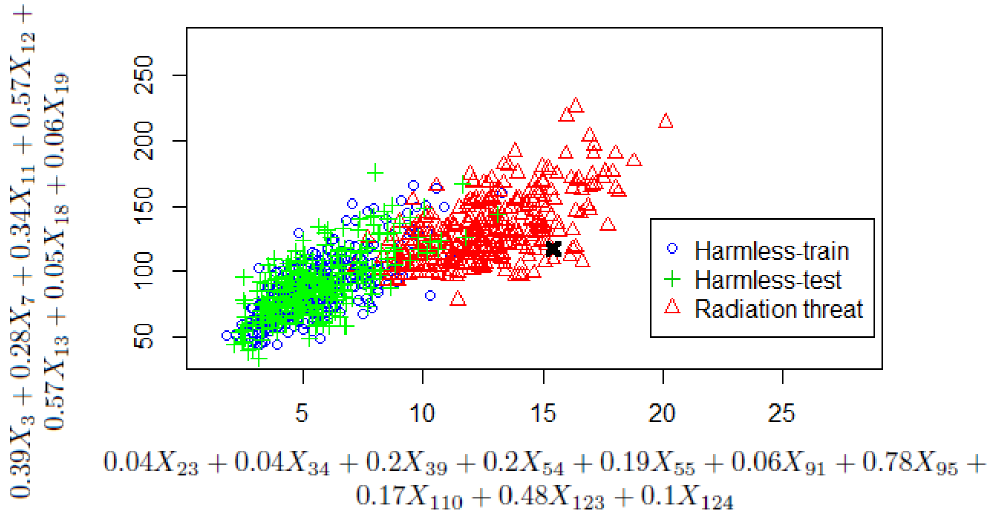


Figure 1.5: Projection of background radiation and threat-injected background radiation measurements onto space determined by one pair of CAA canonical vectors. The equations indicate the multiple-to-multiple linear correlation that defines this projection. Data point labeled with black x corresponds to the individual threat case analyzed in Figure 1.6.

For evaluation purposes, the CAA-based anomaly detector is compared to the PCA-based anomaly detector, a widely used approach in the domain. The PCA-based spectral anomaly detector, based on [41], works by calculating the magnitude of the residual after a background-subtracting projection. The background-subtracting is a strict projection onto the subspace spanned by the top few principal components of the covariance matrix. An alternative way of finding this projection is by a dilation modified projection where the correlation (not covariance) matrix is used to learn the low dimensional projection and then appropriate scaling of the measurement dimensions is performed before projection and scaled back after the projection. In any case, the transformation computes the estimated background contribution to a radiation measurement, assuming that the top few principal components represent expected typical background variation. After projection, the magnitude of the residual essentially provides the PCA-based spectral anomaly score, as it should be negligibly low for spectra consistent with training data distributions. Appendix A.5 contains the PCA spectral anomaly detector algorithm. Additionally, a Sparse PCA anomaly detector was designed and implemented for comparison. The algorithm is analogous to the PCA alternative, with some minor modifications that enable the exchange of PCA for Sparse PCA in the pipeline. The algorithm is described in detail in Appendix A.6.

All three models were trained using 10,000 background records, and the resulting models were evaluated on 15 types of radiation threats. Each of the 15 test sets contained 10,000 samples of injected data corresponding to a particular threat type, combined with the remaining 10,000 background records. Three performance comparison metrics were used: area under the ROC curve (AUC), recall at a fixed low false discovery rate, and false discovery rate at a fixed recall rate of 50%. Figure 1.7 and Tables 1.1 and 1.2 summarize the results, and Appendix A.8 contains ROC curves for all 15 threat types used in the experiments, with the false positive rate axis in logarithmic scale to enhance view at low false positive rates, where most applications tend to reside. For ten of these fifteen threats CAA performs significantly better than PCA and Sparse PCA according to all three performance metrics, and only for one type of threat is another method significantly better than CAA according to all metrics. This is threat type A, where PCA performs best, which can be potentially explained by the fact that sparsity is apparently not very useful in this case and a larger number of bins is necessary to detect this particular type of threat. When the algorithms are applied to all threats combined into a single batch, CAA is by far better than the other two competitors.

Figure 1.5 shows an example of the mapping onto the space determined by a pair of CAA canonical vectors (u, v). The data points corresponding to background radiation, as well as those corresponding to a particular threat, are mapped onto this projection. This particular pair (u, v) is used most often to score the data points belonging to this particular type of threat (listed as type "L" in table labels), meaning the one where the maximum Mahalanobis distance to the Gaussian characterizing background radiation is found most often, as defined in Equation 1.17.

$$(u, v) = \underset{u_i, v_i}{\arg\max}\, D_{M_i}((u_i^T x^T, xv_i))|_{i=1}^{k} \tag{1.17}$$

As Figure 1.5 shows, in this example the threat-injected data distribution visibly diverges from the test set distribution of benign data. For this threat type, CAA model achieves the AUC of 0.995, while PCA-based detector has the AUC of 0.821.

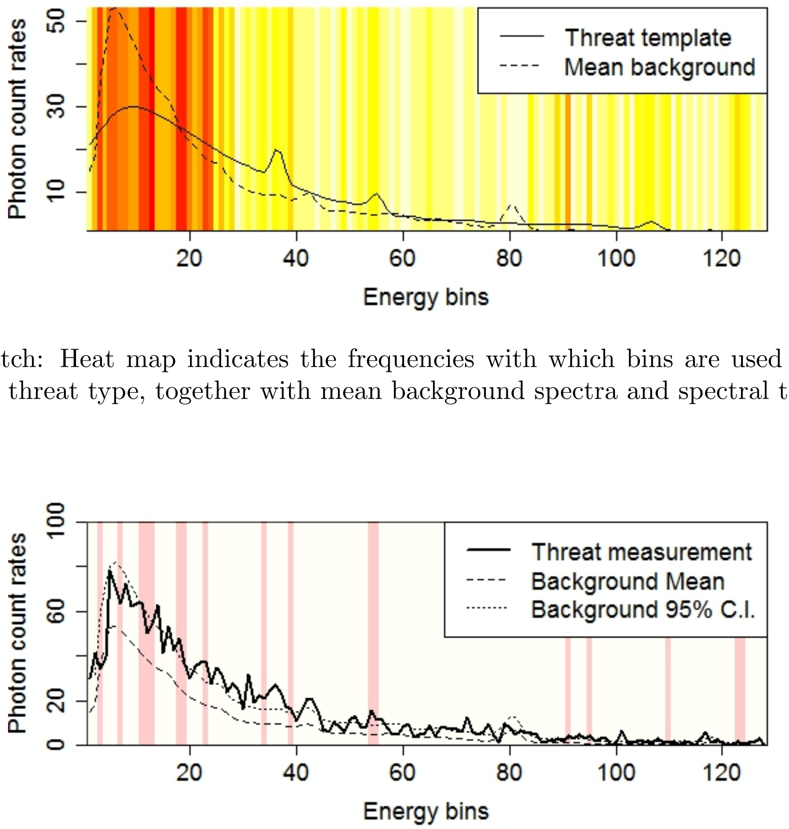
Th	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	**all**
cor.	.72	.79	.79	.81	.87	.87	.88	.88	.88	.91	.92	.92	.94	.94	.96	NA
caa	7	**4***	**4***	**14***	**7***	**7***	**2**	**15***	**100**	**14***	**87***	**14**	**20***	**13***	5	**21***
pca	**59***	1	1	1	2	2	1	3	**100**	7	8	13	2	1	4	14
Spca	6	1	1	2	3	3	1	4	**100**	5	9	6	2	2	**6**	10

Table 1.1: Performance of CAA, PCA and Sparse PCA in terms of recall rate (given in %) at fixed false discovery rate of 0.01. Radiation threat types are ordered according to the strength of the correlation between mean background spectrum and threat template. Asterisks mark cases when the winning method performs significantly better than the second-best.

Th	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	**all**
cor.	.72	.79	.79	.81	.87	.87	.88	.88	.88	.91	.92	.92	.94	.94	.96	NA
caa	11	**21***	**21***	**6***	**13***	**13***	**41**	**6***	**0**	**7***	**0***	**7***	**4***	**7***	18	**8***
pca	**1***	43	39	33	42	41	53	33	**0**	23	12	15	28	32	42	27
Spca	14	46	49	34	27	27	43	20	**0**	17	9	15	36	35	**16***	23

Table 1.2: Performance of CAA, PCA and Sparse PCA in terms of false discovery rate (given in %) at a fixed recall of 50%. Radiation threat types are ordered according to the strength of the correlation between mean background spectrum and threat template. Asterisks mark cases when the winning method performs significantly better than the second-best.

In addition to its good empirical performance, the proposed method yields readily interpretable outputs. When a spectral measurement is identified as a possible threat, the energy bins on which it fails to follow background patterns can be pointed out. This has two main advantages: first, when analyzing an individual data point the user knows which energy bins the algorithm used to make its decision, for easy adjudication of the results (Figure 1.6b). Secondly, when applied to a batch of data associated to a particular threat type, it is possible to identify the bins on which the threat's appearance systematically differs from the background behavior, providing the way to characterize this type of threat (1.6a).



(a) Threat batch: Heat map indicates the frequencies with which bins are used by CAA to flag one particular threat type, together with mean background spectra and spectral template for that threat.

(b) Adjudication of an individual measurement: Radiation spectrum the method correctly labels as inclusive of threat signatures compared to background radiation distribution. Colored bins were used to flag measurement as anomalous, corresponding to the support of the CAA canonical vectors that define the projection where the maximum Mahalanobis distance to the baseline distribution was found. This corresponds to CAA projection shown in Figure 1.5, where this individual measurement is labeled with a black x.

Figure 1.6: Visualization of energy bins that are used to label gamma-ray measurements as likely inclusive of threats.

Figure 1.6 shows the frequency with which energy bins are used to identify anomalies. The top plot shows the usage frequency of bins for a threat-infused data batch associated to the threat type used in Figure 1.5. The bottom plot shows an example of a radiation spectrum the method correctly labels as representative of a threat, mean of the background radiation spectra used for training, and colors the energy bins that were used to label the

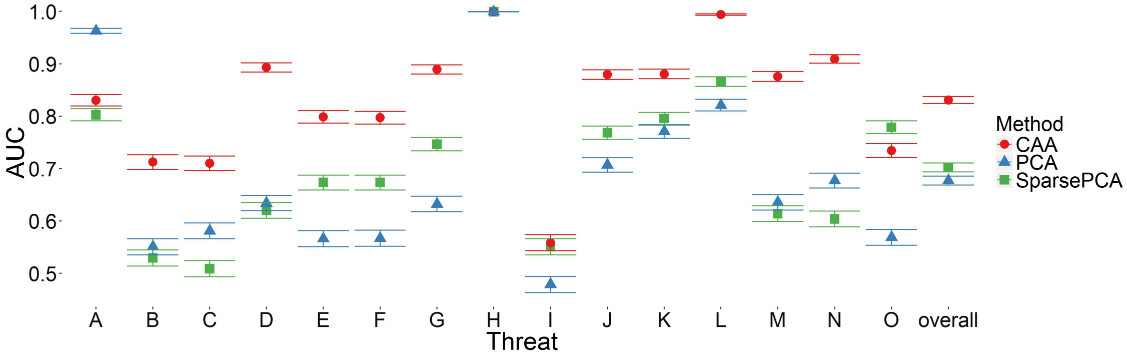


Figure 1.7: AUC and confidence intervals for CAA, PCA and Sparse PCA applied to detecting radiation threats of various types. Radiation threat types are ordered according to correlation between mean background spectrum and threat template spectrum. The right-most column shows performance when all threat types are combined in one batch.

data as anomalous. This individual threat measurement is labeled with a black x in Figure 1.5. It is interesting to see that even though the method is fully unsupervised, such bins correspond to spikes in the injected threat template.

1.4.2 Prediction of neurological recovery

The principal goal in this domain is to help improve care given to comatose survivors of cardiac arrest through a decision support system that can boost the accuracy and timeliness of clinical prognosis. To do so, Canonical Autocorrelation Embeddings is proposed as a new way of characterizing patients through their latent multivariate structures of correlation, and of using the resulting featurization of data as a way to build predictive models.

The first fundamental decision to make is what data and labels to use for training. As it can be seen in Figure 1.3, the main cause of death for patients in our data set is the withdrawal of life-sustaining therapy due to perceived poor neurological prognosis. However, as mentioned in Section 1.1, it is possible that in some cases treatment might be withdrawn too early. Including this data in the training would risk introducing bias, as the model could learn and replicate the mistakes clinicians may be making, leading to a self-fulfilling prophecy. Considering this and the fact that our goal is to predict positive neurological outcome rather than survival alone, the model is trained using only those patients who lived, making our target label whether they had a good or a poor neurological outcome.

For each patient, their entire EEG record is available, with lengths varying from less than an hour to more than a week. In the present experiment, those patients with at least 36 hours of EEG data are considered. When focusing the study on patients who survived till hospital discharge and who were monitored for at least 36 hours, the resulting dataset

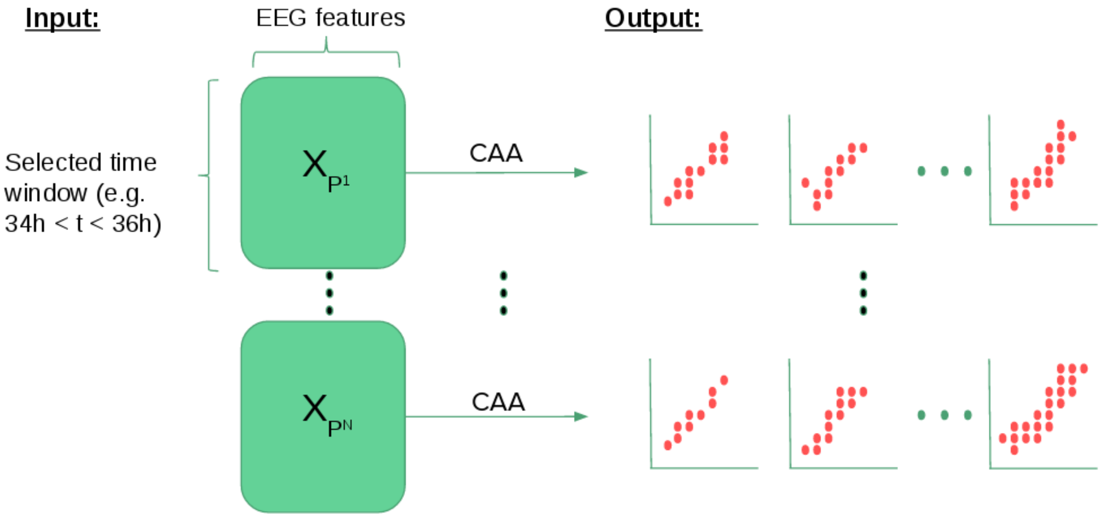


Figure 1.8: Diagram illustrating CAA patient characterization using EEG features as input data.

is composed of 80 patients, 40 of whom had a positive neurological recovery and 40 who did not. CAA is used to characterize a two hour epoch between hours 34 and 36. The specific question the proposed model answers is: can the correlations present during this epoch predict whether the patient will have a good neurological outcome? The reason why only two hours are considered is because it can be expected that the state of the patient fluctuates during their stay, and the resulting variance could obfuscate important patterns of correlation, if observed for prolonged periods of time. Identifying trends over time, or inferring meta-correlation structures that describe these temporal trends, is an important subject of future work beyond the scope of current analysis. Figure 1.8 illustrates the process of characterization of multiple patients' EEG data with CAA.

In order to avoid spurious results, only CAA canonical projections that yield correlations with $R^2 > 0.25$ are considered. Moreover, to ensure that only reasonably close neighbors are used for matching, connections are pruned by only considering distances smaller than $\sqrt{2}$, a threshold that corresponds to a 90° rotation over one axis. Using the resulting pruned distance matrix, k-nearest neighbor algorithm among CAA embeddings is applied. Empirical results obtained through 10-fold cross-validation, with tuning parameters k and t in an internal 10-fold cross-validation loop within each training fold, are presented in Figures 1.9, 1.10, 1.11, 1.12.

A logistic regression with lasso regularization [42] is also considered to predict recovery 36 hours after admission. Given that logistic regression is not suited for sets, but rather takes as input individual data points, two avenues are explored. The first approach takes the last data point after 36 hours of monitoring, that is, the recording at one time step.

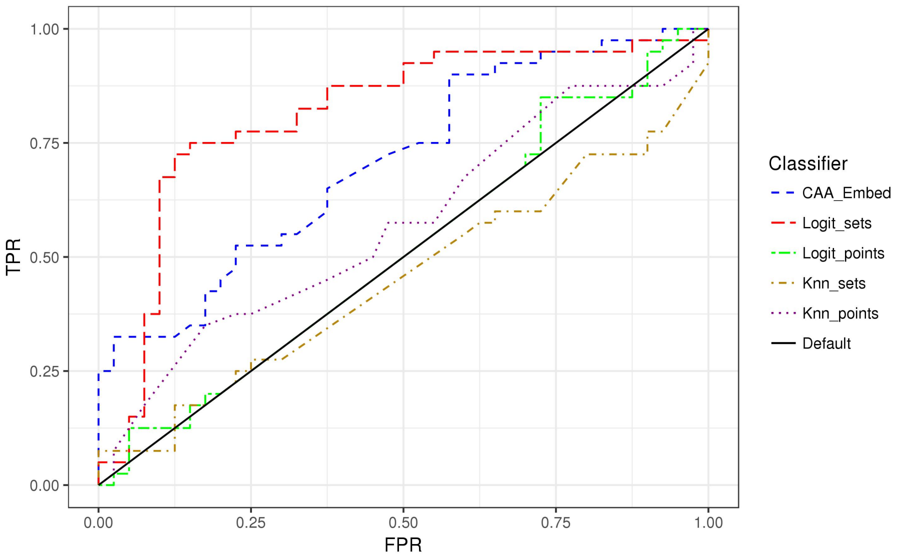


Figure 1.9: ROC curves showing performance of CAA Embeddings, logistic regression on sets, logistic regression on points, k-nn on sets and k-nn on points.

For the remainder of the Chapter, this approach is termed *logistic regression on points*. In the second approach, quartiles for each input feature are calculated over two hours preceding the 36-hour mark, and provided as features to the logistic regression model. This approach is refered to as *logistic regression on sets*. The choice of the lasso regularization parameter λ is made through 10-fold cross-validation. The results are included in Figures 1.9, 1.10, 1.11, 1.12. Furthermore, to be able to better assess the role of the CAA Embeddings, results from a direct application of k-nn algorithm using Euclidean distance and taking the same inputs as the logistic regression models are also included, and referred to as *k-nn on points* and *k-nn on sets*, respectively.

These results show that the proposed methodology has predictive power, and the comparison to k-nn using Euclidean distance on points and set-aggregated features highlights the role of CAA Embeddings. The performance of CAE at low false positive rates is particularly promising, with a true positive rate of 0.25 and corresponding 95% confidence interval $[0.125, 0.46]$ at a false positive rate lower than 0.03. This means that with very low probability of making a Type I error, it is possible to confidently identify at least 12.5% of the patients who will go on to have a positive neurological recovery.

The clinical utility of a prognostic tool in our example application would be determined to a lesser extent by its overall discriminatory power, but more so by its ability to confidently identify patients with essentially either nil or a substantial possibility of recovery. Thus, while in our experiment logistic regression shows an overall better discriminatory

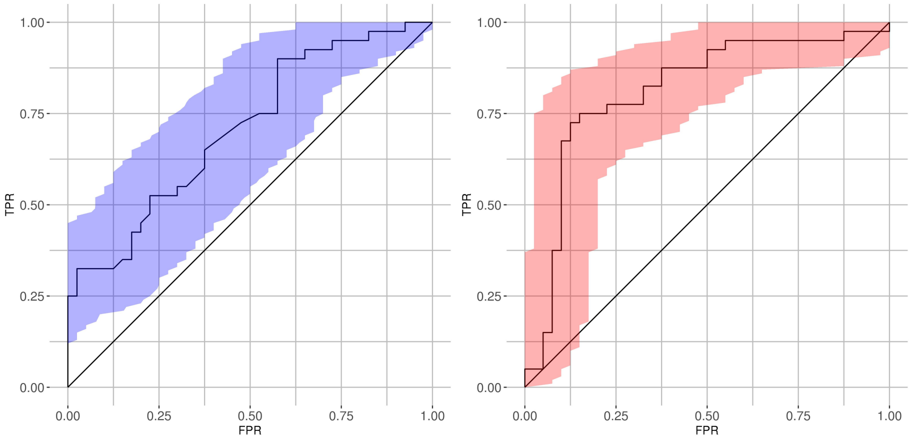


Figure 1.10: ROC curves with 95% confidence intervals. Left: CAA Embeddings, AUC = 0.71 with 95% confidence interval of $[0.6, 0.82]$. Right: Logistic regression on sets, AUC = 0.81 with 95% confidence interval of $[0.71, 0.91]$.

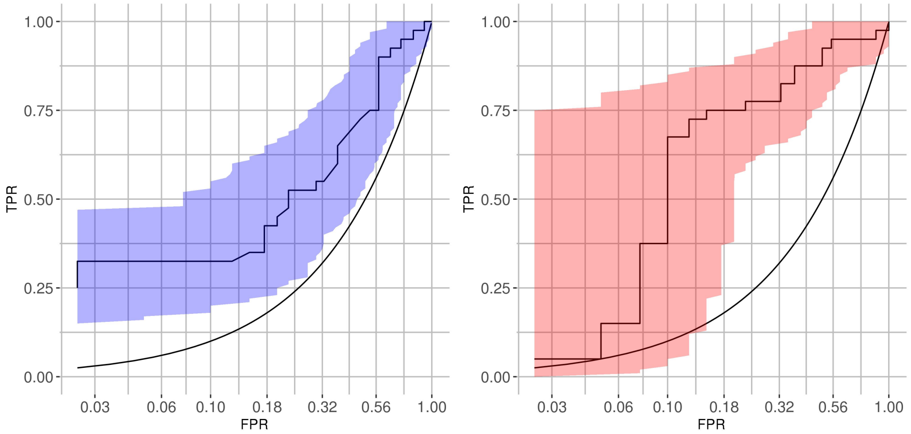


Figure 1.11: ROC curve with 95% confidence intervals displaying false positive rate in x-axis and true positive rate in y-axis, with x-axis in log-scale to emphasize area of low-false positive rate. Left: CAA Embeddings. Right: Logistic regressing on sets.

power than CAE, which can be observed by comparing the Area Under the ROC Curve (AUC), it is important to observe the performance at low false positive rates and low false negative rates, given that these are the operational ranges of the models that would be used in practice. The ROC curves with the x-axes in logarithmic scales to emphasize the low false positive and low false negative rates are shown in Figure 1.11 and Figure 1.12,

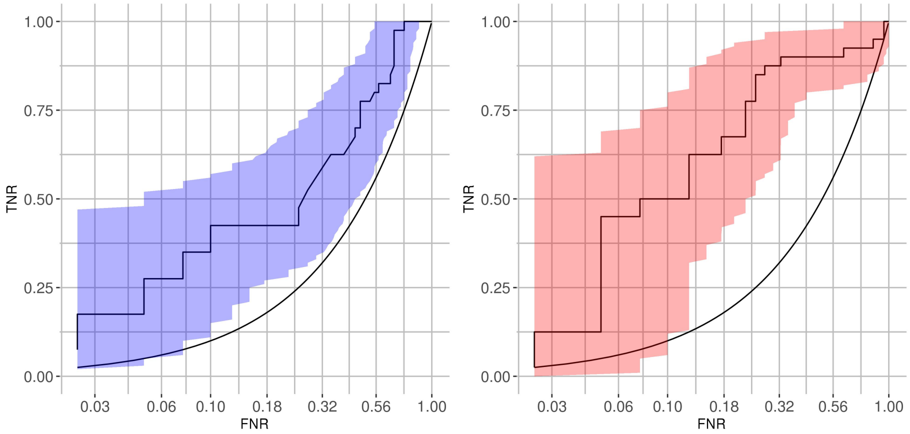


Figure 1.12: ROC curves with 95% confidence intervals displaying false negative rate in x-axis and true negative rate in y-axis, with x-axis in log-scale to emphasize area of low-false negative rate. Left: CAA Embeddings. Right: Logistic regressing on sets.

respectively. The performance of CAE at low false positive rates is promising, while the performance of logistic regression at both low false positives and low false negative rates is not significantly better than random. The proposed CAE model can identify with high confidence a substantial number of patients who will likely go on to have a good neurological outcome.

Even though consensus guidelines advocate maintaining life-sustaining therapies for at least 72 hours after cardiac arrest [18, 19], the burden associated to continuing life-support for patients who will not have a positive neurological recovery still often leads clinicians to withdraw treatment earlier [11]. Hence, the ability to identify with high confidence patients that will likely recover with a good outcome has the potential to save lives. Figure 1.10 shows that the proposed methodology can identify 25% of the patients that will recover with little chance of making such determination in error. And even if the lower bound of the confidence interval is considered, that would correspond to 12.5% of patients that go on to recover. In order to maximize overall performance in addition to optimizing the performance at low false positive rates, an ensemble model including CAE and logistic regression could be used to draw benefits from both of its components: high recall at low false positive rates of CAE, and overall good separability between outcome classes of logistic regression.

It is hard to evaluate the immediate medical impact of these findings in the absence of clinical context. To appropriately estimate the potential impact of such a decision support system in terms of lives saved, it is necessary to compare against physicians' assessments to guarantee that the predictions made with the proposed approach are non-redundant to what doctors already know. Each patient in our dataset is classified by Pittsburgh

Cardiac Arrest Category, a 4-level, validated prognostic indicator assigned in the first six hours of their stay [43]. This classification indicates whether the patient is awake with mild brain injury (category *i*), in a mild to moderately deep coma with good cardiac and pulmonary function (category *ii*), in a mild to moderately deep coma without evidence of severe brain injury but poor cardiac and pulmonary activity (category *iii*), or deeply comatose with loss of some brainstem reflexes (category *iv*). While patients in category *i* have an associated probability of survival of 80%, and 60% probability of having a positive neurological recovery, patients in category *iv* have an associated survival probability of 10%, and only 5% of having a positive neurological outcome. At a false positive rate lower than 0.03, the proposed methodology correctly identified a category *iv* patient who later went on to have a positive recovery. This constitutes a preliminary indication that the patterns of correlations in neurological activity measured with EEG, that are found to be indicative of positive recovery, constitute novel findings and have the potential to improve reliability of prognostication.

From the ROC curves in Figures 1.11 and 1.12, it can be seen that it is easier for the model to determine if patients will have a positive neurological recovery than if they will not. However, this observation should be taken with a grain of salt, and it cannot be concluded that the correlation structures in EEG signals are more useful to predict positive than negative outcomes. The available labeled data encodes positive/negative outcomes, but these are not limited to just neurological activity. A patient could have a positive neurological recovery but have other medical complications that might limit function and thus would result in a *bad outcome* label. Meanwhile, the positive recovery label is potentially much more homogeneous and is sure to indicate positive neurological recovery (as well as positive recovery in other areas). The fact that some patients who had positive neurological recovery could be labeled as having a bad outcome might be adding noise, and it is possible that a cleaner dataset would increase predictive power for those patients who will not go on to have a good neurological outcome.

A limitation of the presented approach is that the analysis is done for a two hour interval after 34 hours of monitoring. Taking into account the results presented in the literature [15], the power of the model could be enhanced by incorporating trajectory modeling. While our model captures correlations observed within an interval of time, and in that sense it goes beyond a purely stationary approach, leveraging the sequential structures in data and using all data collected during a patients' stay, has the potential of further improving performance. Methodologically, this calls for the development of models for trajectory modeling of multivariate correlation structures. This could also encompass further exploration of additional distance metrics that could incorporate other types of information. By leveraging more information, such an approach would have the potential of providing earlier and more specific predictions.

An additional direction for performance enhancement comes from the fact that our characterization of brain activity with CAA is motivated by the importance clinicians place on correlations. However, the correlations they know to be informative are across raw EGG channel measurements, and it is likely that at the current level of data aggregation, a big portion of the information may be to some extent obfuscated. This does not constitute a risk in terms of the validity of the results presented in this Chapter, but it means that if correlations are informative even at this level of aggregation, further promising results may be expected from characterizing correlations in raw EEG signals. In addition to the potentially improved predictive power, such models could lead to biological insights that may not be easily derived with the current approach.

Another limitation of the present model (as well as other relevant approaches) is the selective labels problem [2]. Selective labels is a common yet understudied problem that often arises in decision support, whenever historical decision-making blinds us to the true outcome for certain instances. In the case of predicting neurological recovery, the true outcome is only observed when the clinicians decide to extend life sustaining therapy, while there is no available counterfactual for what would have happened in those cases where life sustaining therapy is withdrawn early. Currently, humans are not always certain of their decision to withdraw life support; therefore, assuming that any case in which doctors decided to stop life-sustenance is an example of a patient with negative neurological outcome might lead to a self-fulfilling prophecy. At the same time, when the predictive model is trained by only considering those cases where the true outcome is observed–that is when life-sustaining therapies were extended and it is possible to observe if the patient had a positive neurological recovery or not–there is a chance our model will not perform well when deployed on the entire population. Currently, this model may only be used to make predictions for the portion of the population for whom life sustaining therapy is being extended. If patients for whom treatment was stopped early are significantly different from those for whom it was not, which is very likely the case, our model could systematically misdiagnose that group if used for the entire population. Chapter 3 discusses the selective labels problem in more detail and introduces novel methodology to tackle this challenge.

Chapter 2

Algorithmic fairness

2.1 Compounding injustices in allocation harms

Section based on:
M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, A. Kalai. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting, In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT*)*, 2019.

When considering the deployment of automated decision-making systems, it is important to acknowledge the increasingly active role these systems play in shaping our future. Far from being passive players that consume information, automated decision-making systems are participating actors: their predictions today affect the world we live in tomorrow. In particular, they determine many aspects of how we experience the world, from the news we read and the products we shop for to the job postings we see. The increased prevalence of machine learning has therefore been accompanied by a growing concern regarding the circumstances and mechanisms by which such systems may reproduce and augment the various forms of discrimination and injustices that are present in today's society.

One domain in which the use of machine learning is growing in popularity—and in which unfair practices can lead to particularly negative consequences—is that of online recruiting and automated hiring. Maintaining an online professional presence has become increasingly important for people's careers, and this information is often used as input to automated decision-making systems that advertise open positions and recruit candidates for jobs and other professional opportunities. In order to perform these tasks, a system must be able to accurately assess people's current occupations, skills, interests, and "potential." However, even the simplest of these tasks—determining someone's current occupation—

can be non-trivial. Although this information may be provided in a structured form on some professional networking platforms, this is not always the case. As a result, recruiters often browse candidates' websites in an attempt to manually determine their current occupations. Machine learning promises to reduce this burden; however, as we will explain in this Section, occupation classification is susceptible to gender bias, stemming from existing gender imbalances in occupations.

To study gender bias in occupation classification, we created a new dataset of hundreds of thousands of online biographies, written in English, from the Common Crawl corpus. Because biographies are typically written in the third person by their subjects (or people familiar with their subjects) and because pronouns are gendered in English, we were able to extract (likely) self-identified binary gender from the biographies. We note, though, that this binary model is a simplification that fails to capture important aspects of gender and erases people who do not fit within its assumptions.

Using this dataset, we predicted people's occupations by performing multi-class classification using three different semantic representations: bag-of-words, word embeddings, and deep recurrent neural networks. For each representation, we considered two scenarios: (1) where explicit gender indicators are available to the classifier, (2) where explicit gender indicators are "scrubbed" to promote fairness or to comply with regulations or laws. We define explicit gender indicators to be information, such as first names and gendered pronouns, that make it possible to determine gender. We note that the practice of "scrubbing" explicit gender indicators and other sensitive attributes is not unique to machine learning, and is often used as a way to mitigate the effects of implicit and explicit bias on decisions made by humans. For example, gender diversity in orchestras was significantly improved by the introduction of "blind" auditions, where candidates play behind a curtain [44].

To quantify gender bias, we compute the true positive rate (TPR) gender gap—i.e., the difference in TPRs between genders—for each occupation. The TPR for a given gender and occupation is defined as the proportion of people with that gender and occupation that are correctly predicted as having that occupation. We also compute the correlation between these TPR gender gaps and existing gender imbalances in occupations, and show how this may compound these imbalances; we connect this finding with an existing notion of indirect discrimination in political philosophy. We show that "scrubbing" explicit gender indicators reduces the TPR gender gaps, while maintaining overall classifier accuracy. However, we also show that significant TPR gender gaps remain in the absence of explicit gender indicators, and that these gaps are correlated with existing gender imbalances. For orchestra auditions, the sounds made by candidates' shoes mean that a curtain is not sufficient to make an audition "blind." It is therefore common practice to additionally roll out a carpet or to ask candidates to remove their shoes [44]. By analogy, "scrubbing" explicit gender indicators is like introducing a curtain—the sounds made by the candidates' shoes remain.

This Section has two main takeaways: First, "scrubbing" explicit gender indicators is not sufficient to remove gender bias from an occupation classifier. Second, even in the absence of such indicators, TPR gender gaps are correlated with existing gender imbalances in occupations; occupation classifiers may therefore compound existing gender imbalances. Although we focus on gender bias, we note that other biases, such as those involving race or socioeconomic status, may also be present in occupation classification or in other tasks related to online recruiting and automated hiring. We structure our analysis so as to inform discussions about these biases as well.

2.1.1 Related work

Recent work has studied the ways in which stereotypes and other human biases may be reflected in semantic representations such as word embeddings [45, 46, 47]. Natural language processing researchers have also studied gender bias in coreference resolution [48, 49], showing that systems perform better when linking a gender pronoun to an occupation in which that gender is overrepresented than to an occupation in which it is underrepresented. Gender bias has also been studied in YouTube's autocaptioning [50], where researchers found a higher word error rate for female speakers. In the context of language identification, researchers have also investigated racial bias, showing that African-American English is often misclassified as non-English [51]. Finally, machine learning methods for identifying toxic comments exhibit disproportionately high false positive rates for words like *gay* and *homosexual* [52].

In the context of structured data, there have been extensive discussions about proxy behavior that may occur when sensitive attributes are not explicitly available but can be determined from other attributes [53, 54, 55]. Related discussions have focused on the phenomenon of differential subgroup validity [56], where the choice of attributes may disadvantage groups for whom the chosen attributes are not equally predictive of the target label [57]. Barocas and Selbst [54] discussed these issues in the context of automated hiring; Kim [58] explained how data-driven decisions that systematically bias people's access to opportunities relate to existing antidiscrimination legislation, identifying voids that may need to be filled to account for potential risks stemming from automated decision-making systems. Researchers have also discussed making available sensitive attributes as a means to improve fairness [59], as well as various ways to use these attributes [60, 53]. Finally, although this Section does not directly consider ranking scenarios, fairness in ranking is particularly relevant to discussions about gender bias in online recruiting and automated hiring [61, 62, 63, 64, 65].

We quantify gender bias by computing the TPR gender gap—i.e., the difference in TPRs between genders—for each occupation. This notion of bias is closely related to the equality of opportunity fairness metric of Hardt et al. [66]. We choose to focus on TPR gender gaps because they enable us to study the ways in which gender imbalances may be compounded; in turn, we relate this to compounding injustices [1]—an existing notion of indirect discrimination in political philosophy that holds that it is a general moral duty to refrain from taking actions that would harm people when those actions are informed by, and would compound, prior injustices suffered by those people. We show that the TPR gender gaps are correlated with existing gender imbalances in occupations. As a result, occupation classifiers compound injustices when existing gender imbalances are attributable to historical discrimination.

This work is also closely related to research on gender bias in hiring [67, 68, 69, 70]. In particular, Bertrand and Mullainathan [71] conducted an experiment in which they responded to help-wanted ads using fictitious resumes, varying names so as to signal gender and race, while keeping everything else the same. They were therefore able to measure the effect of (inferred) gender and race on the likelihood of being called for an interview. Similarly, we study the effect of explicit gender indicators on occupation classification.

Computational linguistics researchers have explored the use of lexical and syntactic features to infer authors' genders [72, 73]. Given that our dataset consists of online biographies, our research is also related to research on differences between the ways that men and women represent themselves. In the context of online professional presences, Altenburger et al. [74] analyzed self-promotion in LinkedIn, finding that women are more modest than men in expressing accomplishments and are less likely to use free-form fields. Researchers have also studied differences in volubility between men and women [75], showing that women's fear of being highly voluble is justified by the fact that both men and women negatively evaluate highly voluble women. Moving beyond self-representation, Niven and Zilber [76] analyzed congressional websites and found that differences between the ways that the media portray men and women in Congress cannot be explained by differences between the ways that they portray themselves. Meanwhile, Smith et al. [77] analyzed attributes used to describe men and women in performance evaluations, showing that negative attributes are more often used to describe women than men. This research on representation by others relates to our work because we cannot be sure that the online biographies in our dataset were actually written by their subjects.

2.1.2 Data collection process

To study gender bias in occupation classification, we created a new dataset using the Common Crawl. Specifically, we identified online biographies, written in English, by filtering for lines that began with a name-like pattern (i.e., a sequence of two capitalized words) followed by the string "is a(n) (xxx) *title*," where *title* is an occupation from the BLS Standard Occupation Classification system.[1] We identified the twenty-eight most frequent occupations based on their appearance in a small subset of the Common Crawl. In a few cases, we merged occupations. For example, we created the occupation *professor* by merging occupations that consist of *professor* and a modifier, such as *economics professor*. Having identified the most frequent occupations, we processed WET[2] files from sixteen distinct crawls from 2014 to 2018, extracting online biographies corresponding to those oc-

[1] https://www.bls.gov/soc/

[2] WET is a special file format containing cleaned text extracted from webpages.

cupations only. Finally, we performed de-duplication by treating biographies as duplicates if they had the same first name, last name, and occupation, and either no middle name was present or one middle name was a prefix of the other. The resulting dataset consists of 397,340 biographies spanning twenty-eight different occupations. Of these occupations, *professor* is the most frequent, with 118,400 biographies, while *rapper* is the least frequent, with 1,406 biographies (see Figure 2.1). The longest biography is 194 tokens, while the shortest is eighteen; the median biography length is seventy-two tokens. We note that the demographics of online biographies' subjects differ from those of the overall workforce, and that our dataset does not contain all biographies on the Internet; however, neither of these factors is likely to undermine our findings.

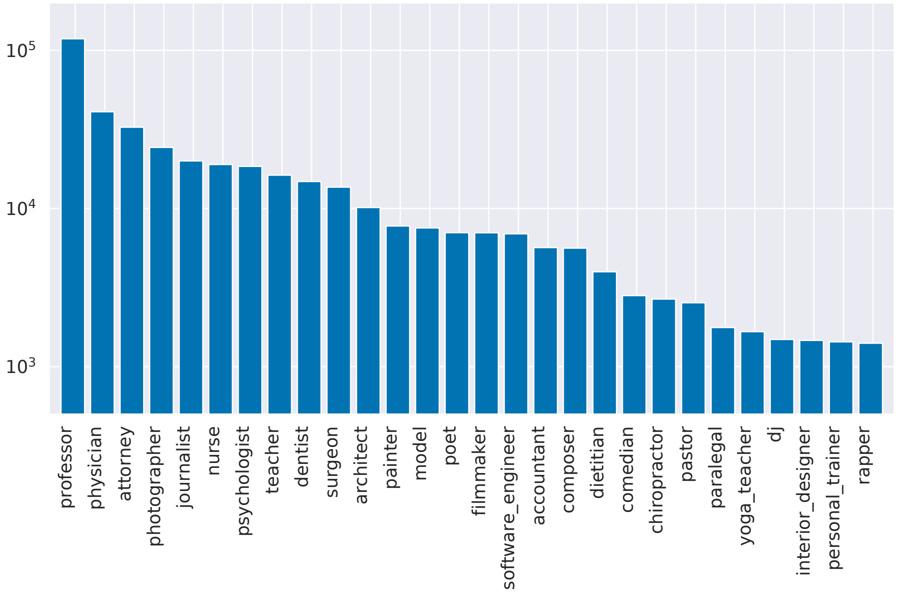


Figure 2.1: Distribution of the number of biographies for the twenty-eight different occupations, shown on a log scale.

Because some occupations have a high gender imbalance, our validation and testing splits must be large enough that every gender and occupation are sufficiently represented. We therefore used stratified-by-occupation splits, with 65% of the biographies (258,370) designated for training, 10% (39,635 biographies) designated for validation, and 25% (99,335 biographies) designated for testing.

A complete implementation that reproduces the dataset can be found in the source code available at http://aka.ms/biasbios.

2.1.3 Methodology

We used our dataset to predict people's occupations, taken from the first sentence of their biographies as described in the previous section, given the remainder of their biographies. For example, consider the hypothetical biography *Nancy Lee is a registered nurse. She graduated from Lehigh University, with honours in 1998. Nancy has years of experience in weight loss surgery, patient support, education, and diabetes.* The goal is to predict *nurse* from *She graduated from Lehigh University, with honours in 1998. Nancy has years of experience in weight loss surgery, patient support, education, and diabetes.*

We used three different semantic representations of varying complexity: bag-of-words (BOW), word embeddings (WE), and deep recurrent neural networks (DNN). When using the BOW and WE representations, we used a one-versus-all logistic regression as the occupation classifier; to construct the DNN representation, we started with word embeddings as input and then trained a DNN to predict occupations in an end-to-end fashion. For each representation, we considered two scenarios: (1) where explicit gender indicators—e.g., first names and pronouns—are available to the classifier, (2) where explicit gender indicators are "scrubbed." For example, these scenarios correspond to predicting the occupation *nurse* from the text *[She] graduated from Lehigh University, with honours in 1998. [Nancy] has years of experience in weight loss surgery, patient support, education, and diabetes,* with and without the bracketed words.

Semantic representations

Bag-of-words The BOW representation encodes the i^{th} biography as a sparse vector x_i^{BOW}. Each element of this vector corresponds to a word type in the vocabulary, equal to 1 if the biography contains a token of this type and 0 otherwise. Despite recent successes of using more complex semantic representations for document classification, the BOW representation provides a good baseline and is still widely used, especially in scenarios where interpretability is important. To predict occupations, we trained a one-versus-all logistic regression with L_2 regularization using our dataset's training split represented using the BOW representation.

Word embeddings The WE representation encodes the i^{th} biography as a vector x_i^{WE}, obtained by averaging the `fastText` word embeddings [78, 79] for the word types present in that biography.[3] The WE representation is surprisingly effective at capturing non-trivial semantic information [80]. To predict occupations, we trained a one-versus-all logistic regression with L_2 regularization using our dataset's training split represented using the WE representation.

[3] We note that the `fastText` word embeddings were trained using the Common Crawl, albeit using a different subset than the one we used to create our dataset.

Deep recurrent neural networks To construct the DNN representation, we started with the `fastText` word embeddings as input and then trained a DNN to predict occupations in an end-to-end fashion. We used an architecture similar to that of Yang et al. [81], but with just one bi-directional recurrent neural network at the level of words and with gated recurrent units (GRUs) [82] instead of long short-term memory units; this model uses an attention mechanism—an integral part of modern neural network architectures [83]. Our choice of architecture was motivated by a desire to use a relatively simple model that would be easy to interpret.

Formally, given the i^{th} biography represented as a sequence of tokens $w_i^1, \dots, w_i^T$, we start by replacing each token w_i^t with the `fastText` word embedding for that word type to yield $e_i^1, \dots, e_i^T$. The DNN then uses a GRU to process the biography in both forward and reverse directions and concatenates the corresponding hidden states from both directions to re-represent the t^{th} token as follows:

$$\overrightarrow{h_i^t} = \overrightarrow{GRU}(e_i^t, h_i^{t-1}) \tag{2.1}$$

$$\overleftarrow{h_i^t} = \overleftarrow{GRU}(e_i^t, h_i^{t+1}) \tag{2.2}$$

$$h_i^t = [\overleftarrow{h_i^t}; \overrightarrow{h_i^t}]. \tag{2.3}$$

Next, the DNN projects each hidden state h_i^t to the attention dimension k_a via a fully connected layer with weights W_a and b_a, and transforms the result into an unnormalized scalar u_i^t via a vector w_a:

$$\hat{u}_i^t = \tanh\left(W_a\, h_i^t + b_a\right) \tag{2.4}$$

$$u_i^t = w_a^\intercal \hat{u}_i^t. \tag{2.5}$$

Each scalar is then normalized to yield an attention weight:

$$\alpha_i^t = \frac{\exp\left(u_i^t\right)}{\sum_{t'=1}^{T} \exp\left(u_i^{t'}\right)}. \tag{2.6}$$

Finally, we obtain the DNN representation via a weighted sum:

$$x_i^{\text{DNN}} = \sum_{t=1}^{T} \alpha_i^t\, h_i^t. \tag{2.7}$$

The DNN makes predictions as follows:

$$\hat{y}_i = \text{softmax}(W_0\, x_i^{\text{DNN}} + b_0), \tag{2.8}$$

where $\hat{y}_i$ is the predicted occupation for the i^{th} biography.

We trained the DNN using our dataset's training split and a standard cross-entropy loss applied to the output of the last layer.

Explicit gender indicators

For each semantic representation, we considered two scenarios. In the first scenario, the representation included all word types, meaning that explicit gender indicators are available to the occupation classifier. In the second scenario, we "scrubbed" explicit gender indicators prior to creating the representation, meaning that these indicators are not available to the occupation classifier. Specifically, we deleted the subject's first name, along with the words *he*, *she*, *her*, *his*, *him*, *hers*, *himself*, *herself*, *mr*, *mrs*, and *ms* from each biography.

2.1.4 Analysis and results

In this section, we analyze the potential allocation harms that can result from semantic representation bias. To do this, we study the performance of the occupation classifier for each semantic representation, with and without explicit gender indicators, as described in the previous section. The classifiers' overall accuracies are shown in Figure 2.2. We start by analyzing gender bias for the scenario in which the semantic representations include all word types, including explicit gender indicators. We then analyze gender bias in the scenario in which explicit gender indicators are "scrubbed," and use the DNN's per-token attention weights to understand proxy behavior that occurs in the absence of explicit gender indicators.

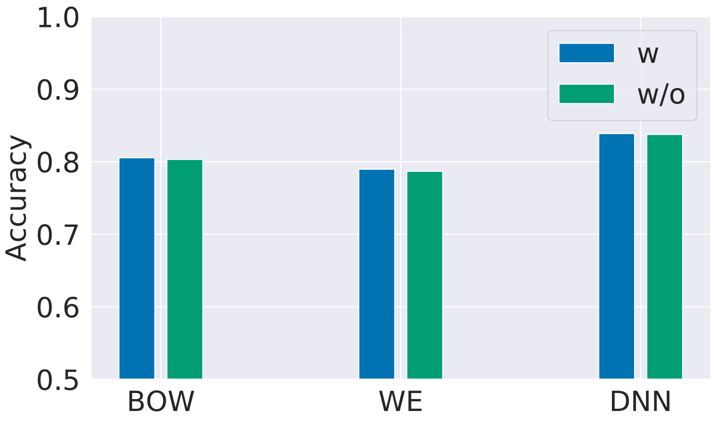


Figure 2.2: Occupation classifier accuracy for each semantic representation, with and without explicit gender indicators.

With explicit gender indicators

True positive rate gender gap For each semantic representation, we quantify gender bias by using our dataset's testing split to calculate the occupation classifier's TPR gender gap—i.e., the difference in TPRs between binary genders g and $\sim g$—for each occupation y:

$$\mathrm{TPR}_{g,y} = P\,[\hat{Y} = y \,|\, G = g, Y = y] \tag{2.9}$$

$$\mathrm{Gap}_{g,y} = \mathrm{TPR}_{g,y} - \mathrm{TPR}_{\sim g,y}, \tag{2.10}$$

where $\hat{Y}$ and Y are random variables representing the predicted and target labels (i.e., occupations) for a biography and G is a random variable representing the binary gender of the biography's subject.

Defining the percentage of people with gender g in occupation y as $\pi_{g,y} = P\,[G = g \,|\, Y = y]$, Figure 2.3 shows $\mathrm{Gap}_{\mathrm{female},y}$ versus $\pi_{\mathrm{female},y}$ for each occupation y for the BOW representation with explicit gender indicators; Figure 2.4 depicts the same information for all three representations, with and without explicit gender indicators.

Compounding imbalance We define the gender imbalance of occupation y as $\frac{\pi_{g,y}}{\pi_{\sim g,y}}$; gender g is underrepresented if $\frac{\pi_{g,y}}{\pi_{\sim g,y}} < 1$ or, equivalently, if $\pi_{g,y} < 0.5$. The gender imbalance is compounded if the underrepresented gender has a lower TPR than the over-represented gender—e.g., if $\mathrm{Gap}_{g,y} < 0$ and g is underrepresented.

Theorem 1. *If $\pi_{g,y} < 0.5$ and $\mathit{Gap}_{g,y} < 0$, then*

$$P\,[G = g \,|\, Y = \hat{Y} = y] < \pi_{g,y}. \tag{2.11}$$

Proof. Via Bayes theorem,

$$P\,[G = g \,|\, Y = \hat{Y} = y] = \frac{\pi_{g,y}\,\mathrm{TPR}_{g,y}}{P\,[\hat{Y} = y \,|\, Y = y]}. \tag{2.12}$$

If $\pi_{g,y} < \pi_{\sim g,y}$ and $\mathrm{TPR}_{g,y} < \mathrm{TPR}_{\sim g,y}$, then

$$\frac{P\,[G = g \,|\, Y = \hat{Y} = y]}{P\,[G = \sim g \,|\, Y = \hat{Y} = y]} = \frac{\pi_{g,y}\,\mathrm{TPR}_{g,y}}{\pi_{\sim g,y}\,\mathrm{TPR}_{\sim g,y}} < \frac{\pi_{g,y}}{\pi_{\sim g,y}}, \tag{2.13}$$

so the gender imbalance for the true positives in occupation y is larger than the initial gender imbalance in that occupation. □

As explained in Section 2.1.1, if the initial gender imbalance is due to prior injustices, an occupation classifier will compound these injustices, which may correspond to indirect discrimination [1].

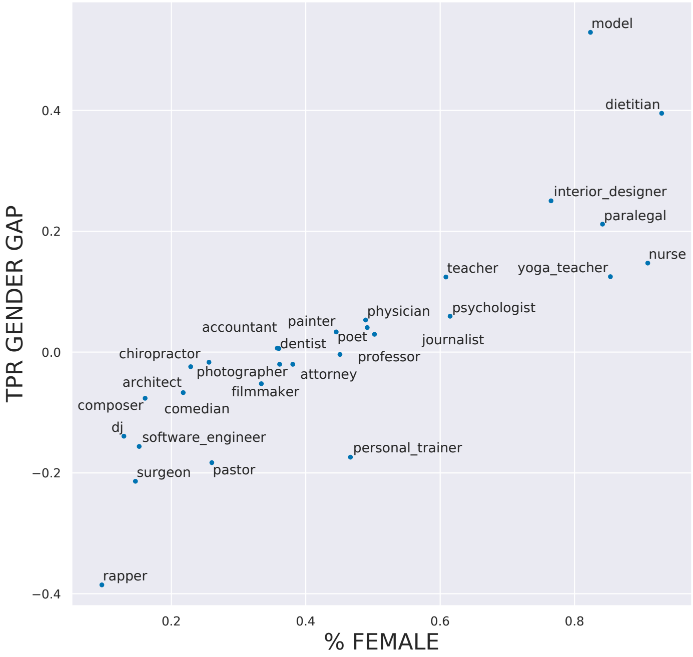


Figure 2.3: $\text{Gap}_{\text{female},y}$ versus $\pi_{\text{female},y}$ for each occupation y for the BOW representation with explicit gender indicators.

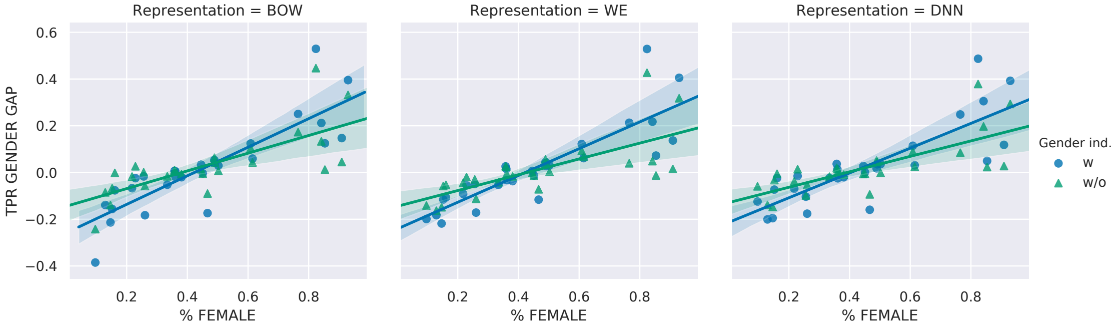


Figure 2.4: $\text{Gap}_{\text{female},y}$ versus $\pi_{\text{female},y}$ for each occupation y for all three semantic representations, with and without explicit gender indicators. Correlation coefficients: BOW-w 0.85; BOW-wo 0.74; WE-w 0.86; WE-wo 0.71; DNN-w 0.82, DNN-wo 0.74.

It is clear from Figure 2.3 that there are few occupations with an equal percentage of men and women—i.e., almost all occupations have a gender imbalance—and that for that for occupations in which women (conversely men) are underrepresented, $\text{Gap}_{\text{female},y} < 0$ (conversely $\text{Gap}_{\text{male},y} < 0$). In other words, there is a positive correlation between the TPR gender gap for an occupation y and the gender imbalance in that occupation. (Figure 2.4 illustrates that this is also the case for the WE and DNN representations.) As a result, if the occupation classifier for the BOW representation were used to recruit candidates for jobs in occupation y, it would compound the gender imbalance by a factor of $\frac{\text{TPR}_{g,y}}{\text{TPR}_{\sim g,y}}$, where g is the underrepresented gender. For example, 14.6% of the surgeons in our dataset's testing split are women—i.e., $\pi_{\text{female,surgeon}} < 0.5$. The classifier for the BOW representation is able to correctly predict that 71% of male surgeons and 54.5% of female surgeons are indeed surgeons—i.e., $\text{Gap}_{\text{female,surgeon}} < 0$. Consequently, only 11.6% of the true positives are women, so the gender imbalance is compounded.

Counterfactuals To isolate the effects of explicit gender indicators on the representations' occupation classifiers, we examined differences between the classifiers' predictions on our dataset's testing split as described above and their predictions on our dataset's testing split with first names removed and other explicit gender indicators (see Section 2.1.3) swapped for their complements, keeping everything else the same. This analysis is similar in spirit to the experiment of Bertrand and Mullainathan [71], in which they responded to help-wanted ads using fictitious resumes in order to measure the effect of gender and race on the likelihood of being called for an interview. By analyzing the counterfactuals obtained by swapping gender indicators, we can answer the question, "Which occupation would this classifier predict if this biography had used indicators corresponding to the other gender." This question is interesting because we would expect an occupation classifier to predict the same occupation for a man and a woman with identical biographies. We note that this question is not the same as the question, "Which occupation would this classifier predict if this biography's subject were the other gender." Although the latter question is arguably more interesting, it cannot be answered without additionally changing all other factors that are correlated with gender [84].

For the BOW representation, we find that the classifier's predictions for 5.5% of the biographies in our testing split change when their gender indicators are swapped; for the WE and DNN representations, these percentages are 12.2% and 4.6%, respectively. To better understand the effects of explicit gender indicators on the classifiers' predictions, we consider pairs of occupations. Specifically, for each gender g and pair of occupations (y^1, y^2), we identify the set of biographies that are incorrectly predicted as having occupation y^1 with their original gender indicators, but correctly predicted as having occupation y^2 when their gender indicators are swapped:

$$\mathbb{S}_{g,(y^1,y^2)} = \{x_i^R : \hat{y}_i = y^1, \hat{y}_i^{(g\leftrightarrow\sim g)} = y^2, y_i = y^2\}, \tag{2.14}$$

where x_i^R is the i^{th} biography, y_i is the target label (i.e., occupation) for that biography, $\hat{y}_i$ is the predicted label for that biography with its original gender indicators, and $\hat{y}_i^{(g \leftrightarrow \sim g)}$ is the predicted label for that biography when its gender indicators are swapped. For example, $\mathbb{S}_{\text{female},(\text{nurse},\text{surgeon})}$ is the set of biographies for female surgeons who are incorrectly predicted as nurses, but correctly predicted as surgeons when their biographies use male indicators. We also identify the total set of biographies $\mathbb{S}_{g,y^2}$ that are only correctly predicted as having occupation y^2 when their gender indicators are swapped, and then calculate the percentage of these biographies for which the predicted label changes from y^1 to y^2:

$$\Pi_{g,(y^1,y^2)} = \frac{|\mathbb{S}_{g,(y^1,y^2)}|}{|\mathbb{S}_{g,y^2}|} \times 100\%. \tag{2.15}$$

Tables 2.1 and 2.2 list, for the BOW representation, the five pairs of occupations with the largest values of $\Pi_{g,(y^1,y^2)}$. For example, 7.1% of male paralegals whose occupations are only correctly predicted when their gender indicators are swapped are incorrectly predicted as attorneys when their biographies use male indicators. Similarly, 14.7% of female rappers whose occupations are only correctly predicted when their gender indicators are swapped are incorrectly predicted as models when their biographies use female indicators.

Without explicit gender indicators

Remaining gender information If there are no differences between the ways that men and women in occupation y represent themselves in their biographies other than explicit gender indicators, then "scrubbing" these indicators should be sufficient to remove all information about gender from the biographies—i.e.,

$$P\,[\tilde{X}^R = \tilde{x}^R \,|\, G = g, Y = y] = P\,[\tilde{X}^R = \tilde{x}^R \,|\, G = \sim g, Y = y], \tag{2.16}$$

where $\tilde{X}^R$ is a random variable representing a biography without explicit gender indicators, G is a random variable representing the binary gender of the biography's subject, and Y is a random variable representing the biography's target label (i.e., occupation). In turn, this would mean that the TPRs for genders g and $\sim g$ are identical:

$$\begin{aligned} \text{TPR}_{g,y} &= P\,[\hat{Y} = y \,|\, G = g, Y = y] && (2.17)\\ &= P\,[\hat{Y} = y \,|\, G = \sim g, Y = y] && (2.18)\\ &= \text{TPR}_{\sim g,y}, && (2.19) \end{aligned}$$

where $\hat{Y} = f(\tilde{X}^R)$ is a random variable representing the predicted label (i.e., occupation) for $\tilde{X}^R$. Moreover, it would also mean that

$$P\,[G = g \,|\, \tilde{X}^R = \tilde{x}^R, Y = y] = P[G = \sim g \,|\, \tilde{X}^R = \tilde{x}^R, Y = y], \tag{2.20}$$

Table 2.1: Pairs of occupations with the largest values of $\Pi_{\text{male},(y^1,y^2)}$—i.e., the percentage of men's biographies that are only correctly predicted as y^2 when their indicators are swapped for which the predicted label changes from y^1.

y^1	y^2	$\Pi_{\text{male},(y^1,y^2)}$
attorney	paralegal	7.1%
architect	interior designer	4.7%
professor	dietitian	4.3%
photographer	interior designer	3.5%
teacher	yoga teacher	3.3%

Table 2.2: Pairs of occupations with the largest values of $\Pi_{\text{female},(y^1,y^2)}$—i.e., the percentage of women's biographies that are only correctly predicted as y^2 when their indicators are swapped for which the predicted label changes from y^1.

y^1	y^2	$\Pi_{\text{female},(y^1,y^2)}$
model	rapper	14.7%
teacher	pastor	8.5%
professor	software engineer	6.5%
professor	surgeon	4.8%
physician	surgeon	3.8%

making it impossible to predict the gender of a "scrubbed" biography's subject belonging to occupation y better than random.

In order to determine whether "scrubbing" explicit gender indicators is sufficient to remove all information about gender, we used a balanced subsample of our dataset to predict people's gender. We created a subsampled training split by first discarding from our dataset's training split all occupations for which there were not at least $1,000$ biographies for each gender. For each of the remaining twenty-one occupations, we then subsampled $1,000$ biographies for each gender to yield $42,000$ biographies, balanced by occupation and gender. To create a subsampled validation split, we first identified the occupation and gender from those represented in the subsampled training split with the smallest number of biographies in our dataset's validation split. Then, we subsampled that number of biographies from our dataset's validation split for each of the twenty-one occupations represented in the subsampled training split and each gender. We created a subsampled testing split similarly. When using the BOW and WE representations, we used a logistic regression with L_2 regularization as the gender classifier; to construct the DNN representation, we started with word embeddings as input and then trained a DNN to predict gender in an end-to-end fashion, similar to the methodology described in Section 2.1.3.

Using the subsampled testing split, we find that the gender classifier for the BOW representation has an accuracy of 65.5%, while the DNN representation has an accuracy of 68.2%. These accuracies are higher than 50%, so "scrubbing" explicit gender indicators is not sufficient to remove all information about gender. This finding is reinforced by the scatterplot in Figure 2.5, which shows log frequency versus correlation with $G =$ female for each word type in the vocabulary. It is clear from this scatterplot that deleting all words that are correlated with gender would not be feasible.

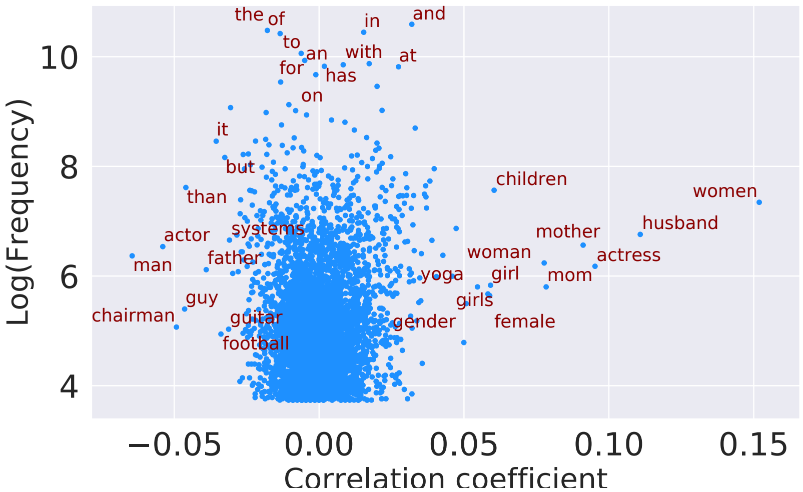


Figure 2.5: Scatterplot of log frequency versus correlation with $G =$ female for each word type in the vocabulary.

True positive rate gender gap and compounding imbalance For each semantic representation, we again quantify gender bias by using our (original) dataset's testing split to calculate the occupation classifier's TPR gender gap for each occupation. Figure 2.4 shows $\text{Gap}_{\text{female},y}$ versus $\pi_{\text{female},y}$ for each occupation y for all three representations, with and without explicit gender indicators. "Scrubbing" explicit gender indicators reduces the TPR gender gaps, while the classifiers' accuracies (shown in Figure 2.2) remain roughly the same; however, for some occupations, $\text{Gap}_{\text{female},y}$ is still very large. Moreover, because there is still a positive correlation between the TPR gender gap for an occupation y and the gender imbalance in that occupation, "scrubbing" explicit gender indicators will not prevent the classifiers from compounding gender imbalances.

We note that compounding imbalances are especially problematic if people repeatedly encounter such classifiers—i.e., if an occupation classifier's predictions determine the data used by subsequent occupation classifiers. Who is offered a job today will affect the gender (im)balance in that occupation in the future. If a classifier compounds existing gender imbalances, then the underrepresented gender will, over time, become even further underrepresented—a phenomenon sometimes referred to as the "leaky pipeline."

To illustrate this phenomenon, we performed simulations using the DNN representation in which the candidate pool at time $t+1$ is defined by the true positives at time t. Defining the percentage of people with gender g in occupation y at time t as $\pi_{g,y}^{(t)}$, we fit a linear regression to the TPR gender gaps for different values of $\pi_{g,y}^{(t)}$:

$$\widehat{\text{Gap}}_{g,y}^{(t)} = \pi_{g,y}^{(t)}\,\beta_1 + \beta_0. \tag{2.21}$$

Using this regression model, we are then able to calculate the percentage of people with gender g in occupation y at time $t+1$:

$$\pi_{g,y}^{(t+1)} = \frac{\pi_{g,y}^{(t)}\,\text{TPR}_{g,y}^{(t)}}{\pi_{\sim g,y}^{(t)}\,(\text{TPR}_{g,y}^{(t)} + \text{Gap}_{g,y}^{(t)}) + \pi_{g,y}^{(t)}\,\text{TPR}_{g,y}^{(t)}}. \tag{2.22}$$

Figure 2.6 shows $\pi_{g,y}^{(t)}$ for $t = 0, \ldots, 10$; each subplot corresponds to a different initial gender imbalance. Over time, the gender imbalances compound. We note that there are many different TPR pairs $\text{TPR}_{g,y}^{(t)}$ and $\text{TPR}_{\sim g,y}^{(t)}$ that can result in a given TPR gender gap $\text{Gap}_{g,y}^{(t)}$. For example, a TPR gender gap of -0.2 might correspond to $0.6 - 0.8$ or to $0.7 - 0.9$. Moreover, different TPR pairs will result in different percentages of people with gender g in occupation y at time $t+1$. The bands in Figure 2.6 therefore reflect these differences.

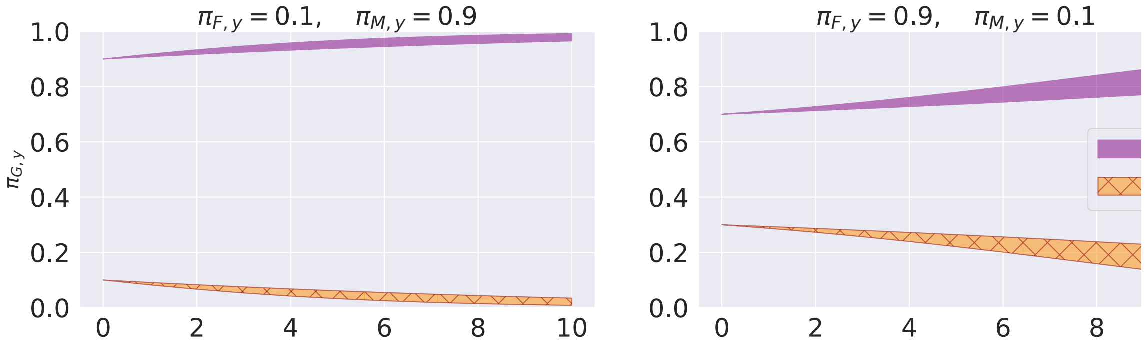
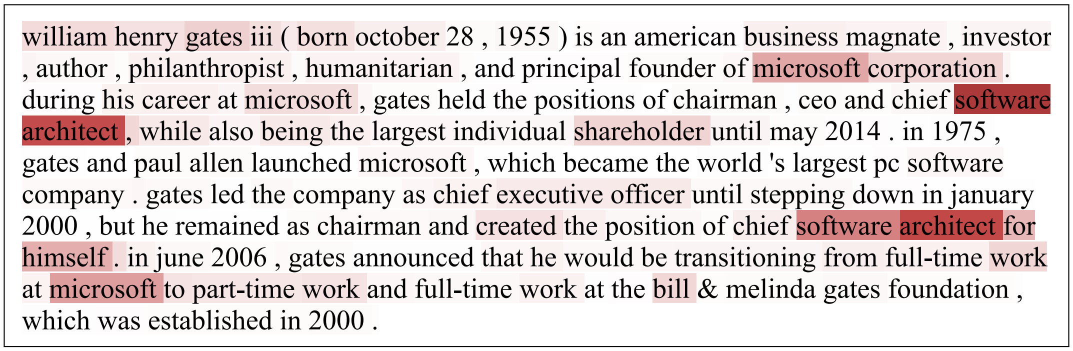


Figure 2.6: Simulations of compounding imbalances using the DNN representation. Each subplot corresponds to a different initial gender imbalance and shows $\pi_{g,y}^{(t)}$ for $t = 0, \ldots, 10$.

Attention to gender The DNN's per-token attention weights allow us to understand proxy behavior that occurs in the absence of explicit gender indicators. The attention



william henry gates iii (born october 28 , 1955) is an american business magnate , investor , author , philanthropist , humanitarian , and principal founder of microsoft corporation . during his career at microsoft , gates held the positions of chairman , ceo and chief software architect , while also being the largest individual shareholder until may 2014 . in 1975 , gates and paul allen launched microsoft , which became the world 's largest pc software company . gates led the company as chief executive officer until stepping down in january 2000 , but he remained as chairman and created the position of chief software architect for himself . in june 2006 , gates announced that he would be transitioning from full-time work at microsoft to part-time work and full-time work at the bill & melinda gates foundation , which was established in 2000 .

Figure 2.7: Visualization of the DNN's per-token attention weights. Predicted label (i.e., occupation): *software engineer*.

weights indicate which tokens are most predictive. For example, Figure 2.7 depicts the per-token attention weights from the occupation classifier for the DNN representation when predicting Bill Gates' occupation from an excerpt of his biography on Wikipedia; the larger the weight, the stronger the color. The attention weights for the words *software* and *architect* are very large, and the DNN predicts *software engineer*.

In order to understand proxy behavior that occurs in the absence of explicit gender indicators, we first used the subsampled testing split, described above, to obtain per-token attention weights from the gender classifier for the DNN representation. We then used these weights to find "proxy candidates"—i.e., the words that are most predictive of gender in the absence of explicit gender indicators. Specifically, we computed the sum of the per-token attention weights for each word type, and then selected the types with the largest sums as "proxy candidates." Across multiple runs, we found that the words *women*, *husband*, *mother*, *woman*, and *female* (ordered by decreasing total attention) were consistently "proxy candidates."

For each "proxy candidate," we then used our dataset's testing split, with and without explicit gender indicators, to create histograms of the per-token attention weights from the occupation classifier for the DNN representation. These histograms represent the extent to which that "proxy candidate" is predictive of occupation, with and without gender indicators. By comparing the histograms for each "proxy candidate," we are able to identify words that are used as proxies for gender in the absence of explicit gender indicators: if there is a big difference between the histograms, then the "proxy candidate" is likely a proxy. Figure 2.8 shows per-occupation histograms for the word *women*, with (left) and without (right) explicit gender indicators. It is clear that in the absence of explicit gender indicators, the classifier has larger attention weights for the word *women* for all occupations. We see similar behavior for the other "proxy candidates," suggesting that the classifier uses proxies for gender in the absence of explicit gender indicators.

The occupations in Figure 2.8 are ordered by TPR gender gap from negative to positive.

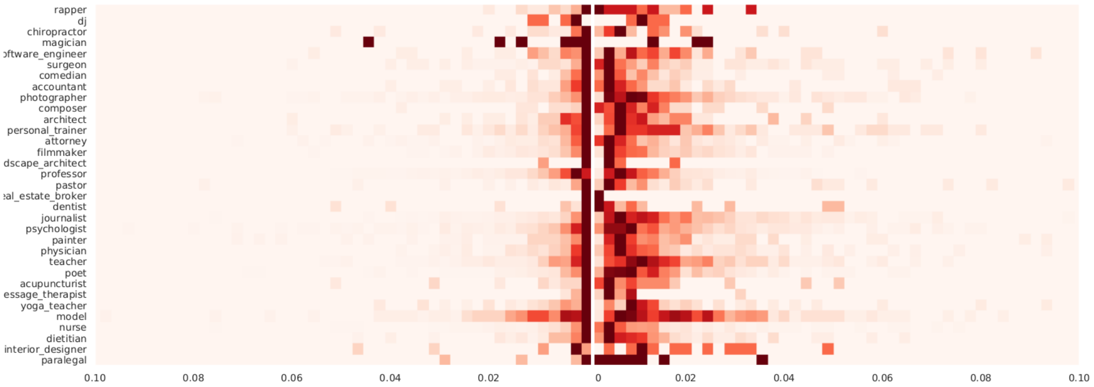


Figure 2.8: Per-occupation histograms of the per-token attention weights from the DNN representation's occupation classifier for the word *women*, with (left) and without (right) explicit gender indicators; occupations are ordered by TPR gender gap.

For occupations in the middle, where there are small or no TPR gender gaps, the classifier still has non-zero attention weights for the word *women*. This means that using gender information does not necessarily lead to a TPR gender gap. We also note that it's possible that the classifier is using gender information to differentiate between occupations with very different gender imbalances that are otherwise similar, such as physician and surgeon.

2.1.5 Discussion

This Section presents a large-scale study of gender bias in occupation classification using a new dataset of hundreds of thousands of online biographies collected for this research and made publicly available. The results show that there are significant TPR gender gaps when using three different semantic representations: bag-of-words, word embeddings, and deep recurrent neural networks. Additionally, theoretical results prove that the correlation between these TPR gender gaps and existing gender imbalances in occupations compounds leads to a compounding imbalance effect. Via simulations, it is shown that compounding imbalances are especially problematic if people repeatedly encounter occupation classifiers because the underrepresented gender will become even further underrepresented.

Recently, Dwork and Ilvento [85] showed that fairness does not hold under composition, meaning that if two classifiers are individually fair according to some fairness metric, then the sequential use of these classifiers will not necessarily be fair according the same metric. One interpretation of our finding regarding compounding imbalances is that unfairness holds under composition. Understanding why this is the case, especially given that fairness does not hold under composition, is an interesting direction for future work.

It is worth highlighting that in the experiments the TPR gender gaps are reduced by "scrubbing" explicit gender indicators, while the classifiers' overall accuracies remain roughly the same. This constitutes an empirical example where there is little tradeoff between promoting fairness—in this case by "scrubbing" explicit gender indicators—and performance. This also constitutes an empirical example where fairness is improved by "scrubbing" sensitive attributes, contrary to other examples in the literature [86]. That said, in the absence of explicit gender indicators, the results show that (1) it is still possible to predict the gender of a biography's subject better than random, even when controlling for occupation; (2) significant TPR gender gaps remain for some occupations; (3) there is still a positive correlation between the TPR gender gap for an occupation and the gender imbalance in that occupation, so existing gender imbalances may be compounded. These findings indicate that there are differences between men's and women's online biographies other than explicit gender indicators. These differences may be due to the ways that men and women represent themselves or due to men and women having different specializations within an occupation. These findings emphasize both the risks of using machine learning in a high-stakes setting and the difficulty of trying to promote fairness by "scrubbing" sensitive attributes.

2.2 What are the biases in my word embedding?

Section based on:
N.Swinger*, M. De-Arteaga*, N.Heffernan, M.Leiserson, A. Kalai. What are the Biases in my Word Embedding?, In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, 2019.

This Section considers the problem of *Unsupervised Bias Enumeration* (UBE): discovering biases automatically from an unlabeled data representation. There are multiple reasons why such an algorithm is useful. First, social scientists can use it as a tool to study human bias, as data analysis is increasingly common in social studies of human biases [47, 87]. Second, finding bias is a natural step in "debiasing" representations [88]. Finally, it can help in avoiding systems that perpetuate these biases: problematic biases can raise red flags for engineers, who can choose to not use a representation or watch out for certain biases in downstream applications, while little or no bias can be a useful green light indicating that a representation is usable. While deciding which biases are problematic is ultimately application specific, UBE may be useful in a "fair ML" pipeline.

We design a UBE algorithm for word embeddings, which are commonly used representations of tokens (e.g. words and phrases) that have been found to contain harmful bias [88]. Researchers linking these biases to human biases proposed the Word Embedding Association Test (WEAT) [89]. The WEAT draws its inspiration from the Implicit Association Test (IAT), a widely-used approach to measure human bias [90]. An IAT $\mathcal{T} = (X_1, A_1, X_2, A_2)$ compares two sets of *target tokens* X_1 and X_2, such as female vs. male names, and a pair of opposing sets of *attribute tokens* A_1 and A_2, such as workplace vs. family-themed words. Average differences in a person's response times when asked to link tokens that have anti-stereotypical vs. stereotypical relationships have been shown to indicate the strength of association between concepts. Analogously, the WEAT uses vector similarity across pairs of tokens in the sets to measure association strength. As in the case of the IAT, the inputs for a WEAT are sets of tokens $\mathcal{T}$ predefined by researchers.

Our UBE algorithm takes as input a word embedding and a list of target tokens, and *outputs* numerous tests $\mathcal{T}_1, \mathcal{T}_2, \ldots$, that are found to be statistically significant by a method we introduce for bounding false discovery rates. A crowdsourcing study of tests generated on three publicly-available word embeddings and a list of names from the Social Security Administration confirms that the biases enumerated are largely consistent with human stereotypes. The generated tests capture racial, gender, religious, and age biases, among others. Table 2.3 shows the name/word associations output by our algorithm that were rated most offensive by crowd workers.

Creating such tests automatically has several advantages. First, it is not feasible to

Word2Vec trained on Google news			fastText trained on the Web			GloVe trained on the Web		
w2v F8	**w2v F11**	**w2v F6**	**fast F10**	**fast F7**	**fast F5**	**glove F8**	**glove F7**	**glove F5**
illegal immigrant	aggravated robbery	subcontinent	n*****	jihad	s******	turban	cartel	pornstar
drug trafficking	aggravated assault	tribesmen	f*****	militants	maid	saree	undocumented	hottie
deported	felonious assault	miscreants	dreads	caliphate	busty	hijab	culpable	nubile

Table 2.3: Terms associated with name groups (see Tables 2.5 and A.1 for name groups **w2v F8**, etc.) generated from three popular pre-trained word embeddings that were rated by crowd workers as both most offensive and aligned with societal biases. These associations do *not* reflect the personal beliefs of the crowd workers or authors of this work. See Appendix A.11 for a discussion of the bleep-censored words.

manually author all possible tests of interest. Domain experts normally create such tests, and it is unreasonable to expect them to cover all possible groups, especially if they do not know which groups are represented in their data. For example, a domain expert based on the United States may not think of testing for caste discrimination, hence biases that an embedding may have against certain Indian last names may go unnoticed. Finally, if a word embedding reveals no biases, this is evidence for lack of bias. We test this by running our UBE algorithm on the supposedly debiased embedding of [88].

Our approach for UBE leverages two geometric properties of word embeddings, which we call the *parallel* and *cluster* properties. The well-known parallel property indicates that differences between two similar token pairs, such as Mary−John and Queen−King, are often nearly parallel vectors. This suggests that among tokens in a similar topic or category, those parallel to name differences may represent biases, as was found by [88] and [89]. The cluster property, which we were previously unaware of, indicates that the (normalized) vectors of names and words cluster into semantically meaningful groups. For names, the clusters capture social structures such as gender, religion, and others. For words, clusters of words include word categories on topics such as food, education, occupations, and sports. We use these properties to design a UBE algorithm that outputs WEATs.

Technical challenges arise around any procedure for enumerating biases. First, the combinatorial explosion of comparisons among multiple groups parallels issues in human IAT studies as aptly described by [91]: "The evaluation of multiple target concepts such as social groups within a multi-ethnic nation [e.g. White vs. Asian Americans, White vs. African Americans, African vs. Asian Americans; 92] requires numerous pairwise comparisons for a complete picture". We alleviate this problem, paralleling that work on human IATs, by generalizing the WEAT to n groups for arbitrary n. The second problem, for any UBE algorithm, is determining statistical significance to account for multiple hypothesis testing. To do this, we introduce a novel rotational null hypothesis specific to word embeddings. Third, we provide a human evaluation of the biases, contending with the difficulty that many people are unfamiliar with some groups of names.

Beyond word embeddings and IATs, related work in other subjects is worth mention.

First, a body of work studies fairness properties of classification and regression algorithms [e.g. 93, 94]. While our work does not concern supervised learning, it is within this work that we find one of our main motivations–the importance of accounting for intersectionality when studying algorithmic biases. In particular, Buolamwini and Gebru [95] demonstrate accuracy disparities in image classification highlighting the fact that the magnitude of biases against an intersectional group may go unnoticed when only evaluating for each protected feature independently. Finally, while a significant portion of the empirical research on algorithmic fairness has focused on the societal biases that are most pressing in the countries where the majority of researchers currently conducting the work are based, the literature also contains examples of biases that may be of particular importance in other parts of the world [96, 97]. UBE can aspire to be useful in multiple contexts, and enable the discovery of biases in a way that relies less on enumeration by domain experts.

2.2.1 Definitions

A d-dimensional word embedding consists of a set of tokens $\mathcal{W}$ with a nonzero vector $\boldsymbol{w} \in \mathbb{R}^d$ associated with each token $w \in \mathcal{W}$. Vectors are displayed in boldface. As is standard, we refer to the *similarity* between tokens v and w by the cosine of their vector angle, $\cos(\boldsymbol{v}, \boldsymbol{w})$. We write $\overline{\boldsymbol{v}} = \boldsymbol{v}/|\boldsymbol{v}|$ to be the vector normalized to unit-length associated with any vector $\boldsymbol{v} \in \mathbb{R}^d$ (or 0 if $\boldsymbol{v} = 0$). This enables us to conveniently write the similarity between tokens v and w as an inner product, $\cos(\boldsymbol{v}, \boldsymbol{w}) = \overline{\boldsymbol{v}} \cdot \overline{\boldsymbol{w}}$. For token set S, we write $\overline{\boldsymbol{S}} = \sum_{v \in S} \overline{\boldsymbol{v}}/|S|$ so that $\overline{\boldsymbol{S}} \cdot \overline{\boldsymbol{T}} = \text{mean}_{v \in S, w \in T} \overline{\boldsymbol{v}} \cdot \overline{\boldsymbol{w}}$ is the mean similarity between pairs of tokens in sets S, T. We denote the set difference between S and T by $S \setminus T$, and we denote the first n whole numbers by $[n] = \{1, 2, \ldots, n\}$.

2.2.2 Generalizing Word Embedding Association Tests

We assume that there is a given set of possible targets $\mathcal{X}$ and attributes $\mathcal{A}$. Henceforth, since in our evaluation all targets are names and all attributes are lower-case words (or phrases), we refer to targets as names and attributes as words. Nonetheless, in principle, the algorithm can be run on any sets of target and attribute tokens. [89] define a WEAT statistic for two equal-sized groups of names $X_1, X_2 \subseteq \mathcal{X}$ and words $A_1, A_2 \subseteq \mathcal{A}$ which can be conveniently written in our notation as,

$$s(X_1, A_1, X_2, A_2) \overset{\text{def}}{=} \left(\sum_{x \in X_1} \overline{\boldsymbol{x}} - \sum_{x \in X_2} \overline{\boldsymbol{x}} \right) \cdot (\overline{\boldsymbol{A}}_1 - \overline{\boldsymbol{A}}_2).$$

In studies of human biases, the combinatorial explosion in groups can be avoided by teasing apart *Single-Category* IATs which assess associations one group at a time [e.g. 98, 99, 91]. In word embeddings, we define a simple generalization for $n \geq 1$, nonempty groups $X_1, \ldots, X_n$ of arbitrary sizes and words $A_1, \ldots, A_n$, as follows:

$$g(X_1, A_1, \ldots, X_n, A_n) \overset{\text{def}}{=} \sum_{i=1}^{n} (\overline{\boldsymbol{X}}_i - \boldsymbol{\mu}) \cdot (\overline{\boldsymbol{A}}_i - \overline{\boldsymbol{\mathcal{A}}})$$

$$\text{where } \boldsymbol{\mu} \overset{\text{def}}{=} \begin{cases} \overline{\boldsymbol{\mathcal{X}}} & \text{for } n = 1, \\ \sum_i \overline{\boldsymbol{X}}_i / n & \text{for } n \geq 2. \end{cases}$$

Note that g is symmetric with respect to ordering and weights groups equally regardless of size. The definition differs for $n = 1$, otherwise $g \equiv 0$.

The following three properties motivate this as a "natural" generalization of WEAT to one or more groups.

Lemma 1. *For any $X_1, X_2 \subseteq \mathcal{X}$ of equal sizes $|X_1| = |X_2|$ and any nonempty $A_1, A_2 \subseteq \mathcal{A}$,*

$$s(X_1, A_1, X_2, A_2) = 2|X_1|\ g(X_1, A_1, X_2, A_2)$$

Lemma 2. *For any nonempty sets $X \subset \mathcal{X}$, $A \subset \mathcal{A}$, let their complements sets $X^c = \mathcal{X} \setminus X$ and $A^c = \mathcal{A} \setminus A$. Then,*

$$g(X, A) = 2g(X, A, \mathcal{X}, \mathcal{A}) = 2\frac{|X^c|}{|\mathcal{X}|}\frac{|A^c|}{|\mathcal{A}|} g(X, A, X^c, A^c)$$

Lemma 3. *For any $n > 1$ and nonempty $X_1, X_2, \ldots, X_n \subseteq \mathcal{X}$ and $A_1, A_2, \ldots, A_n \subseteq \overline{\boldsymbol{\mathcal{A}}}$,*

$$g(X_1, A_1, \ldots, X_n, A_n) = \sum_{i \in [n]} g(X_i, A_i) - \sum_{i,j \in [n]} \frac{g(X_i, A_j)}{n}$$

Lemma 1 explains why we call it a generalization: for $n = 2$ and equal-sized name sets, the values are proportional with a factor that only depends on the set size. More generally, g can accommodate unequal set sizes and $n \neq 2$.

Lemma 2 shows that for $n = 1$ group, the definition is proportional the WEAT with the two groups X vs. all names $\mathcal{X}$ and words A vs. $\mathcal{A}$. Equivalently, it is proportional to the WEAT between X and A and their compliments.

Finally, Lemma 3 gives a *decomposition* of a WEAT into n^2 single-group WEATs $g(X_i, A_j)$. In particular, the value of a single multi-group WEAT reflects a combination of the n association strengths between X_i and A_i, and n^2 disassociation strengths between X_i and A_j. As discussed on the literature on IATs, a large effect could reflect a strong association between X_1 and A_1 or X_2 and A_2, a strong disassociation between X_1 and A_2 or X_2 and A_1, or some combination of these factors. Proofs are deferred to Appendix A.12.

name	meaning	default
WE	word embedding	`w2v`
$\mathcal{X}$	set of names	SSA
n	number of target groups	12
m	number of categories	64
M	number of frequent lower-case words	30,000
t	number of words per WEAT	3
α	false discovery rate	0.05

Table 2.4: Inputs to the UBE algorithm.

2.2.3 Unsupervised Bias Enumeration algorithm

The inputs to our UBE algorithm are shown in Table 2.4. The output is m WEATs, each with n groups with associated sets of words and statistical confidences (p-values) in $[0, 1]$. Each WEAT has words from a single category, but several of the m WEATs may yield no significant associations.

At a high level, the algorithm follows a simple structure. It selects n disjoint groups of names $X_1, \ldots, X_n \subset \mathcal{X}$, and m disjoint categories of lower-case words $\mathcal{A}_1, \ldots, \mathcal{A}_m$. All WEATs share the same n name groups, and each WEAT has words from a single category $\mathcal{A}_j$, with t words associated to each X_i. Thus the WEATs can be conveniently visualized in a tabular structure.

For convenience, we normalize all word embedding vectors to be unit length. Note that we only compute cosines between them, and the cosine is simply the inner product for unit vectors. We now detail the algorithm's steps.

Step 1: Cleaning names and defining groups

We begin with a set of names[4] $\mathcal{X}$, e.g., frequent first names from a database. Since word embeddings do not differentiate between words that have the same spelling but different meanings, we first "clean" the given names to remove names such as "May" and "Virginia", whose embeddings are more reflective of other uses, such as a month or verb and a US state. Our cleaning procedure, detailed in Appendix C, is similar to that of [89].

We then use K-means++ clustering [from scikit-learn, 100, with default parameters] to cluster the normalized word vectors of the names, yielding groups $X_1 \cup \ldots \cup X_n = \mathcal{X}$. Finally, we define $\mu = \sum_i \overline{\boldsymbol{X}}_i / n$.

[4] While the set of names is an input to our system, they could also be extracted from the embedding itself.

Step 2: Defining word categories

To define categories, we cluster the most frequent M lower-case tokens in the word embedding into m clusters using K-means++, yielding clusters of categories $\mathcal{A}_1, \ldots, \mathcal{A}_m$. The constant M is chosen to cover as many recognizable words as possible without introducing too many unrecognizable tokens. As we shall see, categories capture concepts such as occupations, food-related words, and so forth.

Step 3: Selecting words $A_{ij} \subset \mathcal{A}_j$

A test $\mathcal{T}_j = (X_1, A_{1j}, \ldots, X_n, A_{nj})$ is chosen with disjoint $A_{ij} \subset \mathcal{A}_j$, each of size $t = |A_{ij}|$. To ensure disjointness,[5] $\mathcal{A}_j$ is first partitioned into n "Voronoi" sets $V_{ij} \subseteq \mathcal{A}_j$ consisting of the words whose embedding is closest to each corresponding center $\overline{\boldsymbol{X}}_i$, i.e.,

$$V_{ij} = \left\{ w \in \mathcal{A}_j \mid i = \arg\max_{i' \in [n]} \overline{\boldsymbol{w}} \cdot \overline{\boldsymbol{X}}_{i'} \right\}$$

It then outputs A_{ij} defined as the t words maximizing the following:

$$\max_{w \in V_{ij}} (\overline{\boldsymbol{X}}_i - \mu) \cdot (\overline{\boldsymbol{w}} - \overline{\boldsymbol{\mathcal{A}}}_j)$$

The more computationally-demanding step is to compute, using Monte Carlo sampling, the n p-values for $\mathcal{T}_j$, as described next.

Step 4: Computing p-values and ordering

To test whether the associations we find are larger than one would find if there was no relationship between the names X_i and words $\mathcal{A}$, we consider the following "**rotational null hypothesis**": the words in the embedding are generated through some process in which the alignment between names and words is random. This is formalized by imagining that a random rotation was applied (multiplying by a uniformly Haar random orthogonal matrix U) to the word embeddings but not to the name embeddings.

Specifically, to compute p-value p_{ij} for each (X_i, A_{ij}), we first compute a score $\sigma_{ij} = (\overline{\boldsymbol{X}}_i - \mu) \cdot (\overline{\boldsymbol{A}}_{ij} - \overline{\boldsymbol{\mathcal{A}}})$. We then compute $R = 10,000$ uniformly random orthogonal rotations $U_1, \ldots, U_R \in \mathbb{R}^{d \times d}$, drawn according to the Haar measure. For each rotation, we simulate running our algorithm as if the name embeddings were transformed by U (while the word embeddings remain as is). For each rotation U_r, the sets A_{ijr} chosen to maximize $(\overline{\boldsymbol{X}}_i U_r - \mu U_r) \cdot (\overline{\boldsymbol{w}} - \overline{\boldsymbol{\mathcal{A}}}_j)$, and the corresponding V_{ijr} and the resulting σ_{ijr} are computed. Finally, p_{ij} is the fraction of rotations for which the score $\sigma_{ijr} \geq \sigma_{ij}$ (plus an add-1 penalty standard for Monte Carlo p-values).

[5] If multiplicities are desired, the Voronoi sets V_{ij} could be omitted, optimizing $A_{ij} \subset \mathcal{A}_j$ directly.

Furthermore, since the algorithm outputs many (hundreds) of name/word biases, the Benjamini-Hochberg [1995] procedure is used to determine a critical p-value that guarantees an α bound on the rate of false discoveries. Finally, to choose an output ordering on significant tests, the m tests are then sorted by the total scores σ_{ij} over the pairs determined significant.

2.2.4 Evaluation

To illustrate the performance of the proposed system in discovering associations, we use a database of first names provided by the Social Security Administration (SSA), which contains number of births per year by sex (F/M) [102]. Preprocessing details are in Appendix C.

We use three publicly available word embeddings, each with $d = 300$ dimensions and millions of words: `w2v`, released in 2013 and trained on approximately 100 billion words from Google News [103], `fast`, trained on 600 billion words from the Web [79], and `glove`, also trained on the Web using the GloVe algorithm [104].

While it is possible to display the three words in each A_{ij}, the hundreds or thousands of names in each X_i cannot be displayed in the output of the algorithm. Instead, we use a simple greedy heuristic to give five "illustrative" names for each group, which are displayed in the tables in this Section and in our crowdsourcing experiments. The $k + 1^{\text{st}}$ name shown is chosen, given the first k names, so as to maximize the average similarity of the first $k + 1$ names to that of the entire set X_i. Hence, the first name is the one whose normalized vector is most central (closest to the cluster mean), the second name is the one which when averaged with the first is as central as possible, and so forth.

The WEATs can be evaluated in terms of the quality of the name groups and also their associations with words. A priori, it was not clear whether clustering name embeddings would yield any name groups or word categories of interest. For all three embeddings we find that the clustering captures latent groups defined in terms of race, age, and gender (we only have binary gender statistics), as illustrated in Table 2.5 for $n = 12$ clusters. While even a few clusters suffice to capture some demographic differences, more clusters yield much more fine-grained distinctions. For example, with $n = 12$ one cluster is of evidently Israeli names (see column I of table 2.5), which one might not consider predefining a priori since they are a small minority in the U.S. Table A.1 in the Appendix shows demographic composition of clustering for other embeddings. Note that, although we do not have religious statistics for the names, several of the words in the generated associations are religious in nature, suggesting religious biases as well.

Table A.2 in the appendix shows the biases found in the "debiased" `w2v` embedding of [88]. While the name clusters still exhibit strong binary gender differences, many fewer statistically significant associations were generated for the most gender-polarized clusters.

w2v F1	**w2v F2**	**w2v F3**	**w2v F4**	**w2v F5**	**w2v F6**	**w2v F7**	**w2v F8**	**w2v F9**	**w2v F10**	**w2v F11**	**w2v F12**
Amanda	Janice	Marquisha	Mia	Kayla	Kamal	Daniela	Miguel	Yael	Randall	Dashaun	Keith
Renee	Jeanette	Latisha	Keva	Carsyn	Nailah	Lucien	Deisy	Moses	Dashiell	Jamell	Gabe
Lynnea	Lenna	Tyrique	Hillary	Aislynn	Kya	Marko	Violeta	Michal	Randell	Marlon	Alfred
Zoe	Mattie	Marygrace	Penelope	Cj	Maryam	Emelie	Emilio	Shai	Jordan	Davonta	Shane
Erika	Marylynn	Takiyah	Savanna	Kaylei	Rohan	Antonia	Yareli	Yehudis	Chace	Demetrius	Stan
+581	+840	+692	+558	+890	+312	+391	+577	+120	+432	+393	+494
98% F	98% F	89% F	85% F	78% F	65% F	59% F	56% F	40% F	27% F	5% F	4% F
1983	1968	1978	1982	1993	1991	1985	1986	1989	1981	1984	1976
4% B	8% B	48% B	10% B	2% B	7% B	4% B	2% B	5% B	10% B	32% B	6% B
4% H	4% H	3% H	9% H	1% H	4% H	9% H	70% H	10% H	3% H	5% H	3% H
3% A	3% A	1% A	11% A	1% A	32% A	4% A	8% A	5% A	4% A	3% A	5% A
89% W	84% W	47% W	69% W	95% W	56% W	83% W	21% W	79% W	83% W	59% W	86% W

Table 2.5: Illustrative first names (greedily chosen) for $n = 12$ groups on the `w2v` embedding. Demographic statistics (computed a posteriori) are also shown though were not used in generation, including percentage female (at birth), mean year of birth, and percentage Black, Hispanic, Asian/Pacific Islander, and White.

Emb.	# significant	% accurate	% offensive
`w2v`	235	72%	35%
`fast`	160	80%	38%
`glove`	442	48%	24%

Table 2.6: Summary statistics for the WEATs generated using the three embeddings ($n = 12$, $m = 64$). The total number of significant name/word associations, the fraction with which the crowd's choice of name group agreed with that of the generated WEAT (accuracy) among the top-12 WEATs, and the fraction rated as offensive.

w2v F1	w2v F2	w2v F3	w2v F4	w2v F5	w2v F6	w2v F7	w2v F8	w2v F9	w2v F11	w2v F12
	cookbook, baking, baked goods	sweet potatoes, macaroni, green beans			saffron, halal, sweets	mozzarella, foie gras, caviar	tortillas, salsa, tequila	kosher, hummus, bagel	fried chicken, crawfish, grams	beef, beer, hams
herself, hers, moms	husband, homebound, grandkids	aunt, niece, grandmother	hubby, socialite, cuddle	twin sister, girls, classmate	elder brother, dowry, refugee camp			bereaved, immigrated, emigrated	younger brother, twin brother, mentally r********	buddy, boyhood, fatherhood
hostess, cheerleader, dietitian	registered nurse, homemaker, chairwoman		supermodel, beauty queen, stripper	helper, getter, snowboarder	shopkeeper, villager, cricketer		translator, interpreter, smuggler		cab driver, jailer, schoolboy	pitchman, retired, pundit
	log cabin, library, fairgrounds	front porch, carport, duplex	racecourse, plush, tenements	picnic tables, bleachers, concession stand	locality, mosque, slum	prefecture, chalet, sauna		synagogues, constructions, hilltop	apartment complex, barbershop, nightclub	
	parish, church, pastoral	pastor, baptized, mourners	goddess, celestial, mystical		fatwa, mosques, martyrs	monastery, papal, convent	rosary, parish priest, patron saint	rabbis, synagogue, biblical		
volleyball, gymnast, setter	athletic director, winningest coach, officiating	leading rebounder, played sparingly, incoming freshman	hooker, footy, stud	sophomore, junior, freshman	leftarm spinner, dayers, leg spinner				cornerback, tailback, wide receiver	
sorority, gymnastics, majoring	volunteer, volunteering, secretarial	guidance counselor, prekindergarten, graduate		seventh grader, eighth grade, seniors	lecturers, institutes, syllabus		bilingual, permanent residency, occupations		incoming freshmen, schoolyard, recruiting	fulltime, professional, apprenticeship
		civil rights, poverty stricken, nonviolent			subcontinent, tribesmen, miscreants	xenophobia, anarchist, oligarchs	leftist, drug traffickers, undocumented	disengagement, intifada, settlers	blacks, segregation, lynching	
tiara, blonde, sparkly	knitting, sewing, beaded	brown eyes, cream colo..., wore	girly, feminine, flirty	brown hair, pair, skates	sari, turban, hijab				dreadlocks, shoulderpads, waistband	mullet, gear, helmet
					dirhams, lakhs, rupees	rubles, kronor, roulette	pesos, remittances, gross receipts	shekels, settlements, corpus		
		grandjury indicted, degree murder, violating probation		child endangerment, vehicular homicide, unlawful possession	chargesheet, absconding, interrogation	absentia, tax evasion, falsification	illegal immigrant, drug trafficking, deported		aggravated robbery, aggravated assault, felonious assault	
	volunteers, crafters, baby boomers	caseworkers, evacuees, attendants	beauties, celebs, paparazzi	setters, helpers, captains	mediapersons, office bearers, newsmen				recruits, reps, sheriffs	

Table 2.7: The top-12 WEATs output by our UBE algorithm on the `w2v` embedding. Columns represent name groups X_i from Table 2.5, rows represent categories A_j (e.g., a cluster of food-related words). Orange indicate associations where the crowd's most commonly chosen name group agrees with that of the generated WEAT. No significant biases generated for **w2v F10**.

Crowdsourcing evaluation

We solicited ratings on the biases generated by the algorithm from US-based crowd workers on Amazon's Mechanical Turk[6] platform. The aim is to identify whether the biases found by our UBE algorithm are consistent with (problematic) biases held by society at large. To this end, we asked about society's stereotypes, *not* personal beliefs.

We evaluated the top 12 WEATs generated by our UBE algorithm for the three embeddings, considering $n = 12$ first name groups. Our approach was simple: after familiarizing participants with the 12 groups, we showed the (statistically significant) words and name groups of a WEAT and asked them to identify which words would, stereotypically, be most associated with which names group. A bonus was given for ratings that agreed with most other worker's ratings, incentivizing workers to provide answers that they felt corresponded to widely held stereotypes.

This design was chosen over a simpler one in which WEATs are shown to individuals who are asked whether or not these are stereotypical. The latter design might support confirmation bias as people may interpret words in such a way that confirms whatever stereotypes they are being asked about. For instance, someone may be able to justify associating the color red with almost any group, a posteriori.

Note that the task presented to the workers involved fine-grained distinctions: for each of the top-12 WEATs, at least 18 workers would each be asked to match the significant $c \leq 12$ word triples to the c name groups (each identified by five names each). For example, workers faced the triple of "registered nurse, homemaker, chairwoman" with $c = 8$ groups of names, half of which were majority female, and the most commonly chosen group matched the one generated: "Janice, Jeanette, Lenna, Mattie, Marylynn." Across the top-12 WEATs over the three embeddings, the mean number of choices c was 8.1, yet the most commonly chosen group (plurality) agreed with the generated group 65% of the time (see Table 2.6). This is significantly more than one would expect from chance. The top-12 WEATs generated for `w2v` are shown in Table 2.7.

One challenge faced in this process was that, in pilot experiments, a significant fraction of the workers were not familiar with many of the names. To address this challenge, we first administered a qualification exam (common in crowdsourcing) in which each worker was shown 36 random names, 3 from each group, and was offered a bonus for each name they could correctly identify the group from which it was chosen. Only workers whose accuracy was greater than 1/2 (which happened 37% of the time) evaluated the WEATs. Accuracy greater than 50% on a 12-way classification indicates that the groups of names were meaningful and interpretable to many workers.

Finally, we asked 13-15 workers to rate associations on a scale of 1-7 of *political incorrectness*, with 7 being "politically incorrect, possibly very offensive" and 1 being "politically correct, inoffensive, or just random." Only those biases for which the most commonly cho-

[6] http://mturk.com

sen group matched the association identified by the UBE algorithm were included in this experiment. The mean ratings are shown in Table 2.6 and the terms present in associations deemed most offensive are presented in Table 2.3.

2.2.5 Potential indirect biases and proxies

Naively, one may think that removing names from a dataset will remove all problematic associations. However, as suggested by [88], indirect biases are likely to remain. For example, consider the `w2v` word embedding, in which *hostess* is closer to *volleyball* than to *cornerback*, while *cab driver* is closer to *cornerback* than to *volleyball*. These associations, taken from columns **F1** and **F11** of Table 2.7, might serve as a proxy for gender and/or race. For instance, if someone is applying for a job and their profile includes college sports words, such associations encoded in the embedding may lead to racial or gender biases in cases in which there is no professional basis for these associations. In contrast, *volunteer* being closer to *volunteers* than *recruits* may represent a definitional similarity more than a proxy, if we consider proxies to be associations that mainly have predictive power due to their correlation with a protected attribute. While defining proxies is beyond the scope of this work, we do say that $A_{ij}, A_{i'j}, A_{ij'}, A_{i'j'}$ is a *potential indirect bias* if,

$$(\overline{\boldsymbol{A}}_{ij} - \overline{\boldsymbol{A}}_{i'j}) \cdot (\overline{\boldsymbol{A}}_{ij'} - \overline{\boldsymbol{A}}_{i'j'}) > 0. \tag{2.23}$$

One way to interpret this definition is that if the embedding were to match the pair of word sets $\{A_{ij}, A_{i'j}\}$ to the pair of word sets $\{A_{ij'}, A_{i'j'}\}$, it would align with the way in which they were generated. For example, does the embedding predict that *hostess-cab driver* better fits *volleyball-cornerback* or *cornerback-volleyball* (but this question is asked with sets of $t = 3$ words)? Downstream, this would mean that a replacing a the word *cornerback* with *volleyball* on a profile would make it closer to *hostess* than *cab driver*

We consider all possible fourtuples of significant associations, such that $1 \leq i < i' \leq n$ and $1 \leq j < j' \leq m$. In the case of `w2v`, 99% of 2,713 significant fourtuples lead to potential indirect biases according to eq. (2.23). This statistic is of 98% of 1,125 fourtuples and 97% of 1,796 fourtuples for the `fast` and `glove` embeddings, respectively. Hence, while names allow us to capture biases in the embedding, removing names is unlikely to be sufficient to debias the embedding.

2.2.6 Limitations

Absent clusters show the limitations of our approach and data. For example, even for large n, no clusters represent demographically significant Asian-American groups. However, if instead of names we use surnames [U.S. Census, 105], a cluster "Yu, Tamashiro, Heng, Feng, Nakamura, +393" emerges, which is largely Asian according to Census data (see Table A.3 in the Appendix). This distinction may reflect naming practices among Asian Americans [106]. Similarly, our approach may miss biases against small minorities or other groups whose names are not significantly differentiated. For example, it is not immediately clear to what extent this methodology can capture biases against individuals whose gender identity is non-binary, although interestingly terms associated with transgender individuals were generated and rated as significant and consistent with human biases.

2.2.7 Discussion

This Section introduces the problem of Unsupervised Bias Enumeration (UBE). It proposes a UBE algorithm that outputs Word Embedding Association Tests, and evaluates it via crowdsourcing. Unlike humans, where implicit tests are necessary to elicit socially unacceptable biases in a straightforward fashion, word embeddings can be directly probed to output hundreds of biases of varying natures, including numerous offensive and socially unacceptable biases. The racist and sexist associations exposed in publicly available word embeddings raise questions about their widespread use. An important open question is how to reduce these biases. In the next Section we take one step towards answering this question.

2.3 What's in a name? Reducing bias without assuming access to protected attributes

Section based on:
A. Romanov, M. De-Arteaga, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, A. Rumshisky, A. Kalai. What's in a Name? Reducing Bias in Bios Without Assuming Access to Protected Attributes, In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

When the performance of a machine learning system differs substantially for different groups of people, a number of concerns arise [54, 58]. First and foremost, there is a risk that the deployment of such a method may harm already marginalized groups and widen existing inequalities. Chapter 2.1 highlights this concern in the context of online recruiting and automated hiring. Recall that the results show that when predicting an individual's occupation from their online biography, if occupation-specific gender gaps in true positive rates are correlated with existing gender imbalances in those occupations, then those imbalances will be compounded over time—a phenomenon sometimes referred to as the "leaky pipeline." Second, the correlations that lead to performance differences between groups are often irrelevant. For example, while an occupation classifier should predict a higher probability of software engineer if an individual's biography mentions coding experience, there is no good reason for it to predict a lower probability of software engineer if the biography also mentions softball.

Prompted by such concerns about bias in machine learning systems, there is a growing body of work on fairness in machine learning. Some of the foundational papers in this area highlighted the limitations of trying to mitigate bias using methods that are "unaware" of protected attributes such as race, gender, or age [e.g., 93]. As a result, subsequent work has primarily focused on introducing fairness constraints, defined in terms of protected attributes, that reduce incentives to rely on undesirable correlations [e.g., 66, 107]. This approach is particularly useful if similar performance can be achieved by slightly different means—i.e., fairness constraints may aid in model selection if there are many near-optima.

In practice, though, any approach that relies on protected attributes may stand at odds with anti-discrimination law, which limits the use of protected attributes in domains such as employment and education, even for the purpose of mitigating bias. And, in other

"What's in a name? That which we call a rose by any other name would smell as sweet." – *William Shakespeare, Romeo and Juliet.*

domains, protected attributes are often not available [108]. Moreover, even when they are, it is usually desirable to simultaneously consider multiple protected attributes, as well as their intersections. For example, Buolamwini [109] showed that commercial gender classifiers have higher error rates for women with darker skin tones than for either women or people with darker skin tones overall.

We propose a method for reducing bias in machine learning classifiers without relying on protected attributes. In the context of occupation classification, this method discourages a classifier from learning a correlation between the predicted probability of an individual's occupation and a word embedding of their name. Intuitively, the probability of an individual's occupation should not depend on their name—nor on any protected attributes that may be inferred from it. We present two variations of the method—i.e., two loss functions that enforce this constraint—and show that they simultaneously reduce both race and gender biases with little reduction in classifier accuracy. Although we are motivated by the need to mitigate bias in online recruiting and automated hiring, this method can be applied in any domain where individuals' names are available at training time.

Instead of relying on protected attributes, the proposed method leverages the societal biases that are encoded in word embeddings [45, 46]. In particular, we build on the work presented on Section 2.2, which shows that word embeddings of names typically reflect the societal biases that are associated with those names, including race, gender, and age biases, as well encoding information about other factors that influence naming practices such as nationality and religion. By using word embeddings of names as a tool for mitigating bias, the approach is conceptually simple and empirically powerful. Much like the "proxy fairness" approach of Gupta et al. [110], it is applicable when protected attributes are not available; however, it additionally eliminates the need to specify which biases are to be mitigated, and allows simultaneous mitigation of multiple biases, including those that relate to group intersections. Moreover, under the proposed approach it is only necessary to have access to proxy information (i.e., names) at training time and not at deployment time, which avoids disparate treatment concerns and extends fairness gains to individuals with ambiguous names. For example, a method that explicitly or implicitly infers protected attributes from names at deployment time may fail to correctly infer that an individual named Alex is female and, in turn, fail to mitigate gender bias for her. Methodologically, our work is also similar to that of Zafar et al. [111], which promotes fairness by requiring that the covariance between a protected attribute and a data point's distance from a classifier's decision boundary is smaller than some constant. However, unlike our method, it requires access to protected attributes, and does not facilitate simultaneous mitigation of multiple biases.

2.3.1 Method

The proposed methodology discourages an occupation classifier from learning a correlation between the predicted probability of an individual's occupation and a word embedding of their name. This section presents two variations of the method—i.e., two penalties that can be added to an arbitrary loss function and used when training any classifier.

We assume that each data point corresponds to an individual, with a label indicating that individual's occupation. We also assume access to the names of the individuals represented in the training set. The first variation, which we call Cluster Constrained Loss (CluCL), uses k-means to cluster word embeddings of the names in the training set. Then, for each pair of clusters, it minimizes between-cluster disparities in the predicted probabilities of the true labels for the data points that correspond to the names in the clusters. In contrast, the second variation minimizes the covariance between the predicted probability of an individual's occupation and a word embedding of their name. Because this variation minimizes the covariance directly, we call it Covariance Constrained Loss (CoCL). The most salient difference between these variations is that CluCL only minimizes disparities between the latent groups captured by the clusters. For example, if the clusters correspond only to gender, then CluCL is only capable of mitigating gender bias. However, given a sufficiently large number of clusters, CluCL is able to simultaneously mitigate multiple biases, including those that relate to group intersections. For both variations, individual's names are not used as input to the classifier itself; they appear only in the loss function used when training the classifier. The resulting trained classifier can therefore be deployed without access to individuals' names.

Formulation

We define $x_i = \{x_i^1, \ldots, x_i^M\}$ to be a data point, y_i to be its corresponding (true) label, and n_i^f and n_i^l to be the first and last name of the corresponding individual. The classification task is then to (correctly) predict the label for each data point:

$$p_i = H(x_i) \tag{2.24}$$

$$\hat{y}_i = \underset{1 \leq j \leq |C|}{\arg\max}\, p_i[j], \tag{2.25}$$

where $H(\cdot)$ is the classifier, C is the set of possible classes, $p_i \in \mathbb{R}^{|C|}$ is the output of the classifier for data point x_i—e.g., $p_i[j]$ is the predicted probability of x_i belonging to class j—and $\hat{y}_i$ is the predicted label for x_i. We define p_i^y to be the predicted probability of y_i—i.e., the true label for x_i.

The conventional way to train such a classifier is to minimize some loss function $\mathcal{L}$, such as the cross-entropy loss function. We propose to add an additional penalty to this loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L} + \lambda \cdot \mathcal{L}_{\text{CL}}, \tag{2.26}$$

where $\mathcal{L}_{\text{CL}}$ is either $\mathcal{L}_{\text{CluCL}}$ or $\mathcal{L}_{\text{CoCL}}$ (defined in Sections 2.3.1 and 2.3.1, respectively), and λ is a hyperparameter that determines the strength of the penalty. This loss function is only used during training, and plays no role in the resulting trained classifier. Moreover, it can be used in any standard setup for training a classifier—e.g., training a deep neural network using mini-batches and the Adam optimization algorithm [112].

Cluster Constrained Loss

This variation represents each first name n_i^f and last name n_i^l as a pair of low-dimensional vectors using a set of pretrained word embeddings E. These are then combined to form a single vector:

$$n_i^e = \frac{1}{2}\left(E[n_i^f] + E[n_i^l]\right). \tag{2.27}$$

Using k-means [113], CluCL then clusters the resulting embeddings into k clusters, yielding a cluster assignment k_i for each name (and corresponding data point). Next, for each class $c \in C$, CluCL computes the following average pairwise difference between clusters:

$$\begin{aligned} l_c = {} & \frac{1}{k(k-1)} \times \\ & \sum_{u,v=1}^{k} \left(\frac{1}{N_{c,u}} \sum_{\substack{i:y_i=c,\\ k_i=u}} p_i^y - \frac{1}{N_{c,v}} \sum_{\substack{i:y_i=c,\\ k_i=v}} p_i^y \right)^2, \end{aligned} \tag{2.28}$$

where u and v are clusters and $N_{c,u}$ is the number of data points in cluster u for which $y_i = c$. CluCL considers each class individually because different classes will likely have different numbers of training data points and different disparities. Finally, CluCL computes the average of $l_1, \dots l_{|C|}$ to yield

$$\mathcal{L}_{\text{CluCL}} = \frac{1}{|C|} \sum_{c \in C} l_c. \tag{2.29}$$

Covariance Constrained Loss

This variation minimizes the covariance between the predicted probability of a data point's label and the corresponding individual's name. Like CluCL, CoCL represents each name as a single vector n_i^e and considers each class individually:

$$l_c = \mathbb{E}_{i:y_i=c}\left[\left(p_i^y - \mu_p^c\right) \cdot \left(n_i^e - \mu_n^c\right)\right], \tag{2.30}$$

where $\mu_p^c = \mathbb{E}_{i:y_i=c}\,[p_i^y]$ and $\mu_n^c = \mathbb{E}_{i:y_i=c}\,[n_i^e]$. Finally, CoCL computes the following average:

$$\mathcal{L}_{\text{CoCL}} = \frac{1}{|C|} \sum_{c \in C} \|l_c\|,$$

where $\|\cdot\|$ is the ℓ_2 norm.

2.3.2 Evaluation

One of the strengths of the proposed method is its ability to simultaneously mitigate multiple biases without access to protected attributes; however, this strength also poses a challenge for evaluation. We are unable to quantify this ability without access to these attributes. To facilitate evaluation, we focus on race and gender biases only because race and gender attributes are more readily available than attributes corresponding to other biases. We further conceptualize both race and gender to be binary ("white/non-white" and "male/female") but note that these conceptualizations are unrealistic, reductive simplifications that fail to capture many aspects of race and gender, and erase anyone who does not fit within their assumptions. We emphasize that we use race and gender attributes only for evaluation—they do not play a role in our method.

Datasets

We use two datasets to evaluate the proposed method: the adult income dataset from the UCI Machine Learning Repository [114], where the task is to predict whether an individual earns more than $50k per year (i.e., whether their occupation is "high status"), and the dataset of online biographies described in Section 2.1, where the task is to predict an individual's occupation from the text of their online biography.

Each data point in the *Adult* dataset consists of a set of binary, categorical, and continuous attributes, including race and gender. We preprocess these attributes to more easily allow us to understand the classifier's decisions. Specifically, we normalize continuous attributes to be in the range $[0, 1]$ and we convert categorical attributes into binary indicator variables. Because the data points do not have names associated with them, we generate synthetic first names using the race and gender attributes. First, we use the dataset of Tzioumis [115] to identify "white" and "non-white" names. For each name, if the proportion of "white" people with that name is higher than 0.5, we deem the name to be "white;" otherwise, we deem it to be "non-white."[7] Next, we use Social Security Administration data about baby names [2018] to identify "male" and "female" names. For each name, if the proportion of boys with that name is higher than 0.5, we deem the name to be "male;" otherwise, we deem it to be "female."[8] We then take the intersection of these two sets of names to yield a single set of names that is partitioned into four non-overlapping categories by (binary) race and gender. Finally, we generate a synthetic first name for each data point by sampling a name from the relevant category.

Each data point in the *Bios* dataset consists of the text of an individual's biography, written in the third person. We represent each biography as a vector of length V, where

[7] For 90% of the names, the proportion of "white" people with that name is greater than 0.7 or less than 0.3, so there is a clear distinction between "white" and "non-white" names.

[8] For 98% of the names, the proportion of boys with that name is greater than 0.7 or less than 0.3, so there is an even clearer distinction between "male" and "female" names.

V is the size of the vocabulary. Each element corresponds to a single word type and is equal to 1 if the biography contains that type (and 0 otherwise). We limit the size of the vocabulary by discarding the 10% most common word types, as well as any word types that occur fewer than twenty times. Unlike the *Adult* dataset, each data point has a name associated with it. And, because biographies are typically written in the third person and because pronouns are gendered in English, we can extract (likely) self-identified gender. We infer race for each data point by sampling from a Bernoulli distribution with probability equal to the average of the probability that an individual with that first name is "white" (from the dataset of Tzioumis [115], using a threshold of 0.5, as described above) and the probability that an individual with that last name is "white" (from the dataset of Comenetz [105], also using a threshold of 0.5).[9] Finally, as in Section 2.1, we consider two versions of the *Bios* dataset: one where first names and pronouns are available to the classifier and one where they are "scrubbed."

Throughout the evaluation, we use the fastText word embeddings, pretrained on Common Crawl data [116], to represent names.

Classifier and loss function

Our method can be used with any classifier, including deep neural networks such as recurrent neural networks and convolutional neural networks. However, because the focus of this work is mitigating bias, not maximizing classifier accuracy, we use a single-layer neural network:

$$h_i = W_h \cdot x_i + b_h$$
$$p_i = \text{softmax}(h_i)$$

where $W_h \in \mathbb{R}^{|C| \times M}$ and $b_h \in \mathbb{R}^{|C|}$ are the weights. This structure allows us to examine individual elements of the matrix W_h in order to understand the classifier's decisions for any dataset.

Both the *Adult* dataset and the *Bios* dataset have a strong class imbalance. We therefore use a weighted cross-entropy loss as $\mathcal{L}$, with weights set to the values proposed by King and Zeng [117].

Quantifying bias

To quantify race bias and gender bias, we follow the same approach used in Section 2.1 and compute the true positive rate (TPR) race gap and the TPR gender gap—i.e., the differences in the TPRs between races and between genders, respectively—for each occupation.

[9] We note that, in general, an individual's race or gender should be directly reported by the individual in question; inferring race or gender can be both inaccurate and reductive.

The TPR race gap for occupation c is defined as follows:

$$\mathrm{TPR}_{r,c} = P\left[\hat{Y} = c \,|\, R = r, Y = c\right] \tag{2.31}$$

$$\mathrm{Gap}_{r,c} = \mathrm{TPR}_{r,c} - \mathrm{TPR}_{\sim r,c}, \tag{2.32}$$

where r and $\sim r$ are binary races, $\hat{Y}$ and Y are random variables representing the predicted and true occupations for an individual, and R is a random variable representing that individual's race. Similarly, the TPR gender gap for occupation c is

$$\mathrm{TPR}_{g,c} = P\left[\hat{Y} = c \,|\, G = g, Y = c\right] \tag{2.33}$$

$$\mathrm{Gap}_{g,c} = \mathrm{TPR}_{g,c} - \mathrm{TPR}_{\sim g,c}, \tag{2.34}$$

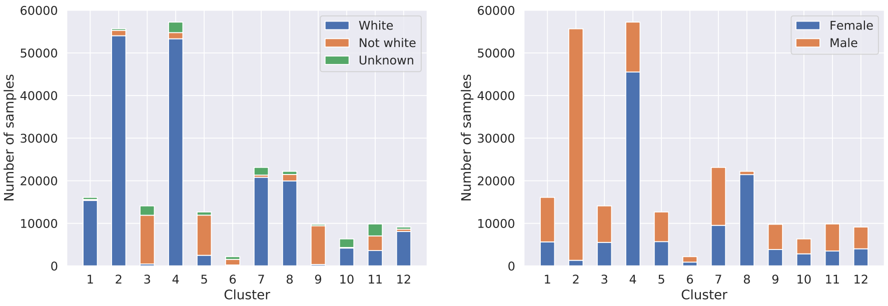
where g and $\sim g$ are binary genders and G is a random variable representing an individual's gender.

To obtain a single score that quantifies race bias, thus facilitating comparisons, we calculate the root mean square of the per-occupation TPR race gaps:

$$\mathrm{Gap}_r^{\mathrm{RMS}} = \sqrt{\frac{1}{|C|}\sum_{c \in C} \mathrm{Gap}_{r,c}^2}. \tag{2.35}$$

We obtain a single score that quantifies gender bias similarly. The motivation for using the root mean square instead of an average is that larger values have a larger effect and we are more interested in mitigating larger biases. Finally, to facilitate worst-case analyses, we calculate the maximum TPR race gap and the maximum TPR gender gap.

We again emphasize that race and gender attributes are used only for evaluating our method.



(a) Race. (b) Gender.

Figure 2.9: Number of data points (from the *Bios* dataset) in each cluster that correspond to each race and gender.

2.3.3 Results

We first demonstrate that word embeddings of names encode information about race and gender. We then present the main results, before examining individual elements of the matrix W_h in order to better understand our method.

Word embeddings of names as proxies

We cluster the names associated with the data points in the *Bios* dataset, represented as word embeddings, to verify that such embeddings indeed capture information about race and gender. We perform k-means clustering (using the k-means++ algorithm) with $k = 12$ clusters, and then plot the number of data points in each cluster that correspond to each (inferred) race and gender. Figures 2.9a and 2.9b depict these numbers, respectively.

Clusters 1, 2, 4, 7, 8, and 12 contain mostly "white" names, while clusters 3, 5, and 9 contain mostly "non-white names." Similarly, clusters 4 and 8 contain mostly "female" names, while cluster 2 contains mostly "male" names. The other clusters are more balanced by race and gender. Manual inspection of the clusters reveals that cluster 9 contains mostly Asian names, while cluster 8 indeed contains mostly "female" names. The names in cluster 2 are mostly "white" and "male," while the names in cluster 4 are mostly "white" and "female." This suggests that the clusters are capturing at least some intersections. Together these results demonstrate that word embeddings of names do indeed encode at least some information about race and gender, even when first and last names are combined into a single embedding vector.

Adult **dataset**

Method	λ	Balanced TPR	$\text{Gap}_g^{\text{RMS}}$	$\text{Gap}_r^{\text{RMS}}$	$\text{Gap}_g^{\max}$	$\text{Gap}_r^{\max}$
None	0	**0.795**	0.299	0.120	0.303	0.148
CluCL	1	0.788	0.278	0.121	0.297	0.145
CluCL	2	0.793	0.259	0.085	0.282	0.114
CoCL	1	0.794	0.215	0.091	0.251	0.119
CoCL	2	0.790	**0.163**	**0.080**	**0.201**	**0.109**

Table 2.8: Results for the *Adult* dataset. Balanced TPR (i.e., per-occupation TPR, averaged over occupations), gender bias quantified as $\text{Gap}_g^{\text{RMS}}$, race bias quantified as $\text{Gap}_r^{\text{RMS}}$, maximum TPR gender gap, and maximum TPR race gap for different values of hyperparameter λ. Results are averaged over four runs with different random seeds.

The results for the *Adult* dataset are shown in Table 2.8. The task is to predict whether an individual earns more than \$50k per year (i.e., whether their occupation is "high status"). Because the dataset has a strong class imbalance, we report the balanced TPR—i.e.,

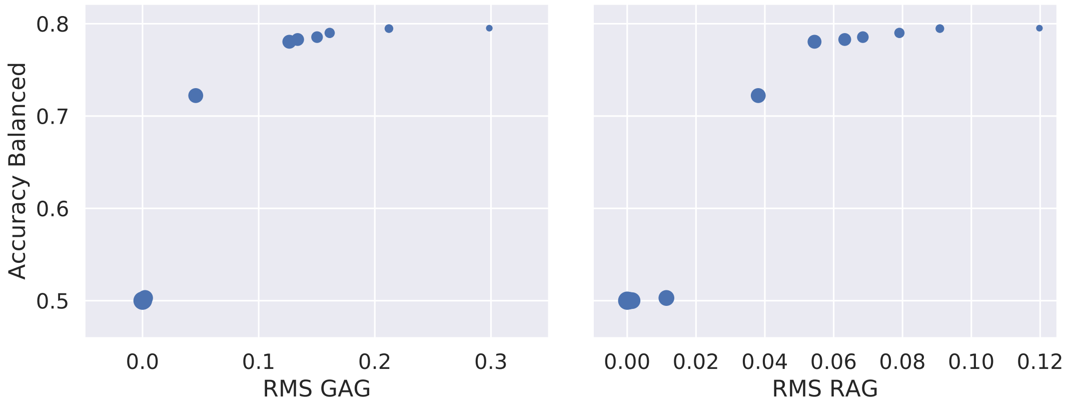


Figure 2.10: Gender bias quantified as $\mathrm{Gap}_g^{\mathrm{RMS}}$ (left) and race bias quantified as $\mathrm{Gap}_r^{\mathrm{RMS}}$ (right) versus balanced TPR for the CoCL variation of our method with different values of hyperparameter λ (a larger dot means a larger value of λ) for the *Adult* dataset. Results are averaged over four runs with different random seeds.

we compute the per-class TPR and then average over the classes. We experiment with different values of the hyperparameter λ. When $\lambda = 0$, the method is equivalent to using the conventional weighted cross-entropy loss function. Larger values of λ increase the strength of the penalty, but may lead to a less accurate classifier. Using $\lambda = 0$ leads to significant gender bias: the maximum TPR gender gap is 0.303. This means that the TPR is 30% higher for men than for women. We emphasize that this does *not* mean that the classifier is more likely to predict that a man earns more than $50k per year, but means that the classifier is more likely to *correctly* predict that a man earns more than $50k per year. Both variations of our method significantly reduce race and gender biases. With CluCL, the root mean square TPR race gap is reduced from 0.12 to 0.085, while the root mean square TPR gender gap is reduced from 0.299 to 0.25. These reductions in bias result in less than one percent decrease in the balanced TPR (79.5% is decreased to 79.3%). With CoCL, the race and gender biases are further reduced: the root mean square TPR race gap is reduced to 0.08, while the root mean square TPR gender gap is reduced to 0.163, with 0.5% decrease in the balanced TPR.

We emphasize that although the proposed method significantly reduces race and gender biases, neither variation can completely eliminate them. In order to understand how different values of hyperparameter λ influence the reduction in race and gender biases, we perform additional experiments using CoCL where we vary λ from 0 to 10. Figure 2.10 depicts these results. Larger values of λ indeed reduce race and gender biases; however, to achieve a root mean square TPR gender gap of zero means reducing the balanced TPR to 50%, which is unacceptably low. That said, there are a wide range of values of λ that sig-

Method	λ	Balanced TPR	$\text{Gap}_g^{\text{RMS}}$	$\text{Gap}_r^{\text{RMS}}$	$\text{Gap}_g^{\max}$	$\text{Gap}_r^{\max}$
None	0	**0.788**	0.173	0.051	0.511	0.121
CluCL	1	0.784	0.168	0.048	0.494	0.120
CluCL	2	0.781	**0.165**	**0.047**	**0.486**	0.114
CoCL	1	0.785	0.168	0.048	0.507	**0.109**
CoCL	2	0.779	0.169	0.048	0.512	0.116

Table 2.9: Results for the original *Bios* dataset. Balanced TPR (i.e., per-occupation TPR, averaged over occupations), gender bias quantified as $\text{Gap}_g^{\text{RMS}}$, race bias quantified as $\text{Gap}_r^{\text{RMS}}$, maximum TPR gender gap, and maximum TPR race gap for different values of hyperparameter λ. Results are averaged over four runs with different random seeds.

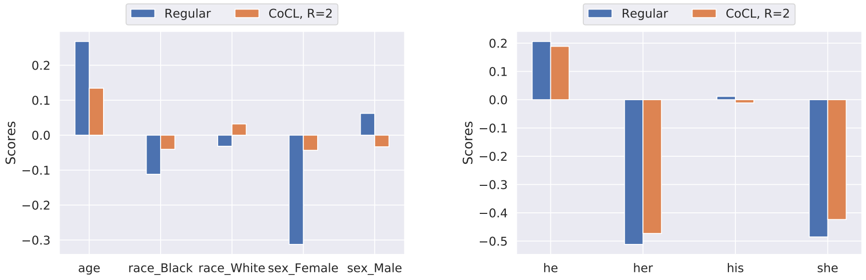
Method	λ	Balanced TPR	$\text{Gap}_g^{\text{RMS}}$	$\text{Gap}_r^{\text{RMS}}$	$\text{Gap}_g^{\max}$	$\text{Gap}_r^{\max}$
None	0	**0.785**	0.111	0.049	0.385	0.123
CluCL	1	0.782	**0.107**	0.048	**0.383**	0.112
CluCL	2	0.778	0.112	0.046	0.395	**0.107**
CoCL	1	0.780	0.109	0.047	0.388	0.117
CoCL	2	0.775	0.108	**0.046**	0.387	0.109

Table 2.10: Results for the "scrubbed" *Bios* dataset. Balanced TPR (i.e., per-occupation TPR, averaged over occupations), gender bias quantified as $\text{Gap}_g^{\text{RMS}}$, race bias quantified as $\text{Gap}_r^{\text{RMS}}$, maximum TPR gender gap, and maximum TPR race gap for different values of hyperparameter λ. Again, results are averaged over four runs.

nificantly reduce race and gender biases, while maintaining an acceptable balanced TPR. For example, $\lambda = 6$ results in a root mean square TPR race gap of 0.038 and a root mean square TPR gender gap of 0.046, with only a 7.3% decrease in the balanced TPR.

Bios dataset

The results of the evaluation using the original and "scrubbed" (i.e., names and pronouns are "scrubbed") versions of the *Bios* dataset are shown in Tables 2.9 and 2.10, respectively. The task is to predict an individual's occupation from the text of their online biography. Because the dataset has a strong class imbalance, we again report the balanced TPR. CluCL and CoCL reduce race and gender biases for both versions of the dataset. For the original version, CluCL reduces the root mean square TPR gender gap from 0.173 to 0.165 and the maximum TPR gender gap by 2.5%. Race bias is also reduced, though to a lesser extent. These reductions reduce the balanced TPR by 0.7%. For the "scrubbed" version, the reductions in race and gender biases are even smaller, likely because most of the



(a) *Adult* dataset. (b) *Bios* dataset, occupation "surgeon."

Figure 2.11: Classifier weight values for several attributes for the conventional weighted cross-entropy loss function (i.e., $\lambda = 0$) and for CoCL with $\lambda = 2$. Results are averaged over four runs with different random seeds.

information about race and gender has been removed by "scrubbing" names and pronouns. We hypothesize that these smaller reductions in race and gender biases, compared to the *Adult* dataset, are because the *Adult* dataset has fewer attributes and classes than the *Bios* dataset, and contains explicit race and gender information, making the task of reducing biases much simpler. We also note that each biography in the *Bios* dataset is represented as a vector of length V, where V is over 11,000. This means that the corresponding classifier has a very large number of weights, and there is a strong overfitting effect. Because this overfitting effect increases with λ, we suspect it explains why CluCL has a larger root mean square TPR gender gap when $\lambda = 2$ than when $\lambda = 1$. Indeed, the root mean square TPR gender gap for the training set is 0.05 when $\lambda = 2$. Using dropout and ℓ_2 weight regularization lessened this effect, but did not eliminate it entirely.

Understanding the method

The proposed method mitigates bias by making training-time adjustments to the classifier's weights that minimize the correlation between the predicted probability of an individual's occupation and a word embedding of their name. Because of our choice of classifier (a single-layer neural network, as described in Section 2.3.2), we can examine individual elements of the matrix W_h to understand the effect of our method on the classifier's decisions. Figure 2.11a depicts the values of several weights for the conventional weighted cross-entropy loss function (i.e., $\lambda = 0$) and for CoCL with $\lambda = 2$ for the *Adult* dataset. When $\lambda = 0$, the attributes "sex_Female" and "sex_Male" have large negative and positive weights, respectively. This means that the classifier is more likely to predict that a man earns more than \$50k per year. With CoCL, these weights are much closer to zero. Similarly, the weights for the race attributes are also closer to zero. We note that the weight for the attribute "age" is also reduced, suggesting that CoCL may have mitigated some form of age bias.

Figure 2.11b depicts the values of several weights specific to the occupation "surgeon" for the conventional weighted cross-entropy loss function (i.e., $\lambda = 0$) and for CoCL with $\lambda = 2$ for the original version of the *Bios* dataset. When $\lambda = 0$, the attributes "she" and "her" have large negative weights, while the attribute "he" has a positive weight. This means that the classifier is less likely to predict that a biography that contains the words "she" or "her" belongs to a surgeon. With CoCL, these magnitudes of these weights are reduced, though these reductions are not as significant as the reductions shown for the *Adult* dataset.

2.3.4 Discussion

This Section proposes a method for reducing bias in machine learning classifiers without relying on protected attributes. In contrast to previous work, this method eliminates the need to specify which biases are to be mitigated, and allows simultaneous mitigation of multiple biases, including those that relate to group intersections. The proposed methodology leverages the societal biases that are encoded in word embeddings of names. Specifically, it discourages an occupation classifier from learning a correlation between the predicted probability of an individual's occupation and a word embedding of their name. Two variations of the method are presented and evaluated using a large-scale dataset of online biographies. Results show that both variations simultaneously reduce race and gender biases, with almost no reduction in the classifier's overall true positive rate. The method is conceptually simple and empirically powerful, and can be used with any classifier, including deep neural networks.

Chapter 3

Limits of available labels and leveraging human consistency

3.1 Introduction

In many domains, humans are routinely tasked with making predictions to inform decisions their job requires them to make. Examples are judges who predict the likelihood of recidivism when determining bail, doctors who predict the likelihood of neurological recovery of comatose patients when deciding whether to extend life support, and recruiters who evaluate the likelihood of a candidate succeeding at a job when hiring. Increasingly, machine learning is being used to aid humans in those predictions. While research has shown that machine learning and actuarial models are better at making predictions than humans [118, 119, 120], the available data frequently presents challenges that limit what can be learned from observed outcomes alone, undermining a model's performance and leading to deceivingly optimistic evaluation metrics.

In this work, we propose using human consistency as a source of information for what cannot be learned from observational data alone. Human consistency has long been used as a source of information. Popularized by Cohen's Kappa Coefficient [121, 122], inter-rater agreement is considered an indicator of reliability [123, 124]. In order to leverage such consistency, we first propose a way of estimating it in cases where a single decision-maker observes each sample. While existing methodology to estimate consistency requires that multiple humans label each case, historical data of high-stakes decisions made by domain experts (e.g. judges, physicians and social workers) usually contains a single human's assessment for each case. The proposed method to identify consistency in these settings works as follows: A predictive model f_h is used to predict the human decisions, and influence functions are used to estimate each expert's influence on a prediction. This yields a metric of robustness for the model's predictions of human decisions by identifying whether a prediction is driven by the historical decisions of multiple experts, or by those of a single or very few experts. For cases in which the model's predictions indicate high certainty–measured in terms of calibrated probability–, this metric allows us to identify *consistency* across decision makers.

Furthermore, we propose a *label amalgamation* strategy to incorporate human knowledge into a model. In applications for which consistency is believed to stem from expertise, the approach allows us to obtain labels for censored cases, and to learn from experts in instances in which they are consistent and at odds with the observed label, while learning from observational data elsewhere. Our approach brings the construct we optimize for closer to the construct that humans care about, without incorporating individual biases, errors, or noise.

Section 3.2 we present related work, in Section 3.3 we describe challenges of algorithmic decision support. In Section 3.4 we introduce the methodology to estimate consistency via influence functions and the label amalgamation approach. In Section 3.5 we present and analyze the results, both on semi-synthetic data in which we consider different scenarios of decision making, and on real data from child maltreatment hotline screenings.

3.2 Related work

Our work draws inspiration from extensive literature that uses inter-rater agreement metrics as an indicator of reliability [121, 122, 123, 124]. Such metrics have been popular in applied psychology literature for decades, and have recently been popularized in computer science through the crowdsourcing literature [125, 126]. With the emergence of an online workforce as an inexpensive source of data labeling, metrics of agreement have been very useful to aggregate and assess the quality of crowd-sourced labels. Unlike in crowdsourcing, in this work we aim to learn from domain experts making high-stakes decisions. Obtaining labels is time consuming in such a setting– data is often sensitive, and qualified labelers are scarce– so we cannot collect multiple assessments for each case and must find ways to leverage the historical decisions available. The proposed approach allows us to estimate consistency across experts from historical data. Here, a note must be made regarding consistency. Experts' consistency–or the lack thereof–has been a subject of study for a long time [127, 128], with results indicating that experts tend to exhibit low overall consistency. We highlight that we are not assuming experts will display overall consistency, but rather that we are leveraging experts' consistency when it is displayed for subsets of cases.

Most closely related to our work is [129, 2], where modeling of human decisions is used to improve evaluation of predictive models in the presence of the selective labels problem and unobservables. While [129, 2] improve *validation* by leveraging *heterogeneity* of human decisions, our focus is instead on improving *training* by leveraging *homogeneity.* Their work proposes a way of evaluating a model meant to assist bail decisions, trained on observed outcomes of recidivism, by making use of the fact that there are more and less lenient judges. This allows the authors to compare the performance of detaining everyone a strict judge would detain, vs. detaining everyone a lenient judge would detain, and matching the detention rate of the strict judge based on who the algorithm predicts to be at highest risk of recidivism. The authors focus are those cases for which humans disagree in their assessment. Instead, we focus on cases for which all humans agree in their assessment, by proposing methodology to identify such cases and incorporating this knowledge into the training of a predictive model.

In terms of the discussion of the risks of the selective labels problem, our work differs from [129, 2] in that by focusing on the portion of cases where there is agreement, we concentrate on the violation of the positivity assumption. Meanwhile, they focus on the simultaneous presence of both selective labels and unobservables, but do not consider the violation of the positivity assumption that stems from homogeneity across human decisions. The selective labels problem is a special case of sample selection bias, which concerns learning in a setting where training and test data are drawn from different distributions [130, 131]. Statistics and quantitative methods literature on missing data has also addressed this problem [132, 133]. However, in addition to assuming conditional ignorability, which fails in the presence of unobservables, a common assumption to the different approaches that have been presented to tackle sample selection bias is positivity, which assumes that every individual has a non-zero probability of being part of the training sample, i.e., $P(d_i = 1|\boldsymbol{x}_i) > 0 \; \forall i = 1,..,n$, where d_i refers to $\boldsymbol{x}_i$ being selected for the sample. As we will discuss, this assumption is easily violated under the selective labels problem. The fairness-related risks of learning from censored data are explored in [134].

The risks of omitted payoff bias are briefly described in [129, 135]; we add to this work by proposing ways of mitigating this problem. Also related, the concepts of an observed space and a construct space are formalized in [136]. While their focus is on features rather than outcomes, the notion that what we observe does not always capture what we care about is at the heart of omitted payoff bias.

Bringing machine learning models closer to experts' knowledge has been explored in the past. In particular, researchers have proposed ways of doing so by prioritizing features that are more credible [137]. Our work also shares similarities with the literature on learning to defer [138, 139], which seeks to combine human and algorithmic decision making. However, existing techniques in this realm rely on the algorithm's ability to self-assess its performance and confidence, and are therefore not directly applicable. We do note, however, that a framework for learning to defer using the criteria presented in this research would be a plausible complement to the proposed methodology.

Finally, a core piece of related work is the literature on influence functions. The local influence method enables the estimation of the influence of minor perturbations of a model over a certain functional [140]. This fundamental work in the field of robust statistics has been widely applied in the literature of semi-parametric and nonparametric estimation [141], and causal inference [142, 143, 144]. It has also been used in machine learning to derive estimators for information theoretic quantities [145] and as a way to explain black-box predictions and generate adversarial attacks [146]. To the best of our knowledge, ours is the first work that proposes the use of influence functions to estimate consistency amongst decision-makers.

3.3 Challenges of algorithmic decision support

In the remainder of this Chapter we assume the data available for learning has the form (X, D, Y), where X corresponds to a set of available covariates, D is an observed human decision that attempts to predict a construct Y^c that is not easily or directly observable, and Y is the observed outcome that proxies for Y^c and is used to train a model. We assume that Y is only observed for one of the values of D, and $D = 1$ whenever humans predict $Y^c = 1$. The diagram in Figure 3.1 illustrates the assumed decision and data generation process. In some instances, humans may have access to an additional set of covariates Z that are unobserved in the data and cannot be used for training.

For example, in the child welfare context, X are available covariates of historical information of the children and adults involved in a call, D is a call-worker's decision to screen-in a call for investigation, Y^c is whether the child is at risk, and Y is whether the investigation leads to out-of-home placement of the child (foster care), which is observed when $D = 1$. Meanwhile, in the bail context X is historical information of a defendant, Y^c corresponds to the risk of societal harm, D is the decision to detain, and Y is rearrest, observed when $D = 0$. Note that while in the child welfare context a human prediction of high-risk leads to an investigation that allows us to observe Y, in the bail context a human prediction of high-risk leads to a detention that does not allow us to observe Y. Our goal is to obtain a model that ***as accurately as possible predicts*** $\mathbf{Y^c}$***.*** Below, we describe the selective labels problem and omitted payoff bias in more detail.

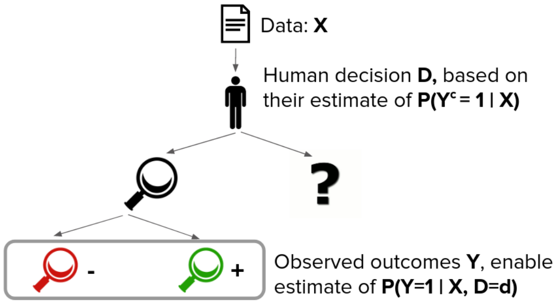
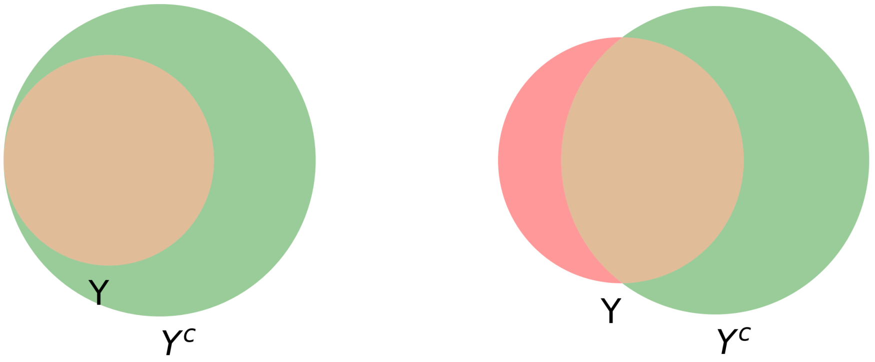

Data: **X**
Human decision **D,** based on their estimate of **P(Y^c = 1 | X)**
?
-
+
Observed outcomes **Y**, enable estimate of **P(Y=1 | X, D=d)**

Figure 3.1: Diagram illustrating experts' decisions and the data generation process of the observed outcomes. The question mark illustrates that observed labels are censored by the human decision.

Selective labels problem Human decisions D often determine whether Y is observed. In this setting, if machine learning algorithms are trained using the observed outcomes, the resulting models are not answering the question "given an individual $\boldsymbol{x}_i$, is situation Y likely to occur?", but rather, "given an individual $\boldsymbol{x}_i$ for whom a human predicts that situation Y is likely to occur, is situation Y indeed likely to occur?". Thus, rather than estimating the probability $P(Y = 1|X)$, the learning algorithms estimate $P(Y|X, D = a)$, where a denotes the decision under which we observe Y. Strategies such as inverse probability weighting are frequently used in such settings to correct for sampling bias in the data. However, in many decision support contexts, such corrective strategies are not available for two reasons. (1) Humans may be making use of unobservables Z that are predictive of Y [129], meaning that $Y \not\perp D \mid X$; and (2) humans may display consistency that violates the positivity assumption, which assumes a non-negligible probability of observing the outcome for all x.

For example, in the task of predicting neurological recovery of comatose patients, one can think of building a model using data from a hospital where the human decision makers are the best in their field, for use in medical centers that lack such expertise. This model might be constructed to predict the likelihood of neurological recovery. However, the outcome would be censored in cases where the physicians withdrew life support on the basis of their assessment that a "good outcome" is unlikely. In such cases it would be desirable to incorporate the certainty of those human decisions in the model. The proposed approach aims to do exactly this.

Omitted payoff bias This type of bias can have different origins. First, there can be unobserved treatment effects, as the human decision may constitute a form of intervention. Without knowing the effect of the intervention, there is a missing component of a known payoff function. For example, a social worker's visit to a home may itself reduce the risk a child is exposed to. Second, there can be mismeasured outcomes linked to issues of construct validity. Often, the humans' objective accounts for factors that are not observed, in which case the prediction loss function is misaligned with the true payoff function, which depends on components not being captured by the loss. This is particularly common in public policy settings where the objective depends on social welfare, which is frequently challenging to estimate from observed outcomes alone. For example, in the child welfare context the goal is to screen-in for investigation all cases that involve a child at risk. However, not all types of risk lead to out-of-home placement, which is the objective optimized for in current deployment setups [147]. Solely optimizing for out-of-home placement may fail to screen in cases in which services are offered as a result of the investigation and the well-being of the child and the family improves. Examples of possible relationships between Y and Y^c under omitted payoff bias as shown in Figure 3.2. By incorporating experts' knowledge into a predictive model, we aim to attenuate the effects of omitted payoff bias.



(a) Y is a proxy that has perfect precision but not perfect recall of Y^c.

(b) Y is a proxy that neither has perfect precision nor perfect recall of Y^c.

Figure 3.2: Examples of possible relationships between Y and Y^c under omitted payoff bias.

3.4 Methodology

3.4.1 Expert consistency estimation via influence functions

There are two main challenges that make the use of human consistency in decision support systems non-trivial. First, it is often the case that a single human expert assesses each sample, and therefore agreeability cannot be measured directly in available data with traditional inter-rater agreeability metrics. Second, many times there is a non-random assignment of experts to cases, and therefore predictive models of human decisions may be able to predict decisions with high confidence, but such confidence does not necessarily imply consistency across humans. We propose to use influence functions to estimate experts' influence over a prediction of human decisions, which serves as a metric of robustness of a given prediction when we shift the importance of individual experts.

Influence of a single decision-maker

Let $f_h(\boldsymbol{x}) = \hat{P}(D = 1|\boldsymbol{x})$ be a predictive model of the human decisions. Influence functions allow us to estimate the effect on an individual prediction $f_h(\boldsymbol{x})$ of performing a small perturbation of our training data by shifting it ϵ in a direction $\boldsymbol{w}$, where $\boldsymbol{w}$ refers to the weight given to the training points. Given a decision-maker h, let $\boldsymbol{w}_h \in \mathbb{R}^m$ be defined as:

$$w_{h_i} \quad = \quad \begin{cases} 1+\varepsilon & \text{for} \quad h_i == h \\ 1 & \text{for} \quad h_i \neq h \end{cases} , \tag{3.1}$$

where w_{h_i} denotes to the ith entry of the weight vector $\boldsymbol{w}_h$, and h_i denotes the human that observed sample $\boldsymbol{x}_i$ and made decision d_i. Perturbations of the model in direction $\boldsymbol{w}_h$

correspond to assuming we up-weight the importance of decision-maker h. The influence function $\mathcal{I}_{up,f_h}(\boldsymbol{w}_h, \boldsymbol{x}_{\text{test}})$ estimates the influence on the predicted probability $f_h(\boldsymbol{x}_{\text{test}})$ of perturbing the training set by $\boldsymbol{w}_h$, and can be derived in analogous fashion to the way [146] derives the influence function on the loss of perturbing a single point. We can define the influence function in terms of ϵ, as specified in Equation 3.2, where the empirical risk minimizer is $\hat{\theta} := \text{argmin}_{\theta\in\Theta}\sum_{i=1}^{n} L(\boldsymbol{x}_i, d_i, \theta)$, and the empirical risk minimizer after the training data has been perturbed by $\boldsymbol{w}_h$ is $\hat{\theta}_{\boldsymbol{w}_h} := \text{argmin}_{\theta\in\Theta}\sum_{i=1}^{n} w_{h_i} L(\boldsymbol{x}_i, d_i, \theta)$, where $L(\boldsymbol{x}_i, d_i, \theta)$ is the loss function and $\theta \in \mathbb{R}^p$ are the model parameters.

$$\mathcal{I}_{up,f_h}(\boldsymbol{w}_h, \boldsymbol{x}_{\text{test}}) \quad := \left.\frac{\partial P(y_{test}|\boldsymbol{x}_{test},\hat{\theta}_{\boldsymbol{w}_h})}{\partial\epsilon}\right|_{\epsilon=0} = \nabla_\theta P(y_{test}|\boldsymbol{x}_{test},\hat{\theta}_{\boldsymbol{w}_h})^T \left.\frac{\partial\hat{\theta}_{\boldsymbol{w}_h}}{\partial\epsilon}\right|_{\epsilon=0} \tag{3.2}$$

$\frac{\partial\hat{\theta}_{\boldsymbol{w}_h}}{\partial\epsilon}$ can be expressed as $\hat{\theta}_{\boldsymbol{w}_h} = \text{argmin}_{\theta\in\Theta}R(\theta) + \frac{1}{n_h}\sum_{i\in J_h}\epsilon L(\boldsymbol{x}_i, d_i, \theta)$, where J_h is the set of cases observed by expert h, $n_h = |J_h|$, and $R(\theta)$ is the empirical risk $R(\theta) := \frac{1}{n}\sum_{i=1}^{n} L(\boldsymbol{x}_i, d_i, \theta)$. From the first order condition we obtain that:

$$0 = \nabla_\theta R(\hat{\theta}_{\boldsymbol{w}_h}) + \frac{1}{n_h}\sum_{i\in J_h}\epsilon\nabla_\theta L(\boldsymbol{x}_i, d_i, \hat{\theta}_{\boldsymbol{w}_h}) \tag{3.3}$$

As $\epsilon \to 0$, $\hat{\theta}_{\boldsymbol{w}_h} \to \hat{\theta}$, so the Taylor expansion centered around $\hat{\theta}$, defining $\boldsymbol{\Delta w}_h = \hat{\theta}_{\boldsymbol{w}_h} - \hat{\theta}$, yields:

$$0 = \nabla_\theta R(\hat{\theta}) + \frac{1}{n_h}\sum_{i\in J_h}\epsilon\nabla_\theta L(\boldsymbol{x}_i, d_i, \hat{\theta}) + [\nabla_\theta^2 R(\hat{\theta}) + \frac{1}{n_J}\sum_{i\in J_h}\epsilon\nabla_\theta^2 L(\boldsymbol{x}_i, d_i, \hat{\theta})]\boldsymbol{\Delta w}_h + \text{h.o.t.} \tag{3.4}$$

Solving for $\boldsymbol{\Delta w}_h$ and making use of the fact that $\hat{\theta}$ minimizes $R(\theta)$ hence $\nabla_\theta R(\hat{\theta}) = 0$, we obtain:

$$\boldsymbol{\Delta w}_h \approx -[\nabla_\theta^2 R(\hat{\theta}) + \frac{1}{n}\sum_{i\in J_h}\epsilon\nabla_\theta^2 L(\boldsymbol{x}_i, d_i, \hat{\theta})]^{-1}[\frac{1}{n_h}\sum_{i\in J_h}\epsilon\nabla_\theta L(\boldsymbol{x}_i, d_i, \hat{\theta})]. \tag{3.5}$$

Let $A = \nabla_\theta^2 R(\hat{\theta})$, $B = \frac{1}{n_h}\sum_{i\in J_h}\nabla_\theta^2 L(\boldsymbol{x}_i, d_i, \hat{\theta})$, $C = \frac{1}{n_h}\sum_{i\in J_h}\nabla_\theta L(\boldsymbol{x}_i, d_i, \hat{\theta})$. Then,

$$\begin{aligned}
\boldsymbol{\Delta w}_h &\approx & -[A+\epsilon B]^{-1}\epsilon C = -(I+\epsilon A^{-1}B)^{-1}A^{-1}\epsilon C = -[\textstyle\sum_{n=0}^{\infty}(-1)^n\epsilon^n(A^{-1}B)^n]A^{-1}\epsilon C\\
&= & -(I-\epsilon A^{-1}B)A^{-1}\epsilon C + \text{h.o.t.} = -\epsilon A^{-1}C + \text{h.o.t.}\\
\Rightarrow \boldsymbol{\Delta w}_h &\approx & -[\nabla_\theta^2 R(\hat{\theta})]^{-1}[\tfrac{1}{n_h}\textstyle\sum_{i\in J_h}\nabla_\theta L(\boldsymbol{x}_i, d_i, \hat{\theta})]\epsilon
\end{aligned} \tag{3.6}$$

Since $\boldsymbol{\Delta w}_h = \hat{\theta}_{\boldsymbol{w}_h} - \hat{\theta}$, and $\hat{\theta}$ does not depend on ϵ, we get that

$$\frac{\partial\hat{\theta}_{\boldsymbol{w}_h}}{\partial\epsilon} = \frac{\partial\boldsymbol{\Delta w}_h}{\partial\epsilon} = -[\nabla_\theta^2 R(\hat{\theta})]^{-1}[\frac{1}{n_h}\sum_{i\in J_h}\nabla_\theta L(\boldsymbol{x}_i, d_i, \hat{\theta})]. \tag{3.7}$$

Replacing this in Equation 3.2 yields

$$\mathcal{I}_{up,f_h}(\boldsymbol{w}_h, x_{\text{test}}) = -\nabla_\theta P(y_{test}|x_{test}, \hat{\theta})^T [\nabla^2_\theta R(\hat{\theta})]^{-1} [\frac{1}{n_h} \sum_{i \in J_h} \nabla_\theta L(\boldsymbol{x}_i, d_i, \hat{\theta})]. \tag{3.8}$$

Now, the influence is fully defined in terms of θ, instead of ϵ, and can be easily calculated. Note that the most computationally intensive component is the Hessian of the empirical risk, which is $O(np^2 + p^3)$, but approaches to compute it efficiently for complex models have been proposed [146].

Logistic regression influence function For our experiments we will use logistic regression models. Logistic regression models happen to perform comparably to more complex models on the problems we consider. The gradient of the predicted probability is $\nabla_\theta f_h(\boldsymbol{x}_i) = \nabla_\theta \sigma(\hat{\theta}^T \boldsymbol{x}_i) = \sigma(\hat{\theta}^T \boldsymbol{x}_i)(1 - \sigma(\hat{\theta}^T \boldsymbol{x}_i))\boldsymbol{x}_i$, where σ is the sigmoid function $\sigma(x) = \frac{1}{1+\exp(-x)}$. The hessian of the empirical risk can be written as $\nabla^2_\theta R(\hat{\theta}) = X \, diag_i[\frac{1}{n}\sigma(\theta^T \boldsymbol{x}_i)(1 - \sigma(\theta^T \boldsymbol{x}_i))]X^T$.

Estimating consistency

Once the influence of each expert is estimated, it can be determined whether a model's confidence in the prediction of human decisions is robust to perturbations over the weight given to experts. For a given data point $\boldsymbol{x}$, we can analyze the distribution over the influence functions $\mathcal{I}_{up,f_h}(\boldsymbol{w}_h, x_{\text{test}})$, $\forall h$. Considering the sorted influence of humans over a prediction, the following two metrics are useful to construct a notion of consistency.

Center of mass The first moment, or center of mass of influence, allows us to measure if the influence is spread across experts or if very few experts have a disproportionate influence. Let k be the number of experts, and $\boldsymbol{s}(\boldsymbol{x})$ be a sorted vector of absolute influence of each decision maker over $f_h(\boldsymbol{x})$, sorted in decreasing order, such that $\boldsymbol{s}(\boldsymbol{x}) = sort([|\mathcal{I}_{up,f_h}(\boldsymbol{w}_h, x)| \text{ for } h = 1, 2, ...k])$. The center of mass $m_1(\boldsymbol{x}, f_h) = \frac{\sum_i i \cdot \boldsymbol{s}_i(\boldsymbol{x})}{\sum_i \boldsymbol{s}_i(\boldsymbol{x})}$, where $\boldsymbol{s}_i(\boldsymbol{x})$ is the ith entry of $\boldsymbol{s}(\boldsymbol{x})$, indicates that the $\lfloor m_1 \rfloor$ experts with the most influence have as much influence as the rest, where $\lfloor \cdot \rfloor$ is the floor function.

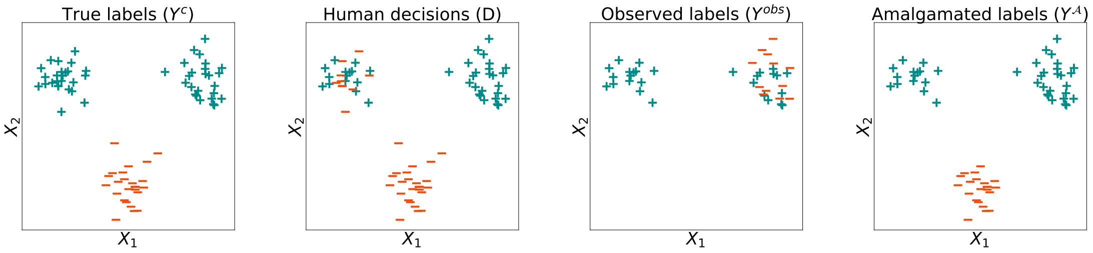
Aligned influence The center of mass allows us to capture the *concentration of influence*, but does not take into account the direction. The second metric $m_2(\boldsymbol{x}, f_h)$ shows if there are *opposing influences.* $m_2(\boldsymbol{x}, f_h) = \frac{\max\left(\sum_{i:\boldsymbol{s}_i(\boldsymbol{x})>0} \boldsymbol{s}_i(\boldsymbol{x}), \sum_{i:\boldsymbol{s}_i(\boldsymbol{x})<0} \boldsymbol{s}_i(\boldsymbol{x})\right)}{\sum_i \boldsymbol{s}_i(\boldsymbol{x})}$ indicates the portion of influence going in the direction of most influence. If all experts influence the prediction in the same direction, then $m_2 = 1$, while if the magnitudes of the influence in both the positive and negative directions are equal, then $m_2 = 0.5$.

Maximum influence This metric is equal to the maximum influence over a given prediction, $m_3(\boldsymbol{x}, f_h) = \max(\boldsymbol{s}(\boldsymbol{x}))$. In cases where the maximum influence is negligible it means that no matter how the weight given to experts is perturbed, the prediction would remain unchanged.

3.4.2 Label amalgamation

Figure 3.3 contains a toy example of the label amalgamation approach. In this setting, we assume that there are roughly three clusters of points, and humans are good at making decisions in two of those clusters, while being uncertain in one of them, as shown in Figure 3.3(b). The human decisions censor the data, meaning that the label Y is only observed for cases where humans predict the label to be $(+)$. Moreover, there is a mismatch between Y and Y^c as a result of omitted payoff bias.

Recall f_h denotes a predictive model of the human decisions, and let , $f_y(\boldsymbol{x}) = P(Y = 1|\boldsymbol{x}, D = a)$ be a predictive model of the observed outcome, where $a \in \{0, 1\}$ depends on the application domain and denotes the decision under which we observe Y. If linear models f_h and f_y are learned from the data, these would have the form $f_h = \alpha_1 X_1 + \alpha_2 X_2$, and $f_y = -\beta_1 X_1$, where $\alpha_1, \alpha_2, \beta_1 \in \mathbb{R}^+$. The goal of the label amalgamation is to learn from humans in the settings where they are consistent and from observed data elsewhere. As displayed in Figure 3.3(d), doing this enables us to recover the true relationship $f_{\mathcal{A}} = \zeta_2 X_2$, where $f_{\mathcal{A}}$ denotes the model learned via label amalgamation, and $\zeta \in \mathbb{R}^+$.



(a) (b) (c) (d)

Figure 3.3: Toy example illustrating how the label amalgamation works.(a) True labels Y^c. (b) Human decisions D. (c) Observed outcomes Y. (d) Amalgamated labels $Y^{\mathcal{A}}$. Selective labels problem prevents us from seeing the outcome whenever the human makes the decision (-), and omitted payoff bias leads to the observed outcome not always corresponding to the true label. Label amalgamation allows us to learn the correct relationship for the true labels.

We define the amalgamated label $Y^{\mathcal{A}}$ as the label that incorporates experts' knowledge in cases where the model of human decisions displays high confidence and high consistency.

Let $\mathcal{A}$ be the amalgamation set, one way of defining this set is:

$$\mathcal{A} = \{\boldsymbol{x}_i \in X : |f_h(\boldsymbol{x}_i) - D_i| < \delta, m_1(\boldsymbol{x}_i, f_h) > \gamma_1, m_2(\boldsymbol{x}_i, f_h) > \gamma_2\}, \tag{3.9}$$

Y	D	f_h	$m_1(\boldsymbol{x}_i, f_h) > \gamma_1$	$m_2(\boldsymbol{x}_i, f_h) > \gamma_2$	$Y^{\mathcal{A}}$
1	0	0.02	Yes	Yes	**0**
1	0	0.01	Yes	No	1
0	1	0.65	Yes	No	0
0	1	0.97	Yes	Yes	**1**
1	1	0.98	Yes	Yes	1

Table 3.1: Example of label amalgamation, for $\delta = 0.05$. Boxed rows highlight cases for which the amalgamated label is different to the observed label.

where f_h, m_1, m_2 are estimated via cross-validation, and $\delta, \gamma_1, \gamma_2$ are parameters. Importantly, f_h must correspond to a calibrated probability. The amalgamated label $Y^{\mathcal{A}}$ is then defined as:

$$Y_i^{\mathcal{A}} = \begin{cases} D_i & \textit{if} \quad \boldsymbol{x}_i \in \mathcal{A} \\ Y_i & \textit{if} \quad \boldsymbol{x}_i \notin \mathcal{A} \end{cases} \tag{3.10}$$

Note that when $Y^{\mathcal{A}} \neq Y$ it may be because Y was missing due to the selective labels problem, or because the observed label is counter to the decision humans consistently make, which indicates that there may be an omitted payoff bias. The new amalgamated label can be used to train a model that incorporates experts' knowledge. Table 3.1 shows an example of how label amalgamation works. We refer to the model trained to predict $Y^{\mathcal{A}}$ as $f_{\mathcal{A}}$.

In general, assume you have an amalgamation set $\mathcal{A}$ and any label amalgamation process that amalgamates labels for this set and learns from observed outcomes elsewhere. Label amalgamation will improve performance with respect to Y^c whenever $Y^{\mathcal{A}}$ is a better approximation to Y^c in this set than Y. More formally,

Theorem 2. *Given* $Y, Y^c, Y^{\mathcal{A}}$*, such that* $Y^{\mathcal{A}} = Y$ *if* $\neg\mathcal{A}$*, if* $P(Y^c = Y^{\mathcal{A}}|\mathcal{A}) \geq P(Y^c = Y|\mathcal{A})$ *then* $E(|Y^c - Y^{\mathcal{A}}||X) \leq E(|Y^c - Y||X)$.

Proof.

$$\begin{aligned} E(|Y^c - Y^{\mathcal{A}}||X) &= E(|Y^c - Y^{\mathcal{A}}|\mathbb{1}_{\mathcal{A}}|X) + E(|Y^c - Y^{\mathcal{A}}|\mathbb{1}_{\neg\mathcal{A}}|X) && (3.11) \\ &= E(|Y^c - Y^{\mathcal{A}}|\mathbb{1}_{\mathcal{A}}|X) + E(|Y^c - Y|\mathbb{1}_{\neg\mathcal{A}}|X) && (3.12) \\ &\leq E(|Y^c - Y|\mathbb{1}_{\mathcal{A}}|X) + E(|Y^c - Y|\mathbb{1}_{\neg\mathcal{A}}|X) && (3.13) \\ &= E(|Y^c - Y|\mathbb{1}_{\mathcal{A}}|X) + E(|Y^c - Y|\mathbb{1}_{\neg\mathcal{A}}|X) && (3.14) \\ &= E(|Y^c - Y||X) && (3.15) \end{aligned}$$

□

In particular, when $Y^{\mathcal{A}} = D$ in $\mathcal{A}$, if consistency is indicative of correctness and it is perfectly estimated, then $P(Y^c = D|\mathcal{A})$, which means that $E(|Y^c - Y^{\mathcal{A}}|\mathbb{1}_{\mathcal{A}}|X) = 0$ and the above result will always hold, with the strict inequality holding true if $\mathcal{A} \neq \emptyset$.

3.4.3 Robustness to model misspecification

Naturally, modeling choices will impact our ability to accurately predict human decisions. This means that if we assume a model that does not match the true functional form of the human decisions, then we may fail to identify cases in which humans are consistent. For example, imagine we are trying to predict whether doctors will prescribe a test or not. Assume the ground truth to be that all doctors prescribe a test whenever a patient has persistent cough *and* suffers from asthma. If we choose to model this decision with a decision tree of depth one, it will appear as if humans never agree, which is not true. This illustrates why the proposed methodology to estimate consistency has strict limitations regarding its ability to infer *lack of* agreement. Note that in this work we do not make any conclusions regarding disagreement. Instead, we identify consistency for subsets of cases, and make no determination regarding all other cases.

Therefore, the relevant question is: can model misspecification undermine our conclusions regarding consistency? In other words, for the subset of cases for which we infer humans are consistent, is this conclusion dependant on the model specification? The answer is no. If $\hat{P}$ is very high or very low, this allows us to infer information about the true probability, P.

Assume $\hat{P}$ is a calibrated probability, and define $H_{\mathcal{S}} = \{D_i \,|\, i \in \mathcal{S}, \hat{P}(D = 1|X_i) > 1-\epsilon\}$. If $\mathcal{S}$ corresponds to a set of datapoints not used during training of the algorithm that yields $\hat{\mathbb{P}}$, we can then obtain a confidence interval for the true probability P, as shown in Equation 3.16.

$$\mathsf{CI}(P(D = 1|X) > 1 - \epsilon; C) \sim \left(\overline{D}_{H_{\mathcal{S}}} - z^* \frac{\sigma}{\sqrt{n}}, \overline{D}_{H_{\mathcal{S}}} + z^* \frac{\sigma}{\sqrt{n}}\right), \tag{3.16}$$

where $\overline{D}_{H_{\mathcal{S}}}$ and σ are the mean and standard deviation of $H_{\mathcal{S}}$, respectively; $n = |H_{\mathcal{S}}|$; and $z^* = \Phi^{-1}(1 - \frac{\alpha}{2})$, for $\alpha = \frac{1-C}{2}$.

Note that if $\hat{P}$ is a calibrated probability and the distribution of the training set is representative of the distribution of $\mathcal{S}$, then $\overline{D}_{H_{\mathcal{S}}} = 1 - \epsilon$. Under this assumptions, if $H_{\mathcal{S}}$ is large enough, this will yield a tight confidence interval. For example, if $n = 1000$, $\mathsf{CI}(P(D = 1|X) > 1 - \epsilon; 99\%) \sim (1 - \epsilon - 0.018, 1 - \epsilon + 0.018)$. When the confidence interval is calculated empirically, no assumptions about the calibration of $\hat{P}$ or the distribution of $\mathcal{S}$ are required.

3.5 Applications

3.5.1 Learning under omitted payoff bias in prediction of child maltreatment risk

Semi-synthetic data generation

Before showing our results on real data, we construct semi-synthetic datasets to simulate the challenges described in Section 3.3 and illustrate how the proposed approach works under simple settings of decision making.

Let X be a subset of the features of the real-world child welfare data, where we remove those features that have low-variance ($\mathrm{Var}[X_i] < p(1-p)$, for $p = 0.9$), greedily remove those with strong pairwise correlations (Pearson correlation coefficient > 0.5), and introduce an intercept term. This yields a dataset $X \in \mathbb{R}^{46544\times 217}$ that is standardized to be centered with unit variance. Unless stated otherwise, we assume the samples-to-experts assignment is that of the real data, where the number of experts is $k = 32$. Semi-synthetic labels Y, Y^c, D are modeled as follows:

- **Y**: Let $\boldsymbol{\beta}^\circ$ be the learned coefficients of a logistic regression with L_1 penalty fitted to predict out-of-home placement in the observed data. We sample coefficients $\boldsymbol{\beta}$ such that $\beta_i \sim N(\beta_i^\circ, 1)$. The label Y is then sampled according to a logistic regression with coefficients $\boldsymbol{\beta}$, such that $Y \sim \mathrm{Binomial}(\frac{1}{1+\exp{(-X^T\boldsymbol{\beta})}})$.

- $\mathbf{Y^c}$: To model omitted payoff bias, we let $Y^c = Y \vee Y^{blind}$, where $\vee$ denotes the inclusive disjunction, such that $Y^c = 1$ if $Y = 1$ and/or $Y^{blind} = 1$. This corresponds to setting (b) in Figure 3.2. For simplicity, we let $Y^{blind} = \mathbb{1}(X_j \neq 1)$, where X_j is one of the covariates. In the experiments presented, $|X_j \neq 1| = 16,519$. Details can be found in Appendix **??**.

- **D**: Let $\boldsymbol{\beta}_d$ be the learned coefficients of a logistic regression with L_1 penalty fitted to predict Y^c. We assume each human h makes decisions according to a logistic regression model with coefficients $\boldsymbol{\beta}_h$, whose relationship to $\boldsymbol{\beta}_d$ is modified to simulate the different scenarios described below. $D_h = \mathbb{1}[X^T\boldsymbol{\beta}_h + \epsilon > 0]$, where $\epsilon \sim \mathrm{Logistic}(0, 0.5)$.

  I **Uninformative humans** For each expert h, resample all non-zero coefficients, such that $\boldsymbol{\beta}_{d_i} \sim \mathcal{U}(-1, 1)$. This breaks all relationships with Y^c and all relationships across humans.

  II **Oracle humans** Let $\boldsymbol{\beta}_h = \boldsymbol{\beta}_d$, $\forall h = 1, \ldots, 32$. This assumes all humans have access to the true model.

  III **Oracle humans and unobserved covariates** Let $\boldsymbol{\beta}_h = \boldsymbol{\beta}_d$, $\forall h = 1, \ldots, 32$. In addition, assume that the 5 covariates with the largest associated coefficients in $\boldsymbol{\beta}_d$ are unavailable to the machine (note that these will also be unavailable when modeling the human decisions f_h).

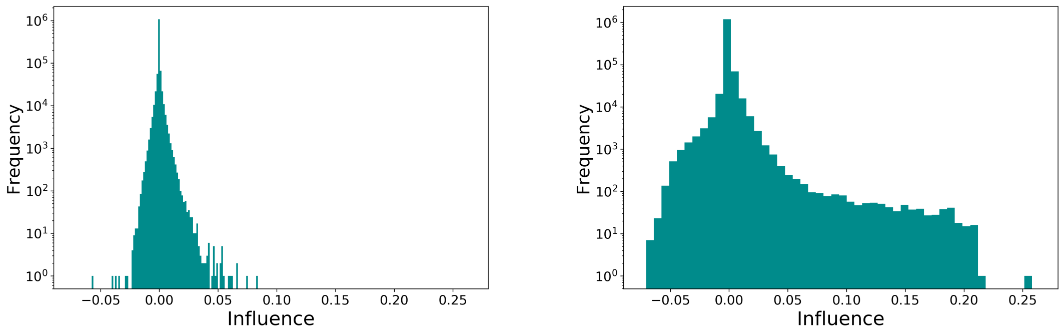
IV **Oracle humans, except for one biased human** Assume all samples-to-experts assignments remain the same, except for one (new) expert who sees all cases for which a binary feature $X_b = 1$, where $|X_b = 1| = 7,045$. Assume that this expert overestimates risk for the population X_b by assigning a coefficient $\beta_{h_b} = 2\max_j(\beta_{d_j})$.

Predictive models and label amalgamation In our experiments, we use 75%-25% train-test splits. Within the training set, we use 3-fold cross-validation to perform the label amalgamation and obtain amalgamated labels $Y^{\mathcal{A}}$ for the entire training set. We use an L_1 logistic regression to model the human decisions for each partition within cross-validation folds, where we tune the L_1 penalty parameter by incrementing it until the condition number of the Hessian of the empirical risk is low, indicating that the Hessian is well defined. This is necessary because calculating the influence function requires the inversion of this Hessian. We denote this model as $f_h^{\mathcal{A}}$ to differentiate it from f_h, a model trained on the entire training set that we will compare against $f_{\mathcal{A}}$. We ensure $f_h^{\mathcal{A}}$ yields a calibrated probability using Platt's scaling, and perform label amalgamation using parameters $\delta = 0.05$, $\gamma_1 = 4.0$, $\gamma_2 = 0.8$. The choice of γ_1, γ_2 is informed by the empirical distribution of consistency metrics when humans all use the same model and are therefore trivially consistent with each other, and discussed in detail in Appendix **??**.

We then train f_y, f_h and $f_{\mathcal{A}}$ as L_2 logistic regression models with target labels Y, D and $Y^{\mathcal{A}}$, respectively. While f_h and $f_{\mathcal{A}}$ are trained on the entire training set, f_y is only trained on the subset where $D = 1$, simulating the selective labels problem. Note that in general $f_h^{\mathcal{A}}$ and f_h may or may not be the same model, even though both predict the same target label. In the present setting, $f_h^{\mathcal{A}}$ is a more constrained model to enable the inversion of the Hessian.

Influence and consistency of decision-makers In setting I, there are no cases where $|f_h^{\mathcal{A}} - D_i| < \delta$, for $\delta = 0.05$. This is what we would expect; if each human is making use of a different model for making decisions, a logistic regression trained to predict all decisions will not have a good performance. As a result, no labels are augmented in this setting.

Meanwhile, comparing the influence of individual decision-makers for settings II and IV provides interesting insights. Intuitively, a good way to think about the influence of a decision maker over the predicted probability is: how sharply and in what direction would the predicted probability change if the importance given to decision-maker h was up-weighted by ϵ. The first stark difference that arises is reflected in the magnitude of the influences. As described in Section 3.5.1, we calculate the influence of each decision-maker over each prediction for all points in the training set, via cross-validation in this set. Figure 3.4 displays histograms of the influences for setting II and IV. When all decision-makers use the same model for making decisions, the influences are tightly concentrated around zero and no decision-maker has a very large influence over any prediction. Meanwhile, if one decision-maker overestimates risk for a group, the influence of an individual over one of the model's prediction may be much larger.



(a) Setting II.

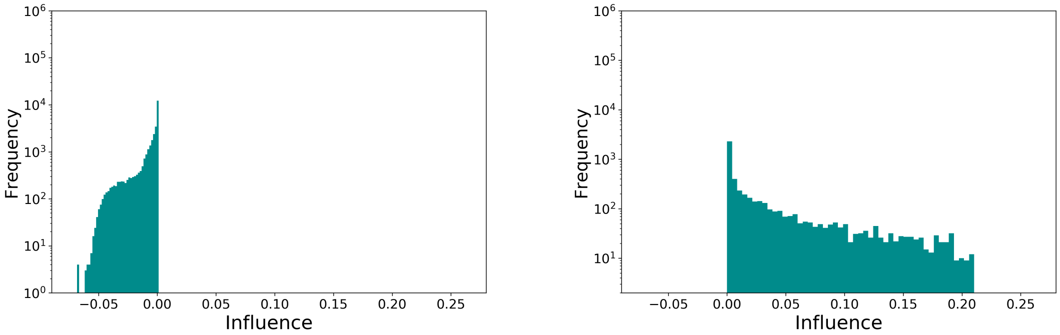
(b) Setting IV.

Figure 3.4: Histogram of influences $\mathcal{I}_{up,f_h}(\boldsymbol{w}_h, \boldsymbol{x})$, for all decision makers h, and $\boldsymbol{x}$ in the training set, with y-axis in log-scale. In setting II all decision-makers use the same model, so the influences are tightly concentrated around zero and no decision-maker has a very large influence over any prediction. Meanwhile, under setting IV there is a decision-maker the overestimates risk for a group, so the influence of an individual over one of the model's prediction may be much larger.

When looking at the influence of the biased decision-maker over the different points, a stark contrast emerges. For those members of the disadvantaged group, this decision-maker has an influence in the positive direction–increasing the weight given to this decision-maker would increase the predicted probability for members of this group. Meanwhile, for those who are not members of the disadvantaged group, it has an influence in the negative direction. Figure 3.5 shows the breakdown of the influence over this two populations.

For each data point, we can compare the influence across decision-makers. Figure 3.6 shows scatterplots for the sorted influence for $x = \text{argmin}_x m_2(x, f_h)$ under setting II and IV. This is the point where there is the least aligned influence across decision-makers. In setting IV the case displaying the minimum aligned influence across decision-makers shows much more drastic discrepancies than in setting II.

Evaluation For each scenario of human decision-making, we evaluate the performance of f_y, f_h and $f_{\mathcal{A}}$ with respect to the construct label Y^c, but also show what the (incorrect) "naive" evaluation looks like, in which the model is evaluated only on the subset of samples for which we observe the label (D=1), and is evaluated with respect to observed label Y. This is generally the only evaluation possible from observed data, but as shown in Figure 3.7 it may distort the performance estimates of the different methods, and in particular it is prone to overestimate the performance of f_y.



(a) All cases except disadvantaged group.

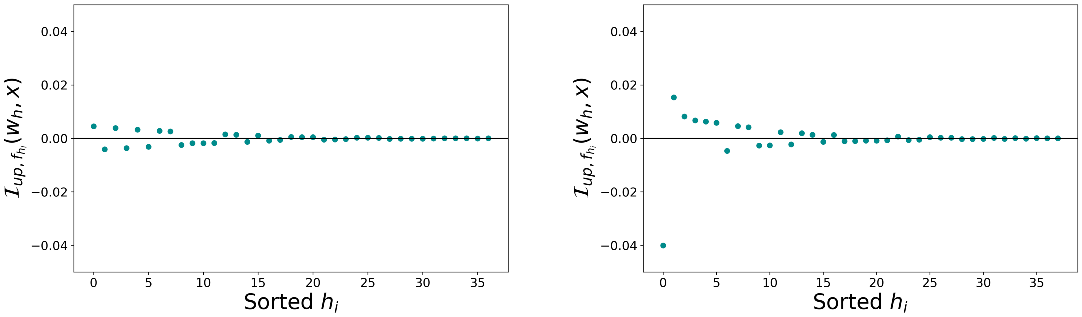
(b) Disadvantaged group.

Figure 3.5: Histograms of influence $\mathcal{I}_{up,f_h}(\boldsymbol{w}_h, \boldsymbol{x})$, for the biased decision maker h in setting IV, with y-axis in log-scale. If importance of this decision-maker was increased, predicted probabilities for members of disadvantaged group would increase, while decreasing for all others.

Figure 3.7 shows performance in terms of Area under the ROC Curve (AUC) across all four settings of decision-making. In setting I, where humans are uninformative, $f_{\mathcal{A}}$ is not affected by this and defaults to f_y. When the human decisions constitute an oracle (II), $f_{\mathcal{A}}$ improves substantially in comparison to f_y, which is also the case in the presence of unobservables (III). Naturally, in cases where humans constitute an oracle, learning from humans alone–or not automating anything–would be the gold standard, and these settings are only considered to illustrate the behavior in simple cases. The performance of f_h, however, can be easily derailed by a single biased decision-maker. Setting IV shows that the performance of f_h in such a setting drops more sharply than that of $f_{\mathcal{A}}$. Most interestingly, it is important in this setting to evaluate the performance of the models on the group for which one decision maker overestimates the risk. Table **??** shows the screen-in rate and true positive rate (TPR) for the three models in the top 25% highest-risk cases. Scenario IV is particularly interesting because it highlights that under non-random assignment of samples-to-experts f_h is susceptible to the bias of a single human, whereas $f_{\mathcal{A}}$ is able to incorporate expert knowledge in a robust fashion, not learning that bias.

Child maltreatment hotline screenings

The Allegheny County child maltreatment hotline receives over 15,000 calls a year, and call workers are tasked with deciding which cases should be screened-in for further investigation. Efforts to increase the availability of historical information about children and adults involved in a call have been accompanied by an interest in the use of risk assessment tools to aid call workers in identifying high-risk cases. Allegheny County has already implemented



(a) Setting II.

(b) Setting IV.

Figure 3.6: Scatterplots of influence $\mathcal{I}_{up,f_h}(\boldsymbol{w}_h, \boldsymbol{x})$, for $x = \text{argmin}_x m_2(x, f_h)$. In the x-axis, the decision-makers are sorted according to the magnitude of the influence for this particular datapoint. In setting IV the case displaying the minimum aligned influence across decision-makers shows more drastic discrepancies than in setting II.

one such system [147], which predicts probability of out-of-home placement. The selective labels problem arises because the result of an investigation is only observed for calls that are screened in. Omitted payoff bias is also a concern, as there may be treatment effects of the visitation, as well as indicators of risk that are not well captured in out-of-home placement. Finally, unobservables are present since the information communicated in the call is not used by the risk assessment model, which is instead trained only on data available in administrative records. The data used in our experiments corresponds to $46,544$ referrals (i.e. calls) between 2010 and 2014. This subset includes the first call associated to each child in this period of time, during which no risk assessment model was deployed. Over 800 variables are available which include information regarding demographics, behavioral health, and past interactions with county prison and public welfare for all adults and children associated to a referral. We perform the same feature selection described in 3.5.1 yielding 217 features. The observed label Y records whether out-of-home placement is observed in the 730 days following a call. We estimate models $f_h, f_y, f_\mathcal{A}$ using the same experimental setup and parameters described in Section 3.5.1.

In the present setting, we do not have ground truth Y^c to evaluate performance. However, when a social worker investigates a case there are other outcomes that are recorded and that are not used by the predictive model of out-of-home placement, f_y. Investigators record if the claims in a call are substantiated and if services are offered to the family. One of the most pressing concerns of optimizing for out-of-home placement alone would be if it fails to identify high-risk cases for which less aggressive interventions (e.g. services) change the outcome and improve the well-being of a child. Therefore, we evaluate if the label amalgamation incorporates human knowledge that makes it easier to identify these

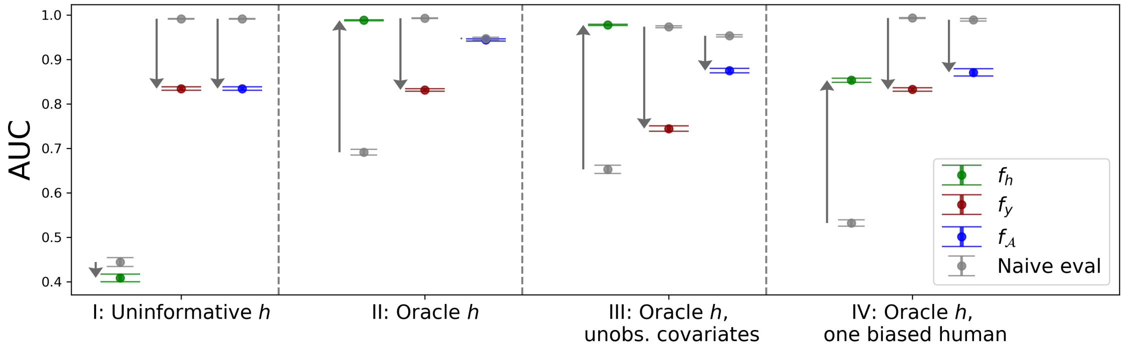
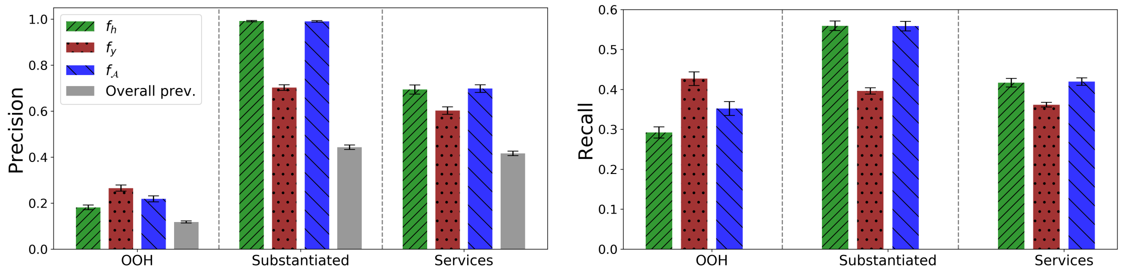


Figure 3.7: Mean±std AUC over 20 runs of $75-25\%$ train-test splits. Each setting (I-IV) represents a different scenario of human decisions. For each scenario, performance for the different models is shown: f_h trained on human decisions, f_y trained on observed outcomes, $f_{\mathcal{A}}$ trained on amalgamated labels. Grey markers indicate 'naive' evaluation, evaluated only on the samples for which Y is observed ($D = 1$). Arrows indicate the change between the naive evaluation and the correct evaluation with respect to Y^c. The results demonstrate that label amalgamation is not misguided by uninformative humans (I), while successfully incorporating human knowledge when informative (II-IV).

cases. We do so by measuring precision and recall of the highest scored screened in cases per model. Note that we only consider screened-in cases as we are blind to the outcomes of all others due to selective labels.

Figure 3.8 shows both precision and recall in the top 25% highest scores for the outcome Y (out-of-home placement) and additional outcomes that are not observed by the models (substantiation and services). Note that–due to the selective labels problem–recall is only useful in relative terms, since the absolute number may be over-estimated if not all cases are captured in the denominator. Precision does not have issues of this sort. As seen in Figure 3.8, there are cases whose risk does not seem to be captured by out-of-home placement, as indicated by the high rates of substantiation and services offered. This can be seen by the fact that recall and precision of humans is substantially higher for these labels. $f_{\mathcal{A}}$ successfully incorporates this human knowledge, considerably improving recall and precision of cases that are substantiated and to which services are offered, while still having a higher recall of out-of-home placement than f_h.

One concern of label amalgamation in this setting would be if widespread stereotypes held by call workers are incorporated into the model. In particular, it would be worrying if racial disparities arise. However, when looking at the prevalence of non-white families/children in the top 25% highest scored screen-ins, we find that f_y, f_h and $f_{\mathcal{A}}$ lead to similar screen-in rates, which also correspond to the overall proportion of non-white families. The results are shown in Table 3.2. This provides preliminary indication that label



(a) Precision in top 25% highest scored screen-in cases.

(b) Recall in top 25% highest scored screen-in cases.

Figure 3.8: Precision and recall on child maltreatment risk assessment for top 25% highest scored screened-in cases by model. Error bars show mean±std over 100 runs of $75 - 25\%$ train-test splits. Results show that (1) there are elements of risk that are not captured by the model trained to predict out-of-home placement label f_y, but are optimized for by humans, and (2) label amalgamation improves recall and precision for these cases, while having a better performance on out-of-home-placement than a model trained on human decisions alone.

amalgamation in this setting would not exacerbate nor mitigate racial biases. We note, however, that this is only one of many ethical concerns that arise, some of which cannot be evaluated through statistical tests alone and instead require a careful analysis of the sociotechnical context in which the system is embedded. As such, this evaluation does not constitute a recommendation that the proposed approach should be deployed in the child welfare context. We expand this discussion in Section **??**.

Model	Screen-in rate
f_h	0.594 ± 0.013
f_Y	0.594 ± 0.013
$f_{\mathcal{A}}$	0.589 ± 0.015
Overall	0.585 ± 0.007

Table 3.2: Screen-in rate for non-white children if threshold for screen-in is set to top 25% highest risk cases. This result provides preliminary indication that label amalgamation does not introduce racial biases.

3.5.2 Selective labels problem in prediction of neurological recovery of comatose patients

As discussed in Chapter 1, developing machine learning methodologies to assist physicians tasked with deciding whether to extend life-sustaining therapies of comatose survivors of cardiac arrest has the potential to save lives. Identifying complex patterns in qEEG recordings and other sources of clinical data could motivate the continuation of life-sustaining therapies for patients who do not exhibit previously known markers of good prognosis. However, the nature of the data available to train such models presents several challenges. One of the core challenges, relevant to the present Chapter, is the selective labels problem. Neurological outcomes are only observed for patients who are kept on life sustaining therapies, while there is no counterfactual available for patients who are withdrawn from life sustaining therapies.

Many machine learning approaches have been proposed to assist experts in predicting the probability of recovery [148, 15, 149], such as the one described in Chapter 1. For any of these approaches, it must be decided what data is used to train the model. One option is to assume that the physicians always made the correct choices, and therefore their decisions should be used as labels. This could be reasonable if we knew that physicians only choose to withdraw life-sustaining therapies when there is exhaustive evidence showing that the patient will not recover. However, that is not the case. This decision has a high degree of uncertainty associated to it, and is the result of a complex interaction between the family's and patient's previously expressed will, the physician's clinical assessment and individual physician's idiosyncrasies.

Alternatively, a model can be trained using only those cases for whom life sustaining therapies are continued. This has the advantage of only relying on observed outcomes, but has the problem of not being trained on a representative distribution of patients. If it can be assumed that there are no unobservables, and that all cases have a non-negligible probability of having life-sustaining therapies continued, the shift in the distribution can be accounted for via sample selection bias correction methodologies, such as inverse sampling weighting. It is known that unobservables influence this decision, since this is not exclusively a clinical decision but rather a sociomedical decision that also depends on the will of the family and the patient. In this section, the first question tackled is whether the positivity assumption is violated, meaning whether consistency in the decision to withdraw life sustaining therapies for subsets of patients leads to systematic blindness for subsets of the population. If this is the case, it means that learning from observed outcomes alone and extrapolating to the entire population requires strong assumptions about the model specification.

Data

Since 2010, comatose survivors of cardiac arrest admitted at an academic medical center had electroencephalography (EEG) measures brain activity continuously monitored. The data used in this Section has the same origin as the data used in Chapter 1, but corresponds to a longer period of time. While Chapter 1 uses data collected between 2010 and 2015, this Section considers data between 2010 and 2019. In addition to quantitative EEG (qEEG) summary measures at one-per-second resolution, static features such as test results, responsiveness metrics, are available.

EEG data collection and processing At the hospital, electrodes for EEG collection are adhered in standard positions, according to the 10-20 International System of electrode placement (10–20 system (EEG)), and data are typically recorded at 256Hz from 22 electrodes. Features are then extracted from these waveform data using FDA-approved clinical software (Persyst(R) Version 12, Persyst Development Corp, Prescott AZ). Part of this processing includes artifact detection and rejection. These artifacts include both physiological artifacts, such as electromyographic (EMG) artifact from muscle activity and shivering, and non-physiological artifacts, such as 60Hz interference from ambient alternating current electrical devices. The feature categories extracted from the EEG signal are listed in Table 3.3, and include many features known to be predictive of brain injury. For example, rythmicity measures can help predict poor prognosis as very rhythmic activity is often a sign of brain dysfunction. Similarly, asymmetry indices that summarize regional variations in EEG signal compared to the rest of the brain can be helpful, as certain type of asymmetries–such as "posterior dominant rythm"–are characteristic of healthy brains, while other types of assymetries may indicate brain injury. The resulting dataset contains 6036 timeseries per patient, where each timeseries corresponds to one of the features listed in Table 3.3, was made available for this research.

As a next step of featurization, basic statistics are computed for each feature over the time window of the first two hours of EEG monitoring for each patient, which corresponds to the first 7200 points of each time-series, as the data is collected at per-second resolution. The statistics computed are minimum, maximum, mean, median and standard deviation. This yields 30,180 features per patient. To avoid making assumptions of the nature of missing data, which is likely not missing at random, we drop all features for which there are missing values, yielding 8,914 features.

Demographics and static clinical features In addition to EEG features, 215 features corresponding to demographics and clinical information are available. Of these, we consider 27 features that physicians working with these patients deem as relevant for the task at hand. These include age, gender, features concerning potentially relevant clinical history, features collected via CT scans if performed, etiology of the cardiac arrest, and Pittsburgh Cardiac Arrest Category assigned to the patient by the treating physician in the first hours

Number of features	Measurement category
3	Artifact Intensity
24	Electrode Signal Quality
3	Seizure Probability
1,560	FFT (Fast Fourier Transformation) Spectrogram
195	aEEG
39	Peak Envelope, 0 - 25 Hz
3,783	Rhythmicity Spectrogram
12	Asymmetry EASI/REASI
170	Relative Asymmetry Spectrogram
15	Spikes
39	Suppression ratio
190	aEEG+ (0.16 - 25Hz) (LFF 1 sec, HFF 25 Hz, custom (off))

Table 3.3: Quantitative EEG (qEEG) features.

upon admission. This last feature, already discussed in Chapter 1, is a 4-level ordinal illness severity score that summarizes the physician's assessment of the patient's status.

Observed decisions and outcomes For each patient, it is observed whether life-sustaining therapies are *withdrawn*, and whether this decision is made for neurological or non-neurological reasons. For patients for whom life-sustaining therapies are not withdrawn, it is observed whether they suffer from brain death or not. Among patients who survive, there is very high variability on brain function recovered, and some patients continue to require intensive care. Discharge destination provides information about the patient's *disposition*, as the patient may be discharged to care facilities or be discharged home if they require less care. Patients are considered to have a positive outcome when they have been discharged home, although it should be noted that this is not a perfect proxy as this could also be influenced by family support and financial resources available for at-home care. Disposition and survival are therefore combined to summarize if recovery was positive or negative. Figure 3.9 shows the pipeline of the different decisions and outcomes observed.

Physician on-call At any given time, there is a physician on-call. This physician is responsible for making decisions concerning patients, including the decision to withdraw life-sustaining therapies. This yields the same setup that has been considered throughout this Chapter, in which the decision of a single expert is observed for each case. However, decisions to extend life-sustaining therapies are less easily assigned to a physician, as this is a decision that should be continuously reassessed. Guidelines stipulate that life-sustaining therapies should be extended for at least 72 hours. Therefore, choosing to extend life-

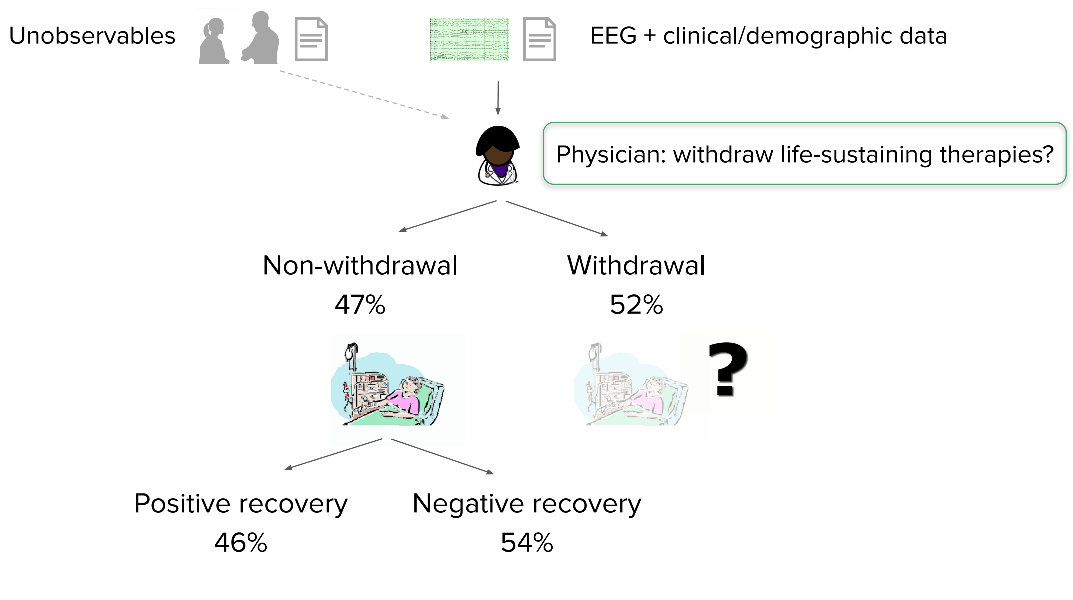


Figure 3.9: Pipeline of decision and outcomes observed, with percentage of data that falls into each category.

sustaining therapies at that threshold carries particular significance and could influence future decisions. Therefore, we leverage this to determine the physician-to-patient mapping, as specified in Equation 3.17, where $h^{(t)}$ denotes the physician on-call at time t_i, where t_i is the time after admission for patient $\mathbf{x}_i$. Note that patients who are not withdrawn and take less than 72 hours to be awake and follow commands do not have a physician assigned, as no decision was made for them.

$$h(\boldsymbol{x}_i) \quad = \quad \begin{cases} h^{(t_i = \text{final})} & \textit{if} \quad d_i = \text{withdrawn} \\ \text{None} & \textit{elif} \quad \text{hours to awake} < 72 \\ h^{(t_i = 72)} & \textit{elif} \quad d_i = \text{non-withdrawn} \end{cases} \tag{3.17}$$

In the 10 year period during which the data was collected, 20 physicians had shifts during which they were responsible for making decisions regarding the continuation of life-sustaining therapies. Figure 3.10 shows the count of decisions attributed to each physician.

Population 2,518 survivors of cardiac arrest were admitted to the hospital after resuscitation. This number is reduced to 1,810 when excluding patients who were withdrawn for non-neurological reasons and those who died from rearrest. These patients are excluded as (1) the cause of death for this group of patients is not neurological, (2) rearrest often

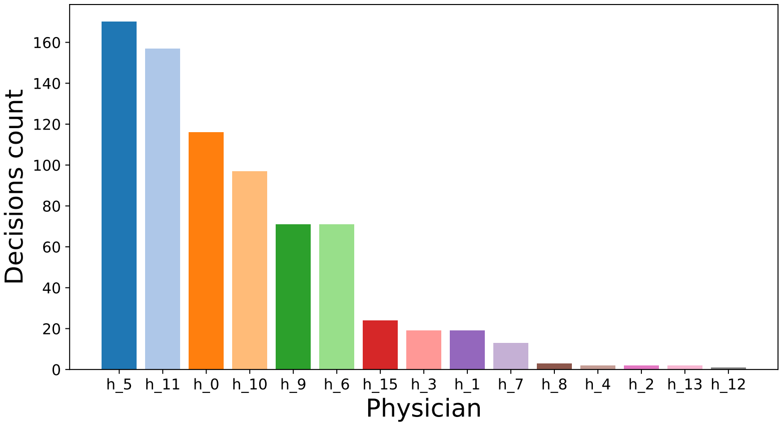


Figure 3.10: Number of decisions attributed to each physician, following the criteria specified in 3.17.

happens soon after admission so this group of patients are rarely part of the population for whom physicians need to decide whether to extend life sustaining therapies. Not all patients have EEG recordings available, Of the remaining patients, 930 had EEG data available for two hours upon the moment when monitoring began. This is the population considered in this research.

Systematic blindness in prediction of neurological recovery

To analyze whether consistency in human decisions leads to systematic blindness due to the selective labels problem, a predictive model of the human decisions f_h is used. Then, the (calibrated) predicted probabilities and the True Negative Rate (TNR) at low False Positive Rate (FPR) are analyzed, where a positive label corresponds to the decision to extend life sustaining therapies.

The calibrated probability of continuation of life sustaining therapies indicates the probability that an outcome will be observed for a given case. The positivity assumption requires that $P(D = 1|x) > \epsilon$, $\forall x$, so calculating $P(D = 1|x)$ allows us to test this assumption. Moreover, while the dataset contains all admissions of comatose survivors of cardiac arrest over a period of 10 years in a large hospital, the dataset is relatively small, with around 2,000 points. For this reason, ϵ must be large enough to allow for significant inference. Another way to think about this is in terms of TNR at low FPR. If the FPR is fixed at 2%, what is the recall? A high recall at low FPR indicates that there is a portion of cases that can be correctly identified as withdrawn from life-sustaining therapies. The FPR incurred provides a sense of how many cases of this set would actually have a label observed.

Featurization and predictive models The combination of EEG and static features yields a total of 8,941 features. While the number of patients in the dataset is large considering that each data point corresponds to EEG recordings of a comatose survivor of cardiac arrest, 930 data points is relatively small from a statistical perspective. Feature dimensionality is reduced by applying Principal Component Analysis (PCA) and keeping the top k components so that 98% of the variance is explained. It is possible to keep this much variance while significantly reducing dimentionality given that there is a lot of redundancy and strong correlation across many features. The final dataset used in the experiments contains 157 features per patient.

Four different predictive models are considered: a logistic regression with an L_1 penalty, a logistic regression with an L_2 penalty, a neural network with strong regularization and a neural network with weak regularization. The neural network considered is a single-layer perceptron with a logistic sigmoid activation function and n layers, where n corresponds to the number of features, i.e. $n = 157$. A constrained version with regularization $\alpha = 0.0001$ is considered, as well as a weakly constrained version with $\alpha = 0.1$. This second version is considered because rather than being concerned with overall performance, we are concerned with correctly approximating human decisions in some subsets of high-consistency, for which reducing regularization could be helpful. Experiments were conducted using a 75%-25% random train-test partition, parameter tuning was performed using a 5-fold cross-validation inside the training set, and confidence bounds on the ROC curves are obtained using Wilson score interval. Figure 3.11 shows the reverse ROC curve of all four models when trained to predict the decision to withdraw life-sustaining therapies.

The results shown in Figure 3.11 and Table 3.4 highlight the fact that there is a subset of cases for which it can be predicted with high probability that life-sustaining therapies will be withdrawn. In particular, Table 3.4 shows that using a logistic regression with L_1 penalty it is possible to identify at least 19% of withdrawals while only incurring in 2% false negatives. This means that there is a portion of cases that have a very low probability of having life-sustaining therapies extended, and therefore it is extremely unlikely to observe a “true label” of neurological recovery for them. Moreover, given that the size of the data is relatively small, this further constrains the possibility to make inference from observed

Model (f_h)	TNR @ 0.02 FNR
Logit L_1	$.23 \pm 0.04$
Logit L_2	$.23 \pm 0.05$
Neural net ($\alpha = 0.1$)	$.18 \pm 0.03$
Neural net reg. ($\alpha = 0.0001$)	$.10 \pm 0.03$

Table 3.4: TNR at fixed 2% FNR across models trained to predict physicians' decisions to withdraw life-sustaining therapies, f_h.

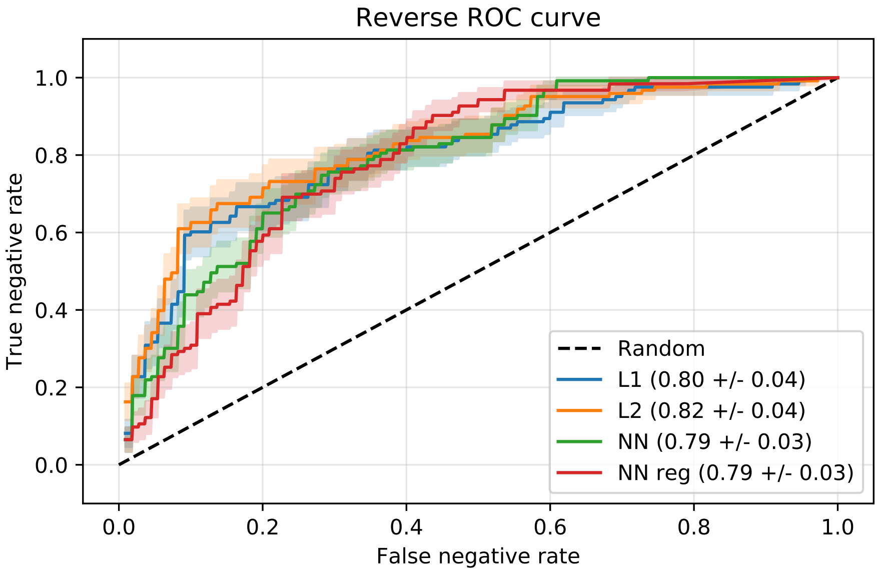


Figure 3.11: Reverse ROC curves of models trained to predict physicians' decisions to withdraw life-sustaining therapies (f_h) evaluated on test set. Performance at low false negative rate indicates there is a portion of cases that are consistently withdrawn from life-sustaining therapies.

outcomes of patients who are likely to be withdrawn. For example, in our data 2% false negatives corresponds to 9 patients.

Moving forward we will use the logistic regression with L_1 penalty to model f_h. Figure 3.12 shows the calibration plot and histogram of predicted probabilities for this classifier.

Prediction of neurological recovery under different assumptions

In the absence of observed outcomes for all patients, training a predictive model of neurological recovery requires assumptions about the data generating process. Often, these assumptions are implicit. A common implicit assumption in this domain–and in clinical settings in general–is that models trained on the subset of the population for whom there are observed outcomes will generalize to the rest of the population. This is an underlying assumption whenever a method is trained only on the portion of cases for which there are observed outcomes. In this Section we leverage the label amalgamation approach proposed in Section 3.4 to explore how predictions differ if we train from observed outcomes alone vs. from amalgamated outcomes. That is equivalent to making different assumptions. In the former, it is assumed that the model learned from observed outcomes will generalize to

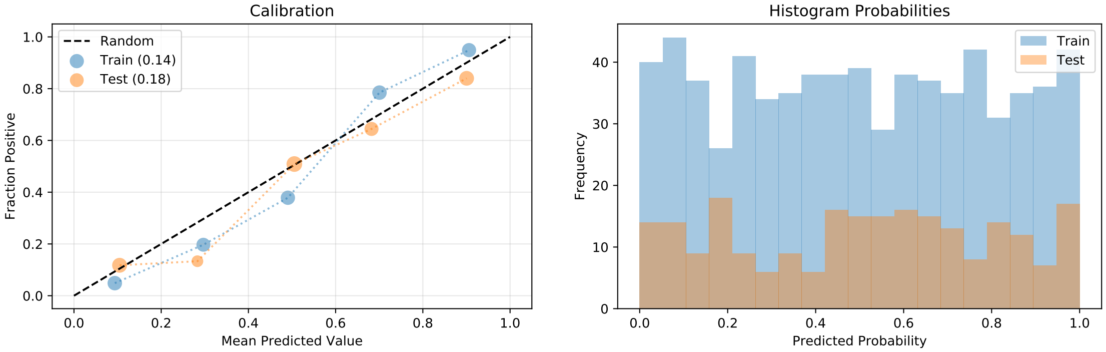


Figure 3.12: Calibration and predicted probabilities of L_1 logistic regression f_h. Good calibration (left) shows that predicted probabilities f_h also indicate the probability that an outcome will be observed. Histogram (right) shows that there are cases with very low probability of having their label observed.

the entire population–this has implicit assumptions about the nature of the missing data and/or the correctness of the model specification. Meanwhile, label amalgamation assumes that estimated consistency across experts is indicative of correctness.

Assuming that estimated consistency across experts is indicative of correctness is reasonable given that there is substantial clinical knowledge that informs physicians' decisions. However, there is a risk of incorporating self-fulfilling prophecies. Therefore, we propose label amalgamation as a way to test how robust the predictions of a model learned from observed outcomes alone are to this hypothesis, rather than as a replacement of the predictive model. If a prediction of very high probability of recovery drops substantially when label amalgamation is performed, it is possible that the initial prediction was an extrapolation to an unseen portion of the feature space and should be taken with a grain of salt.

For consistency and simplicity, we use an L_1 logistic regression to model f_y–the predicted probability of the observed outcome (neurological recovery). Figure 3.13 shows the ROC curve of this model and Figure 3.14 shows a scatter plot displaying f_y in the x-axis and f_h in the y-axis. The y-axis shows the physicians' believes, while also showing the probability that an outcome was observed for similar cases. The color in the legend indicates whether an outcome was observed and what the outcome was. Those cases in the bottom of the plot are almost always withdrawn from life-sustaining therapies. Therefore the estimated probability of recovery obtained via f_y corresponds to an extrapolation.

Label amalgamation The amalgamation set $\mathcal{A}$ is defined as indicated in Equation 3.18. This is almost identical as the way this set was defined in the previous application but also amalgamates points for which the maximum influence is negligible. The parameters are set empirically to $\gamma_1 = 2$, $\gamma_2 = 0.8$, $\gamma_3 = 0.005$. These values were chosen considering the

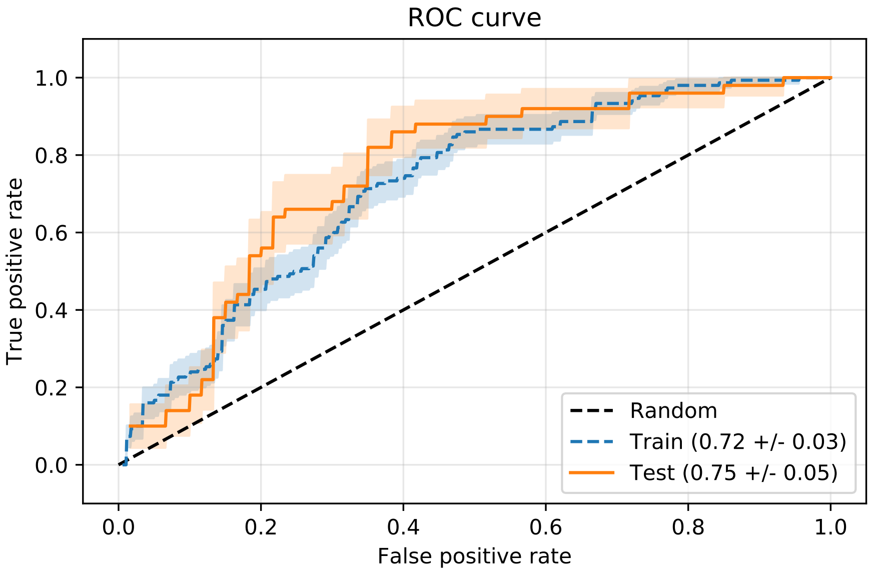


Figure 3.13: ROC curve of models trained to predict probability of positive recovery using observed outcomes (f_y) evaluated on test set.

meaning of each metric and the fact that there is a total of 15 physicians whose decisions are recorded in the data, only 10 of which observe more than 10 cases each. A value of $\gamma_1 = 2$ means that at least two physicians need to have as much influence as the rest, $\gamma_2 = 0.7$ means that at least 70% of the influence should be in the same direction, and $\gamma_3 = 0.005$ allows us to include cases that are robust to perturbations over the weight assigned to physicians because giving more weight to any physician would have a negligible impact of the prediction. Finally, δ is set to $\delta = 0.1$. This value is higher than in the previous application because given this data is smaller a case needs to have a higher probability of having a label observed in order to enable meaningful inference.

$$\mathcal{A} = \{\boldsymbol{x}_i \in X : |f_h(\boldsymbol{x}_i) - D_i| < \delta, (m_1(\boldsymbol{x}_i, f_h) > \gamma_1, m_2(\boldsymbol{x}_i, f_h) > \gamma_2) \vee m_3(\boldsymbol{x}_i, f_h) < \gamma_3)\} \tag{3.18}$$

Figure 3.15 shows the cases that are amalgamated. As in the previous application and as explained in the methodology section, for the purpose of amalgamation $f_h(\boldsymbol{x}_i)$ is estimated via cross validation (3-fold) performed inside the training set, with the L_1 penalty parameter chosen via a grid search to ensure the Hessian is invertible (if it is not, a higher penalty is enforced). The maximum value of the grid search is chosen in a transductive manner via cross-validation in the training set. This figure shows how accounting for consistency across experts limits the amalgamated set, since many points that have very

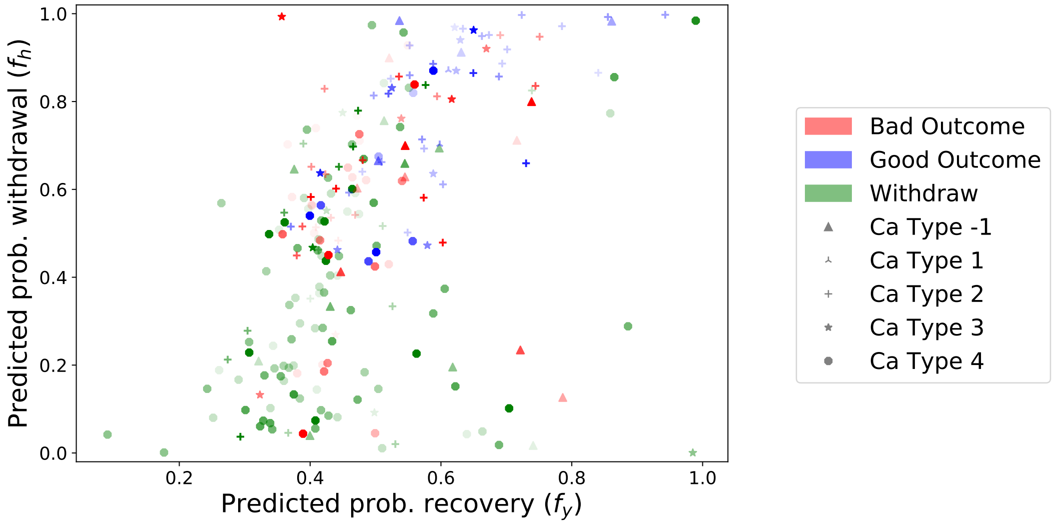


Figure 3.14: Scatter plot displaying test set predictions f_y in the x-axis and f_h in the y-axis. The y-axis shows the physicians' believes, while also showing the probability that an outcome was observed for similar cases. The legend shows whether the outcome for that case was not observed (withdrawn), and if it was observed it shows what the outcome was (bad/good outcome). Shapes in the legend show the ca type assigned to the patient, where ca type = 0 indicates a missing value. The lower region of the plot shows that cases with very low probability of continuation of life-sustaining therapies have a very high variance in the estimated probability of recovery, which can be either a result of (1) physicians being consistently wrong in the decision to withdraw life-sustaining therapies, or (2) the model f_y incorrectly extrapolating to an unseen portion of the feature space.

high-probability of withdrawal still exhibit disagreement among experts.

When amalgamation is performed a model $f_{\mathcal{A}}$ can be trained using the amalgamated set $\mathcal{A}$. The resulting test set predictions are shown in Figure 3.16. The shift in predicted probabilities between f_y and $f_{\mathcal{A}}$ can be see in Figure 3.17. While some predictions shift substantially, some remain relatively the same. This provides a way to test the robustness of the predictions of f_y under the assumption that human consistency is indicative of correctness.

Influence-driven second opinion recommendation

The influence function approach can be used to augment physicians' knowledge in different ways. It can be used to augment the information accompanying a prediction of the probability of neurological recovery, as shown above. It can also be used to suggest who to ask for a second opinion, as will be shown below.

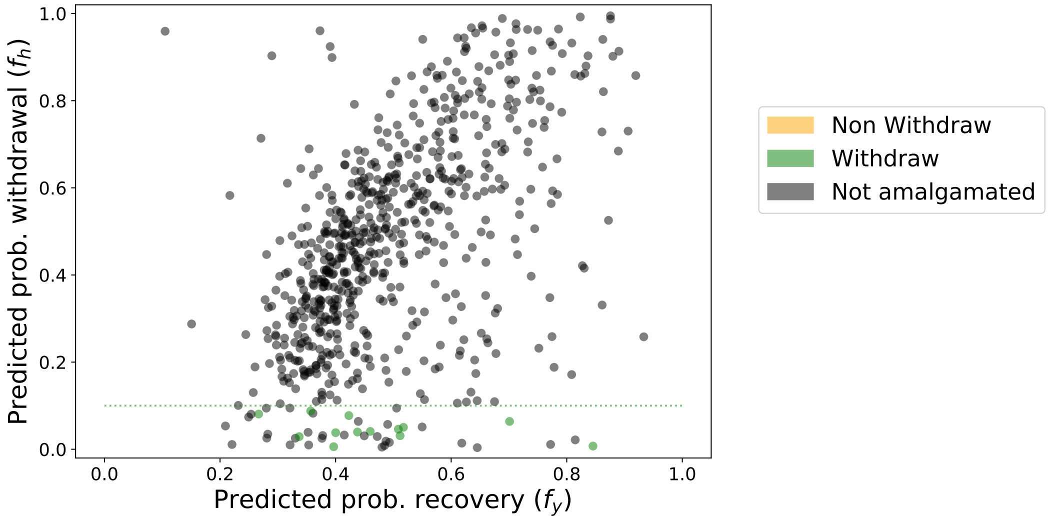


Figure 3.15: Scatter plot displaying f_y in the x-axis and f_h in the y-axis (equivalent as Figure 3.14) for the training set, where f_h is estimated via cross-validation in this set to enable meaningful amalgamation. Legend indicates if the point has estimated high consistency and is chosen for amalgamation or not. Among those selected for amalgamation the legend indicates if the case was withdrawn. It can be seen that all cases with estimated consistency in decision to withdraw were indeed withdrawn.

In this Chapter it has been shown that (1) partial consistency in historical expert decisions leads to systematic blindness that cannot be addressed through existing sample selection correction methodology, (2) such consistency exists in historical decisions to withdraw life-sustaining therapies–this is a real problem and not simply an edge-case that in theory could arise, (3) learning from data that suffers from this bias requires assumptions, and different assumptions can lead to vastly different predictions. Moreover, in prediction of neurological recovery–as in many other high-stakes tasks–none of the assumptions considered under the different hypothesis can be expected to hold true, and incorrect models could have fatal consequences. That raises the question: what should the algorithm do with the portion of cases that lack enough observed outcomes to learn from? We propose to use machine learning to recommend which physician to ask for a second opinion.

The use of machine learning to decide *when* to ask for a second opinion has been explored by [150]. Here, we propose a method to decide *who* to ask for a second opinion. Influence functions allow us to identify physician(s) whose influence on the estimated probability opposes the general consensus on withdrawal. This means that while patterns in the historical data indicate that a patient is very likely to be withdrawn, patterns on the decisions made by a particular physician deviate from this.

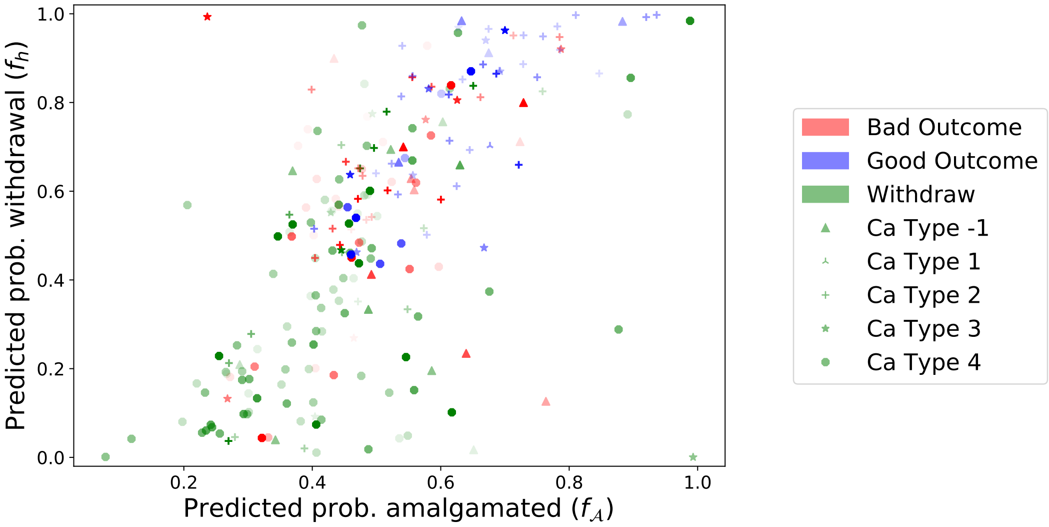


Figure 3.16: Scatter plot displaying test set predictions $f_{\mathcal{A}}$ in the x-axis and f_h in the y-axis. The legend shows whether the outcome for that case was not observed (withdrawn), and if it was observed it shows what the outcome was (bad/good outcome). Shapes in the legend show the ca type assigned to the patient, where ca type $= 0$ indicates a missing value. When compared to Figure 3.14 it can be seen that under amalgamation points shift towards the diagonal. Interestingly, the case that remains an outlier in the bottom right belongs to category 3, meaning the patient was in a mild to moderately deep coma, as opposed to category 4 which corresponds to a deep coma.

In general, the second opinion on the algorithmic prediction can be selected as shown in Equation 3.19, where $\tau_{decision}$ is the decision threshold. As it can be noted in the equation, it is possible for there to be no second opinion recommended, if no expert influences the prediction in an opposing direction.

$$h_{ask}(\boldsymbol{x}_i) \quad = \quad \begin{cases} \text{argmin}_h(\{\mathcal{I}_{up,f_h}(\boldsymbol{w}_h, \boldsymbol{x}_i) : \mathcal{I}_{up,f_h}(\boldsymbol{w}_h, x) < 0\}) & \text{if} \quad f_h(\boldsymbol{x}_i) > \tau_{decision} \\ \text{argmax}_h(\{\mathcal{I}_{up,f_h}(\boldsymbol{w}_h, \boldsymbol{x}_i) : \mathcal{I}_{up,f_h}(\boldsymbol{w}_h, x) > 0\}) & \text{if} \quad f_h(\boldsymbol{x}_i) < \tau_{decision} \end{cases} \tag{3.19}$$

When providing second opinions for high probability withdrawals, the second opinion would be requested from the expert with the largest influence in the positive direction, as indicated in Equation 3.19. This corresponds to the physician who, if given more weight, would steer the prediction away from a high probability of withdrawal. Naturally, the magnitude of the influence matters, and if negligibly small influences want to be discarded, it is possible to change the threshold defining the set in Equation 3.19 from 0 to γ_3 (or a similar value). This would effectively constraint second opinions to those who have a

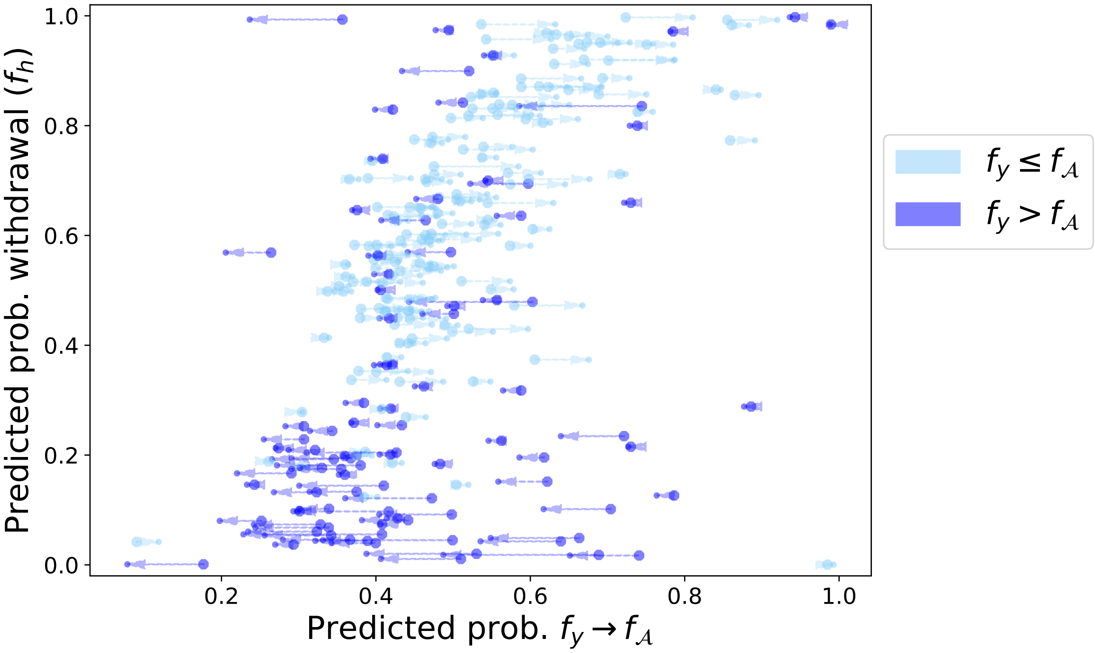


Figure 3.17: Scatter plot displaying the shift in test set predictions between f_y and $f_{\mathcal{A}}$ in the x-axis, and f_h in the y-axis. It can be seen that some predictions remain relatively unchanged, while others experience significant shifts. Predictions that experience a big shift under amalgamation should be taken with a grain of salt, as the prediction f_y may be the result of an extrapolation to an unseen region of the feature space.

sufficiently large opposing influence.

Semi-synthetic validation What is a desirable behavior of the proposed second opinion recommender algorithm? The algorithm should be able to identify experts who would provide a different perspective. To validate this, we construct semi-synthetic data in which we have ground truth for whether a physician would indeed be more likely to recommend extending life-sustaining therapies (or any other decision that is under consideration).

Let X be the set of features used throughout this Section. To reduce the noise and increase the strength of the signal, we use bootstrapping to double the size of the data, so that there are 1,860 data points. We standardize the data to be centered and with unit variance.

As in the previous semi-synthetic data created in this Chapter, let $\boldsymbol{\beta}^{\circ}$ be the learned coefficients of a logistic regression with L_1 penalty fitted to predict the observed decision to extend life-sustaining therapies. The L_1 penalty is enforced to simplify the synthetic

Group (i)	$P(D=1\|g_i=1)$	$P(D=1\|g_i=1,h_i=1)$
g_0	0.36	0.54
g_1	0.45	0.64
g_2	0.43	0.62
g_3	0.42	0.61
g_4	0.41	0.51

Table 3.5: Rates of $D = 1$ for each group, comparing overall rates and rates for physician that estimates a higher probability of recovery for that group. The semi-synthetic data is modeled such that each physician h_i estimates a higher probability of recovery for members of group g_i.

data, as we drop all features with a corresponding zero coefficient, keeping a dataset $X \in \mathrm{R}^{1860\times 32}$. We then sample coefficients $\boldsymbol{\beta}$ such that $\beta_i \sim N(\beta_i^\circ, 1)$.

We then assume that each patient belongs to one of five groups $\{g_0, g_1, g_2, g_3, g_4\}$, which we assign randomly, and represent the membership to each group as boolean features. We simulate that there are five physicians, $\{h_0, h_1, h_2, h_3, h_4\}$, and that the assignment of cases-to-physicians is random. Consistent with the real-world case, we assume that we only observe the assessment of a single physician per case. Finally, we simulate the decisions by modeling each physician's decisions h_i according to a logistic regression with coefficients $\boldsymbol{\beta}_{h_i}$. We assume that $\boldsymbol{\beta}_{h_i}$ are identical to $\boldsymbol{\beta}$ except for the coefficients that concern synthetic group membership, which are equal to -1 except for the group i, to which we assign the coefficient $\max(\boldsymbol{\beta})$.

The decisions are then $D_h = \mathbb{1}[X^T\boldsymbol{\beta}_h + \epsilon > 0]$, where $\epsilon \sim \text{Logistic}(0, 0.5)$. This means that we are assuming all experts make decisions according to logistic regression models that are identical except that for each physician h_i there is a group g_i for whom they systematically estimate a higher probability of recovery. Table 3.5 shows the rate at which each physician extends life sustaining therapies for each group ($P(D = 1|g_i = 1, h_i = 1)$) vs. the overall rate for that group ($P(D = 1|g_i = 1)$). As it can be seen in the Table, for each i physician h_i extends life sustaining therapies for group g_i more than the average.

For evaluation, we perform three-fold cross validation to estimate $\mathcal{I}_{up,f_h}(\boldsymbol{w}_h, \boldsymbol{x}_i)$ for all x_i, using an L_2 logistic regression to model f_h. Figure 3.18 shows the results of who would be recommended by the algorithm for a second opinion, , for the subset of patients who have a predicted probability of withdrawal of life-sustaining therapies greater than 50%. The recommendation is given as specified in Equation 3.19, for $\tau_{decision} = 0.5$. As desired, for each group g_i the physician who is more likely to extend life-sustaining therapies (h_i) is the one that is most often recommended by the algorithm.

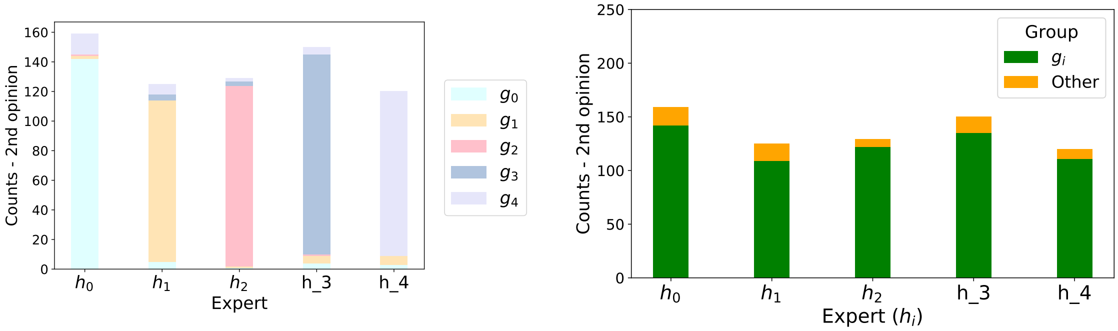


Figure 3.18: Results of who would be recommended by the algorithm for a second opinion, for the subset of patients who have a predicted probability of withdrawal of life-sustaining therapies greater than 50%. As desired, for each group g_i the physician who is more likely to extend life-sustaining therapies (h_i) is the one that is most often recommended by the algorithm.

Second opinion recommendations in predictions of neurological recovery Now that the methodology has been validated in semi-synthetic data, we move to see what the recommendations would look like in the real-world data, where we do not have ground truth for what the second opinions would be.

For consistency with the first part of this Section, we use the exact same models and data partitions. Table 3.6 presents a summary of basic statistics for each physician, including number of decisions assigned to each physician, percentage of cases for which they extend life sustaining therapies, and the mean and standard deviation of the influence over the cases in the set $\{\boldsymbol{x} : f_h(\boldsymbol{x}) < 0.1\}$, where $f_h(\boldsymbol{x})$ is calculated via cross-validation in the training set. The Pearson correlation coefficient between the average influence and the rate at which physicians extend life sustaining therapies is -0.059, which shows that the influence is not merely capturing whether the physician is on average more likely to extend life sustaining therapies.

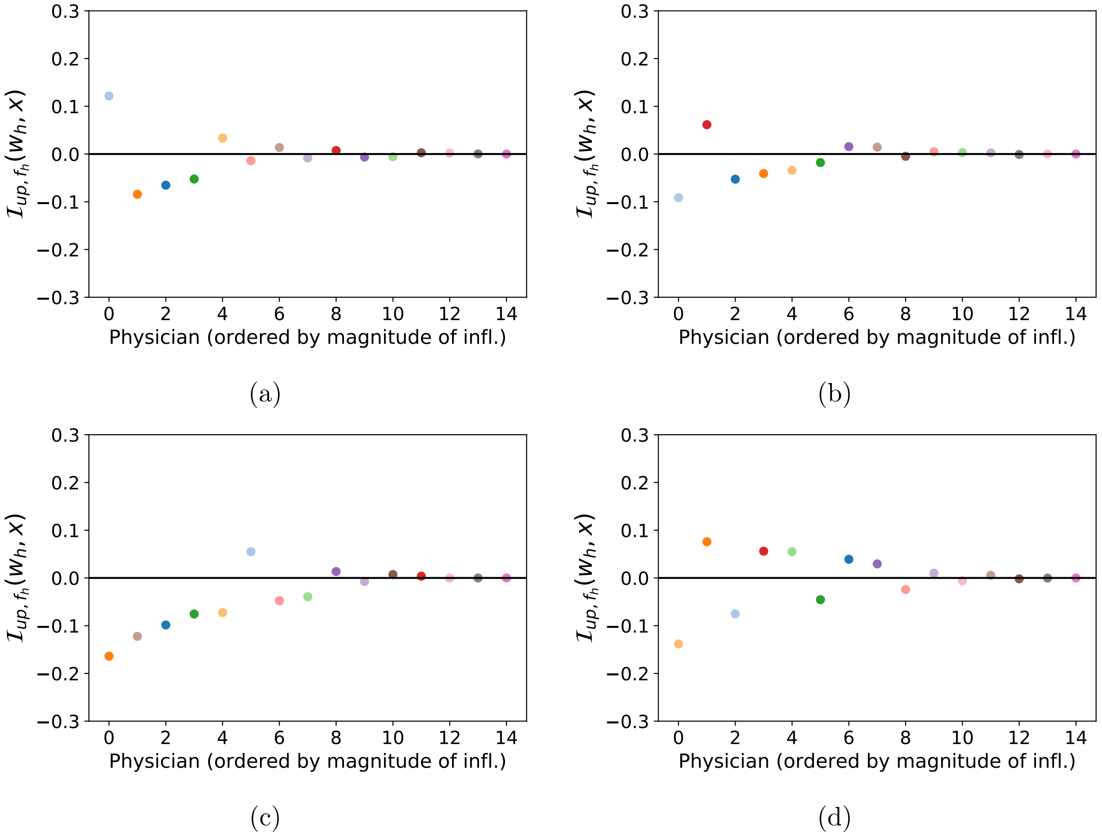
Figure 3.19 shows examples of cases that are predicted to be withdrawn with high likelihood ($f_h(\boldsymbol{x}) < 0.1$) and for which there is a suggested second opinion. Physicians with the largest positive influence would be asked for their opinion, as they influence the predicted probability away from the estimated (low) probability of extending life-sustaining therapies. However, not all cases have a suggested second opinion, since there are instances where no physician has a non-negligible influence, as shown in Figure 3.20a. The intuition behind these cases is that perturbations of the training data shifting the importance given to each physician would not change the predicted probability. Similarly, there are cases where all non-negligible influence is negative, meaning that there are perturbations that would make the predicted probability of non-withdrawal even lower, but there are no perturbations that would make it higher, as seen in the example in Figure 3.20.

h	count	rate extend	mean infl.	std infl.
h_5	127	0.37	0.0039	0.0282
h_{11}	115	0.41	0.0035	0.0344
h_0	91	0.45	-0.0096	0.0440
h_{10}	70	0.50	0.0070	0.0205
h_6	55	0.24	0.0042	0.0203
h_9	55	0.25	0.0013	0.0180
h_{15}	18	0.28	0.0053	0.0212
h_1	13	0.31	0.0019	0.0105
h_3	13	0.38	0.0032	0.0139
h_7	11	0.27	0.0001	0.0082
h_8	3	0.67	-0.0003	0.0017
h_2	2	0.50	-0.0014	0.0180
h_{12}	1	0.00	-0.0002	0.0009
h_{13}	1	1.00	-0.0003	0.0021
h_4	1	0.00	0.0000	0.0008

Table 3.6: Summary of basic statistics for each physician, including number of decisions assigned to each physician, percentage of cases for which they extend life sustaining therapies, and the mean and standard deviation of the influence over the cases in the set $\{\boldsymbol{x} : f_h(\boldsymbol{x}) < 0.1\}$, where $f_h(\boldsymbol{x})$ is calculated via cross-validation in the training set.

Naturally, the next question that arises is: for how many cases would there be a recommended second opinion? Figure3.21 shows a scatterplot of the influence of all physicians for each point in the set $\{\boldsymbol{x} : f_h(\boldsymbol{x}) < 0.1\}$, for $\boldsymbol{x}$ in the training set and $f_h(\boldsymbol{x})$ estimated via cross-validation. The set of physicians considered is restricted to the 10 physicians who observe more than 10 cases, since the influence of the others may not be too sensitive to outliers. The x-axis corresponds to the data points, ordered according to the variance in the influence. While some cases have very low variance, for most cases there is at least one physician who, if given more weight, would increase the predicted probability of extending life-sustaining therapies. Physicians with positive influence are candidates for being consulted for a second opinion.

Another important question to ask is: who is recommended for a second opinion? If there is a single physician that is always recommended by the algorithm, this would not be useful in practice, as it would overburden this person and would always be recommending the same perspective. Figure 3.22a shows the frequency with which each physician would be consulted for a second opinion, according to the proposed methodology. Moreover, Figure 3.22a shows the rate at which the physicians extend life-sustaining therapies. The comparison between these two plots shows that the second opinion is not simply



(a)

(b)

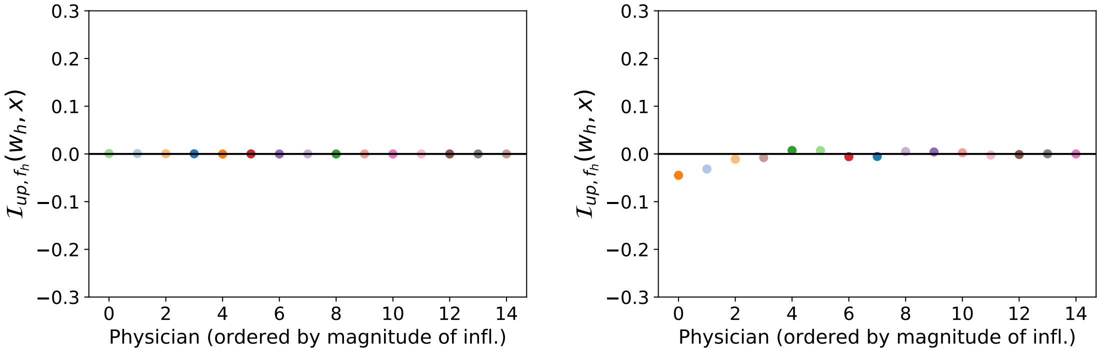
(c)

(d)

Figure 3.19: Examples of distribution of influence for cases that are predicted to be withdrawn with high likelihood ($f_h(\boldsymbol{x}) < 0.1$) and for which there is a suggested second opinion. Physicians with the largest positive influence would be asked for their opinion, as they influence the predicted probability away from the estimated (low) probability of extending life-sustaining therapies.

capturing who is more likely to extend life-sustaining therapies. For example, physician h_6 is frequently recommended for a second opinion, even though their overall rate of non-withdrawal is quite low. This highlights an attribute of the influence-driven second opinion: unlike summary statistics of the overall behavior, the influence-driven approach provides recommendations for *individual cases*.

Finally, analyzing the distribution of the influence for each physician can tell us about their behavior with respect to the rest of their colleagues and, even more so, to the predictions made by the algorithm. Figure 3.23 shows the influence of each physician for the set $\{\boldsymbol{x} : f_h(\boldsymbol{x}) < 0.1\}$, for $\boldsymbol{x}$ in the training set, for $f_h(\boldsymbol{x})$ estimated via cross-validation. For each physician, the datapoints were ordered according to the magnitude of the influence.



(a)

(b)

Figure 3.20: Examples of distribution of influence for cases without suggested second opinions. In (a) all physicians have virtually null influence, and in (b) no physician has a non-negligible influence on the predicted probability in the positive direction.

As it can be seen in the Figure, the distribution of influence is similar across physicians, with a large concentration around 0 and tails of varying lengths. However, there are physicians, e.g. h_6, for whom the influence skews towards the positive direction, while others, e.g. h_0, skew towards a higher frequency of negative influences. Additionally, it can be seen how some, e.g. h_7 rarely have a non-negligible influence.

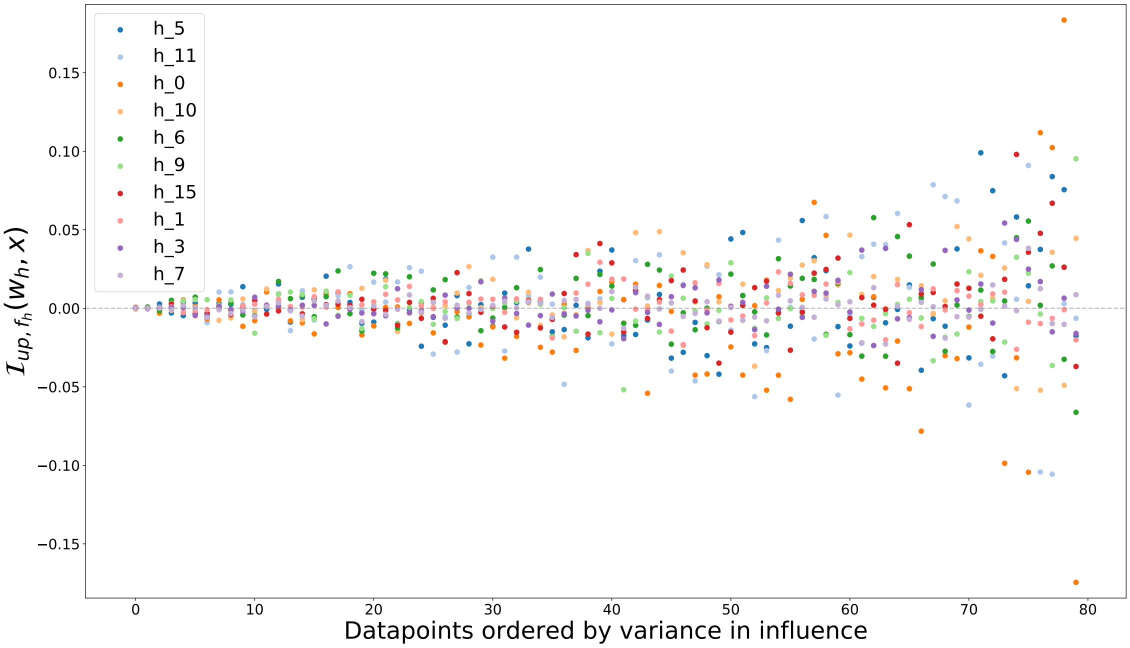
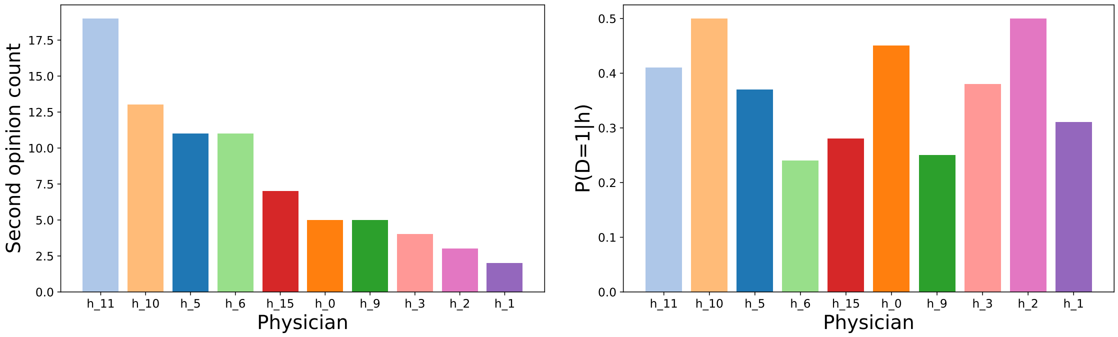


Figure 3.21: Influence of physicians for each case in the set $\{\boldsymbol{x} : f_h(\boldsymbol{x}) < 0.1\}$, for $\boldsymbol{x}$ in the training set and $f_h(\boldsymbol{x})$ estimated via cross-validation. X-axis corresponds to cases, ordered by the variance in the influence. While some cases have very low variance, for many cases there is at least one physician who, if given more weight, would increase the predicted probability of extending life-sustaining therapies. Physicians with positive influence are candidates for being consulted for a second opinion.



(a) (b)

Figure 3.22: (a) Frequency with which each physician would be consulted for a second opinion, according to the proposed methodology, and (b) rates at which each physician extends life-sustaining therapies. On (a) it can be observed that different physicians would be recommended for second opinions, and the comparison between both plots shows that the second opinion is not simple capturing who is more likely to extend life-sustaining therapies.

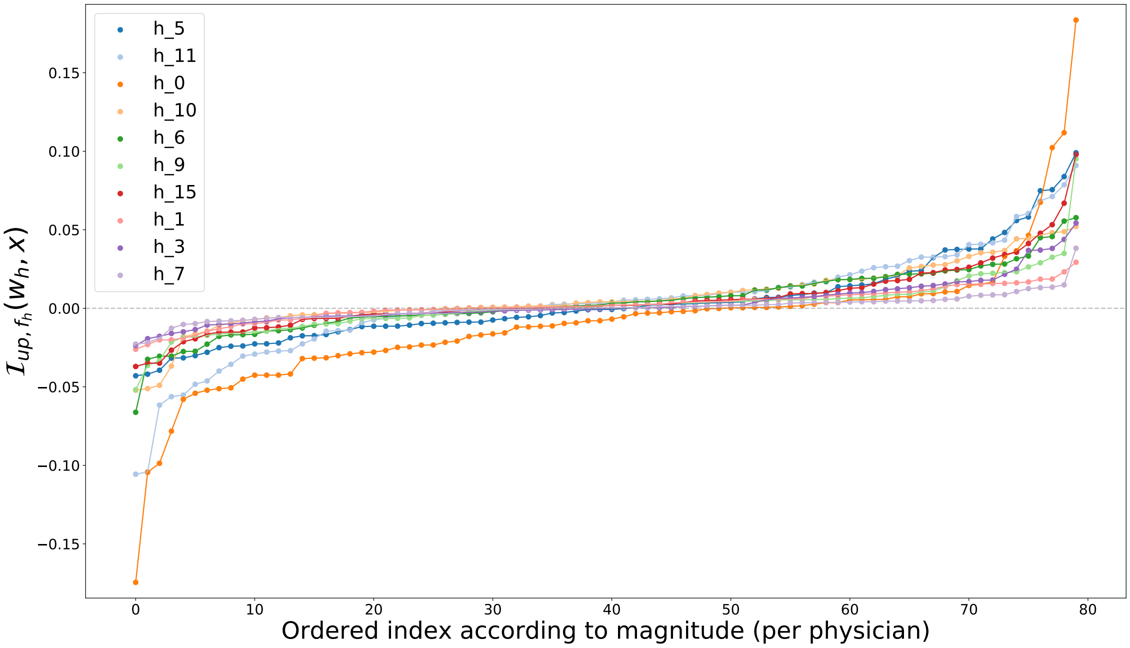


Figure 3.23: Influence of each physician for the set $\{\boldsymbol{x} : f_h(\boldsymbol{x}) < 0.1\}$, for $\boldsymbol{x}$ in the training set, for $f_h(\boldsymbol{x})$ estimated via cross-validation. X-axis corresponds to ordered index according to magnitude of influence, per physician. While the distribution of influence is similar across physicians, some (e.g. h_6) skew towards a positive influence, whereas others (e.g. h_0) skew towards a negative influence, and some (e.g. h_7) have influence very close to zero for most cases.

Chapter 4

Conclusions

4.1 Summary of contributions

4.1.1 Methodological contributions

Unraveling complex structures to inform decisions Chapter 1 introduces ***Canonical Autocorrelation Analysis*** (CAA), a method for automated discovery of multiple-to-multiple correlation structures within a set of features. This method, which builds on sparse Canonical Correlation Analysis (CCA) and Principal Component Analysis (PCA), can be useful when looking for hidden parsimonious structures in data, each involving only a small subset of all features. The utility of CAA for anomaly detection is demonstrated in Chapter 1, which also introduces a distance metric between CAA correlation structures in Section 1.3.4, enabling us to obtain a feature space embedding termed ***Canonical Autocorrelation Embeddings (CAE)***. In this embedding, each individual/object is represented by the set of its multivariate correlation structures. This methodology is particularly fitting to supervised learning tasks where each individual or object of study has a batch of data points associated to it, as in for instance patients for whom several vital signs or other health related parameters are recorded over time.

Exposing and penalizing structures that may bias decisions Chapter 2 introduces algorithmic fairness methodology for discovering and mitigating biases. Section 2.2 introduces an ***Unsupervised Bias Enumeration Algorithm*** (UBE) for word embeddings. The associations are identified by geometric patterns in word embeddings that run parallel between people's names and common lower-case tokens. The algorithm is highly unsupervised as it does not require the sensitive features to be pre-specified. This is desirable because: (a) many forms of discrimination–such as racial discrimination–are linked to social constructs that may vary depending on the context, rather than to categories with fixed definitions; and (b) it makes it easier to identify biases against intersectional groups, which depend on combinations of sensitive features. The application of this algorithm exposes a large number of offensive associations related to sensitive features such as race and gender on publicly available embeddings, including a supposedly ”debiased” embedding. Section 2.2 proposes one of the first methodologies to ***reduce bias in predictive models without requiring access to protected attributes***. This method leverages the societal biases that are encoded in word embeddings, eliminating the need for access to protected attributes. Crucially, it only requires access to individuals' names at training time and not at deployment time. Two variations of the proposed method are evaluated using a semi-synthetic as well as a large-scale dataset of online biographies. The results show that both variations can simultaneously reduce race and gender biases.

Leveraging structures in humans' historical decisions Drawing inspiration from the literature on crowd-sourcing and wisdom of the crowds, Chapter 3 proposes methodology to tackle some of the limits of learning from observed outcomes alone by also learning from consistency amongst experts. However, while in crowd-sourcing the same instance is assessed by multiple people, in historical data of experts' decisions it is often the case that each instance is assessed by a single expert, such as a physician or a judge. I propose an ***influence-function-based method to estimate human consistency***. The proposed method identifies cases for which the human decisions can be predicted with high confidence and for which the prediction is influenced by the historical decisions of several experts. Under the assumption that human consistency is indicative of correctness, this human knowledge can then be incorporated into a model trained to predict observed labels through a proposed ***label amalgamation*** approach. When it cannot be assumed that expert consistency is indicative of correctness, influence functions can be used to answer the question “who should the expert ask for a second opinion”. Chapter 3 closes with a proposed approach for ***influence-driven second opinion recommendation***.

4.1.2 Domain-specific contributions

Nuclear physics In Chapter 1 the proposed CAA method is applied to perform anomaly detection to identify potential nuclear threats. The results show that this method can help detect sources of radiation embedded in noisy background by finding multi-energy-bin combinations that reflect correlations between subsets of bins characteristic to background gamma-ray spectra. We show that such characterization of multiple-to-multiple bin correlations can be used as a powerful alternative to popular spectral anomaly detection methods which represent a null-space of expected background variance using linear combinations of photon counts observed in all energy bins. CAA's ability to focus on the most informative subsets of bins allows it to more effectively characterize background radiation variability, enabling higher threat sensitivity at lower false detection rates, when compared to the standard PCA-based approach.

Automated recruiting Chapter 2.1 presents a large-scale study of gender bias in automated recruiting. Maintaining an online professional presence has become increasingly important for people's careers, and this information is often used as input to automated decision-making systems that advertise open positions and recruit candidates for jobs and other professional opportunities. In order to perform these tasks, a system must be able to accurately assess candidate's current occupations, skills, interests, and "potential." However, even the simplest of these tasks–determining someone's current occupation–can be non-trivial. The results in Chapter 2.1 show that occupation classification is susceptible to gender bias, stemming from existing gender imbalances in occupations, and that removing explicit gender indicators (e.g. gender pronouns) is not enough to remove this bias. Additionally, the theoretical results show that whenever differences in true positive rates are correlated with pre-existing imbalances–as shown to happen in the large-scale study conducted in this thesis–, the imbalances will be compounded.

Child welfare Efforts to increase the availability of historical information about children and adults involved in calls received at the child abuse and maltreatment hotline have been accompanied by an interest in the use of risk assessment tools to aid call workers in identifying high-risk cases. Chapter 3 studies the risk of learning under omitted payoff bias in this context. The empirical results indicate that there are elements of risk that are optimized for by call workers but that are not wholly captured in the target label optimized for in currently deployed models, which could lead the risk assessment tool to underestimate risk for some cases. The proposed label amalgamation methodology successfully incorporates some of this information, bringing the construct optimized by the algorithm closer to the construct that call workers care about. This result highlights the importance of considering the construct validity of the target optimized for by the algorithm and proposes a path forward.

Prediction of neurological recovery of comatose patients Chapter 1 present a proof of concept to illustrate the potential utility of CAE by applying it to characterize electroencephalographic recordings from 80 comatose survivors of cardiac arrest, aiming to identify patients who will survive to hospital discharge with favorable functional recovery. The results show that at a low false positive rate the approach is able to identify a significant subset of patients who are likely to have a good neurological outcome, some of whom have otherwise unfavorable clinical characteristics. Importantly, some of these patients had 5% predicted chance of favorable recovery based on initial illness severity measures alone. This proof of concept shows that leveraging multivariate correlation structures present in the EEG data could help unravel patterns that are indicative of a positive prognosis and motivate the continuation of life-sustaining therapies for these patients. However, there are limitations to the straight-forward application of machine learning to predict neurological recovery. Chapter 3 uses EEG data collected over 10 years at a medical hospital to show that there exist a subset of patients who are consistently withdrawn from life-sustaining therapies, which means that there is a subpopulation for whom there is no information available of what would have happened if life-sustaining therapies had been continued. This Chapter also introduces methodology and provides empirical evidence to show how predictions vary when learning from observed outcomes alone and when also learning from consistent decisions historically made by physicians. Finally, the Chapter closes proposing an approach to use machine learning to recommend physicians who to ask for a second opinion when deciding if life-sustaining therapies should be extended.

4.2 Broader impact

Machine learning as a tool to increase usability of complex data As Chapter 1 shows, machine learning provides an opportunity to support experts' decisions by summarizing and discovering signals contained in complex sources of data. While experts usually have access to this data and routinely make use of it, the raw data is often hard for them to parse, leading to and under-utilization of information. For example, bedside monitoring continuously records multiple timeseries data of patients, but physicians can only consume and interpret a portion of it. Similar situations arise when analysts are interpreting spectral measurements in an effort to detect potential radioactive threats, as discussed in Chapter 1, or when call workers are provided with hundreds of features of historical information concerning a call received at the child abuse hotline, as discussed in Chapter 3. The methodology proposed in Chapter 1, and the empirical results shown, contribute to a growing body of literature on algorithmic-assisted decision making in high-stakes settings.

Algorithms, compounding injustices, and "leaky pipelines" Chapter 2.1 shows that whenever gaps in true positive rates are correlated with previous class imbalances, the imbalances will be compounded. I relate this effect to compounding injustices—an existing notion of indirect discrimination in the political philosophy literature that holds that it is a general moral duty to refrain from taking actions that would harm people when those actions are informed by, and would compound, prior injustices suffered by those people [1]. If a classifier compounds existing imbalances (e.g. gender imbalances in occupations), then the underrepresented will become even further underrepresented over time–a phenomenon sometimes referred to as the "leaky pipeline." This result has important practical relevance at a time when algorithmic decision support is increasingly adopted by organizations, many of which are simultaneously trying to improve diversity and inclusivity of their workplaces and products.

Algorithmic fairness: beyond fixed and known protected attributes Most methodologies to mitigate algorithmic bias require access to protected features. However, those deploying ML technologies often lack access to such features, and their use in certain domains may constitute disparate treatment under some legislations. For example, under Title VII employers can be found liable for employment discrimination if membership in a protected class is part of the input features of a model used to classify employees or potential hires, since this could constitute disparate treatment [151]. Moreover, discrimination is linked to complex social constructs that interact with each other, vary depending on the context, and cannot be reduced to binary encodings. Chapter 2 introduces some of the first methodologies to enumerate and mitigate biases without access to protected attributes. The proposed work can inform the design and deployment of methods that better suits the reality of operational and societal contexts.

Optimizing for constructs that matter: beyond observed labels In most public policy settings, humans are optimizing for complex constructs that are not easily quantifiable, such as social welfare. Chapter 3 studies this in the context of child welfare, where call workers are concerned with assessing whether a child is at risk of adverse outcomes, while deployed models estimate the probability of out-of-home placement, an imperfect proxy for risk. A similar situation arises in criminal justice, where risk assessment models meant to assist judges in bail decisions estimate the probability of recidivism, while judges are concerned with more complex constructs. This is illustrated by the fact that while youth is predictive of a higher risk of recidivism, it is also considered by judges as a mitigator, as it implies a lower level of culpability [152]. As a result, predictive models trained on quantifiable proxies of these outcomes may appear effective according to evaluation metrics, but deviate from what experts care about during deployment. Chapter 3 characterizes this disconnect and proposes ways to mitigate it. This work informs the potential risks of deploying algorithmic tools in sensitive policy domains, and gives a first step towards bridging the gap between what the algorithm optimizes and what the experts–and society–care about.

Bibliography

[1] Deborah Hellman. Indirect discrimination and the duty to avoid compounding injustice. *Foundations of Indirect Discrimination Law, Forthcoming*, 2018. 2, 32, 38, 114

[2] Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 275–284. ACM, 2017. 3, 29, 73, 74

[3] Alexander Smith. Hiroshima 70th anniversary: What to know about nuclear weapons in 2015. 2015. 6

[4] Nrc: Fact sheet on dirty bomb. http://www.nrc.gov/reading-rm/doc-collections/fact-sheets/fs-dirty-bombs.html. Accessed: 2015-10-21. 6

[5] Douglas Holdstock and Lis Waterston. Nuclear weapons, a continuing threat to health. *The Lancet*, 355(9214):1544–1547, 2000. 6

[6] Tucker Reals. Radioactive material stolen in mexico, apr 2015. URL http://www.cbsnews.com/news/mexico-on-alert-after-radioactive-iridium-192-stolen-from-truck-in-tabasco-state/. 6

[7] Swedish Physicians against Nuclear Weapons. ”nuclear terrorism”. URL ”http://laromkarnvapen.se/en/nuclear-weapons-politics/nuclear-terrorism/”. 6

[8] Prateek Tandon. *Bayesian Aggregation of Evidence For Detection and Characterization of Patterns in Multiple Noisy Observations.* PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, September 2015. 7, 134

[9] Rafael Lozano, Mohsen Naghavi, Kyle Foreman, Stephen Lim, Kenji Shibuya, Victor Aboyans, Jerry Abraham, Timothy Adair, Rakesh Aggarwal, Stephanie Y Ahn, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and

2010: a systematic analysis for the global burden of disease study 2010. *The lancet*, 380(9859):2095–2128, 2012. 7

[10] Emelia J Benjamin, Michael J Blaha, Stephanie E Chiuve, Mary Cushman, Sandeep R Das, Rajat Deo, Sarah D de Ferranti, James Floyd, Myriam Fornage, Cathleen Gillespie, et al. Heart disease and stroke statistics—2017 update: a report from the american heart association. *Circulation*, 135(10):e146–e603, 2017. 7

[11] Jonathan Elmer, Cesar Torres, Tom P Aufderheide, Michael A Austin, Clifton W Callaway, Eyal Golan, Heather Herren, Jamie Jasti, Peter J Kudenchuk, Damon C Scales, et al. Association of early withdrawal of life-sustaining therapy for perceived neurological prognosis with mortality after cardiac arrest. *Resuscitation*, 102:127–135, 2016. 7, 27

[12] Stephen Laver, Catherine Farrow, Duncan Turner, and Jerry Nolan. Mode of death after admission to an intensive care unit following cardiac arrest. *Intensive care medicine*, 30(11):2126–2128, 2004. 7

[13] Clifton W Callaway, Michael W Donnino, Ericka L Fink, Romergryko G Geocadin, Eyal Golan, Karl B Kern, Marion Leary, William J Meurer, Mary Ann Peberdy, Trevonne M Thompson, et al. Part 8: Post–cardiac arrest care. *Circulation*, 132(18 suppl 2):S465–S482, 2015. 7

[14] C Bassetti, Fulvio Bomio, Johannes Mathis, and Christian W Hess. Early prognosis in coma after cardiac arrest: a prospective clinical, electrophysiological, and biochemical study of 60 patients. *Journal of Neurology, Neurosurgery & Psychiatry*, 61(6):610–615, 1996. 7

[15] Jonathan Elmer, John J Gianakas, Jon C Rittenberger, Maria E Baldwin, John Faro, Cheryl Plummer, Lori A Shutter, Christina L Wassel, Clifton W Callaway, Anthony Fabio, et al. Group-based trajectory modeling of suppression ratio after cardiac arrest. *Neurocritical care*, 25(3):415–423, 2016. 7, 8, 11, 28, 90

[16] Christopher M Booth, Robert H Boone, George Tomlinson, and Allan S Detsky. Is this patient dead, vegetative, or severely neurologically impaired?: assessing outcome for comatose survivors of cardiac arrest. *Jama*, 291(7):870–879, 2004. 7

[17] Barbara Gold, Laura Puertas, Scott P Davis, Anja Metzger, Demetris Yannopoulos, Dana A Oakes, Charles J Lick, Debbie L Gillquist, Susie Y Osaki Holm, John D Olsen, et al. Awakening after cardiac arrest and post resuscitation hypothermia: are we pulling the plug too early? *Resuscitation*, 85(2):211–214, 2014. 7

[18] Jonathan Elmer and Clifton W Callaway. The brain after cardiac arrest. In *Seminars in neurology*, volume 37, pages 019–024. Thieme Medical Publishers, 2017. 7, 27

[19] Maximilian Mulder, Haley G Gibbs, Stephen W Smith, Ramnik Dhaliwal, Nathaniel L Scott, Mark D Sprenkle, and Romergryko G Geocadin. Awakening and withdrawal of life-sustaining treatment in cardiac arrest survivors treated with therapeutic hypothermia. *Critical care medicine*, 42(12):2493, 2014. 7, 27

[20] Claudio Sandroni, Alain Cariou, Fabio Cavallaro, Tobias Cronberg, Hans Friberg, Cornelia Hoedemaekers, Janneke Horn, Jerry P Nolan, Andrea O Rossetti, and Jasmeet Soar. Prognostication in comatose survivors of cardiac arrest: an advisory statement from the european resuscitation council and the european society of intensive care medicine. *Intensive care medicine*, 40(12):1816–1831, 2014. 7

[21] Marleen C Cloostermans, Fokke B van Meulen, Carin J Eertman, Harold W Hom, and Michel JAM van Putten. Continuous electroencephalography monitoring for early prediction of neurological outcome in postanoxic patients after cardiac arrest: a prospective cohort study. *Critical care medicine*, 40(10):2867–2875, 2012. 7

[22] Jeannette Hofmeijer, Marleen C Tjepkema-Cloostermans, and Michel JAM van Putten. Burst-suppression with identical bursts: a distinct eeg pattern with poor outcome in postanoxic coma. *Clinical neurophysiology*, 125(5):947–954, 2014. 8

[23] Jonathan Elmer, Jon C Rittenberger, John Faro, Bradley J Molyneaux, Alexandra Popescu, Clifton W Callaway, and Maria Baldwin. Clinically distinct electroencephalographic phenotypes of early myoclonus after cardiac arrest. *Annals of neurology*, 80(2):175–184, 2016. 8

[24] Jeannette Hofmeijer, Tim MJ Beernink, Frank H Bosch, Albertus Beishuizen, Marleen C Tjepkema-Cloostermans, and Michel JAM van Putten. Early eeg contributes to multimodal outcome prediction of postanoxic coma. *Neurology*, 85(2):137–143, 2015. 8

[25] Massimiliano Ignaccolo, Mirek Latka, Wojciech Jernajczyk, Paolo Grigolini, and Bruce J West. The dynamics of eeg entropy. *Journal of biological physics*, 36(2): 185–196, 2010. 8

[26] Harold Hotelling. Relations between two sets of variates. *Biometrika*, pages 321–377, 1936. 9

[27] Ola Friman, Magnus Borga, Peter Lundberg, and Hans Knutsson. Exploratory fmri analysis by autocorrelation maximization. *NeuroImage*, 16(2):454–464, 2002. 9

[28] Wim De Clercq, Anneleen Vergult, Bart Vanrumste, Wim Van Paesschen, and Sabine Van Huffel. Canonical correlation analysis applied to remove muscle artifacts from the electroencephalogram. *Biomedical Engineering, IEEE Transactions on*, 53(12): 2583–2587, 2006. 9

[29] Koby Todros and AO Hero. Measure transformed canonical correlation analysis with application to financial data. In *Sensor Array and Multichannel Signal Processing Workshop (SAM), 2012 IEEE 7th*, pages 361–364. IEEE, 2012. 9

[30] Daniela M Witten and Robert J Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*, 8(1):1–27, 2009. 9

[31] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, page kxp008, 2009. 9, 12, 131

[32] Khalid El-Arini, Andrew W Moore, and Ting Liu. Autonomous visualization. In *Knowledge Discovery in Databases: PKDD 2006*, pages 495–502. Springer, 2006. 9

[33] Madalina Fiterau and Artur Dubrawski. Projection retrieval for classification. In *Advances in Neural Information Processing Systems*, pages 3023–3031, 2012. 9

[34] WJ Krzanowski. Between-groups comparison of principal components. *Journal of the American Statistical Association*, 74(367):703–707, 1979. 10

[35] Bruce Korth and Ledyard R Tucker. Procrustes matching by congruence coefficients. *Psychometrika*, 41(4):531–535, 1976. 10

[36] Wolfgang Förstner and Boudewijn Moonen. A metric for covariance matrices. In *Geodesy-The Challenge of the 3rd Millennium*, pages 299–309. Springer, 2003. 10

[37] Jonathan Elmer, Jon C Rittenberger, Patrick J Coppler, Francis X Guyette, Ankur A Doshi, Clifton W Callaway, et al. Long-term survival benefit from treatment at a specialty center after cardiac arrest. *Resuscitation*, 108:48–53, 2016. 11

[38] Jochen Gorski, Frank Pfeuffer, and Kathrin Klamroth. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical Methods of Operations Research*, 66(3):373–407, 2007. 13

[39] Aleksandar Lazarevic, Levent Ertöz, Vipin Kumar, Aysel Ozgur, and Jaideep Srivastava. A comparative study of anomaly detection schemes in network intrusion detection. In *SDM*, pages 25–36. SIAM, 2003. 15

[40] Louis De Branges. The stone-weierstrass theorem. *Proceedings of the American Mathematical Society*, 10(5):822–824, 1959. 15

[41] Lucas Parra, Gustavo Deco, and Stefan Miesbach. Statistical independence and novelty detection with information preserving nonlinear maps. *Neural Computation*, 8(2):260–269, 1996. 20

[42] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 24

[43] Patrick J Coppler, Jonathan Elmer, Luis Calderon, Alexa Sabedra, Ankur A Doshi, Clifton W Callaway, Jon C Rittenberger, Cameron Dezfulian, et al. Validation of the pittsburgh cardiac arrest category illness severity score. *Resuscitation*, 89:86–92, 2015. 28

[44] Claudia Goldin and Cecilia Rouse. Orchestrating impartiality: The impact of" blind" auditions on female musicians. *American economic review*, 90(4):715–741, 2000. 31

[45] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357, 2016. 32, 61

[46] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186, 2017. 32, 61

[47] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018. 32, 48

[48] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018. 32

[49] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*, 2018. 32

[50] Rachael Tatman. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, 2017. 32

[51] Su Lin Blodgett and Brendan O'Connor. Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061*, 2017. 32

[52] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. 2017. 32

[53] Devin G Pope and Justin R Sydnor. Implementing anti-discrimination policies in statistical profiling models. *American Economic Journal: Economic Policy*, 3(3): 206–31, 2011. 32

[54] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Cal. L. Rev.*, 104: 671, 2016. 32, 60

[55] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013. 32

[56] Ian Ayres. Outcome tests of racial disparities in police practices. *Justice research and Policy*, 4(1-2):131–142, 2002. 32

[57] Toon Calders and Indrė Žliobaitė. Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and privacy in the information society*, pages 43–57. Springer, 2013. 32

[58] Pauline T Kim. Data-driven discrimination at work. *Wm. & Mary L. Rev.*, 58:857, 2016. 32, 60

[59] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012. 32

[60] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Mark DM Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pages 119–133, 2018. 32

[61] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. Fa*ir: A fair top-k ranking algorithm. In *Proceedings of the ACM Conference on Information and Knowledge Management*, pages 1569–1578, 2017. 32

[62] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. Ranking with fairness constraints. In *Proceedings of the International Colloquium on Automata, Languages, and Programming*, 2018. 32

[63] Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, page 22, 2017. 32

[64] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. *arXiv preprint arXiv:1805.01788*, 2018. 32

[65] Sahin Cem Geyik and Krishnaram Kenthapadi. Building representative talent search at LinkedIn. LinkedIn engineering blog post, Available at https://engineering.linkedin.com/blog/2018/10/building-representative-talent-search-at-linkedin, October 2018. 32

[66] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016. 32, 60

[67] Heather Sarsons. Gender differences in recognition for group work. *Harvard University Working Paper*, 2015. 33

[68] Heather Sarsons. Interpreting signals in the labor market: evidence from medical referrals. *Job Market Paper*, 2017. 33

[69] Donna K Ginther and Shulamit Kahn. Women in economics: Moving up or falling off the academic career ladder? *Journal of Economic perspectives*, 18(3):193–214, 2004. 33

[70] Marianne Bertrand and Esther Duflo. Field experiments on discrimination. In *Handbook of Economic Field Experiments*, volume 1, pages 309–393. Elsevier, 2017. 33

[71] Marianne Bertrand and Sendhil Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013, 2004. 33, 40

[72] Na Cheng, Rajarathnam Chandramouli, and KP Subbalakshmi. Author gender identification from text. *Digital Investigation*, 8(1):78–88, 2011. 33

[73] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4): 401–412, 2002. 33

[74] Kristen M Altenburger, Rajlakshmi De, Kaylyn Frazier, Nikolai Avteniev, and Jim Hamilton. Are there gender differences in professional self-promotion? an empirical case study of linkedin profiles among recent mba graduates. In *ICWSM*, pages 460–463, 2017. 33

[75] Victoria L Brescoll. Who takes the floor and why: Gender, power, and volubility in organizations. *Administrative Science Quarterly*, 56(4):622–641, 2011. 33

[76] David Niven and Jeremy Zilber. Do women and men in congress cultivate different images? evidence from congressional web sites. *Political Communication*, 18(4): 395–405, 2001. 33

[77] David G Smith, Judith E Rosenstein, Margaret C Nikolov, and Darby A Chaney. The power of language: Gender, status, and agency in performance evaluations. *Sex Roles*, pages 1–13, 2018. 33

[78] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. 35

[79] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018. 35, 54

[80] Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*, 2016. 35

[81] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, 2016. 36

[82] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 36

[83] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 36

[84] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017. 40

[85] Cynthia Dwork and Christina Ilvento. Fairness under composition. *arXiv preprint arXiv:1806.06122*, 2018. 46

[86] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic fairness. In *AEA Papers and Proceedings*, volume 108, pages 22–27, 2018. 47

[87] Austin C Kozlowski, Matt Taddy, and James A Evans. The geometry of culture: Analyzing meaning through word embeddings. *arXiv preprint arXiv:1803.09288*, 2018. 48

[88] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing

word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357, 2016. 48, 49, 54, 58, 142, 144

[89] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186, 2017. 48, 49, 50, 52, 141

[90] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464, 1998. 48

[91] Matthias Bluemke and Malte Friese. Reliability and validity of the single-target iat (st-iat): assessing automatic affect towards multiple attitude objects. *European journal of social psychology*, 38(6):977–997, 2008. 49, 50

[92] Thierry Devos and Mahzarin R Banaji. American= white? *Journal of personality and social psychology*, 88(3):447, 2005. 49

[93] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, NY, USA, 2012. ACM. 50, 60

[94] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*, 2017. 50

[95] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018. 50

[96] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017. 50

[97] Md Hoque, Rawshan E Fatima, Manash Kumar Mandal, Nazmus Saquib, et al. Evaluating gender portrayal in bangladeshi tv. *arXiv preprint arXiv:1711.09728*, 2017. 50

[98] Andrew Karpinski and Ross B Steinman. The single category implicit association test as a measure of implicit social cognition. *Journal of personality and social psychology*, 91(1):16, 2006. 50

[99] Lars Penke, Jan Eichstaedt, and Jens B Asendorpf. Single-attribute implicit association tests (sa-iat) for the assessment of unipolar constructs. *Experimental Psychology*, 53(4):283–291, 2006. 50

[100] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 52, 141

[101] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246. URL http://www.jstor.org/stable/2346101. 54

[102] Social Security Administration. Baby names from social security card applications - national level data, 2018. URL https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-level-data. Accessed 5 July 2018. 54, 64, 141

[103] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013. 54

[104] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162. 54

[105] Joshua Comenetz. Frequently occurring surnames in the 2010 census. *United States Census Bureau*, 2016. 58, 65, 141

[106] Ellen Dionne Wu. "they call me bruce, but they won't call me bruce jones:" asian american naming preferences and patterns. *Names*, 47(1):21–50, 1999. doi: 10.1179/nam.1999.47.1.21. URL https://doi.org/10.1179/nam.1999.47.1.21. 58

[107] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM, 2018. 60

[108] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudík, and H. Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 2019. 61

[109] Joy Adowaa Buolamwini. *Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers*. PhD thesis, Massachusetts Institute of Technology, 2017. 61

[110] Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. Proxy fairness. *arXiv preprint arXiv:1806.11212*, 2018. 61

[111] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970, 2017. 61

[112] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 63

[113] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007. 63

[114] Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml. 64

[115] Konstantinos Tzioumis. Demographic aspects of first names. *Scientific data*, 5: 180025, 2018. 64, 65

[116] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016. 65

[117] Gary King and Langche Zeng. Logistic regression in rare events data. *Political Analysis*, 9:137–163, 2001. 65

[118] Paul E Meehl. Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. In *In Proceedings of the 1955 Invitational Conference on Testing Problems*, pages 136–141. University of Minnesota Press, 1954. 72

[119] Robyn M Dawes, David Faust, and Paul E Meehl. Clinical versus actuarial judgment. *Science*, 243(4899):1668–1674, 1989. 72

[120] William M Grove, David H Zald, Boyd S Lebow, Beth E Snitz, and Chad Nelson. Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment*, 12 (1):19, 2000. 72

[121] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960. 72, 73

[122] Uchila N Umesh, Robert A Peterson, and Matthew H Sauber. Interjudge agreement and the maximum value of kappa. *Educational and Psychological Measurement*, 49 (4):835–850, 1989. 72, 73

[123] Mousumi Banerjee, Michelle Capozzoli, Laura McSweeney, and Debajyoti Sinha. Beyond kappa: A review of interrater agreement measures. *Canadian journal of statistics*, 27(1):3–23, 1999. 72, 73

[124] Kilem Gwet et al. Inter-rater reliability: dependency on trait prevalence and marginal homogeneity. *Statistical Methods for Inter-Rater Reliability Assessment Series*, 2(1): 9, 2002. 72, 73

[125] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics, 2008. 73

[126] Hossein Amirkhani and Mohammad Rahmati. Agreement/disagreement based crowd labeling. *Applied intelligence*, 41(1):212–222, 2014. 73

[127] James Shanteau. Competence in experts: The role of task characteristics. *Organizational behavior and human decision processes*, 53(2):252–266, 1992. 73

[128] James Shanteau. Why task domains (still) matter for understanding expertise. *Journal of Applied Research in Memory and Cognition*, 4(3):169–175, 2015. 73

[129] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. Technical report, National Bureau of Economic Research, 2017. 73, 74, 76

[130] Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, page 114. ACM, 2004. 74

[131] Jiayuan Huang, Arthur Gretton, Karsten M Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608, 2007. 74

[132] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2014. 74

[133] Shaun R Seaman and Ian R White. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3):278–295, 2013. 74

[134] Nathan Kallus and Angela Zhou. Residual unfairness in fair machine learning from prejudiced data. *arXiv preprint arXiv:1806.02887*, 2018. 74

[135] Aaron Chalfin, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. Productivity and selection of human capital with machine learning. *American Economic Review*, 106(5):124–27, 2016. 74

[136] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016. 74

[137] Jiaxuan Wang, Jeeheh Oh, Haozhu Wang, and Jenna Wiens. Learning credible models. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2417–2426. ACM, 2018. 74

[138] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, pages 67–82. Springer, 2016. 74

[139] David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Increasing fairness by learning to defer. *arXiv preprint arXiv:1711.06664*, 2017. 74

[140] R Dennis Cook. Assessment of local influence. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(2):133–155, 1986. 74

[141] Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Y Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993. 74

[142] James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000. 74

[143] Maria Cuellar and Edward Kennedy. A nonparametric projection-based estimator for the probability of causation, with application to water sanitation in kenya. *Available at SSRN 3257980*, 2018. 74

[144] Edward H Kennedy. Semiparametric theory and empirical processes in causal inference. In *Statistical Causal Inferences and Their Applications in Public Health Research*, pages 141–167. Springer, 2016. 74

[145] Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabas Poczos, Larry Wasserman, and James M Robins. Influence functions for machine learning: Nonparametric estimators for entropies, divergences and mutual informations. *arXiv preprint arXiv:1411.4342*, 2014. 74

[146] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1885–1894. JMLR. org, 2017. 74, 78, 79

[147] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148, 2018. 76, 87

[148] Jonathan Elmer, Bobby L Jones, Vladimir I Zadorozhny, Juan Carlos Puyana, Kate L Flickinger, Clifton W Callaway, and Daniel Nagin. A novel methodological framework for multimodality, trajectory model-based prognostication. *Resuscitation*, 137:197–204, 2019. 90

[149] Paul S Chan, John A Spertus, Harlan M Krumholz, Robert A Berg, Yan Li, Comilla Sasson, Brahmajee K Nallamothu, Get With the Guidelines-Resuscitation Registry Investigators, et al. A validated prediction tool for initial survivors of in-hospital cardiac arrest. *Archives of internal medicine*, 172(12):947–953, 2012. 90

[150] Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Robert Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. Direct uncertainty prediction for medical second opinions. *International Conference on Machine Learning (ICML)*, 2019. 100

[151] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Cal. L. Rev.*, 104: 671, 2016. 115

[152] Megan T Stevenson and Christopher Slobogin. Algorithmic risk assessments and the double-edged sword of youth. *Behavioral sciences & the law*, 36(5):638–656, 2018. 115

[153] Karl Sjöstrand, Line Harder Clemmensen, Rasmus Larsen, and Bjarne Ersbøll. Spasm: A matlab toolbox for sparse statistical modeling. *Journal of Statistical Software Accepted for publication*, 2012. 135

[154] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006. 135

[155] Claudia Bianchi. Slurs and appropriation: An echoic account. *Journal of Pragmatics*, 66:35–44, 2014. 139

[156] Jabari Asim. *The N word: Who can say it, who shouldn't, and why*. Houghton Mifflin Harcourt, 2008. 139

Appendix

A.1 Solution of CAA optimization problem via KKT conditions

Without loss of generality, assuming v is fixed and we are optimizing for u, the optimization problem in Lagrangian form can be written as formulated in Eq. A.1.

$$\begin{aligned} \min_{u} -u^T X^T X v + \sum_{i=1}^{m} (\lambda_1 ||v_i||_1 + \lambda_2) ||u_i||_1 \\ +\lambda_3 ||u||_2^2 - \lambda_2 c_1 - \lambda_3 \end{aligned} \tag{A.1}$$

$$\text{for} \quad 0 \leq c_1 \leq 1, 0 \leq c_2 \leq 10 \leq \lambda_1, 0 \leq \lambda_2, 0 \leq \lambda_3$$

The KKT conditions are:

- **Stationarity:** $0 \in -X^T X v + 2\lambda_3 u + \Gamma$
  for $\Gamma_i = (\lambda_2 + \lambda_1 |v_i|) sgn(u_i) \; \forall i = 1, ..., m$
- **Complementary slackness:**
  $\lambda_1 \sum_{i=1}^{m} |u_i||v_i| = 0; \qquad \lambda_2(||u||_1^2 - c_1) = 0; \qquad \lambda_3(||u||_2^2 - 1) = 0$
- **Primal feasibility:** $\sum_{i=1}^{m} |u_i||v_i| \leq 0; \qquad ||u||_1^2 - c_1 \leq 0; \qquad ||u||_2^2 - 1 \leq 0$
- **Dual feasibility:** $0 \leq \lambda_i$, $i = 1, 2, 3$

From complementary slackness and primal feasibility, either $\lambda_3 = 0$ and $||u||_2 \leq 1$, or $\lambda_3 > 0$ and $||u||_2 = 1$. Assuming $\lambda_3 > 0$ and solving the stationarity condition, we obtain that for $i = 1, ..., m$, Eq. A.2 holds, where $S_\lambda(x)$ is the soft-thresholding operator.

$$2\lambda_3 u_i = S_{(\lambda_1 |v_i| + \lambda_2)}((X^T X v)_i) \tag{A.2}$$

From complementary slackness, λ_3 must be such that $||u||_2 = 1$, therefore, Eq. A.3 is obtained.

$$u = \frac{S_{\Phi(v)}(X^T X v)}{||S_{\Phi(v)}(X^T X v)||_2^2} \tag{A.3}$$

$$\begin{aligned}\Phi(v, \lambda_1, \lambda_2) : \mathbb{R}^m &\longrightarrow \mathbb{R}^m \\ v_i &\longrightarrow \lambda_1 |v_i| + \lambda_2\end{aligned}$$

Additionally, λ_1 must be such that $\sum_{i=1}^{m} |u_i||v_i| = 0$, which will be guaranteed by setting $\lambda_1 = \max_i \frac{|(X^T X v)_i|}{|v_i|}$. Finally, either $\lambda_2 = 0$ results in a feasible solution, or λ_2 is chosen such that $||u||_1 = c_1$, which can be done through a binary search.

A.2 Relationship between CAA and Sparse PCA

As mentioned in Section 1.1, CAA and Sparse PCA have fundamentally different objectives, but given that Sparse CCA applied to identical matrices ($X = Y$) results in Sparse PCA components, it is worth taking a look at the details of how CAA and Sparse PCA differ. While Sparse PCA finds one-dimensional projections of data that maximize *variance* of data, CAA finds two-dimensional projections where *correlation* between the two sets is maximized. Furthermore, it is easy to see how the variables retrieved by CAA differ from those retrieved by Sparse PCA. As previously mentioned, applying Sparse CCA to matrices $X = Y$ results in Sparse PCA solutions $u = v$ [31]. Therefore, Sparse PCA can be written as

$$\begin{aligned} & max_{u,v} u^T X^T X v \\ ||u||_2^2 \leq 1, ||v||_2^2 \leq 1 \quad & ||u||_1 \leq c_1, ||v||_1 \leq c_2 \end{aligned} \tag{A.4}$$

In the following, we analyze how the objective values retrieved by CAA (Eq. 1.2) and Sparse PCA (Eq. A.4) differ.

- Sparse PCA optimal criterion value retrieved:

$$\begin{aligned} u^T X^T X u &= (\textstyle\sum_{i \in P} u_i X_i)^T (\textstyle\sum_{j \in P} u_j X_j) \\ &= \sum_{i \in P} \sum_{j \in P} u_i u_j X_i^T X_j \end{aligned}$$

$$\text{for } P = \{i | u_i \neq 0\}$$

- CAA optimal criterion value retrieved:

$$
\begin{aligned}
lu^T X^T X v - \lambda u^t v = (\sum_{i \in P_1} u_i X_i)^T (\sum_{i \in P_2} v_i X_i) - \lambda u^T v && \text{(A.5)} \\
\text{for } P_1 = \{i | u_i \neq 0\}, P_2 = \{i | v_i \neq 0\} && \text{(A.6)} \\
&& \text{(A.7)}
\end{aligned}
$$

Accounting for the constraint that the vectors u, v in the solution are orthogonal, we can rewrite this as:

$$
\begin{aligned}
& (\sum_{i \in P_1} u_i X_i)^T (\sum_{j \in P_2} v_j X_j) \text{ s.t. } P_1 \cap P_2 = \emptyset \\
= \;& (\sum_{i \in P_1} \sum_{j \in P_2} u_i v_j X_i^T X_j) \text{ s.t. } P_1 \cap P_2 = \emptyset.
\end{aligned}
$$

Notice that in the case of Sparse PCA, all interactions between variables in the subset P are considered, while CAA only considers interactions across two disjoint groups. Therefore, there are two types of interactions Sparse PCA considers that CAA does not: the variance of each variable and correlation/covariance between variables in the same subset. Note the first one is only relevant for Sparse PCA when using the covariance matrix, but the second is relevant both when Sparse PCA is applied to the correlation matrix or to the covariance matrix. As a result, CAA and Sparse PCA optimize different objectives, retrieving vectors that involve different subsets of features, and such subsets correspond to different types of structures in data.

A.3 Proof: CAA distance metric

In this section we prove that the metric defined to measure the distance between CAA canonical spaces satisfies the necessary conditions to be a well-defined distance.

$$
d(C_1, C_2) = \min(||u_1 - u_2||_2 + ||v_1 - v_2||_2 \;,\; ||u_1 - v_2||_2 + ||v_1 - u_2||_2) \tag{A.8}
$$

- Non-negativity: stems directly from the non-negativity of the ℓ_2 norm, together with the fact that the set of non-negative real numbers is closed under the summation and minimum operations.

- Identity:

$$\begin{aligned} 0 &= \min(||u_1 - u_2||_2 + ||v_1 - v_2||_2 \ , \ ||u_1 - v_2||_2 + ||v_1 - u_2||_2) \\ \Leftrightarrow 0 &= ||u_1 - u_2||_2 + ||v_1 - v_2||_2 \ \vee \ 0 = ||u_1 - v_2||_2 + ||v_1 - u_2||_2 \\ &\Leftrightarrow (0 = ||u_1 - u_2||_2 \wedge 0 = ||v_1 - v_2||_2) \\ &\vee (0 = ||u_1 - v_2||_2 \wedge 0 = ||v_1 - u_2||_2) \\ &\Leftrightarrow (u_1 = u_2 \wedge v_1 = v_2) \\ &\vee (u_1 = v_2 \wedge v_1 = u_2) \end{aligned}$$

  Given that we are dealing with these as non-ordered pairs, $d(C_1, C_2) = 0 \Leftrightarrow C_1 = C_2$.

- Symmetry: Stems directly from the fact that we define C_1 and C_2 as non-ordered pairs, hence the definition of the distance for each is exactly the same.

- Triangle inequality: The triangle inequality comes as a result of the triangle inequality of the ℓ_2 norm. We want to show that

$$d(C_1, C_3) \leq d(C_1, C_2) + d(C_2, C_3)$$

$$\begin{aligned} d(C_1, C_3) &\leq ||u_1 - u_3||_2 + ||v_1 - v_3||_2 \\ &= ||u_1 - u_3 + u_2 - u_2||_2 + ||v_1 - v_3 + v_2 - v_2||_2 \\ &\leq ||u_1 - u_2||_2 + ||u_2 - u_3||_2 + ||v_1 - v_2||_2 + ||v_2 - v_3||_2 \\ &= ||u_1 - u_2||_2 + ||v_1 - v_2||_2 + ||u_2 - u_3||_2 + ||v_2 - v_3||_2 \end{aligned}$$

  Through an analogous process,

$$\begin{aligned} d(C_1, C_3) &\leq ||u_1 - u_3 + v_2 - v_2||_2 + ||v_1 - v_3 + u_2 - u_2||_2 \\ &\leq ||u_1 - v_2||_2 + ||v_1 - u_2||_2 + ||v_2 - u_3||_2 + ||u_2 - v_3||_2 \end{aligned}$$

  Additionally, the following is also true:

$$d(C_1, C_3) \leq ||u_1 - v_3||_2 + ||v_1 - u_3||_2$$

  Therefore, through analogous reasoning, we derive the following two sets of inequalities:

$$\begin{aligned} d(C_1, C_3) &\leq ||u_1 - v_3 + u_2 - u_2||_2 + ||v_1 - u_3 + v_2 - v_2||_2 \\ &\leq ||u_1 - u_2||_2 + ||v_1 - v_2||_2 + ||u_2 - v_3||_2 + ||v_2 - u_3||_2 \end{aligned}$$

$$\begin{aligned} d(C_1, C_3) &\leq ||u_1 - v_3 + v_2 - v_2||_2 + ||v_1 - u_3 + u_2 - u_2||_2 \\ &\leq ||u_1 - v_2||_2 + ||v_1 - u_2||_2 + ||v_2 - v_3||_2 + ||u_2 - u_3||_2 \end{aligned}$$

The four inequalities we have derived span the four possible cases for $d(C_1, C_2) + d(C_2, C_3)$, which concludes our proof.

A.4 Principal angles and CAA

Although principal angles might initially seem like a good alternative to measure distances between CAA canonical spaces, note that this is not a viable option. Even though each pair of vectors defining a CAA canonical space constitute an orthonormal basis of a subspace, two orthogonal basis defining the same subspace do not represent the same correlation structure. This can be derived from the fact that, as shown in Section 1.3.4, two different pairs of vectors cannot represent the same correlation structure. It is also easy to understand why this would not be the case with a simple counterexample in $\mathbb{R}^3$. Consider the following two pairs of vectors:

$$\begin{cases} u_1 = (1, 0, 0) \\ v_1 = (0, 1, 0) \end{cases} \qquad \begin{cases} u_2 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0) \\ v_2 = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0) \end{cases}$$

Even though they are both orthonormal bases of the same subspace, $u_1 v_1^T \neq u_2 v_2^T$.

A.5 PCA spectral anomaly detector

The PCA spectral anomaly detector used for comparison purposes in this paper is frequently used in the radiation threat detection domain, and the description provided in this abstract can also be found in [8]. The algorithm first filters the energy data and performs smoothing via a 10s rolling window. It then computes the special covariance matrix shown in Equation A.9, where we assume the background data is a matrix $X \in \mathbb{R}^{n \times q}$. This covariance matrix retains 0.01 of the mean, instead of fully centering the data.

$$\begin{aligned} \Sigma &= \frac{XX^T}{q} - 0.99mm^T \\ m_j &= \sum_{i=1}^{n} X_{i,j} \end{aligned} \tag{A.9}$$

The correlation matrix $C = A\Sigma A$ is later calculated, where A is the design matrix

$$A = diag(\frac{1}{diag(\Sigma + 1)})$$

Finally, the Singular Value Decomposition is performed on the correlation matrix and the basis matrix T is created as

$$T = I_q - A^{-1}U_{PC}U_{PC}^T A$$

where U_{PC} contains the top principal component eigenvectors. Finally, the residuals can be obtained as $\sigma =|| TX_{test}^T ||_2$.

A.6 Sparse PCA spectral anomaly detector

Even though Sparse PCA is not generally used in the radiation threat detection domain, a Sparse PCA spectral anomaly detector was designed and implemented for comparison purposes in this paper. As in the case of the PCA-based approach, the algorithm first filters the energy data and performs smoothing via a 10s rolling window. It then normalizes the data to have mean of 0.01 the original mean and and Euclidean length of each column equal to 1, where each column corresponds to a variable.

Top sparse principal components, U_{Spc}, are extracted using the algorithm in [153], which is itself based on the formulation done in [154].

The basis matrix T is created as

$$T = I_q - U_{Spc}U_{Spc}^T$$

Given a new data set X_{test}, it is first normalized using the mean and Euclidean length of the training data, and then residuals are obtained as $\sigma =|| TX_{test}^T ||_2$.

A.7 EEG features

EEG features.

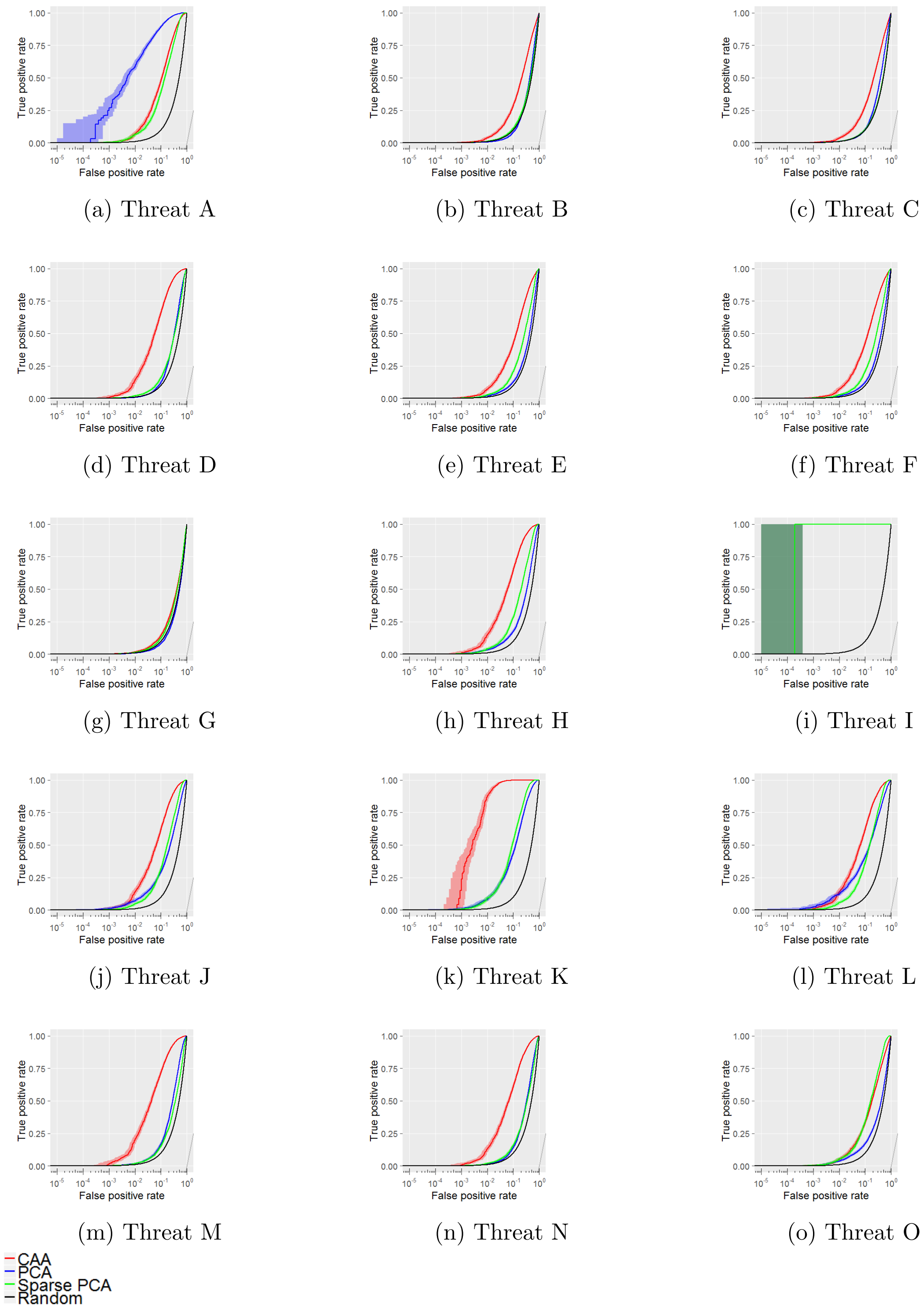
Below is the complete list of the EEG features available and used in Chapter 1.

Feature	Details
Artifact Intensity	Muscle
Artifact Intensity	Chew
Artifact Intensity	V-Eye
Artifact Intensity	L-Eye
Artifact Detector (Signal quality -Electrode 1)	
Artifact Detector (Signal quality -Electrode 2)	
Artifact Detector (Signal quality -Electrode 3)	
Artifact Detector (Signal quality -Electrode 4)	
Artifact Detector (Signal quality -Electrode 5)	
Artifact Detector (Signal quality -Electrode 6)	
Artifact Detector (Signal quality -Electrode 7)	
Artifact Detector (Signal quality -Electrode 8)	
Artifact Detector (Signal quality -Electrode 9)	
Artifact Detector (Signal quality -Electrode 10)	
Artifact Detector (Signal quality -Electrode 11)	
Artifact Detector (Signal quality -Electrode 12)	
Artifact Detector (Signal quality -Electrode 13)	
Artifact Detector (Signal quality -Electrode 14)	
Artifact Detector (Signal quality -Electrode 15)	
Artifact Detector (Signal quality -Electrode 16)	
Artifact Detector (Signal quality -Electrode 17)	
Artifact Detector (Signal quality -Electrode 18)	
Seizure Probability	
aEEG, Left Hemisphere	Max
aEEG, Left Hemisphere	Min
aEEG, Left Hemisphere	Median
aEEG, Left Hemisphere	Q75%
aEEG, Left Hemisphere	Q25%
aEEG, Right Hemisphere	Max
aEEG, Right Hemisphere	Min
aEEG, Right Hemisphere	Median
aEEG, Right Hemisphere	Q75%
aEEG, Right Hemisphere	Q25%
aEEG+(filt)(LFF0.16sec,HFF(off), aEEG2-20 512), L	Max

aEEG+(filt)(LFF0.16sec,HFF(off),aEEG2-20 512), L	Min
aEEG+(filt)(LFF0.16sec,HFF(off),aEEG2-20 512), L	Median
aEEG+(filt)(LFF0.16sec,HFF(off),aEEG2-20 512), L	Q75%
aEEG+(filt)(LFF0.16sec,HFF(off),aEEG2-20 512), L	Q25%
aEEG+(filt)(LFF0.16sec,HFF(off),aEEG2-20 512), R	Max
aEEG+(filt)(LFF0.16sec,HFF(off),aEEG2-20 512), R	Min
aEEG+(filt)(LFF0.16sec,HFF(off),aEEG2-20 512), R	Median
aEEG+(filt)(LFF0.16sec,HFF(off),aEEG2-20 512), R	Q75%
aEEG+(filt)(LFF0.16sec,HFF(off),aEEG2-20 512), R	Q25%
PeakEnvelope, 1 - 20 Hz, Left Hemisphere	
PeakEnvelope, 1 - 20 Hz, Right Hemisphere	
Spike Detections	
Suppression Ratio, Left Hemisphere	
Suppression Ratio, Right Hemisphere	
FFT Power, 1 - 4 Hz, Left Hemisphere	
FFT Power, 1 - 4 Hz, Right Hemisphere	
FFT Power, 4 - 8 Hz, Left Hemisphere	
FFT Power, 4 - 8 Hz, Right Hemisphere	
FFT Power, 8 - 13 Hz, Left Hemisphere	
FFT Power, 8 - 13 Hz, Right Hemisphere	
FFT Power, 13 - 20 Hz, Left Hemisphere	
FFT Power, 13 - 20 Hz, Right Hemisphere	
FFT Alpha/Delta, 8-13/1-4 Hz, Left Hemisphere	
FFT Alpha/Delta, 8-13/1-4 Hz, Right Hemisphere	
Rhythmicity Spectrogram, Left Hemisphere	1-4Hz
Rhythmicity Spectrogram, Left Hemisphere	4-8Hz
Rhythmicity Spectrogram, Left Hemisphere	8-13Hz
Rhythmicity Spectrogram, Left Hemisphere	13-20Hz
Rhythmicity Spectrogram, Right Hemisphere	1-4Hz
Rhythmicity Spectrogram, Right Hemisphere	4-8Hz
Rhythmicity Spectrogram, Right Hemisphere	8-13Hz
Rhythmicity Spectrogram, Right Hemisphere	13-20Hz

A.8 Radiation threat detection: ROC curves

Receiver operating characteristic (ROC) curves for all 15 types of threats for which CAA was tested are shown below. All PCA, Sparse PCA and CAA were trained using benign background radiation, and the resulting model was evaluated on similar background data inclusive of signatures of 15 different types of threats. In the ROC curves, the false positive rate axis is shown in logarithmic scale, to enhance view at low false positive rates.



(a) Threat A

(b) Threat B

(c) Threat C

(d) Threat D

(e) Threat E

(f) Threat F

(g) Threat G

(h) Threat H

(i) Threat I

(j) Threat J

(k) Threat K

(l) Threat L

(m) Threat M

(n) Threat N

(o) Threat O

A.9 True positive rate gender gaps across representations

Figure A.2 shows TPR gender gaps for BOW trained without gender indicators. Figures A.3 and A.4 show the results for WE, with and without gender indicators, respectively. Figures A.5 and A.6 show the results for DNN, with and without gender indicators, respectively.

A.10 Attention to gender

A.10.1 Attention to gender proxies

Figure A.7 shows the aggregated attention of the DNN model to words “wife” and “husband”. As with the word “women”, the model trained without gender indicators places more attention on these words. Notice, however, that the shift in attention weights, while it exists, is smaller than for the word “women”, which is consistent with the lower aggregate attention in the gender prediction model.

A.10.2 Attention to gender indicators

Figure A.8 shows the attention of the model, trained with and without gender indicators, on the word “she” during the prediction of the occupation based on biographies *with* gender indicators. One may expect that in the latter case the model would not attend to this word as it has not seen it during the training. However, the results indicate quite the opposite. In fact, the model puts *much more* attention to it. This can be attributed to the use of word embeddings, which enables the model to learn about words even if it has not explicitly seen them. Interestingly, when exposed to the word “she” during prediction, the model seems to receive a stronger gender signal than it has seen during training, and pays a significant amount of attention to it.

A.11 Offensive Stereotypes and Derogatory Terms

The authors consulted with colleagues whether to display the offensive terms and stereotypes that emerged from the embedding using our algorithms. First, regarding derogatory terms, people we consulted found the explicit inclusion of some of these terms offensive. We are also sensitive to the fact that, even in investigating them, we are ourselves using them. The terms we bleep-censor in the tables include slurs regarding race, homosexuality, transgender, and mental ability [155]. In particular, these include three variants on “the n word” [156], *shemale*, *faggot*, *twink*, *mentally retarded*, and *rednecks*. It is not obvious that such slurs would be generated given common naming conventions. Nonetheless, many of these terms were in groups of words that matched stereotypes indicated by crowd workers.

Of course, the associations of words and groups are also offensive, but unfortunately, it is impossible to convey the nature of these associations without presenting the words in the tables associated with the groups. In an attempt to soften the effect, we use group letters rather than illustrative names or summary statistics in our tables. While this decreases the transparency, it gives the reader a choice about whether or not to examine the associated names. Some colleagues were taken aback by an initial draft, in which names and associations were displayed in the same table, and it was noted that it that may be especially offensive to individuals whose name appeared on top of a column of offensive stereotypes. For the names, we restrict our selection of names to those that had at least 1,000 occurrences in the data so that the name would not be uniquely identified with any individual.

In addition, we considered withholding the entire tables and merely presenting the rating statistics. However, we decided that, given that our concern in the analysis is uncovering that such troubling associations are being made by these tools, it was important to be clear and unflinching about what we found, and not risk obscuring the very phenomenon in our explanation.

A.12 Proofs of Lemmas

Proof of Lemma 1. For $n = 2$, using our $\overline{\boldsymbol{X}}$ notation and their assumption $|X_1| = |X_2|$, simple algebra shows that,

$$(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2) \cdot (\overline{\boldsymbol{A}}_1 - \overline{\boldsymbol{A}}_2) = \frac{1}{|X_1|} s(X_1, A_1, X_2, A_2).$$

Since $\boldsymbol{\mu} = (\overline{\boldsymbol{X}}_1 + \overline{\boldsymbol{X}}_2)/2$, we have that $\overline{\boldsymbol{X}}_1 - \boldsymbol{\mu} = (\overline{\boldsymbol{X}_1} - \overline{\boldsymbol{X}}_2)/2 = -(\overline{\boldsymbol{X}}_2 - \boldsymbol{\mu})$, and:

$$\begin{aligned} g(X_1, A_1, X_2, A_2) &= (\overline{\boldsymbol{X}}_1 - \boldsymbol{\mu}) \cdot (\overline{\boldsymbol{A}}_1 - \overline{\boldsymbol{\mathcal{A}}}) + (\overline{\boldsymbol{X}}_2 - \boldsymbol{\mu}) \cdot (\overline{\boldsymbol{A}}_2 - \overline{\boldsymbol{\mathcal{A}}}) \\ &= \frac{\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2}{2} \cdot \left(\overline{\boldsymbol{A}}_1 - \overline{\boldsymbol{\mathcal{A}}} - (\overline{\boldsymbol{A}}_2 - \overline{\boldsymbol{\mathcal{A}}})\right) \\ &= \frac{1}{2}(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2) \cdot (\overline{\boldsymbol{A}}_1 - \overline{\boldsymbol{A}}_2), \end{aligned}$$

which when combined with the previous equality establishes the first equation in Lemma 1. □

Proof of Lemma 2. Since we have shown that $(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2) \cdot (\overline{\boldsymbol{A}}_1 - \overline{\boldsymbol{A}}_2) = 2g(X_1, A_1, X_2, A_2)$ above, we immediately have that $g(X, A) = 2g(X, A, \mathcal{X}, \mathcal{A})$. Moreover, simple algebra shows that $g(X, A, \mathcal{X}, \mathcal{A})$ and $g(X, A, X^c, A^c)$ are proportional because $\overline{\boldsymbol{X}} - \overline{\boldsymbol{\mathcal{X}}} = \frac{|X^c|}{|\mathcal{X}|}(\overline{\boldsymbol{X}} - \overline{\boldsymbol{X^c}})$ and similarly $\overline{\boldsymbol{A}} - \overline{\boldsymbol{\mathcal{A}}} = \frac{|A^c|}{|\mathcal{A}|}(\overline{\boldsymbol{A}} - \overline{\boldsymbol{A^c}})$. □

Proof of Lemma 3. Follows simply from the definition of g and μ for $n \geq 2$ and $n = 1$. □

A.13 Preprocessing names and words for Chapter 2

A.13.1 Preprocessing first names from SSA dataset

The SSA dataset [102] has partial coverage for earlier years and includes all names with at least 5 births, we use only years 1938-2017 and select only the names that appeared at least 1,000 times, which cover more than 99% of the data by population. From this data, we extract the fraction of female and male births for each name as well as the mean year of birth. Of course, we select only the names appearing in the embedding.

Note that the mean of the fraction of females among our names is significantly greater than 50%, even though the US population is nearly balanced in binary gender demographics. The subtle reason is there is greater variability in female names in the data, whereas the most common names are more often male. That is, the data have fewer predominantly male first names in total with more people being given those names on average. Since we are including each name only once, this increases the female representation in the population.[1]

A.13.2 Preprocessing last names from U.S. Census

A dataset of last names is made publicly available by the Census Bureau of the United States and contains last names occurring at least 100 times in the 2010 census [105], broken down by percentage of race, including White, Black, Hispanic, Asian and Pacific Islander, and Native American. Again we filter for names that appear at least 1,000 times and apply the binary classification procedure described in Section 2.2.3 to clean the data.

A.13.3 "Cleaning" names

[89] apply a simple procedure in which they remove the 20% of words whose mean similarity to the other names is smallest. We apply a similar but slightly more sophisticated procedure by training an linear Support Vector Machine [scikit-learn's LinearSVC, 100, with default parameters] to distinguish the input names from an equal number of non-names chosen randomly from the most frequent 50,000 words in the embedding. We then remove the 20% of names with smallest margin in the direction identified by the linear classifier.

Figure A.9 illustrates the effect of cleaning the last names and shows that the names that tend to be removed are those that violate Zipf's law.

[1] We performed similar experiments on a sample of names drawn according to the population and, while the names are gender balanced, the clusters exhibit less diversity and most often simply are split by gender and age – one can even have an entire cluster solely consisting of people named *Michael*.

A.13.4 Preprocessing words

To identify the most frequent M words in the embedding, we first restrict to tokens that consist only of the 26 lower-case English letters or spaces for embeddings that contain phrases. We also omit lower-case tokens when the upper-case version of the token is more frequent. For instance, the lower-case token “john” is removed because “John” is more frequent.

A.14 Biases in different lists/embeddings

Table A.1 shows the names from other embeddings. Table A.2 shows the biases found in the “debiased” `w2v` embedding of [88], while Table A.3 show last-name biases generated from the `w2v` embeddings.

fast F1	**fast F2**	**fast F3**	**fast F4**	**fast F5**	**fast F6**	**fast F7**	**fast F8**	**fast F9**	**fast F10**	**fast F11**	**fast F12**
Nakesha	Carolyn	Tamara	Lillian	Alejandra	Katelyn	Ahmed	Landon	Stephan	Marquell	Greg	Gerardo
Keisha	Nichole	Emi	Lucinda	Maricella	Jayda	Shanti	Keenan	Nahum	Antwan	Willie	Renato
Kandyce	Mel	Isabella	Velda	Ona	Shalyn	Mariyah	Skye	Sabastian	Dakari	Edward	Pedro
Kamilah	Tawnya	Karina	Antoinette	Fabiola	Jaylyn	Siddharth	Courtland	Philippe	Pernell	Jefferey	Genaro
Rachal	Deirdre	Joli	Flossie	Sulema	Evie	Yasmin	Luke	Jarek	Jarred	Russ	Matteo
+702	+821	+622	+478	+400	+851	+288	+576	+312	+440	+474	+234
98% F	98% F	97% F	96% F	93% F	90% F	64% F	22% F	9% F	6% F	4% F	2% F
1980	1972	1987	1972	1984	1993	1992	1991	1987	1984	1973	1987
29% B	4% B	5% B	14% B	2% B	3% B	6% B	5% B	6% B	34% B	8% B	1% B
3% H	2% H	9% H	9% H	64% H	2% H	4% H	1% H	9% H	3% H	3% H	65% H
1% A	2% A	6% A	6% A	8% A	2% A	33% A	3% A	4% A	2% A	5% A	7% A
66% W	91% W	80% W	71% W	25% W	93% W	56% W	90% W	80% W	61% W	84% W	27% W

glove F1	**glove F2**	**glove F3**	**glove F4**	**glove F5**	**glove F6**	**glove F7**	**glove F8**	**glove F9**	**glove F10**	**glove F11**	**glove F12**
Elsie	Brenda	Claudia	Patrica	Kylee	Laticia	Alejandra	Amina	Eldridge	Damion	Kevin	Gustavo
Carlotta	Katie	Tiara	Caren	Shaye	Jayci	Epifanio	Yair	Tad	Ronney	Ernest	Etienne
Elizabeth	Janette	Lena	Mikala	Tayla	Shalanda	Monalisa	Rani	Godfrey	Winford	Haley	Lorenzo
Dovie	Liza	Melina	Cherise	Latasha	Kalynn	Eulalia	Danial	Asa	Tavaris	Matt	Emil
Gladys	Debra	Sasha	Lorine	Jessi	Noelani	Alicea	Safa	Renard	Tylor	Gilbert	Roberto
+263	+396	+359	+889	+520	+1270	+395	+396	+434	+627	+429	+218
99% F	98% F	95% F	94% F	89% F	83% F	68% F	58% F	18% F	11% F	7% F	6% F
1972	1974	1987	1973	1987	1978	1985	1989	1979	1982	1979	1987
15% B	4% B	6% B	7% B	9% B	14% B	1% B	5% B	13% B	11% B	7% B	3% B
11% H	3% H	12% H	3% H	3% H	28% H	67% H	4% H	3% H	2% H	3% H	41% H
6% A	3% A	7% A	2% A	3% A	2% A	9% A	22% A	4% A	2% A	4% A	6% A
68% W	89% W	73% W	88% W	85% W	55% W	22% W	68% W	80% W	84% W	85% W	50% W

deb. F1	**deb. F2**	**deb. F3**	**deb. F4**	**deb. F5**	**deb. F6**	**deb. F7**	**deb. F8**	**deb. F9**	**deb. F10**	**deb. F11**	**deb. F12**
Denise	Kayla	Evelyn	Marquisha	Zoe	Kamal	Nicolas	Luis	Michal	Shaneka	Randall	Brian
Audrey	Lynae	Marquetta	Madalynn	Nana	Nailah	Carmella	Deisy	Astrid	Dondre	Scarlett	Ernie
Maryalice	Gabe	Gaylen	Celene	Crystal	Kalan	Adrien	Alexandro	Ezra	Laquanda	Windell	Matthew
Sonja	Tayla	Gaye	Nyasia	Georgiana	Aisha	Stefania	Elsa	Armen	Tavon	Corrin	Kenny
Glenna	Staci	Eula	Lanora	Sariyah	Rony	Raphael	Eliazar	Juliane	Tanesha	Coley	Wayne
+714	+845	+506	+819	+512	+334	+322	+538	+282	+688	+407	+313
99% F	81% F	80% F	78% F	71% F	62% F	59% F	56% F	54% F	49% F	29% F	5% F
1971	1989	1969	1984	1984	1991	1984	1986	1987	1983	1982	1974
4% B	4% B	17% B	5% B	10% B	6% B	6% B	1% B	2% B	49% B	9% B	5% B
3% H	3% H	6% H	3% H	9% H	5% H	16% H	72% H	6% H	3% H	3% H	3% H
3% A	2% A	4% A	3% A	11% A	32% A	5% A	8% A	3% A	2% A	4% A	5% A
89% W	91% W	72% W	89% W	70% W	56% W	73% W	18% W	88% W	45% W	83% W	87% W

w2v L1	**w2v L2**	**w2v L3**	**w2v L4**	**w2v L5**	**w2v L6**	**w2v L7**	**w2v L8**	**w2v L9**	**w2v L10**	**w2v L11**	**w2v L12**
Moser	Stein	Boyer	Romano	Murphy	Cantrell	Gauthier	Burgess	Gaines	Lal	Mendez	Yu
Persson	Zucker	Lasher	Klimas	Nagle	Wooddell	Medeiros	Willson	Derouen	Haddad	Aguillon	Tamashiro
Pagel	Avakian	Sawin	Pecoraro	Igoe	Maness	Lafrance	Hatton	Gaskins	Mensah	Aispuro	Heng
Runkel	Sobel	Stoudt	Arnone	Crosbie	Newcomb	Lounsbury	Mutch	Aubrey	Vora	Forero	Feng
Wagner	Tepper	Mcintire	Morreale	Dillon	Greathouse	Renard	Patten	Rodgers	Omer	Jurado	Nakamura
+3035	+775	+3013	+1416	+665	+2444	+756	+2818	+1779	+423	+1913	+393
1% B	2% B	3% B	1% B	4% B	8% B	8% B	12% B	34% B	15% B	1% B	1% B
2% H	3% H	2% H	6% H	3% H	2% H	4% H	3% H	3% H	7% H	80% H	3% H
1% A	1% A	1% A	1% A	1% A	1% A	1% A	1% A	1% A	28% A	5% A	79% A
94% W	93% W	92% W	91% W	90% W	86% W	85% W	81% W	60% W	46% W	12% W	11% W

Table A.1: The first name clusters from the `fast`, `glove` and `debiased` embeddings, followed by last name clusters from the `w2v` embedding. Demographic statistics (computed a posteriori) are also shown though were not used in generation, including percentage female (at birth), mean year of birth, and percentage Black, Hispanic, Asian/Pacific Islander, and White.

deb. F1	deb. F2	deb. F3	deb. F4	deb. F5	deb. F6	deb. F7	deb. F8	deb. F9	deb. F10
professor emeritus, registered nurse, adjunct professor	eighth grader, seventh grader, sixth grader	lifelong resident, postmaster, homemaker	granddaughter, grandson, daughter	bloke, chap, hubby	shopkeeper, villager, elder brother	mobster, chef, restaurateur	translator, interpreter, notary	mathematician, physicist, researcher	cousin, jailer, roommate
volunteering, homebound, nurse practitioner	seniors, eighth grade, boys	grandparents, aunts, elderly	graduated, grandchildren, siblings	bedtime, marital, bisexual	expatriate, hostels, postgraduate		undocumented, farmworkers, bilingual		blacks, academically, mentally r********
	medley, solo, trio	bluegrass, bandleader, banjo	trombone, percussionist, clarinet		artiste, verse, remix	maestro, accordion, operas	flamenco, tango, vibes	avant garde, violinist, techno	rapper, gospel, hip hop
	volleyball, softball, roping	bass fishing, rodeo, deer hunting	wearing helmet, horseback riding, snorkeling	racecourse, footy, footballing	cricket, badminton, cricketing	peloton, anti doping, gondola		luge, biathlon, chess	basketball, sprints, lifting weights
		rural, fairgrounds, tract	westbound, southbound, eastbound	foreshore, tenements, tourist attraction	slum, headquarter, minarets	seaside, boutiques, countryside	barangays, squatters, plazas	settlements, prefecture, inhabitants	
		supper, barbecue, chili	macaroni, green beans, pancakes		halal, sweets, hummus	pizzeria, mozzarella, pasta	tortillas, salsa, tequila	kosher, vodka, bagel	
					dirhams, emirate, riyals	euros, francs, vintages	peso, reais, nationalized	supervisory board, zloty, ruble	
		pastor, church, parish	baptized, sisters, brothers	mystical, witch, afterlife	fatwa, mosque, martyrs	nuns, papal, monastery		rabbis, synagogue, commune	
	captains, bridesmaids, grads	caretakers, grandmothers, superintendents	cousins, helpers, friends	punters, blokes, celebs	mediapersons, office bearers, shopkeepers				rappers, recruits, officers
			clan, overthrow, starvation		subcontinent, rulers, tribals		leftist, indigenous peoples, peasants	rightist, disengagement, oligarchs	civil rights, segregation, racial
					rupees, dinars, crores		pesos, remittances, cooperatives	shekels, rubles, kronor	
		convicted felon, felony convictions, probate	child endangerment, unlawful possession, vehicular homicide		chargesheet, absconding, petitioner	absentia, annulment, penitentiary			aggravated robbery, aggravated assault, felonious assault

Table A.2: The top-12 WEATs output by our UBE algorithm on the "debiased" `w2v` embedding of [88], again with $n = 12$. Despite being debiased, demographic statistics (again computed a posteriori) reveal names still cluster by gender, but the extreme gender clusters have many fewer statistically significant associations. For instance, the most male groups **deb. F11** and **deb. F12** are not shown because no significant associations were generated.

w2v L1	**w2v L2**	**w2v L3**	**w2v L4**	**w2v L5**	**w2v L6**	**w2v L7**	**w2v L8**	**w2v L9**	**w2v L10**	**w2v L11**	**w2v L12**
potato salad, pretzels, chocolate cake	kosher, bagel, hummus	pumpkin, brownies, donuts	mozzarella, pasta, deli	pint, whiskey, cheddar	pecans, grits, watermelon	maple syrup, syrup, foie gras	cider, lager, malt	fried chicken, crawfish, sweet potatoes	sweets, saffron, mango	tortillas, salsa, tequila	noodles, dumplings, soy sauce
concentration camp, extermination, postwar	disengagement, neocons, intifada			unionists, sectarian, pedophiles		province, separatist, sovereignty	antisocial behavior, cricket, asylum seekers	blacks, segregation, civil rights	non governmental, miscreants, encroachments	drug traffickers, leftist, undocumented	hyun, bian, motherland
	co founder, venture capitalist, psychotherapist	assessor, wildlife biologist, secretary treasurer	restaurateur, plumber, firefighter	solicitor, selector, handicapper	jailer, rancher, appraiser		schoolboy, barrister, chap	cheerleader, bailiff, recruiter	shopkeeper, aspirant, taxi driver	translator, smuggler, interpreter	villager, vice, housewife
	synagogues, skyscraper, studio	log cabin, zoning ordinance, barn	pizzeria, borough, firehouse	pubs, racecourse, western suburbs	fairgrounds, acre tract, concession stand	rink, cottage, chalet	disused, derelict, leisure		locality, slum, hostel		prefecture, guesthouse, metropolis
	authors, hedgefund managers, creators	crafters, hobbyists, racers	mobsters, restaurateurs, captains	gardai, lads, footballers	sheriffs, folks, appraisers	skaters, premiers, mushers	blokes, householders, solicitors		mediapersons, newsmen, office bearers		migrant workers, maids, civil servants
	rabbis, synagogue, biblical		papal, pontiff, convent	archdiocese, clerical, diocese	denomination, pastor, church		vicar, creationism, traditionalists	pulpit, preaching, preach	fatwa, fasting, sufferings	rosary, parish priest, patron saint	commune, monks, temples
	shekels, settlements, nonprofit	mill levy, assessed valuation, tax abatement			millage, payday lenders, appropriations		unfair dismissal, attendances, takings		rupees, lakhs, dirhams	pesos, remittances, indigent	baht, overseas, income earners
	pollster, liberal, moderates				commissioners, countywide, statewide	ridings, selectmen, byelection		desegregation, uncommitted, voter registration	panchayat, candidature, localities	barangay, immigration reform, congresswoman	plenary session, landslide, multiracial
	insider trading, attorneys, lawsuit	felonious assault, drug paraphernalia, criminal mischief			sheriff, meth lab, jailers	impaired driving, criminal negligence, penitentiary	affray, bailiffs, aggravated burglary	aggravated robbery, racially charged, probation violation	absconding, chargesheet, complainant	illegal immigrant, drug trafficking, deadly weapon	
						loonie, francs, takeovers	sharemarket, credit crunch, gilts		load shedding, microfinance, rupee	peso, reais, nationalization	cross strait, yuan, ringgit
walleye, lakes, aquarium	transatlantic, iceberg, flotilla				crappie, bass fishing, boat ramp			shad, barrier islands, grouper	mangroves, jetty, kite	sardines, tuna, archipelago	mainland, seaweed, island
feedlot, barley, wheat		cornfield, pumpkins, alfalfa			mowing, deer hunting, pasture				agro, saplings, livelihood	farmworkers, coca, sugarcane	bamboo, cassava, palm oil

Table A.3: The top-12 WEATs output by our UBE algorithm on the `w2v` embedding for *last names*. The corresponding name groups are presented in Table A.1.

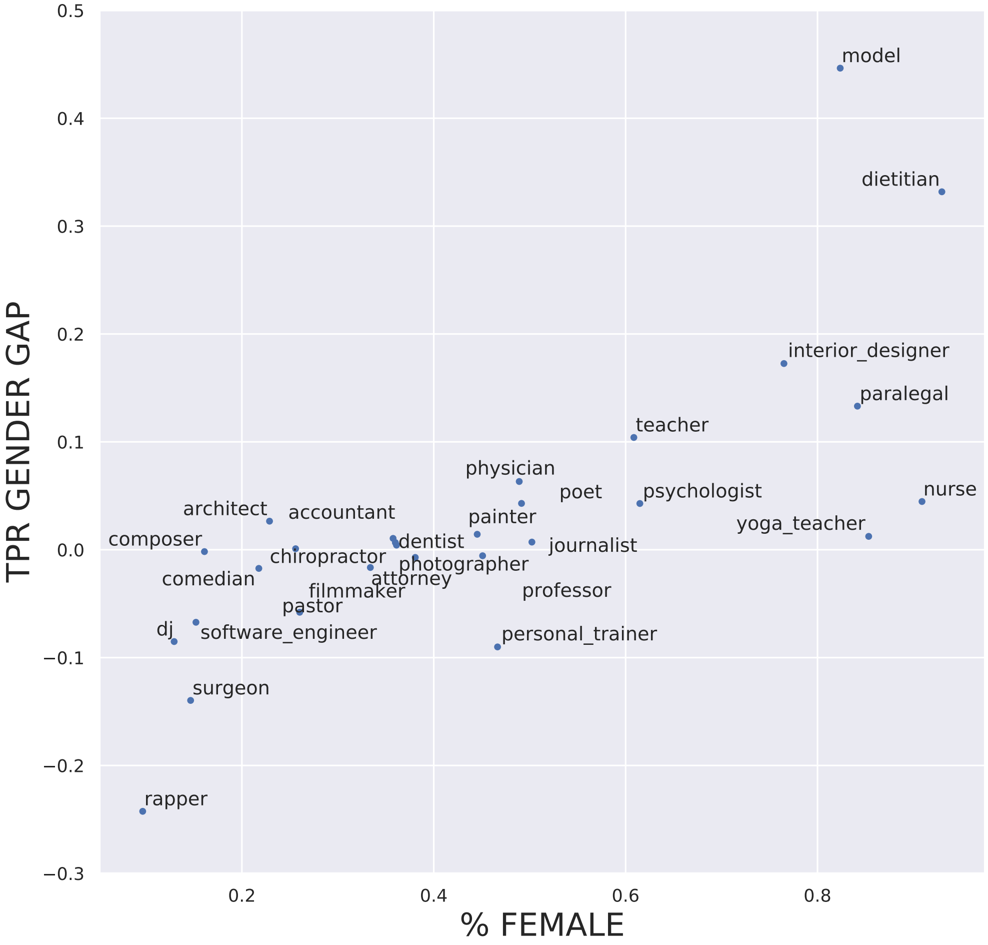


Figure A.2: Gender gap per occupation vs. % females in occupation for BOW trained without gender indicators.

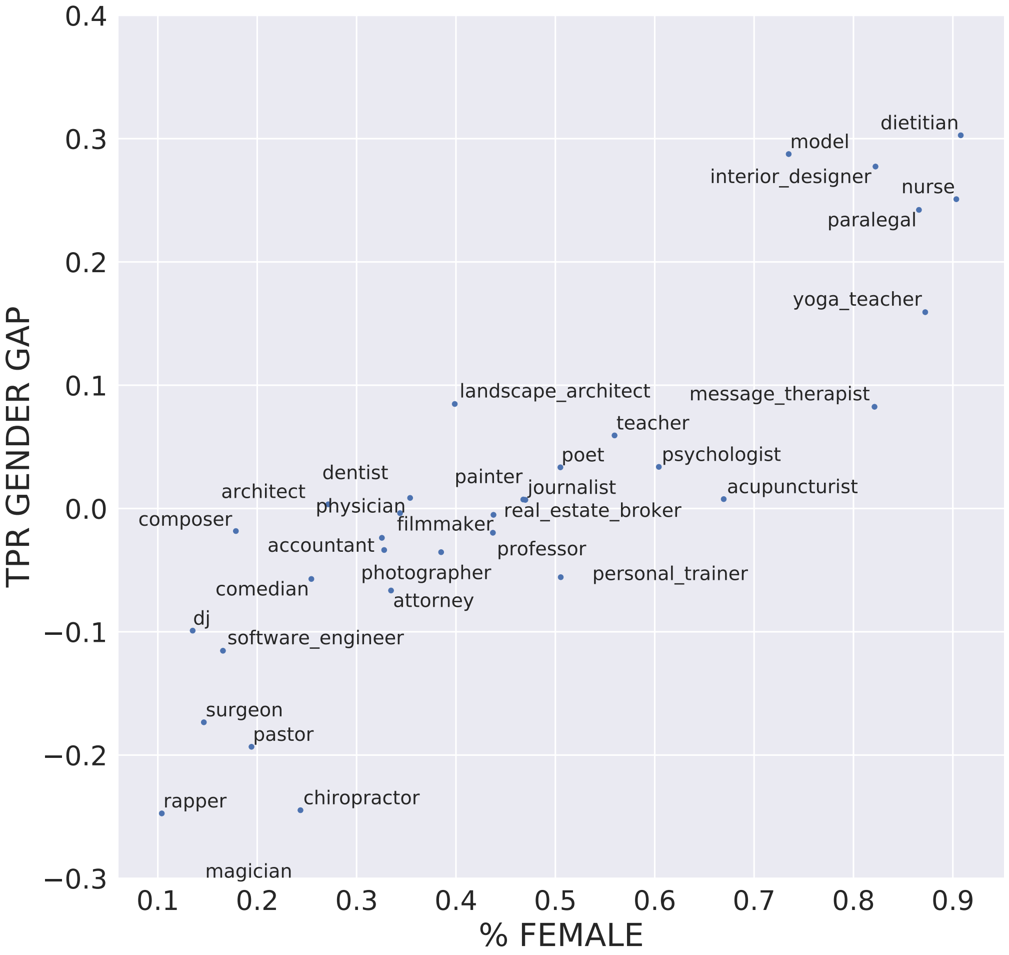


Figure A.3: Gender gap per occupation vs. % females in occupation for WE trained with gender indicators.

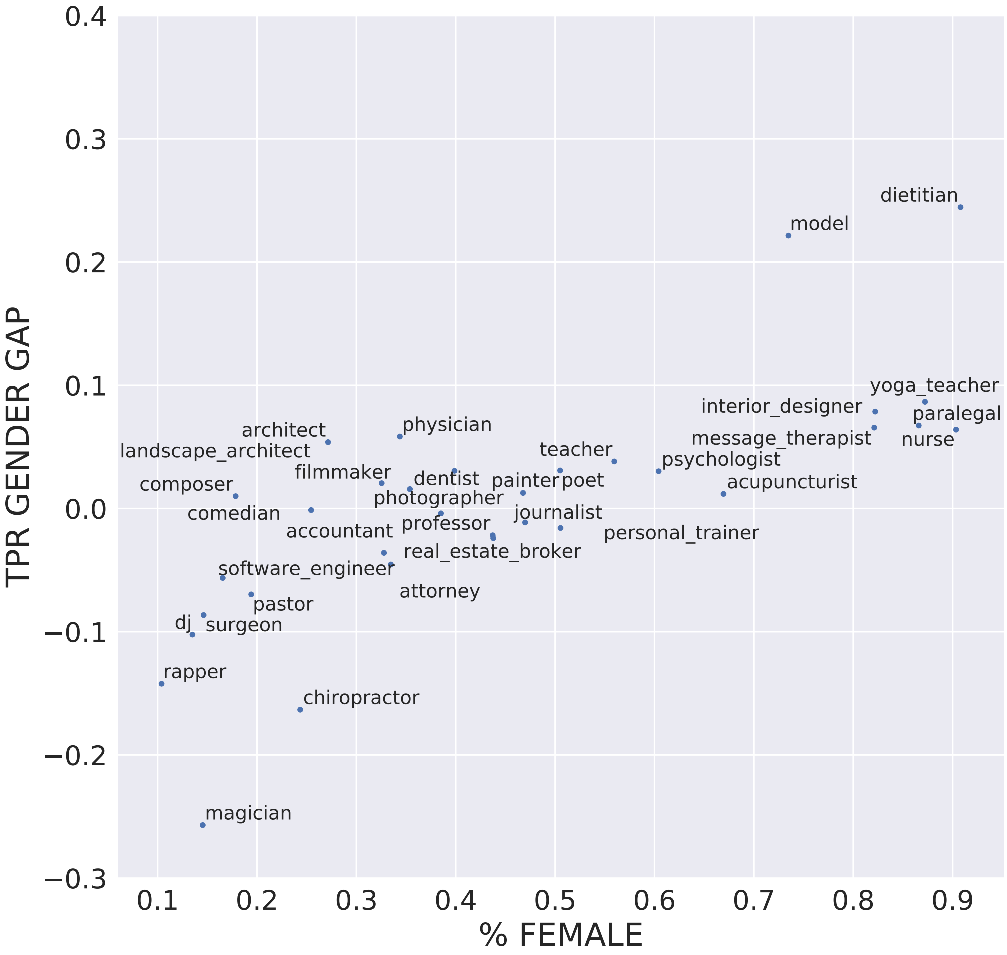


Figure A.4: Gender gap per occupation vs. % females in occupation for WE trained without gender indicators.

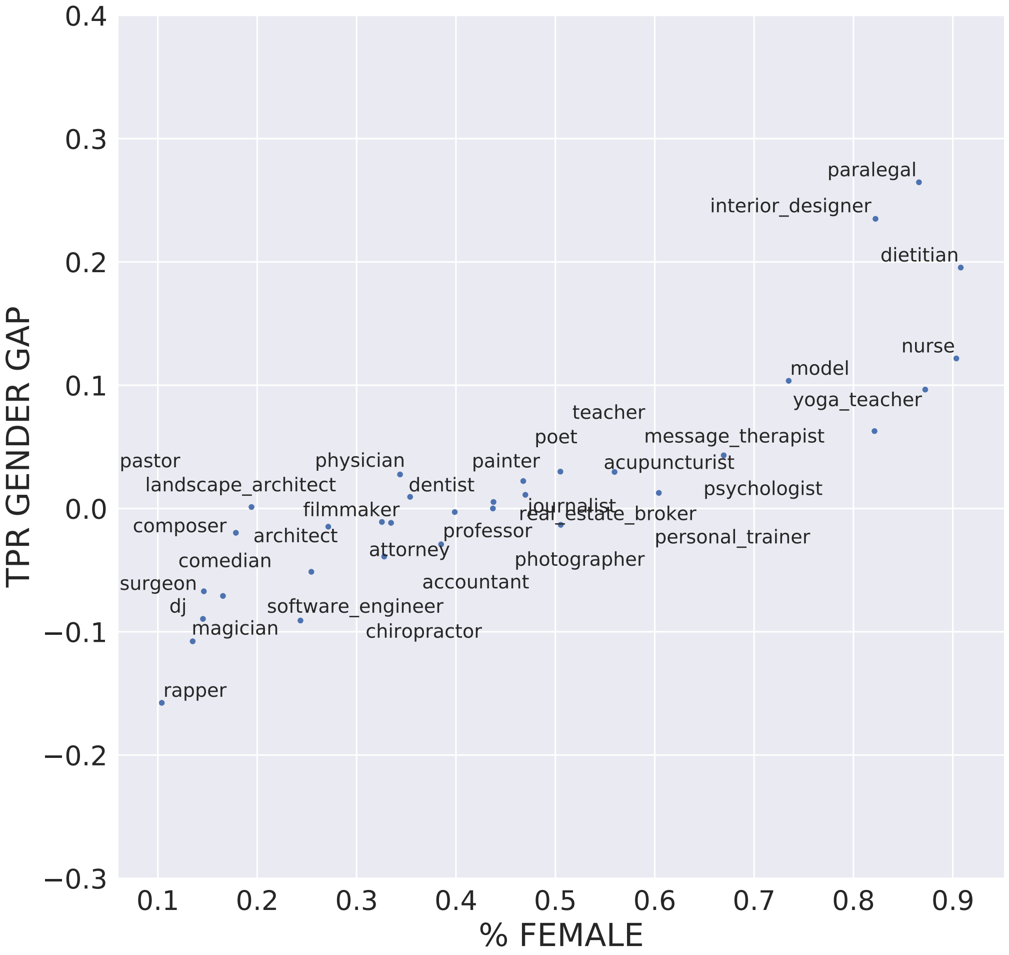


Figure A.5: Gender gap per occupation vs. % females in occupation for DNN trained with gender indicators.

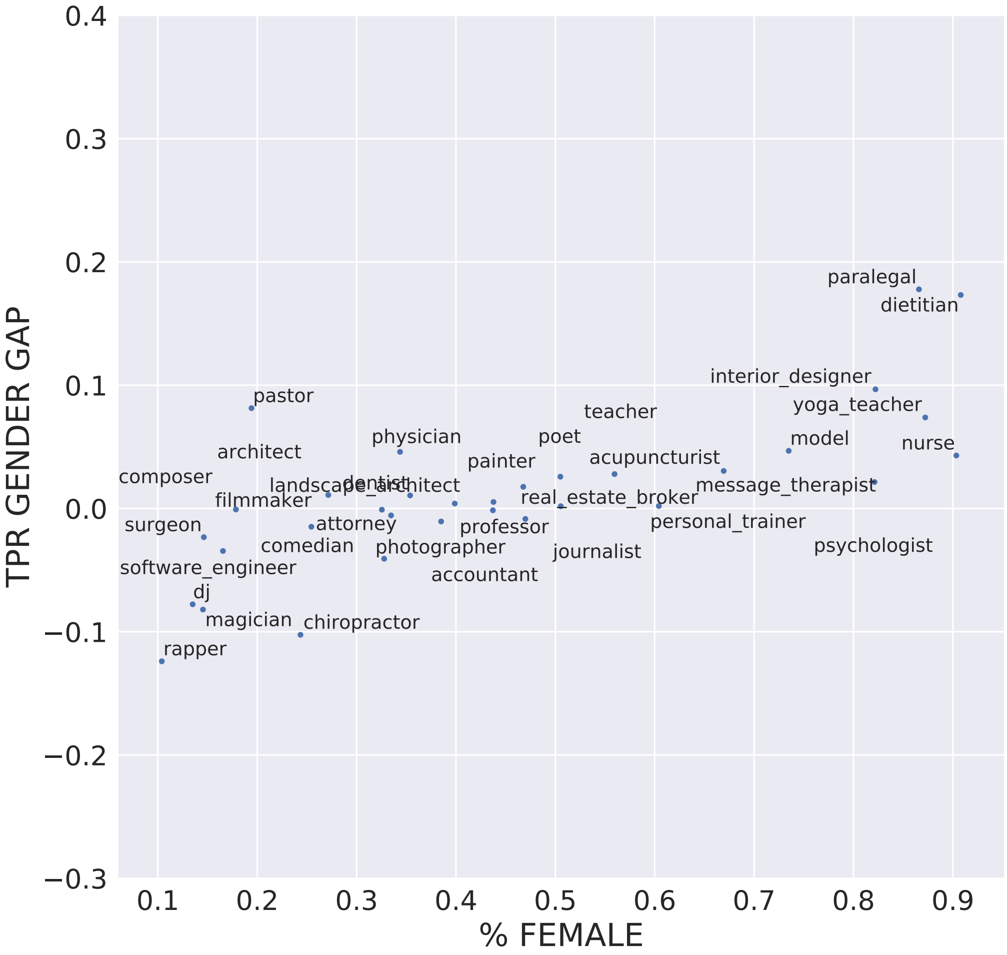
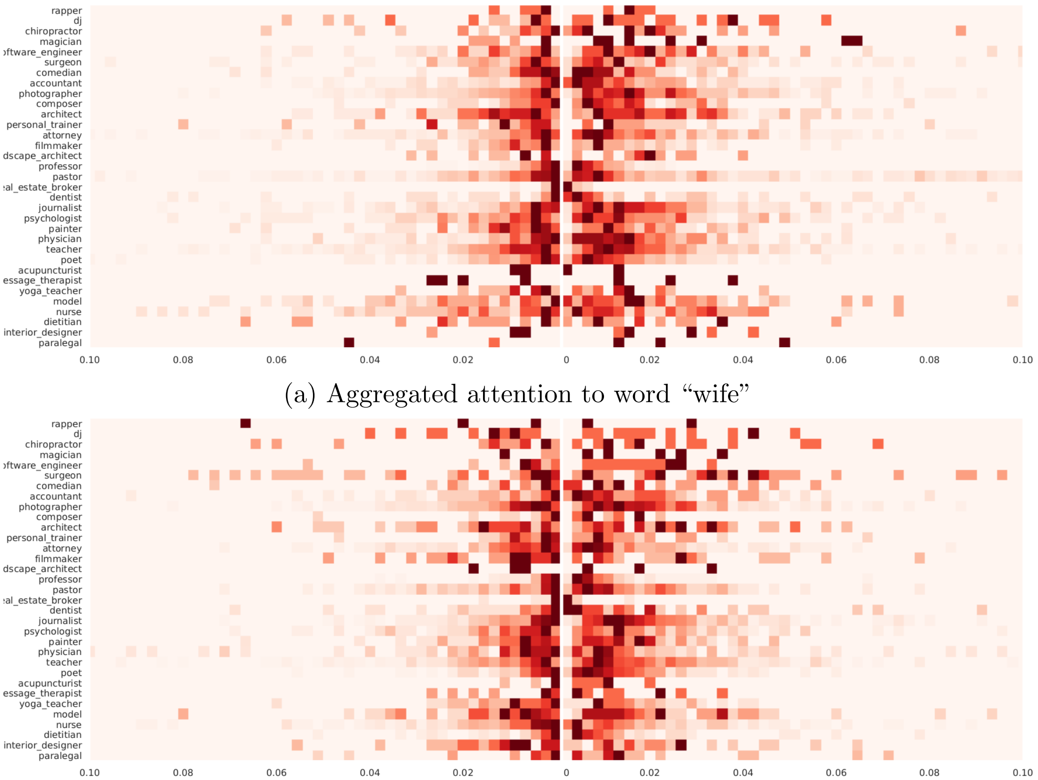


Figure A.6: Gender gap per occupation vs. % females in occupation for DNN trained without gender indicators.



(a) Aggregated attention to word "wife"

(b) Aggregated attention to word "husband"

Figure A.7: Aggregated attention of DNN to words "wife" (A.7a) and "husband" (A.7b). In the left, results when model trained with gender indicators. In the right, results when model trained without gender indicators.

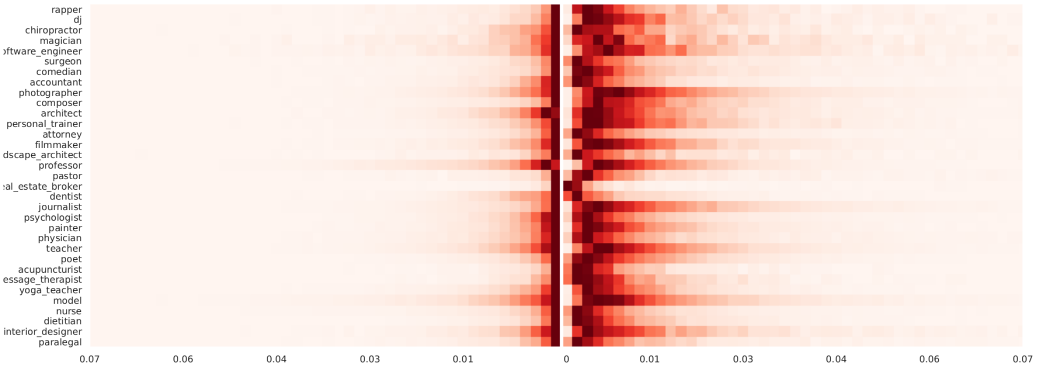


Figure A.8: Aggregated attention of DNN to word "she". In the left, results when model trained with gender indicators. In the right, results when model trained without gender indicators.

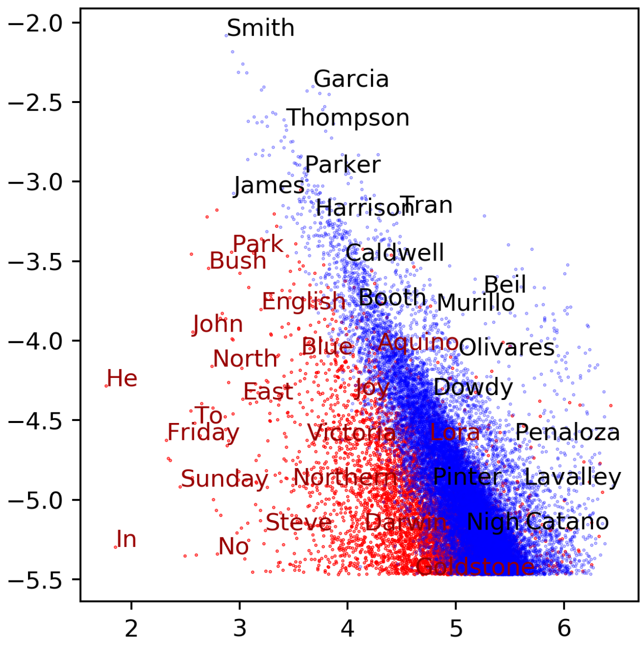


Figure A.9: A plot of log-probability (y-axis) vs. word embedding index (x-axis) for the last name data and the word2vec word embedding. Orange points represent last names we keep and blue points are outliers we remove. As expected from Zipf's law, the probabilities and frequencies exhibit a power-law relationship. Names removed from the data by our classifier, displayed in red, are typically words that have other more common uses than as last names.