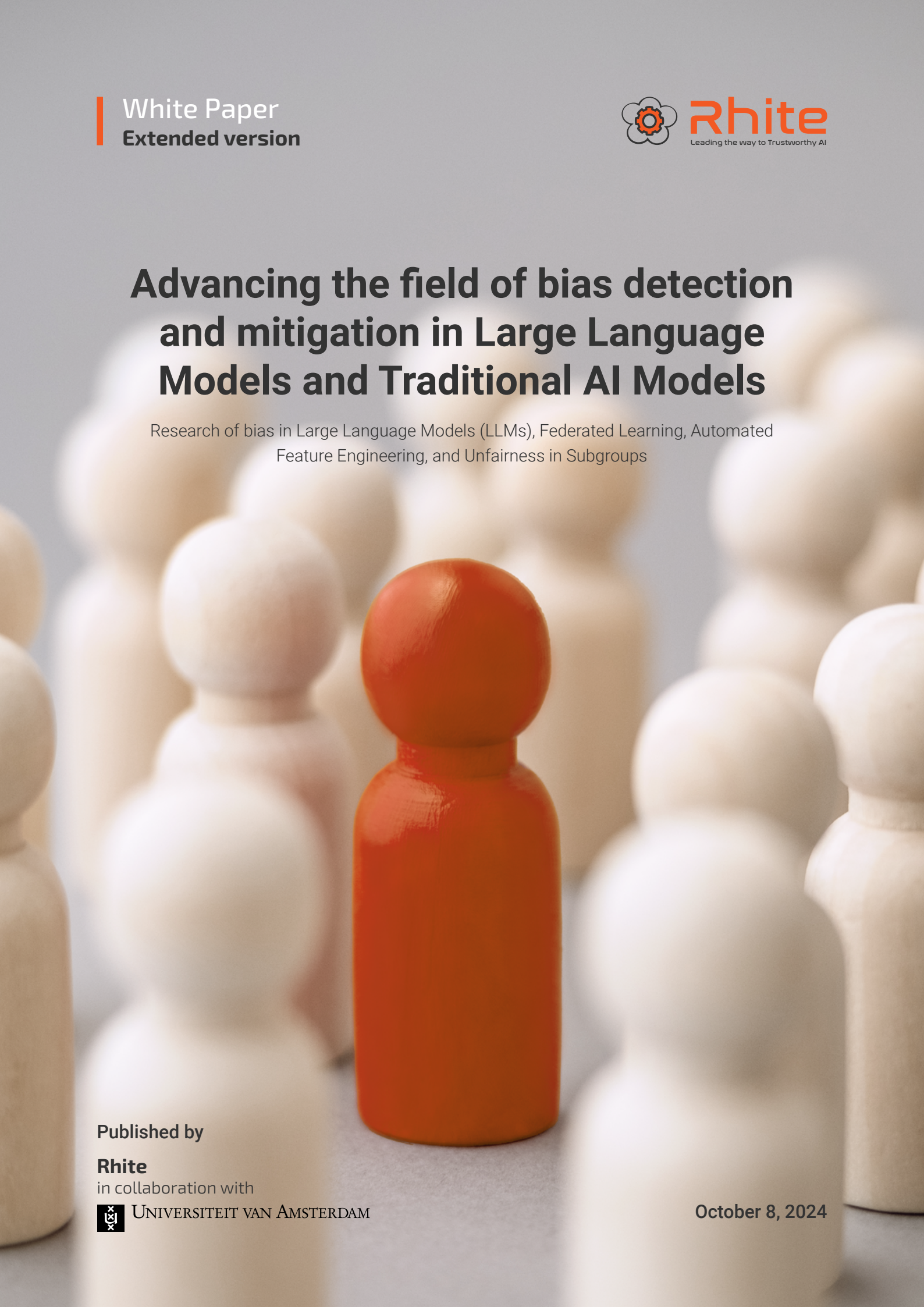
White Paper
**Extended version**

Rhite
Leading the way to Trustworthy AI

# Advancing the field of bias detection and mitigation in Large Language Models and Traditional AI Models

Research of bias in Large Language Models (LLMs), Federated Learning, Automated Feature Engineering, and Unfairness in Subgroups



**Published by**

**Rhite**
in collaboration with
Universiteit van Amsterdam

October 8, 2024

## WE STAND BY
# RESPONSIBLE
## INNOVATION

At Rhite, we foster a collaborative environment that thrives on knowledge exchange and pioneering research. We strongly advocate for the responsible development of new technologies, dedicating significant resources to exploring how to make Trustworthy AI technically achievable.

# WHY
## *THIS* PAPER?

At Rhite, we believe that addressing bias in AI is essential not only for creating fair and responsible technology but also for building trust in AI across industries and communities. We are committed to advancing the understanding of bias detection and mitigation through rigorous research, collaboration, and transparency. This white paper represents a key step in that mission, offering valuable insights and innovative approaches to both practitioners and researchers.
Here's why we've dedicated our efforts to this project:

**Impact on society**
AI systems are increasingly influencing decisions in critical areas like hiring, healthcare, and finance. Ensuring these systems are fair and unbiased is essential to prevent harmful outcomes for individuals and communities.

**Bridging the knowledge gap**
There is a significant lack of real-world understanding regarding how to effectively detect and mitigate bias in AI systems. This white paper seeks to fill that gap by providing actionable insights and guidance for professionals and industries.

**Advancing Responsible AI**
As powerful technologies like LLMs and Federated Learning continue to emerge, staying ahead of the curve in bias mitigation is vital. This white paper introduces novel methods that pave the way for new directions in ethical AI development.

# ABOUT US

Rhite
Leading the way to Trustworthy AI

Rhite helps you navigate the technical and legal aspects of AI while managing risks, minimizing adverse impact, and achieving compliance. Our technical and legal consultancy spans the whole journey: whether it's developing cutting-edge tools, making informed procurement decisions, or navigating usage choices.

## Our expertise

**We offer a unique blend of technical know-how and legal expertise in AI.** Rhite's experienced advisors adopt a holistic, risk-based approach to guide you through the process of ensuring ethical and regulatory compliance.

### WHAT WE DO

- Legal and technical consultancy on AI;
- Guidance to comply with the requirements of the EU AI Act;
- Auditing of algorithms and AI systems;
- Privacy, security, safety and fundamental rights Impact assessments of AI solutions;
- Bespoke trainings on AI Risk Management;
- Implementation of Responsible AI programs.

### HOW WE DO IT

**RHITE** is an acronym representing the principles we believe should underpin the design, development, and use of AI:

- **Responsible**
- **Humane**
- **Ingenious**
- **Transparent**
- **Empathic**



## A holistic approach towards Trustworthy AI


## Our founders

### Isabel Barberá

Co-founder | AI advisor & Privacy Engineer

With a multidisciplinary background in privacy and security, engineering, AI, law and ethics, she guides organisations in the design and implementation of responsible digital solutions. She is an advocate of Trustworthy AI by design and passionate about the protection of human rights.

### Martijn Korse

Co-founder | Privacy & Security Engineer

Martijn has a long career in the field of software engineering, DevSecOps and cybersecurity. Besides that, he also has a background in psychology and philosophy. Like Isabel, Martijn has a passion for privacy and security by design and he is also a strong advocate of responsible human-centered design.

Learn more about us on our website!
**www.rhite.tech**

# Abstract

Bias in AI systems is a well-known yet persistent issue that presents significant risks across diverse applications, such as AI-based hiring, medical diagnostics, financial risk modeling, and workflow automation systems. The emergence of Large Language Models (LLMs) has revolutionized natural language processing (NLP), powering tools like chatbots, translators, and content generation platforms. However, despite their benefits and powerful capabilities, LLMs are also prone to various forms of bias — ranging from gender and racial to ideological and cultural biases. This white paper presents **six focused studies** aimed at addressing bias and fairness in AI systems (see below). Together, these studies highlight the strengths and limitations of existing bias detection techniques, while also introducing novel approaches that open the way for further research.

Study #1
**TOPIC**
### AI-Based Hiring
Applying post-processing bias assessment techniques to explore whether LLMs mitigate or amplify bias in hiring decisions.

Study #2
**TOPIC**
### Unveiling Bias Mechanisms in LLMs
Identifying novel in-processing techniques to examine how bias is encoded in LLMs, offering advanced methods for more effective bias identification and mitigation.

Study #3
**TOPIC**
### Gender Bias in LLMs
Comparing in-processing and post-processing techniques to assess gender bias.

Study #4
**TOPIC**
### Bias Assessment in Federated Learning
Exploring the balance between privacy and fairness in decentralized systems, which is especially crucial in sectors handling sensitive data.

Study #5
**TOPIC**
### Causal Fairness Analysis with Automated Feature Engineering
Finding novel causality-driven approaches to improve bias mitigation. Exploring how causal factors, rather than correlations, can better address bias in fields like law enforcement and healthcare.

Study #6
**TOPIC**
### Profile-Based Subgroup Discovery (PSD)
Providing a new method for uncovering hidden biases within subgroups, providing a more detailed perspective on fairness, particularly in credit scoring use cases.

With this white paper, we at Rhite reaffirm our commitment to advancing research in bias detection and mitigation, contributing to the development of more fair and equitable AI systems. This white paper reflects our dedication to these efforts. The code used in the six studies underlying this white paper is made publicly available on GitHub. We are also offering the community three synthetic datasets (one balanced and two biased) containing résumés with sensitive attributes like gender and ethnicity, along with labels indicating each candidate's suitability for a profession. These datasets are available in CSV format and cover a wide range of personal and professional information typically found in résumés.

# Bias detection and mitigation in LLMs

Traditional AI models like decision trees, logistic regression, and support vector machines have been extensively studied for bias detection and mitigation. They often rely on fairness metrics such as demographic parity and equalized odds and use strategies like pre-processing (modifying training data), in-processing (adjusting learning algorithms), or post-processing (correcting outcomes) to tackle bias. In contrast, LLMs like GPT-4, BERT, and LLaMA, while highly capable in natural language tasks, are far more complex, making bias detection and mitigation significantly more challenging due to their high-dimensional nature and the subtle ways in which bias is embedded in their representations. In the past year, multiple studies have revealed gender and racial biases in models like BERT and GPT based systems.
While in traditional AI models bias is linked to specific features and are easier to detect, LLMs require advanced techniques to uncover and address biases. Although research on LLM bias is emerging, established fairness tools for LLMs are lacking, unlike in traditional models which benefit from robust toolkits. This highlights the need for continued research and development of effective bias mitigation strategies for LLMs.

Through the various studies that were conducted within this research, we explored bias detection in LLMs with **three different approaches**.

## Post-processing bias techniques

Based on Study #1

Full research: **AI-Based Hiring and the Appeal of Novelty: Do LLMs Solve or Exacerbate the Problem of Discrimination?**
Researcher: **Alexia Muresan (UvA)**
Supervisors: **Leonard Bereska, MSc (UvA)**, **Isabel Barberá (Rhite)**

Models: **LLMs (BERT and GPT-3.5 Turbo), Support Vector Classifier (SVC), Logistic Regression (LR), Gradient Boosting (GB) and Random Forest (RF)**
Datasets: **Three synthetic datasets**

This research aims to assess whether transitioning to **LLMs** for hiring decisions offers improvements in fairness and performance compared to traditional AI models. If LLMs do not provide significant benefits in terms of performance, efficiency, or fairness, focusing on mitigation strategies may not be necessary. However, a comprehensive comparison between LLMs and traditional AI models in hiring contexts has not yet been conducted.
The study addresses this gap by comparing traditional machine learning models and LLMs for **résumé classification**, focusing on bias and fairness. It explores key research questions such as how the models compare in terms of bias, their robustness to biased training data in hiring scenarios, and whether they contain inherent bias unrelated to their training data. Due to the lack of available data that met the specific requirements of this study, three synthetic datasets were generated. The first dataset was designed to be completely free of discriminatory bias, ensuring a balanced representation of gender and ethnicity. The other two datasets were derived from this balanced dataset by intentionally introducing bias, gender bias (second dataset) and ethnicity bias (third dataset).

"This research provides better guidance to industries in the field of Human Resources (HR), where fairness in automated decision-making is vital for preventing discrimination. With AI increasingly integrated in hiring applications, understanding whether LLMs help or worsen bias is crucial."

## In-processing bias techniques

Based on Study #2

Full research: **Unveiling the Mechanisms of Bias in Large Language Models by Eliciting Latent Knowledge**
Researcher: **Tarmo Pungas (UvA)**
Supervisors: **Leonard Bereska, MSc (UvA)**, **Isabel Barberá (Rhite)**

Models: **Llama 13B, Llama 3 8B and Llama 3 70B**
Datasets: **StereoSet, CrowS-Pairs, Disambiguation datasets**
Bias Assessments methods: **PCA, Patching, Probing intervention and Probe generalization**



Despite extensive research aimed at detecting and mitigating biases that LLMs exhibit, we still lack a comprehensive understanding of how LLMs encode bias. By leveraging knowledge-eliciting techniques, this study aims to bridge that gap by identifying and manipulating bias directions within model activations. Successfully doing so could pave the way for more effective bias mitigation strategies. The key research questions driving this study are: 1) How can knowledge-eliciting techniques be utilized to identify and understand bias manifestations in LLMs? 2) What are the implications of these mechanisms for the development of more effective bias mitigation strategies? This research hypothesizes the existence of a specific bias direction within LLMs and aims to explore how identifying and adjusting this direction could influence the model's output.

"We focused on this research to explore advanced methods of how bias is encoded and can be manipulated within LLMs at a more technical level, offering industries innovative ways to directly address bias in their AI systems when using LLMs."

## Comparing in-processing and post-processing techniques to assess gender bias

Based on Study #3

Full research: **Assessing and Addressing Gender Bias in Large Language Models**
Researcher: **Dennis Agafonov (UvA)**
Supervisors: **Dr G. Sileno (UvA)**, **Isabel Barberá (Rhite)**

Models: **BLOOM- series LLMs**
Datasets: **Five variations of Tweets**
Bias Assessments methods: **Seat, Disco, CSPS, and Sentiment Analysis**

Using established taxonomies, this research categorizes bias assessment methods for LLMs into three groups: probability-based, embedding-based, and output text-based methods. These methods offer distinct approaches to measuring bias in LLMs, from token probabilities and internal embeddings to sentiment analysis in generated text.
This research focuses on assessing gender bias in autoregressive LLMs, which are extensively used in various applications, including the well-known GPT-series. The study specifically targets four variants of the BLOOM-series LLMs, chosen for their open-source nature, which offers greater accessibility and flexibility for research compared to proprietary models like GPT-3 and GPT-4. To achieve a comprehensive evaluation, four distinct bias assessment methods were selected and, where necessary, adapted to ensure compatibility with autoregressive LLMs. Each method was chosen for its unique approach to quantifying gender bias, allowing for a more holistic and nuanced analysis.

"We focused on this research to deepen our understanding of how gender bias manifests in LLMs, aiming to guide industries with the tools to mitigate these biases in applications like chatbots and automated customer service."

# Bias detection and mitigation in traditional AI Models

Bias remains a critical concern in traditional AI models. Our research tackles these challenges by exploring decentralized systems that balance privacy with fairness, investigating causal factors behind bias, and discovering hidden biases within subgroups. Through these studies, we aim to shed light on the limitations of current methods and explore new ways to enhance fairness in AI applications.

## Bias detection in the Development phase - Aggregation bias

Based on Study #4

Full research: **Bridging Fairness and Privacy: Bias Assessment in Federated Learning**
Researcher: **Jelke Matthijssen (UvA)**
Supervisors: **Dr G. Sileno (UvA)**, **Isabel Barberá (Rhite)**

Federated Learning Framework: **Flower**
Dataset: **ACS PUMS dataset**

To effectively assess bias in Federated Learning, new methods must be developed that detect bias without compromising local data privacy. Current research has proposed an aggregated local bias assessment technique that combines local bias scores using the same aggregation method used for model aggregation (Ezzeldin et al., 2023; Zhang et al., 2020). However, this method lacks theoretical foundation and comprehensive experimental validation. This research aims to analyse bias assessment techniques within Federated Learning, focusing on evaluating the accuracy of the privacy-preserving aggregated local bias assessment and comparing bias in federated models to that in centrally trained models. Additionally, it investigates how client heterogeneity affects bias by experimenting with different types and amounts of data diversity among clients.

"We chose this research to explore the trade-offs between maintaining user privacy and mitigating bias, as well as to investigate the effects of bias in decentralized systems. This is critical for industries that handle sensitive personal data, such as healthcare and finance, where fairness and privacy must both be ensured."

## Bias detection in Data Understanding and Preparation phase - Proxies and Subgroups

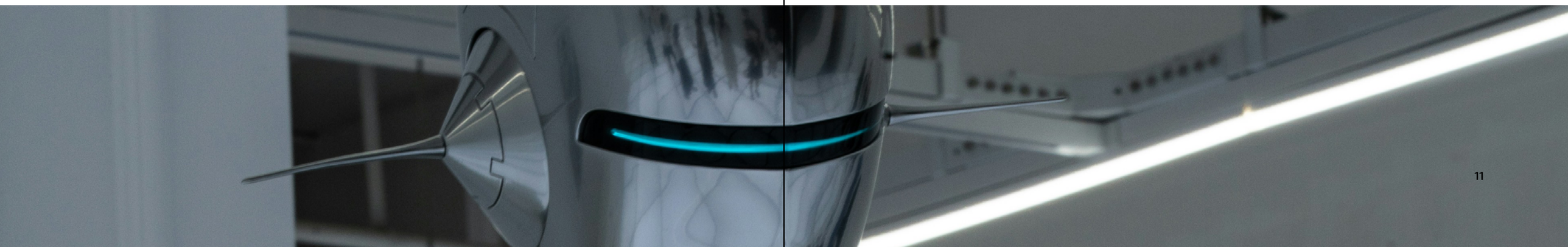
Based on Study #5

Full research: **Causal Fairness Analysis with Automated Feature Engineering**
Researcher: **Wietse van Kooten (UvA)**
Supervisors: **Dr E. Acar (UvA)**, **Isabel Barberá (Rhite)**

Models: **Structural Causal Model (SCM) and Standard Fairness Model (SFM)**
Dataset: **COMPAS**

Causal inference aims to understand how changes in one variable influence another using **Structural Causal Models (SCMs)**. These models help calculate potential outcomes and counterfactuals, which are essential for determining path-specific effects such as direct, indirect, and spurious effects. In **causal fairness analysis**, these effects are decomposed to assess their impact on fairness.
The **Standard Fairness Model (SFM)** is a key tool used to represent causal diagrams and identify biases. For instance, in a hiring decision context, education might have a direct effect on hiring, while prior job performance acts as a mediator, and socio-economic background serves as a confounder, creating potential spurious effects on the education-hiring relationship due to systemic biases.
Automated Feature Engineering is the process of creating new features from existing data to improve model performance and interpretability, particularly useful for detecting trends across subgroups, addressing issues like Simpson's paradox. This research applies automated feature engineering within the SFM to enhance fairness. The experiments use the **COMPAS dataset**, which predicts the likelihood of recidivism, and focus on two scenarios: automated feature engineering on mediators alone and automated feature engineering on both mediators and confounding variables. Our research demonstrates how Automated Feature Engineering can be effective in improving fairness within causal fairness frameworks.

"We chose this research to show how causal relationships can improve both fairness and accuracy in AI models. We believe that measuring causality rather than just correlation is a critical advancement in understanding the true sources of bias. This is important because addressing causal factors allows for more precise bias mitigation, especially in sectors like law enforcement and healthcare, where decisions have significant real-world impacts."



### Bias detection in Data Understanding and Preparation phase - Proxies and Subgroups

Based on Study #6

Full research: **Profile-based subgroup discovery for Fairness Analysis**
Researcher: **Dionne Gantzert (UvA)**
Supervisors: **Dr G. Sileno (UvA)**, **Isabel Barberá (Rhite)**

Models: **Logistic Regression (LR), XGBoost Classifier**
Dataset: **German Credit Risk**
Bias Assessment Methodology: **Profile-based Subgroup Discovery (PSD)**

This research proposes a novel clustering method designed to generate simple and interpretable clusters for subgroup discovery, called Profile-based Subgroup Discovery (PSD), based on previous semi-hierarchical methods for profile extraction. Our methodology involves two steps: first, partitioning the data space based on the target variable and then applying iterative clustering to obtain profiles; second, extracting descriptive rules from these profiles to identify subgroups. Like other Clustering Subgroup Discovery and Subgroup Discovery techniques, PSD relies on discriminative decision rules that can be applied in real-world applications. Our method stands out by integrating the target variable into the clustering process, aligning it closely with subgroup discovery techniques. Our research aims to enhance the understanding of biased relationships within data by discovering subgroups unfairly treated by classifiers. We focus on two aspects: identifying subgroups exhibiting gender bias and identifying subgroups showing bias in general, regardless of sensitive attributes such as gender. Our approach was tested on the well-known German Credit dataset in the context of credit scoring.

> "We chose this research to address the limitations of traditional fairness metrics, which often overlook bias within subgroups. PSD is a methodology that helps uncover these hidden biases offering a granular approach to fairness in AI, essential for equitable decision-making in industries such as credit scoring."

## Conclusions

Combating bias in LLMs and traditional AI models is an ongoing challenge that requires continuous research, innovation and collaboration. The findings of this white paper underscore the importance of **selecting the right tools and strategies** based on specific use cases and bias types. As AI continues to evolve, so too must our approaches to ensuring fairness and equity in these systems. Continued collaboration between academia and industry, along with a commitment to ethical AI practices, will be essential in driving progress and fostering trust in AI technologies.
We invite the AI community, researchers, and developers to help advance the important work of bias detection and mitigation in AI systems. At Rhite, we are committed to an open-source vision. We encourage you to explore and contribute to our **GitHub library**, which contains a growing collection of bias detection code and techniques aimed at enhancing fairness in AI. Whether you're refining existing models, suggesting new features, or developing entirely new approaches, your input is invaluable. Together, we can ensure that bias detection tools are not only effective but accessible to everyone.

WANT TO KNOW MORE?

# KEEP READING!

You've just reached the end of the project overview.
In the next part of the white paper, we'll take a closer look at the research that shaped the work presented here. This section will give you more insight into the process, the data, and the steps taken to reach the findings, helping to paint a fuller picture of the work behind the results.

# DETAILED RESEARCH

# Table of Contents

# Introduction

## BACKGROUND AND PROBLEM STATEMENT

Bias in AI systems is one of the most critical challenges facing the adoption of artificial intelligence across various industries. These biases can lead to unfair and discriminatory outcomes, adversely affecting decision-making processes in areas such as hiring, credit scoring, healthcare, and law enforcement. Despite significant attention and numerous studies, addressing bias remains a complex and unresolved issue.

Bias can arise from various sources, including biased training data, flawed algorithms, and unintentional human biases introduced during model development. This bias can manifest in several ways leading to unfair outcomes for certain demographic groups.

The presence of bias in AI systems can lead to serious consequences, as illustrated by the use cases explored in this white paper:

- **Financial Sector**: Biased models in credit scoring can lead to unfair loan approvals, disadvantaging specific demographic groups.
- **Healthcare**: Bias in predictive models can result in unequal treatment recommendations, exacerbating health disparities.
- **Hiring**: AI-based hiring tools can reinforce existing workplace biases, limiting opportunities for underrepresented groups.
- **Law Enforcement**: Predictive policing models may disproportionately target minority communities, perpetuating systemic inequalities.

## PURPOSE AND OBJECTIVES

This white paper aims to provide a comprehensive analysis of the current state of bias detection tools, offering insights into their applicability and effectiveness across different scenarios. By evaluating various tools and methodologies, we aim to guide industry stakeholders in selecting and implementing the most appropriate solutions to address bias in their specific AI use cases.

## SCOPE

The whitepaper covers:

- **An overview of existing bias detection tools**
- **Classification of these tools based on their applicability to different use cases**
- **Detailed analysis of specific methodologies, including those focused on LLMs and on traditional AI models**
- **Practical recommendations for implementing bias detection and mitigation strategies**



# Research Methodology

## COLLABORATION WITH UNIVERSITY OF AMSTERDAM

This research was conducted in collaboration with the University of Amsterdam, involving six students from their master's program in Artificial Intelligence. The students researched different bias identification and mitigation methodologies, contributing to the comprehensive evaluation presented in this white paper.

## RESEARCH PHASES

The research was conducted in two main phases:

- **Information Gathering**: This phase involved creating an overview of existing bias detection tools and reviewing academic and industry research on their effectiveness.
- **Practical Testing**: In this phase, selected tools were tested based on different use cases to evaluate their performance in real-world scenarios.

PART **1**

# Bias Detection and Mitigation in Large Language Models (LLMs)

**PROJECTS** IN THIS CHAPTER:

- **AI-Based Hiring and the Appeal of Novelty: Do LLMs Solve or Exacerbate the Problem of Discrimination?** by Alexia Muresan
- **Unveiling the Mechanisms of Bias in LLMs by Eliciting Latent Knowledge** by Tarmo Pungas
- **Assessing and Addressing Gender Bias in Large Language Models** by Dennis Agafonov

Post-processing bias techniques

## AI-Based Hiring and the Appeal of Novelty: Do LLMs Solve or Exacerbate the Problem of Discrimination?

Researcher: Alexia Muresan

Link to research: AI-Based Hiring LLMs and Traditional AI.pdf

### INTRODUCTION

As AI continues to dominate and become increasingly integral across various professional fields, it is vital to understand and manage its impact, particularly its potential social ramifications. A key example of AI's growing influence is its extensive use in automated hiring practices, which offer significant cost and time efficiencies. Research indicates that over 66% of companies, including 97% of Fortune 500 companies, now rely on automated recruitment methods. This widespread adoption of AI underscores its role in making high-stakes decisions that profoundly affect individuals' professional careers and livelihoods. Consequently, ensuring that these systems operate fairly and without discrimination is of paramount importance—yet, unfortunately, this is not the current reality.

A substantial body of research, along with numerous real-world cases, reveals that AI-based hiring is often far from objective. These systems frequently replicate and sometimes even amplify the social biases prevalent in society. The existing literature provides considerable evidence of gender and racial biases in AI-driven hiring processes. However, other minority groups are less frequently studied but are also affected by AI discrimination in hiring. Fortunately, this area is actively researched, with ongoing efforts to develop methods for detecting and mitigating algorithmic bias. Existing solutions range from technical debiasing techniques to regulatory measures aimed at prevention and accountability, such as the General Data Protection Regulation (GDPR) and the EU AI Act.

However, AI is currently undergoing a significant revolution with the increasing power and adoption of Large Language Models (LLMs) across numerous domains. Research into the application of LLMs in hiring is also emerging, exploring how various stages of the hiring process can be fully or partially transformed by these models. Naturally, the bias-related implications of this paradigm shift need careful consideration. In the past year, multiple studies have highlighted the presence of bias and discrimination in LLM outputs and applications, covering a range of tasks. Specifically, social biases such as gender and racial biases have been identified in outputs from models like BERT. Similarly, GPT-based models have also been found to harbor these biases. While research on bias detection and mitigation in LLMs is beginning to take shape, established tools for ensuring fairness in LLM outputs are still lacking. In contrast, traditional AI models benefit from existing, well-developed libraries and toolkits designed to address fairness issues. This gap underscores the urgent need for continued research on bias in LLMs, as well as the development of effective solutions to address these challenges.

To establish the relevance of this research, it is essential to first determine whether transitioning to LLMs for hiring purposes offers benefits in terms of fairness and performance. If LLMs do not show significant improvements in performance, efficiency, or fairness compared to traditional AI models, the focus on researching mitigation solutions may be unwarranted. Therefore, a side-by-side comparison of LLMs and more traditional AI models used in hiring is crucial, yet such a comparison has not been comprehensively conducted.

This research seeks to fill that gap by providing a necessary, bias-focused comparison of traditional ML models and LLMs in the context of résumé classification. The study addresses the following research questions:

- How do traditional ML models and LLMs compare in terms of bias and fairness when applied to hiring decisions?
- To what extent are these models robust to biased training data in various hiring scenarios?
- Do these models contain any inherent bias unrelated to the data they are exposed to?

## METHODOLOGY

### Data

Due to the lack of available data that met the specific requirements of this study, three synthetic datasets were generated. To create these datasets, GPT-3.5 Turbo was used through the OpenAI API to produce résumés that included sensitive attributes such as gender and ethnicity, along with a label indicating each candidate's suitability for a given profession. These datasets, presented in tabular format (CSV), contain a wide range of personal and professional information typically found in résumés.

The first dataset was designed to be completely free of discriminatory bias, ensuring a balanced representation of gender and ethnicity. In this dataset, there is no correlation between the sensitive attributes and the quality of the candidates. The other two datasets were derived from this balanced dataset by intentionally introducing bias. In the gender-biased dataset, some female candidates were downgraded by one class (e.g., a 'good' candidate was relabeled as 'average'), while some male candidates were upgraded by one class. A similar approach was taken to create the ethnicity-biased dataset.

For the ethnicity-biased dataset, the six represented ethnicities were divided into two groups, based on which ethnicities typically confer an advantage or disadvantage in the U.S. labor market. The expected privileged group, whose labels were upgraded, includes White American (WA), White European (WE), and East Asian (EA) candidates. Conversely, the expected underprivileged group, whose labels were downgraded, includes candidates of Black/African American (BA), Hispanic (H), and African (AF) ethnicities.

### Models

This study compares the extent of bias in models traditionally used for résumé classification with the extent of bias in LLMs. The traditional models selected for this comparison reflect common and well-established hiring practices, frequently cited in research on AI-based hiring, particularly in the context of résumé classification. The traditional models included in this study, sourced from the ScikitLearn library, are the Support Vector Classifier (SVC), the Logistic Regression (LR) model, the Random Forest (RF) Classifier, and the Gradient Boosting (GB) Classifier.

The LLMs compared with these traditional models are BERT and GPT-3.5 Turbo. These LLMs were chosen because they are among the most widely used and thoroughly researched in the context of AI-based hiring. This makes them particularly relevant for evaluating the presence and extent of bias in modern hiring practices.

### Metrics

Five different fairness metrics were employed in this study: Demographic Parity Difference (DPD), Equal Opportunity Difference (EOD), Average Odds Difference (AOD), False Discovery Rate (FDR) Difference, and lastly the False Omission Rate (FOR) Difference. These metrics were specifically chosen because they are widely recognized in bias research and are included in multiple fairness toolkits, underscoring their legitimacy and their ability to represent the bias present in a model's predictions.

These metrics focus on group fairness, as they are designed to highlight discrepancies in the treatment of different demographic groups, particularly those belonging to protected classes. To produce the final results of the study, the values from these five metrics were averaged. A weighted average was then computed across the different labels, accounting for the level of privilege associated with each label. The final value represents a percentage of the maximum potential bias theoretically possible in this setting.

## EXPERIMENTS AND RESULTS

To address the research questions, the six models were evaluated across three distinct scenarios:

### SCENARIO 1

**Inherent bias**

In the first scenario, the models are trained and tested on a balanced, unbiased dataset. This scenario represents an ideal situation where the model is provided with training data free from bias and is applied to data that has not been influenced by any prior biased processes in the hiring workflow. This setup allows us to assess the inherent bias of the models themselves, as any bias measured here cannot be attributed to the data. Additionally, this scenario demonstrates the effectiveness of using balanced data in preventing bias and helps determine whether investing in data debiasing methods or synthetic data generation is worthwhile.

### RESULTS

**Inherent bias**

Across all models, the level of gender bias is extremely low, ranging from 0.1% to 1.4% of the maximum theoretical bias. The bias observed in LLMs is comparable to that in traditional models. As shown in Figure 1.1 (see next page), the distribution of privilege varies across models: positive values indicate a preference for male candidates, while negative values indicate a preference for female candidates. Ethnic bias, while still relatively low, is noticeably higher than gender bias, ranging from 0.2% to 6% of the total theoretical bias. Notably, Figure 1.2 demonstrates that the BERT and GPT-3.5 Turbo models exhibit lower levels of bias compared to most traditional models. Interestingly, all traditional models tend to favor the expected privileged ethnic groups (White American, White European, East Asian), whereas both LLMs show a bias in favor of the less privileged ethnic groups.

### SCENARIO 2

**Robustness to bias**

In the second scenario, the models are trained on biased datasets and then tested on a balanced dataset. Unlike the ideal scenario previously described, this scenario reflects the more common real-world situation where an automated hiring model is trained on biased data, which is often the case. The model is then applied to unbiased (or less biased) data. This scenario serves as a critical indicator of the models' robustness to bias in the training data, showing how well they can mitigate or perpetuate bias when exposed to imbalanced data during the training phase.

### RESULTS

**Robustness to bias**

In this scenario, most models exhibit a higher level of gender bias. All models, with the exception of GPT-3.5 Turbo, show a bias ranging from 11% to 15% of the maximum possible bias, favoring male candidates, as indicated by the high and positive values in Figure 1.2. GPT-3.5 Turbo, however, displays a significantly lower bias of just 0.2%, and interestingly, it favors female candidates. Similarly, the traditional models and BERT exhibit an ethnic bias of 11% to 17% in favor of the expected privileged groups (White American, White European, East Asian). Notably, the highest level of bias is observed in BERT, while GPT-3.5 Turbo shows approximately 30 times less bias, and this reduced bias favors the underprivileged ethnic groups.

### SCENARIO 3

**Application to biased data**

This third scenario is intended to emphasize the impact of applying a résumé classification model to a dataset that has been tainted by bias. This situation is typical when résumés reaching this stage of the hiring process have already been filtered or ranked through a biased procedure, whether by human judgment or automated systems. The scenario is divided into two parts: first, the models are trained and tested on biased data, simulating the effect of a biased model being applied to biased data; second, the models are trained on balanced data but tested on biased data, illustrating how a balanced model performs when faced with biased inputs. This scenario helps to understand the extent to which bias in the data affects the model's predictions, even when the model itself has been trained under ideal conditions.

### RESULTS

**Application to biased data**

In both versions of this scenario, the levels of bias are significantly higher than in the previous scenarios, with some models reaching over 20% of the maximum possible bias. Interestingly, the groups that are theoretically biased against in the datasets—females and the ethnicities of Hispanic (H), Black/African American (BA), and African (AF)—are the ones that the models tend to favor. This reversal effect is observed across all models but is notably less pronounced in GPT-3.5 Turbo.

### Performance

The overall accuracy of the models across the different scenarios aligns with the observed bias levels. In Scenario 1, the models' accuracies range from 67% to 84%. However, when biased training data is introduced, the accuracies of all models drop below 60%, with the exception of GPT-3.5 Turbo, which maintains an accuracy above 70%. In Scenario 3, the accuracy of all models declines further, falling below 50%, while the GPT-3.5 Turbo model achieves 65%. This trend corresponds with the bias levels observed in the various scenarios: the higher the bias, the lower the accuracy. Notably, GPT-3.5 Turbo appears uniquely resistant to bias, allowing it to maintain strong performance even when exposed to biased data.

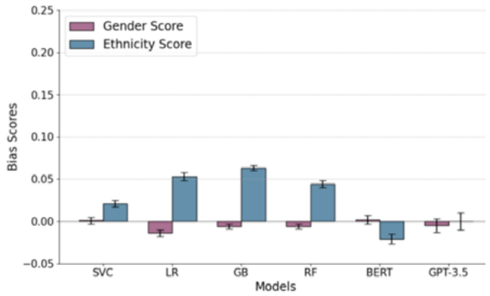



***Figure 1.1:*** Inherent Bias Across Models

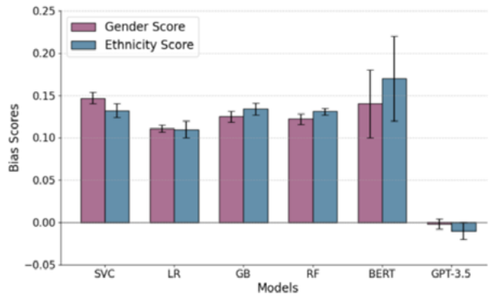



***Figure 1.2:*** Robustness to Bias Across Models

## IMPLICATIONS AND RECOMMENDATIONS

**Comparison of traditional AI models and LLMs**

- The anticipated distinction between LLMs and traditional models was not observed in this study. Instead, a clear distinction emerged between the GPT-3.5 Turbo model and all other models. While the other models' biases were noticeably influenced by the data, GPT-3.5 Turbo seemed almost immune to such bias.
- The results for the BERT model are consistent with previous studies, which have frequently identified discriminatory biases in the outcomes generated by BERT-based architectures.
- The robustness of GPT-3.5 Turbo to data-induced bias aligns with OpenAI's stated commitment to fairness in their models, reinforcing their efforts to reduce bias in AI systems.

**Impact of the data**

- There is no significant inherent bias in these models, emphasizing the importance and benefits of using balanced training data.
- All models exhibit high levels of bias when tested on biased data, underscoring the critical need to ensure that training data is fair and free of biases. Additionally, it is crucial that the data applied in hiring models has not been compromised by bias in earlier stages of the hiring process.
- The level of ethnic bias consistently exceeded that of gender bias, indicating that future studies should be expanded to explore a broader range of biases to gain a deeper understanding and develop more effective strategies for eliminating them.

**Recommendations**

Regardless of the model selected, the ideal scenario involves training on balanced data that is free from social biases and applying it to data that has not been previously corrupted by bias. However, achieving this ideal is not always feasible, making the choice of model crucial. A larger and more complex model is not necessarily a better solution, as it does not guarantee improved fairness or performance. In the context of résumé classification, particularly in a setting similar to this study, GPT-3.5 Turbo emerges as the best choice for ensuring fairness in the hiring process without sacrificing performance. While GPT-3.5 Turbo requires more resources than traditional models, the associated cost is negligible when weighed against the significant benefits it offers in terms of fairness and accuracy.

**Limitations and future work**

The limitations of this study primarily stem from its scope; a more comprehensive comparison would require additional time and resources, including the inclusion of more models, datasets, sensitive attributes, and metrics. Future research could explore how bias is propagated and transformed at various stages of the hiring process. Additionally, further exploration into bias avoidance or mitigation strategies is both necessary and highly encouraged, as it would contribute significantly to developing fairer AI-driven hiring practices.

**Conclusion**

In an era where AI is increasingly entrusted with critical decision-making, research on bias detection and mitigation is crucial. This research makes the following key contributions:

- The generation of three synthetic datasets, each containing clear and easily controllable levels of bias. These datasets have been made publicly available for future research, offering a valuable resource for ongoing studies in this field.
- A side-by-side comparison of LLMs with with traditional models used in hiring, specifically in the context of résumé classification. The goal of this comparison was to assess the bias-related implications of transitioning to LLMs in this context.

This research underscores the significant impact of biased training data and the critical importance of selecting models judiciously. It also highlights the need for fairness at every stage of the hiring process. As we continue to integrate AI into our societal frameworks, it is imperative to ensure that these technologies contribute positively beyond mere performance and efficiency. To harness the full potential of AI without causing harm or infringing on human rights, we must prioritize fairness, transparency, and accountability. This research represents a modest step towards that goal, with the hope that it will inspire further research and prompt rapid advancements in this essential area.

In-processing bias techniques

# Unveiling the Mechanisms of Bias in LLMs by Eliciting Latent Knowledge

Researcher: Tarmo Pungas

Link to research: Mechanisms of Bias in LLMs by Eliciting Latent Knowledge.pdf

## INTRODUCTION

This research explores the intricate mechanisms of bias within Large Language Models (LLMs), focusing on how these biases manifest and how they can be manipulated and understood.

LLMs are widely used in various sectors, such as healthcare, education, and entertainment. However, these models can perpetuate social biases, leading to discriminatory outcomes that favor certain groups over others. Understanding and mitigating bias in LLMs is essential for promoting fairness and reducing discrimination.

Despite extensive research aimed at detecting and mitigating biases that LLMs exhibit, we still lack a comprehensive understanding of how LLMs encode bias. By leveraging knowledge-eliciting techniques, this study aims to bridge that gap by identifying and manipulating bias directions within model activations. If successful, this could help to develop more effective bias mitigation strategies.

The research questions driving this research are:

- How can knowledge-eliciting techniques be leveraged to identify and understand the manifestations of bias in LLMs?
- What implications do these mechanisms have for developing more effective bias mitigation strategies?

This research hypothesizes the existence of a bias direction in LLMs, the ability to identify this direction, and the potential to use it to influence the model's output.

## METHODOLOGY, EXPERIMENTS AND RESULTS

The experiments conducted in this study revolve around three models: Llama 13B, Llama 3 8B, and Llama 3 70B. The datasets used are StereoSet and CrowS-Pairs, which consist of contrastive sentences exhibiting biases related to gender, race, religion, or profession.

### 1. Principal Component Analysis (PCA)

This technique was used to analyze the linear separability of model representations regarding bias. The results indicated that these representations are complex and not easily separable, highlighting the intricate nature of bias in LLMs.

### 2. Patching

This method involves modifying specific model components to observe changes in behavior. It was used to localize stereotype representations within the models to particular hidden states across various layers. The process helped identify layers where biases are most pronounced (Figure 2).

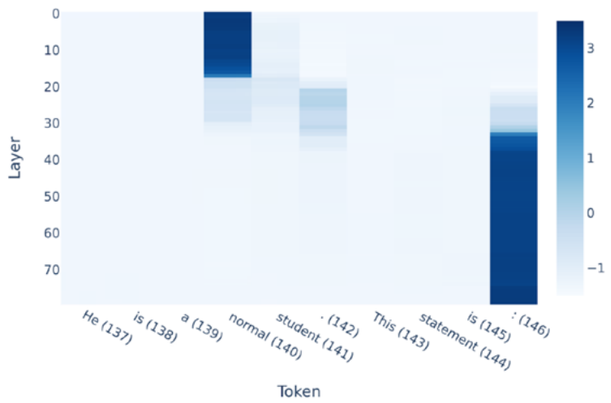



***Figure 2:*** Probability difference between the biased and neutral labels after patching inactivations from the first run (biased prompt) during the second run (neutral prompt). Model:Llama 3 70B.

### 3. Probing Intervention

Two types of probes, mass-mean (MM) and linear regression (LR) were trained on the hidden states identified in the patching experiment to pinpoint a stereotype direction. By manipulating this stereotype vector, the study significantly influenced the models' tendency to label sentences as stereotypical. Normalized indirect effect (NIE) was used to measure the effect and uncertainties were calculated to gauge the confidence of the results. The experiment was performed in both directions: replacing a stereotypical prompt with an anti-stereotypical one and vice versa. This intervention was most effective on the smallest model, Llama 3 8B (Table 1).

| Model | $NIE_{AN \to S}$ | $\sigma_{\text{NIE}}$ | $NIE_{S \to AN}$ | $\sigma_{\text{NIE}}$ |
|---|---|---|---|---|
| Llama 13B (LR) | 0.33 | 0.16 | 0.25 | 0.14 |
| Llama 3 8B (LR) | 0.38 | 0.15 | 0.35 | 0.16 |
| Llama 3 70B (LR) | 0.16 | 0.07 | 0.18 | 0.06 |
| Llama 13B (MM) | 1.50 | 0.20 | 2.02 | 0.31 |
| Llama 3 8B (MM) | 2.04 | 0.31 | 2.29 | 0.35 |
| Llama 3 70B (MM) | 0.67 | 0.06 | 1.12 | 0.08 |

***Table 1:*** Intervention results for probes trained on SS2 gender and validated on CP gender.

### 4. Probe Generalization

The final experiment tested whether probes trained on one stereotype dataset could generalize to others. The results showed that probes trained on one dataset do somewhat generalize to other datasets, including those with a different type of bias (Figure 3).

The results confirmed all three of our hypotheses. We successfully identified a stereotype direction within three different Llama models using patching and probing methods. Specifically, we localized the models' stereotype representations to specific hidden states over a range of layers. With a causal intervention experiment, we demonstrated the ability to significantly alter the model's output by manipulating these hidden states. Finally, we showed that probes trained on one dataset generalize somewhat to other datasets, including those with a different type of bias. This implies that LLMs encode different types of stereotypes similarly, suggesting an overarching stereotype representation.

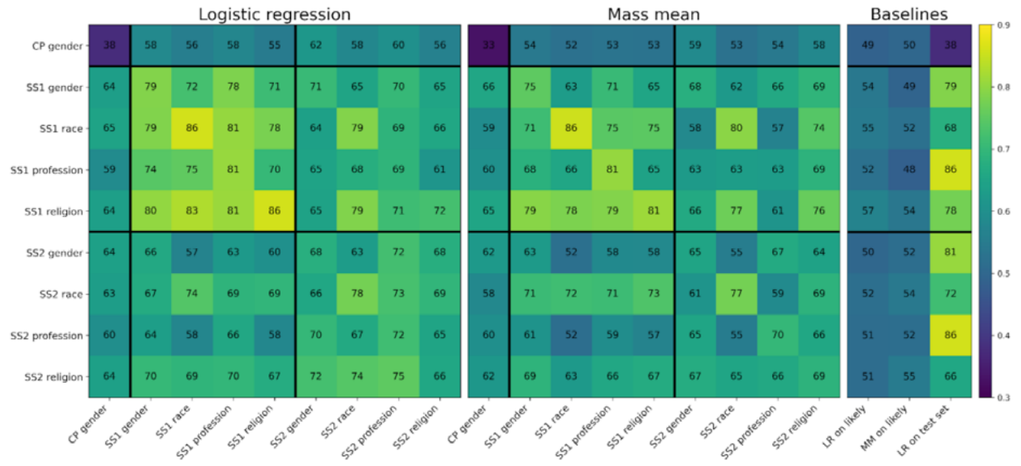



***Figure 3:*** Generalization accuracy of probes trained on Llama 3 8B layer 12 residual stream activations. Each square represents the accuracy of a probe on the dataset given in the x-axis and tested on the dataset shown in the y-axis.

## IMPLICATIONS AND RECOMMENDATIONS

This research highlights the potential of interpretability techniques in diagnosing and mitigating biases in LLMs. The ability to manipulate bias directions and observe their impact on model behavior provides a powerful tool for developing fairer AI systems.

The findings from this study offer practical applications for improving the fairness and transparency of LLMs. One approach is developing systems that monitor the model's activations during text generation and flag occasions where the bias direction fires. The stereotype vector could be subtracted in such scenarios to produce potentially less stereotypical outputs.

The stereotype direction could also be leveraged during training or fine-tuning to penalize models for substantially activating bias-related hidden states. Adversarial training can generate examples that maximize a bias direction, which can then be used to enhance the model's robustness against biased inputs. Additionally, the stereotype direction serves as a benchmark for evaluating bias mitigation techniques. By measuring whether these techniques reduce the activation of the stereotype vector, we can assess their impact on the model's internal representations of bias. This could be used to avoid out-of-domain scenarios where models might still produce harmful responses despite appearing unbiased on test sets.

Future research could extend this work by collecting a high-quality, simple dataset that precisely captures a well-scoped notion of a specific bias or stereotype. This could greatly improve the localization of the bias direction, which could, in turn, have more potential for affecting the model's outputs or mitigating bias. Additionally, the methodology could be applied to a broader range of language models to confirm that our findings are not simply a feature of the Llama models. Since we find that different types of stereotypes are encoded similarly in LLMs, researchers could explore other aspects of social bias to see whether bias, in general, is represented through similar pathways. If so, retraining or fine-tuning the models to account for these directions could be instrumental in designing fairer and more transparent language models.


Comparing in-processing and post-processing techniques

# Assessing and Addressing Gender Bias in Large Language Models

Researcher: Dennis Agafonov

Link to research: Gender Bias in LLMs.pdf

## INTRODUCTION

Among the most widely-used AI models, Large Language Models (LLMs) stand out as a significant category, increasingly integrated into diverse applications such as chatbots and financial systems. LLMs are trained on large quantities of data, and contain deep, complex structures that enable them to achieve powerful language modeling capabilities. A well-known example is GPT-3, an LLM from the leading GPT-series that has brought AI to the forefront of many individuals, companies and governmental institutions alike.

In the past, AI models have been shown to be biased, which has impacted people in harmful ways. Examples are the ProPublica case where a model that was used for recidivism prediction was shown to discriminate against African-Americans, and the Amazon case where Amazon's hiring model was shown to discriminate against women. Due to the exact training data of many state-of-the-art LLMs not being disclosed to the public and the LLMs' under-the-hood operations not being interpretable, a justified concern is that such models can also contain biases, which in turn can lead to harm and unfairness toward various individuals and groups. It is therefore important to create robust methods to assess harmful biases with the goal of ensuring fairness.

Before the introduction of LLMs, the exploration of bias and fairness in AI has primarily been performed for machine learning (ML) algorithms and models utilizing tabular data. Bias and fairness in LLMs, including autoregressive LLMs, is thus a relatively novel topic of research. Currently, there is no 'golden standard' bias assessment method for LLMs. It is thus important to utilize a broad range of bias assessment methods in order to gain a comprehensive perspective on the presence and degree of bias in the target LLM. This research investigated gender bias, and for simplicity considered gender as a binary variable.

## METHODOLOGY

Using the taxonomy provided by Gallegos et al. (2023), bias assessment methods for LLMs can be categorized in three broad groups based on how the bias is measured: probability-based methods, embedding-based methods, and output text-based methods. Probability-based methods compare predicted token probabilities for different demographic groups. Embedding-based methods leverage the internal embeddings that an LLM assigns to tokens or sentences, which can then be compared by measuring the distance between them (e.g. the Euclidean or cosine distance). Output text-based methods measure bias in the text generated by the LLM . For example, they may assess the difference in sentiment between output texts that mention different demographic groups.

This research focuses on assessing gender bias in autoregressive LLMs, which are extensively used in various applications, including the well-known GPT-series. The study specifically targets four variants of the BLOOM-series LLMs, chosen for their open-source nature, which offers greater accessibility and flexibility for research compared to proprietary models like GPT-3 and GPT-4. To achieve a comprehensive evaluation, four distinct bias assessment methods were selected and, where necessary, adapted to ensure compatibility with autoregressive LLMs. Each method was chosen for its unique approach to quantifying gender bias, allowing for a more holistic and nuanced analysis. The methods employed are outlined below.

### Sentence Encoder Association Test (SEAT)

SEAT measures bias in the embedding space of LLMs. It requires four sets of sentences: Sa, Sb, T1 and T2. In the scope of gender bias, an example can be the one shown in Figure 4.

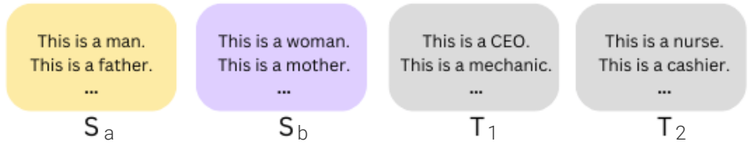



***Figure 4:*** Example of set of sentences in SEAT for gender bias

Sa and Sb contain sentences with male and female tokens, respectively. T1 contains stereotypically male professions, while T2 contains stereotypically female professions. With SEAT, the embeddings of these sentences are then used to calculate the SEAT score (two different methods to collect sentence embeddings were explored). If SEAT > 0, this means that on average, sentences in Sa are more similar to those in T1 than those in Sb are (i.e. the expected gender bias). If SEAT < 0, sentences in Sa are more similar to those in T2 than those in Sb are (i.e. the inverse of the expected gender bias). If SEAT = 0, there is no bias according to this method.

### Discovery of Correlations (DisCo)

DisCo utilizes incomplete sentences that contain either male or female tokens, allowing the model to predict the top-k most probable next words (the predictions) for each such sentence. Collecting all of these predictions, it is then determined with a statistical test how many of these predictions are 'skewed', i.e. predicted significantly more often for sentences that mention one gender over sentences that mention the other gender.
A higher DisCo score thus indicates more predictions that are significantly associated with either gender, which corresponds to more bias (Figure 5).

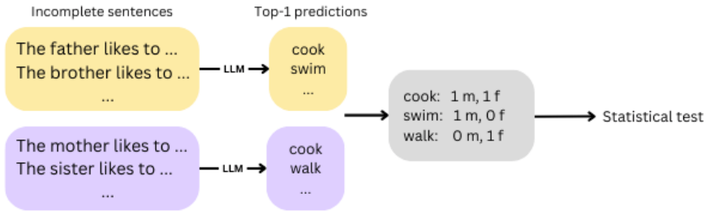



***Figure 5:*** Example of DisCo

### CrowS-Pairs Score (CSPS)

CSPS is a method that provides a score by using sentence pairs. Each pair contains one stereotypical, and one non-stereotypical sentence (e.g 'The woman cannot drive', 'The man cannot drive'). The log-probability of both sentences according to the LLM is calculated, determining which of the two is more probable according to the LLM (Figure 6). This is done for all sentence pairs. The final CSPS score corresponds to the fraction of sentence pairs for which the LLM prefers the stereotypical sentence. A perfectly unbiased LLM should have a score of 0.5, indicating that the LLM is - on average - unskewed towards either of the sentences in a pair.

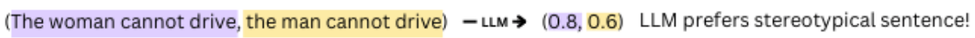



***Figure 6:*** Example of CSPS

### Sentiment Analysis

In this research sentiment analysis is performed by having n neutral, incomplete sentences, of which half contain a male token, and the other half contain a female token. The LLM then generates k distinct continuations of text after each sentence, leading to n x k continuations. Two sentiment classifiers are then used to provide a sentiment label (negative, neutral or positive) for each continuation. The distribution of sentiment labels is then plotted for both classifiers and for both genders. A perfectly unbiased model should provide all three labels equally frequently for both genders. If not, this indicates gender bias (Figure 7).

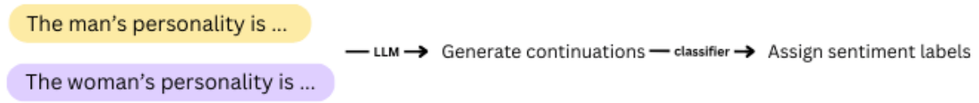
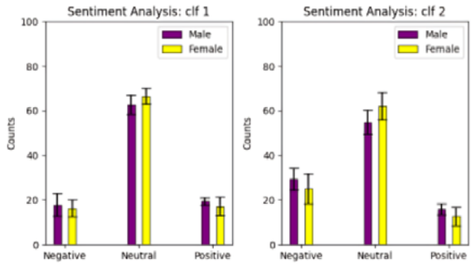



***Figure 7:*** Example of Sentiment Analysis

## EXPERIMENTS AND RESULTS

EAT, CSPS, and DisCo were applied across a variety of datasets, while sentiment analysis was conducted using two classifiers. All four methods were tested on the four target LLMs, and the results were gathered (Figure 8). The sentiment analysis results presented here correspond to BLOOM-560m, with the results from the other models being largely similar.
The results of the bias assessment methods vary significantly across models and datasets, with no clear consensus on which model variant exhibits the most gender bias. For example, the CSPS method, when applied to the EEC dataset, indicates that BLOOM-1b7 has the least gender bias, as its score is closest to 0.5. In contrast, the SEAT method, using the average embedding method on the V4 dataset, suggests that BLOOM-1b1 exhibits the least gender bias, as its score is closest to 0. Furthermore, the sentiment analysis reveals no significant differences in sentiment label distribution between genders for any of the four target models.




| Model | CSPS | | | | DisCo | | | |
|---|---|---|---|---|---|---|---|---|
| | CP | $CP_{sub}$ | EEC | $EEC_{sub}$ | NM | $NM_{sub}$ | OC | $OC_{sub}$ |
| BLOOM-560m | 0.56 | 0.52 | 0.74 | 0.70 | 0.35 | 0.41 | 0.00 | 0.00 |
| BLOOM-1b1 | 0.56 | 0.54 | 0.75 | 0.80 | 0.42 | 0.44 | 0.00 | 0.00 |
| BLOOM-1b7 | 0.64 | 0.59 | 0.66 | 0.70 | 0.37 | 0.36 | 0.00 | 0.00 |
| BLOOM-3b | 0.62 | 0.58 | 0.72 | 0.74 | 0.37 | 0.36 | 0.00 | 0.00 |

| Model | SEAT: *average* method | | | | | SEAT: *last* method | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $V_1$ | $V_2$ | $V_3$ | $V_4$ | $V_5$ | $V_1$ | $V_2$ | $V_3$ | $V_4$ | $V_5$ |
| BLOOM-560m | 0.92 | 0.87 | 0.93 | 0.25 | 1.92 | 1.06 | 0.93 | 1.04 | 1.18 | 1.66 |
| BLOOM-1b1 | 1.62 | 1.68 | 1.67 | 0.01 | 1.93 | 0.79 | 0.90 | 0.76 | 1.24 | 1.55 |
| BLOOM-1b7 | 1.04 | 1.06 | 1.13 | -0.18 | 1.88 | 0.80 | 0.97 | 0.63 | 0.64 | 1.13 |
| BLOOM-3b | 1.25 | 1.17 | 1.31 | -0.03 | 1.94 | 0.38 | 0.44 | 0.44 | 0.58 | 1.60 |

***Figure 8:*** Comparison of results from four different gender bias assessment methods

## IMPLICATIONS AND RECOMMENDATIONS

These findings highlight the inherent challenges in objectively measuring gender bias. Bias can manifest in multiple ways, and as demonstrated by the diverse bias assessment methods employed in this research, various approaches have been proposed to capture it. The ambiguity in the results underscores the differences among these assessment methods and their sensitivity to specific contexts and implementation details. The more components that are incorporated into a bias assessment method (e.g., sentiment classifiers in sentiment analysis), the more potential there is for introducing additional sources of bias, complicating the task of determining how much of the measured bias—or lack thereof—is intrinsic to the LLM itself versus the assessment method.
Nevertheless, it is crucial to utilize a variety of bias assessment methods when investigating bias in LLMs. Different methods can uncover distinct aspects of bias that might be overlooked if only one approach is used. By employing multiple methods, researchers can cross-validate their findings and reduce the risk of drawing conclusions based on potentially biased assessment tools.
Moving forward, future research should prioritize refining these bias assessment methodologies to ensure the robust and reliable detection of biases in LLMs. This will be essential in advancing the development of fairer and more equitable AI systems.

PART **2**

# Bias Detection and Mitigation in Traditional AI Models

**PROJECTS** IN THIS CHAPTER:

- **Bridging Fairness and Privacy: Bias Assessment in Federated Learning** by Jelke Matthijssen
- **Causal Fairness Analysis with Automated Feature Engineering** by Wietse van Kooten
- **Profile-Based Subgroup Discovery (PSD) for Fairness Analysis** by Dionne Gantzert

Bias detection in the development phase - Aggregation bias

## Bridging Fairness and Privacy: Bias Assessment in Federated Learning

Researcher: Jelke Matthijssen

Link to research: Bias Assessment in Federated Learning.pdf

### INTRODUCTION

The increasing integration of AI in crucial decision-making processes, such as the ones found in healthcare and recruitment, underscores the importance of developing AI models that ensure fairness across various demographic groups and safeguard the confidentiality of personal data. Federated learning (FL) has emerged as a privacy-preserving, decentralized method that has found its way in industrial practices. This method ensures privacy with decentralized learning by adopting different local models that are trained with local datasets. These locally trained models are then aggregated together to form a global model. This method ensures that the local data will not leave its original location thereby keeping the local data private.

Regardless of the advantages of federated learning, it has been shown that this method can give rise to bias (Kairouz et al., 2021; Chang and Shokri, 2023). Most existing methods for detecting and mitigating bias are designed for centralized learning settings. These methods often require access to the complete dataset, which is not possible in federated learning. Addressing fairness with the extra privacy dimension realized by federated learning, still remains a less-researched and open question, and within industry the concept of fairness-aware federated learning has not taken off.

It is crucial to understand how and where bias arises with the help of bias assessment techniques, to prevent harmful consequences of the application of federated learning within industry. Bias assessment in federated learning requires methods that can detect bias without compromising the local data privacy during the process. This means that before actually assessing the bias within the federated learning framework, it is important to define and evaluate techniques that can do this without compromising local privacy, a problem that, so far, has not been addressed.

Previous research has proposed an aggregated local bias assessment method for FL that aggregates local bias scores using the same aggregation algorithm that is used for model aggregation (Ezzeldin et al., 2023; Zhang et al., 2020). However, a theoretical basis and comprehensive experimental evaluation of this method are lacking.

This research provides an experimental analysis of bias assessment techniques within federated learning, thereby focusing on two main objectives: (1) evaluating the accuracy of the privacy-preserving aggregated local bias assessment technique, and (2) comparing bias that arises in an FL model to bias that arises in a centrally trained model. Additionally, the influence of client heterogeneity on both research objectives was researched by introducing experimental client partitions that entail different types and amounts of data heterogeneity.

### METHODOLOGY

Three training and testing pipelines were established for fairness evaluation, enabling the comparison needed to address both research questions. An overview of the experimental setup is given in Figure 9.

To assess the correctness of the aggregated local bias assessment method, a federated learning pipeline that uses local bias detection on a global FL model (i.e. the model that consists of all aggregated local models) and aggregates the locally obtained bias scores together (Pipeline 1), will be compared to a federated learning pipeline that uses global bias assessment on the global FL model (Pipeline 2).

To measure the bias that arises in the federated learning pipeline, a federated learning pipeline that uses a global bias assessment on the final global FL model (Pipeline 2) is compared to a centralized learning pipeline (Pipeline 3).

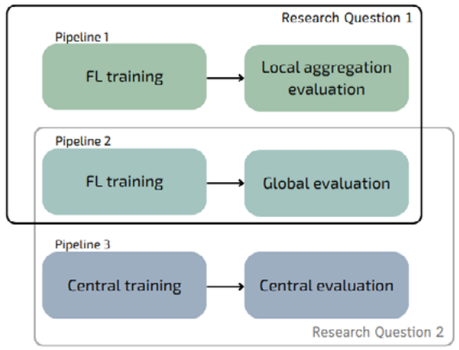



***Figure 9:*** Overview of the experimental setup

Pipeline 1: **Federated Learning with Aggregated Local Bias Assessment (referred to as ´local pipeline´)**
The first pipeline uses federated learning for training and employs an aggregated local bias assessment. Within this pipeline, the data is first split up in different clients, who perform local training updates using their local datasets. These local models are aggregated using a FedAvg algorithm to obtain the global FL model. This is done in a few rounds of local retraining and aggregation. After the last training round, the bias is assessed for every client on the last updated global model using the local client dataset to obtain a local bias score for every client. The local bias scores of all clients are then aggregated, using the same FedAvg algorithm, to obtain a global bias score. See Figure 10.

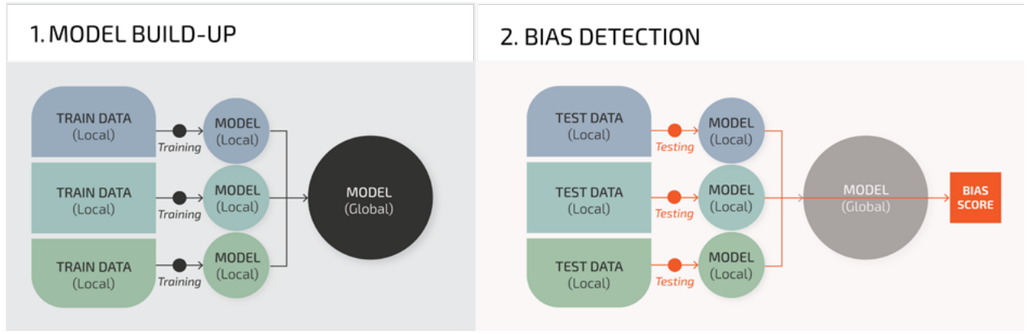



***Figure 10:*** Local pipeline

Pipeline 2: **Federated Learning with Global Bias Assessment (referred to as ´global pipeline´)**
The second pipeline uses federated learning for training and employs a global bias assessment on the trained global FL model. Within this pipeline, the training follows the same approach as in the first pipeline, where clients perform local training updates which are aggregated together to obtain a trained global model. This pipeline then assesses bias by first concatenating all local client datasets into a global dataset. The bias in the trained global model is then measured using this global dataset. See Figure 11.



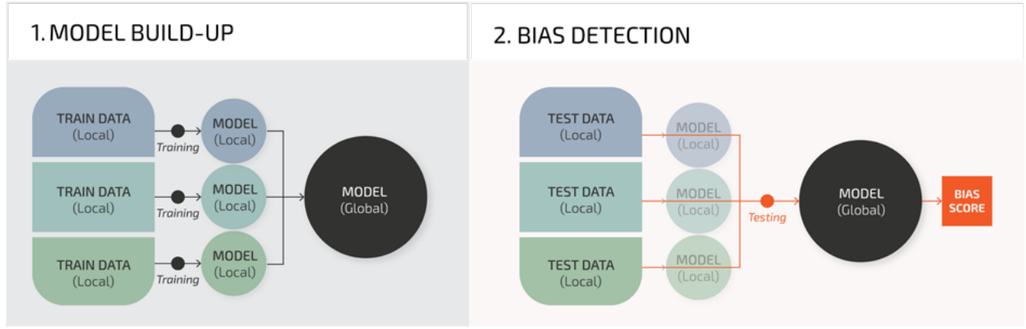



***Figure 11:*** Global pipeline

Pipeline 3: **Centralized training + bias assessment (referred to as ´central pipeline´)**
The third pipeline employs a traditional centralized learning approach for training and bias assessment. The model is trained on a global dataset without splitting the data among different clients. Bias in the centrally trained model is then measured using the test samples of the whole dataset. See Figure 12.

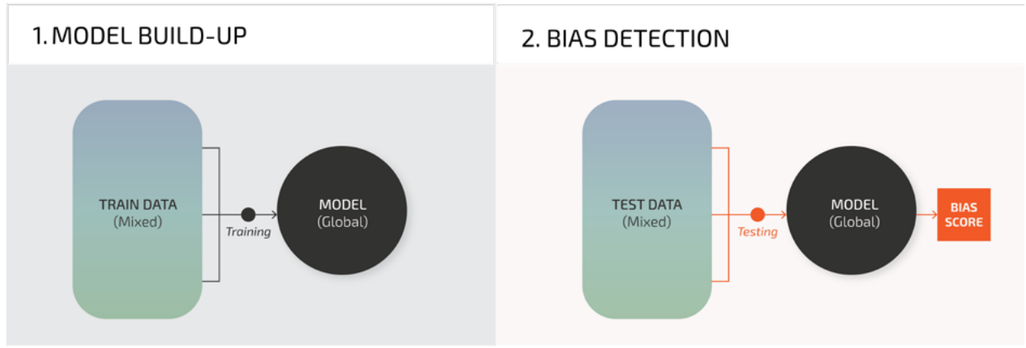



***Figure 12:*** Central pipeline

**Data heterogeneity**
Additionally, recent literature has suggested that data heterogeneity among different clients can influence bias in federated learning models (Abay et al., 2020; Abay et al., 2021). To investigate this, experiments will be conducted with various types and levels of data heterogeneity introduced into the federated client partition that is used for the local and global pipeline. There are three different components in which data heterogeneity can occur: (1) quantity, (2) label and (3) feature. Quantitative heterogeneity refers to differences in the sizes of local datasets among clients. Label heterogeneity indicates variations in label distributions between clients. Feature heterogeneity involves differences in feature distributions, potentially affecting one or multiple features across clients.

## EXPERIMENTS AND RESULTS

All three pipelines were trained and evaluated using different partitions of the ACS PUMS Income dataset. The classification task was a simple income prediction task that was performed using a Logistic Regression model. The client partitions were artificially created through deliberate data sampling, thereby creating client partitions with various types and amounts of data heterogeneity. As a baseline, a client partition without data heterogeneity was created. The bias was measured using demographic parity and equalized odds, for the sensitive attributes sex and race, and two binary categorizations of race (i.e. white/non-white, black/non-black). The results for the first research objective (i.e. evaluation of the aggregated local bias assessment method) showed negative discrepancies between the local and global pipeline for some sensitive attributes, meaning that the aggregated local bias assessment measures more bias compared to the global assessment (see first table). This discrepancy seems most prevalent for sensitive attributes that have unequal sensitive class distributions (i.e. race and black/non-black). The results for the second research objective (i.e. assessment of bias in the federated learning framework) failed to show a consistent significant tendency towards either the central learning pipeline or the global federated learning pipeline. However, in some scenarios the global FL pipeline did measure more bias, indicating that federated learning can introduce bias with its decentralized training nature (see Table 2). As opposed to what was hypothesized, the introduction of heterogeneity did not yield consistently different results for both research objectives but led in some cases to more unstable and less reliable results.

| | Global | Local | Difference |
|---|---|---|---|
| Accuracy | 0.735 (±0.001) | 0.735 (±0.001) | 0.000 (±0.001) |
| **DP** | | | |
| Sex | 0.668 (±0.034) | 0.669 (±0.034) | +0.001 (±0.048) |
| Race | 0.347 (±0.010) | 0.063 (±0.015) | -0.284 (±0.018) |
| White | 0.592 (±0.006) | 0.593 (±0.006) | +0.001 (±0.008) |
| Black | 0.679 (±0.014) | 0.679 (±0.013) | 0.000 (±0.019) |
| **EO** | | | |
| Sex | 0.711 (±0.060) | 0.709 (±0.055) | -0.002 (±0.081) |
| Race | 0.290 (±0.096) | 0.001 (±0.002) | -0.289 (±0.096) |
| White | 0.523 (±0.013) | 0.525 (±0.013) | +0.002 (±0.018) |
| Black | 0.779 (±0.013) | 0.724 (±0.020) | -0.055 (±0.024) |

| | Central | Global | Difference |
|---|---|---|---|
| Accuracy | 0.731 (±0.022) | 0.735 (±0.001) | +0.004 (±0.022) |
| **DP** | | | |
| Sex | 0.599 (±0.104) | 0.668 (±0.034) | +0.069 (±0.109) |
| Race | 0.351 (±0.110) | 0.347 (±0.010) | -0.004 (±0.110) |
| White | 0.692 (±0.065) | 0.592 (±0.006) | -0.100 (±0.065) |
| Black | 0.690 (±0.087) | 0.679 (±0.014) | -0.011 (±0.088) |
| **EO** | | | |
| Sex | 0.581 (±0.110) | 0.711 (±0.060) | +0.130 (±0.125) |
| Race | 0.324 (±0.078) | 0.290 (±0.096) | -0.034 (±0.124) |
| White | 0.672 (±0.080) | 0.523 (±0.013) | -0.149 (±0.081) |
| Black | 0.793 (±0.066) | 0.779 (±0.013) | -0.014 (±0.067) |

*Table 2:* Comparison of results from the three different pipelines

## IMPLICATIONS AND RECOMMENDATIONS

This research has established that the aggregated local bias assessment method does not consistently measure the same bias scores as the accurately established global bias assessment. The next step would be to probe where these differences are coming from and in what scenario's they occur. Hereby, the goal is to obtain a bias assessment methodology that allows for an accurate, privacy-preserving and global measurement of bias within federated learning. Furthermore, this research has shown that federated learning can introduce more bias compared to central learning, although this is not the case for every metric, sensitive attribute and client partition. This occurrence of bias does show the need for a better understanding of bias in federated learning, to ensure safe use of FL in practice and industry. Moving forward, a promising direction is to identify what part of federated learning contributes to this bias, such that effective bias mitigation methods can be established for federated learning. This research has taken promising initial steps towards comprehensively understanding bias assessment in federated learning. It is hoped that these findings will inspire further exploration in this area, thereby contributing to the development of fair, privately-trained AI models.

Bias detection in data understanding and preparation phase - Proxies and Subgroups

# Causal Fairness Analysis with Automated Feature Engineering

Researcher: Wietse van Kooten

Link to research: Causal Fairness Analysis with Automated Feature Engineering

## INTRODUCTION

Machine learning (ML) is becoming crucial in decision-making processes across various fields like hiring, law enforcement, and healthcare. However, ML systems often inherit and amplify biases in their training data, leading to unfair and discriminatory outcomes. This paper addresses these challenges by applying causal fairness analysis, which combines causal inference with fair ML practices to detect, quantify, and mitigate biases in data and decision-making processes.

Understanding causal mechanisms, rather than merely identifying correlations, is fundamental in ML. Correlation can indicate a relationship between variables but does not imply causation. Causal inference, introduced by Judea Pearl, helps in determining how specific factors cause disparities, thereby enabling fairer decision-making processes. For example, causal inference was crucial in disproving the tobacco industry's claim that a genetic factor, rather than smoking itself, caused health issues.

In this research we discuss the importance of fairness in ML, guided by regulations like the General Data Protection Regulation (GDPR) and the EU AI Act. These regulations emphasize the need for transparency, accuracy, and non-discrimination in AI systems. Specifically, this research makes novel contributions to causal fairness analysis in the following ways:

- It extends the Standard Fairness Model (SFM) by incorporating automated feature engineering.
- It demonstrates that features developed through this extension enhance performance in both fairness and accuracy.

## METHODOLOGY

Causal inference aims to understand how changes in one variable influence another, using Structural Causal Models (SCMs). SCMs allow the modelling of causal relationships, enabling the calculation of potential outcomes and counterfactuals. This is needed to calculate various path-specific effects, like direct, indirect and spurious effects.

With causal fairness analysis, we decompose the direct, indirect, and spurious effects to measure their impact on fairness. The Standard Fairness Model (SFM) serves as a template for these analyses representing a range of causal diagrams. This model helps in understanding the causal structure and identifying biases inherent in the data. In Figure 13 we see such a template. The SFM model has been introduced by Drago Plecko, along with the whole concept of causal fairness analysis.

Automated feature engineering involves creating new features from existing data to enhance model performance and interpretability. This process helps in detecting trends across subgroups, addressing issues like Simpson's paradox. This research applies automated feature engineering within the SFM, demonstrating its compatibility and effectiveness in improving fairness. The experiments focus on two scenarios:

1. Automated feature engineering on features of W.
2. Automated feature engineering on features of W and Z.

The experiments use the COMPAS dataset, which assesses the likelihood of recidivism.



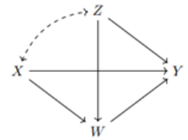



***Figure 13:*** The causal diagram of the SFM (standard fairness model). With X the protected attributes, Z confounding variables (often demographic variables), W mediators, and Y the outcome.

## EXPERIMENTS AND RESULTS

Results indicate that automated feature engineering improves model fairness and accuracy. The AFE model consistently shows slight improvements in accuracy and significant improvements in fairness, measured by metrics like the natural direct effect (NDE), natural indirect effect (NIE) and experimental spurious effect (Exp-SE). Table 3 shows one of the experiments, and how it improves both fairness and accuracy.

| $\lambda$ values | $\lambda = 0.1$ | | $\lambda = 0.5$ | | $\lambda = 1$ | |
|---|---|---|---|---|---|---|
| | Baseline | **AFE** | Baseline | **AFE** | Baseline | **AFE** |
| $\text{Accuracy}_{sd}$ | $0.683_{0.015}$ | $0.687_{0.013}$ | $0.683_{0.012}$ | $0.692_{0.012}$ | $0.665_{0.010}$ | $0.660_{0.010}$ |
| $\text{NDE}_{sd}$ | $0.066_{0.010}$ | $\mathbf{0.018_{0.010}}$ | $0.028_{0.008}$ | $\mathbf{0.015_{0.008}}$ | $0.008_{0.012}$ | $0.007_{0.007}$ |
| $\text{ExpSE}_{x_1}$ | $-0.024_{0.004}$ | $-0.028_{0.004}$ | $\mathbf{-0.015_{0.002}}$ | $-0.019_{0.003}$ | $-0.008_{0.002}$ | $-0.007_{0.002}$ |
| $\text{ExpSE}_{x_0}$ | $0.013_{0.002}$ | $0.012_{0.001}$ | $0.010_{0.001}$ | $\mathbf{0.007_{0.001}}$ | $0.005_{0.001}$ | $\mathbf{0.003_{0.001}}$ |

***Table 3:*** Results from one of the experiments

The top row of the table displays values of λ. Directly below each λ, the results for baseline and Automated Feature Engineering Extension are grouped. There are four metrics presented in the rows: Accuracy, NDE, $\text{ExpSE}_{x_0}$, and $\text{ExpSE}_{x_0}$. The standard deviation is noted in lowercase. If the values of the predictor for either baseline or Automated Feature Engineering (AFE) are significantly higher, the numbers are highlighted in bold.

Key Metrics from the experiments:

- Accuracy: Slight improvements observed in the AFE model across different λ values.
- NDE: Significant reduction in NDE values in the AFE model, indicating improved fairness on the direct effect. So the effect X -> Y.
- IDE: Significant reduction in NDE values in the AFE model, indicating improved fairness on the indirect effect. So the effect X -> W -> Y.
- Exp-SE: AFE model shows improved management of spurious effects, enhancing overall fairness. X -> Z -> Y, and X -> Z -> W -> Y.

## IMPLICATIONS AND RECOMMENDATIONS

Feature engineering has been widely studied and demonstrated to enhance accuracy and interpretability. The work of Salazar and others has shown the benefits of automated feature engineering while upholding fairness. This research builds on these foundations, extending feature engineering within the SFM framework to improve fairness and accuracy in ML models.

Incorporating automated feature engineering in the SFM framework enhances both the fairness and accuracy of ML models. This approach effectively manages biases and improves the transparency and interpretability of outcomes. The results support the potential of automated feature engineering as a valuable preprocessing step to the baseline model, particularly in terms of debiasing. By addressing the causal mechanisms of bias and implementing automated feature engineering, this research contributes to the development of fairer, more interpretable, and more accurate ML systems.

Bias detection in data understanding and preparation phase - Proxies and Subgroups

# Profile-Based Subgroup Discovery (PSD) for Fairness Analysis

Researcher: Dionne Gantzert

Link to research: Profile-Based Subgroup Discovery (PSD)

## INTRODUCTION

Machine learning models are increasingly being utilized in all types of applications, as for instance financial applications like credit scoring, where they play a crucial role in loan approvals and financial inclusion. However, these models can reproduce biases present in the data, leading to unfair outcomes, as for instance reinforcing unacceptable gender bias. In credit scoring, gender bias can result for instance, in qualified women being denied loans, hindering their access to financial resources.

Although fairness in machine learning has gained significant attention in recent years, most existing approaches focus on group fairness metrics, which are limited in their reliance on group-level averages, which can conceal outcome disparities for subgroups within a disadvantaged group. For instance, even if the average loan approval rates for men and women are similar, the model might still disadvantage women in specific subgroups, such as single mothers. This phenomenon is known as fairness gerrymandering, where subgroup fairness is sacrificed to achieve parity across entire groups. From a societal point of view, this type of treatment is clearly unfair, and thus unacceptable. From a technical point of view, it is crucial to incorporate subgroup fairness in the evaluation of AI systems to address fairness gerrymandering.

An essential aspect of subgroup fairness is the identification of subgroups. While recent studies have primarily focused on intersectional bias, which involves the combination of multiple sensitive attributes, defining subgroups should not be limited to this method. Predefining subgroups based solely on sensitive attributes may overlook other important biased relationships. In this context, it seems sound to utilize subgroup discovery (SD). SD aims to describe relationships between independent variables and specific target variable values, extracting significant rules through data mining techniques. This approach avoids predefined subgroups, identifying those that are most relevant to the decision-making process.

Clustering is a method that can be utilized as a subgroup discovery technique by dividing unlabeled data into subgroups based on similarity. Clustering for subgroup discovery (CSD) produces clusters that can be easily distinguished using simple decision rules, resulting in interpretable subgroups. While clustering can be used to identify subgroups, it differs from traditional subgroup discovery, which focuses on finding significant rules related to a target variable. In contrast, CSD aims to find significant rules that describe similar individuals without considering the target variable. This raises questions about how CSD would perform if the target variable was taken into account during the subgroup identification process. Additionally, since subgroup identification is an essential part of subgroup fairness, it also prompts inquiry into how CSD can identify subgroups that are treated unfairly by a classifier compared to conventional SD methods. To address these questions, this research proposes a novel clustering method designed to generate simple and interpretable clusters for subgroup discovery, called Profile-based Subgroup Discovery (PSD), based on previous semi-hierarchical methods for profile extraction. Our methodology involves two steps: first, partitioning the data space based on the target variable and then applying iterative clustering to obtain profiles; second, extracting descriptive rules from these profiles to identify subgroups. Like other CSD and SD techniques, PSD relies on discriminative decision rules that can be applied in real-world applications. Our method stands out by integrating the target variable into the clustering process, aligning it closely with subgroup discovery techniques. Our research aims to enhance the understanding of biased relationships within data by discovering subgroups unfairly treated by classifiers. We focus on two aspects: identifying subgroups exhibiting gender bias and identifying subgroups showing bias in general, regardless of sensitive attributes such as gender. Our approach was tested on the well-known German Credit dataset in the context of credit scoring.



## METHODOLOGY

PSD consists of two steps: (1) adapting the clustering pipeline to incorporate the target variable, and (2) developing a method to interpret and describe the resulting clusters. To incorporate the target variable as in subgroup discovery, we employ a targeted clustering approach. This involves splitting the data into two subsets based on the target variable's values (i.e., positive and negative labels). Subsequently, we apply the chosen clustering algorithm independently to each subset.

Figure 14 illustrates the difference between applying clustering directly to the data and using our approach. This strategy offers several advantages. Firstly, the resulting clusters will only contain instances with the same target label (positive or negative), facilitating a clear comparison between positive and negative profiles. For clarity, we will refer to the positive profiles as PSD+ and the negative profiles as PSD-. This allows us to investigate potential similarities or significant differences in their characteristics. Additionally, this approach enables a more focused analysis of subgroups within each target class, potentially leading to more actionable insights in machine learning fairness. By concentrating on one specific target variable at a time, such as clustering positive instances, this approach aligns with subgroup discovery techniques that also focus on describing rules for a specific target variable. The clustering algorithm used in this research is the variability controlled hierarchical K-medoids (VHK), as proposed by Wilms et al. (2022). Subsequently, we described the clusters using profile descriptions. Each profile description represents a subgroup identified by the VHK algorithm.

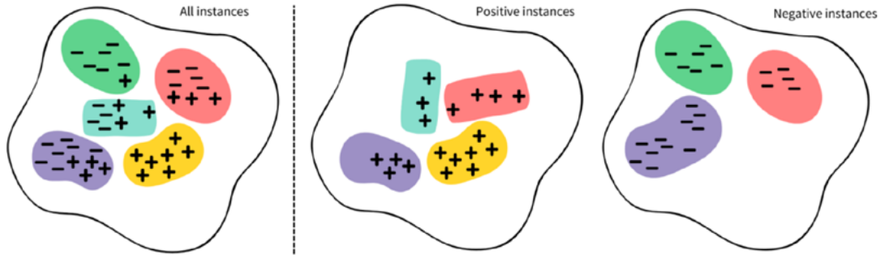
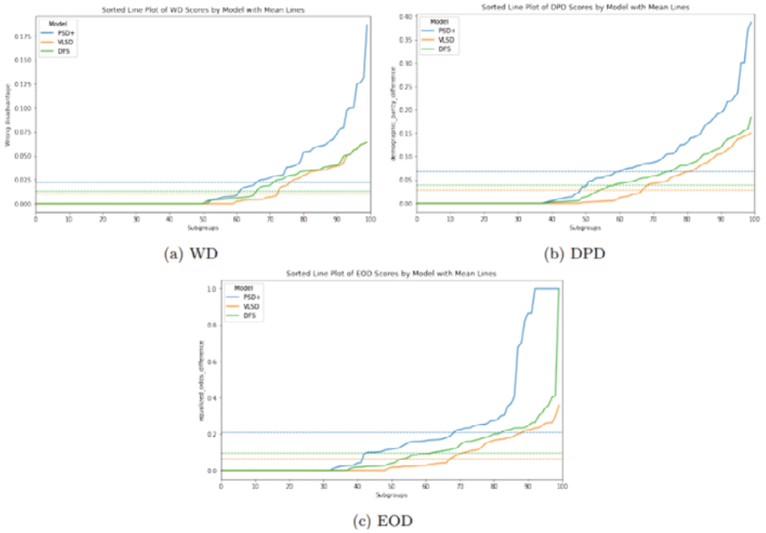



***Figure 14:*** Illustration of a traditional clustering (left) in comparison to our clustering method (right).

## EXPERIMENTS AND RESULTS

To detect gender bias in the subgroups, we applied three distinct fairness metrics: wrong disadvantage, demographic parity difference, and equalized odds difference. Wrong disadvantage is also viewed as the percentage of false negatives. Each identified subgroup was evaluated using these metrics for the logistic regression and XGBoost classifiers. Notably, even though the 'Sex' attribute was excluded from the clustering process, PSD identified more gender-biased subgroups compared to traditional subgroup discovery algorithms such as DFS and VLSD. We can see this in Figure 15, where the area under the curve is bigger for PSD than for VLSD and DFS. A deeper understanding of these biases would require a qualitative analysis of the subgroup descriptions.

We also identified subgroups exhibiting general bias using four distinct fairness metrics: demographic parity difference (DPD) subgroup fairness, equalized odds difference (EOD) subgroup fairness, true positive difference (TPD) subgroup fairness, and false positive difference (FPD) subgroup fairness. These metrics assess the fairness of subgroups relative to the overall data fairness. For DPD and EOD, lower scores indicate subgroups similar to the data distribution, while higher scores indicate more distinct subgroups. Conversely, for TPD and FPD, higher scores indicate subgroups similar to the data distribution. When identifying bias in these subgroups, we focus on higher scores for DPD and EOD, and lower scores for TPD and FPD. As shown in Figure 16, our PSD algorithm identifies subgroups that score lower in DPD and EOD, and higher in TPD and FPD, indicating that our PSD algorithm finds subgroups most similar to the data and thus identifies less biased subgroups in general.




(a) WD (b) DPD

(c) EOD

***Figure 15:*** Gender Bias Subgroup Fairness measures.

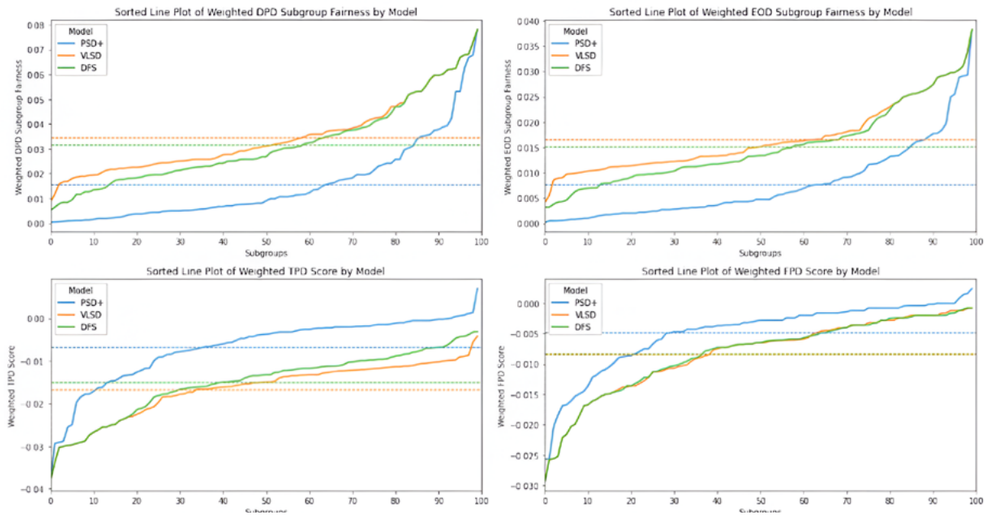



***Figure 16:*** General Bias Subgroup Fairness measures.

## IMPLICATIONS AND RECOMMENDATIONS

Based on the results for both gender bias and general bias identification, we find that PSD effectively identifies gender-biased subgroups, which can help in practice to identify subgroups exhibiting more gender bias. However, while we identify more gender-biased subgroups, this study did not further investigate the characteristics of these subgroups, which is necessary for a comprehensive evaluation. Additionally, PSD finds more subgroups similar to the data distribution, which may not be advantageous in practice, as we seek more distinct subgroups. Overall, this research aids in identifying subgroups, but further investigation is needed to evaluate the fairness of these subgroups comprehensively.



# Limitations of the methods and tools tested

## Methods for Traditional AI models

These methods are effective in controlled environments where biases are well-defined and can be explicitly measured. Techniques like Profile-Based Subgroup Discovery (PSD) and Causal Fairness Analysis have shown success in identifying and mitigating bias in structured data applications like credit scoring and recidivism prediction. In contrast, these methods often struggle with scalability and generalization across different contexts. They are limited by the need for predefined fairness metrics and may not capture complex biases that emerge in unstructured data.

## Methods for LLMs

Tools designed for LLMs, such as SEAT, DisCo, and CSPS, have been effective in detecting biases in text generation tasks. These tools leverage the contextual understanding of LLMs to identify gender, racial, and other biases in generated content. The main limitation of these tools is their dependency on extensive computational resources and the difficulty in interpreting high-dimensional embeddings. Additionally, these tools often require access to large, unbiased datasets for effective fine-tuning, which may not always be available.

# Challenges and shortcomings

**Limited scope of fairness metrics**
Many existing tools focus on a narrow set of fairness metrics, which may not fully capture the multifaceted nature of bias in AI/ML systems. For example, traditional metrics like demographic parity do not account for intersectional biases that affect subgroups within larger demographic categories.

**Underrepresentation of certain bias types**
Research and tools often emphasize gender and racial biases, while other important dimensions, such as socio-economic status, disability, and age, receive less attention. This gap limits the applicability of these tools in diverse real-world scenarios.

**Challenges in Federated Learning**
Bias detection and mitigation in federated learning environments remain underexplored. Existing methods are not well-equipped to handle the decentralized nature of federated learning, where data privacy and heterogeneity present unique challenges.

**Data dependency**
Many bias detection tools require access to large and diverse datasets to function effectively. However, acquiring such datasets is challenging due to privacy concerns, data availability, and the cost associated with data collection and annotation.



**Complexity and interpretability**
Advanced models, particularly LLMs, pose significant challenges in terms of interpretability. Understanding how these models encode and propagate biases requires sophisticated techniques that are still in developmental stages.

**Scalability issues**
Scalability is a significant concern. Implementing bias detection and mitigation across large-scale, real-time systems demands computational efficiency and robustness, which many current tools lack.

# Conclusion

The challenge of bias in AI and machine learning is far from resolved, but our research offers critical insights into both the progress made and the work that remains. While existing bias detection tools have shown promise, their effectiveness varies significantly depending on the context, revealing that no single solution can address all forms of bias. Our findings indicate that combining approaches from both traditional AI and Large Language Models (LLMs) often yields better results, though significant limitations in scalability and coverage persist.
The innovative techniques explored in this white paper point to exciting new possibilities for bias detection and mitigation, particularly in understanding hidden biases and addressing them more effectively. However, this research also highlights the importance of continued experimentation and refinement. The path forward requires not only improved tools but also ongoing collaboration between researchers, industry leaders, and policymakers to ensure that AI systems evolve toward greater fairness and accountability. The need for ethical, transparent, and inclusive AI practices has never been more urgent, and we believe this work provides a foundation for future advancements in this critical area.

# GitHub Library

- Links to Code & Research repository: **Rhite Research Repositories**
- Link to synthetic datasets: **Synthetic datasets**

# Acknowledgements

We would like to extend our sincere gratitude to the following individuals and organizations who contributed to the development of this white paper:

**University of Amsterdam (UvA)**
We deeply appreciate the collaboration with the University of Amsterdam, particularly the six master's students from the AI program (Master in Artificial Intelligence) whose thesis research formed the foundation of our analysis, and their supervisors Dr. E. Acar, Leonard Bereska MSc and Dr G Sileno.

*Student Researchers:*
Alexia Muresan
Dennis Agafonov
Dionne Gantzert
Jelke Matthijsse
Tarmo Pungas
Wietse van Kooten

**Reviewers and Supervisors**
We appreciate the valuable guidance and feedback from our peer reviewer and senior data scientist who ensured the quality and rigor of this white paper.
*External Supervisor*: Shieltaa Rita Dewika Dielbandhoesing

**Industry & Academic contributions**
We are grateful for the insights and feedback from industry and academic organisations who contributed to the research.
*Industry*: IBM Nederlands, Sabiha Majumder from ABN AMRO, Iker Ceballos from Acuratio
*Academia*: Leiden University (BIAS project), Sainyam Galhotra (Cornell University)

**Rhite Team**
Special thanks to the team members at Rhite for their expertise in responsible AI and their dedication, which guided the direction, design and focus of the six studies and this white paper.
*Lead Researchers*: Isabel Barberá
*Support Team*: Lea Magnano

Your collective efforts and commitment to advancing the field of AI ethics and bias mitigation have been invaluable to this project!



# Contacts

For further discussion or questions regarding this white paper, please contact:
**Isabel Barberá** | Head of AI Research at Rhite
Email: **info@rhite.tech**

General Inquiries
**Rhite**
Website: **www.rhite.tech**

We welcome your feedback and look forward to engaging with you on the important topic of bias detection and mitigation in AI systems.

# References

## AI-Based Hiring and the Appeal of Novelty: Do LLMs Solve or Exacerbate the Problem of Discrimination?


- Jaimeen Ahn and Alice Oh. Mitigating language-dependent ethnic bias in bert. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021.
- Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. Measuring implicit bias in explicitly unbiased large language models, 2024.
- Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. Investigating gender bias in bert. Cognitive Computation, 13(4):1008–1018, May 2021.
- Ishita Chakraborty, Khai Chiong, Howard Dover, and K. Sudhir. AI and AI-human based salesforce hiring using interview videos. SSRN Electronic Journal, 2022.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. Association for Computing Machinery, 2024.
- Zhisheng Chen. Collaboration among recruiters and artificial intelligence: Removing human prejudices in employment. Cognition, Technology amp; Work, 25(1):135–149, 2022.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, 2018.
- Xiangjue Dong, Yibo Wang, Philip S. Yu, and James Caverlee. Disclosure and mitigation of gender bias in llms. 2024.
- Eleanor Drage and Kerry Mackereth. Does AI debias recruitment? race, gender, and AI's "eradication of difference". Philosophy Technology, pages 35–89, 2022.
- Fairlearn. Common Fairness: https://fairlearn.org/main/userguide/assessment/commonfairnessmetrics.html. Accessed :2024
- Li Lucy and David Bamman. Gender and representation bias in gpt-3 generated stories. Proceedings of the Third Workshop on Narrative Understanding, 2021.
- Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. ACM Computing Surveys, 2021.
- Sushruta Mishra, Pradeep K Mallik, Hrudaya K Tripathy, Lambodar Jena, and Gyoo-Soo Chae. Stacked knn with hard voting predictive approach to assist hiring process in it organizations. International Journal of Electrical Engineering Education, 2021.
- Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. Hate speech detection and racial bias mitigation in social media based on bert model. PLOS ONE, 15(8), Aug 2020.
- Rizwan Qureshi Abbas Shah amgad muneer Muhammad Irfan Anas Zafar Muhammad
- Bilal Shaikh Naveed Akhtar Jia Wu Seyedali Mirjalili Mubarak Shah Muhammad Usman Hadi, qasem al tashi. A survey on large language models: Applications, challenges, limitations, and practical usage. 2023.
- Sydney Myers. 2023 applicant tracking system (ats) usage report: Key shifts and strategies for job seekers, May 2024.
- Selin E. Nugent and Susan Scott-Parker. Recruitment ai has a disability problem: Anticipating and mitigating unfair automated hiring decisions. Intelligent Systems, Control and Automation: Science and Engineering, page 85–96, 2022.
- Devah Pager, Bruce Western, and Bart Bonikowski. Discrimination in a low-wage labour market: A field experiment. American Sociological Review, 74(5):777–799, 2009.
- Jahan Mohan Reddy, Sirisha Regella, and Srinivasa Reddy Seelam. Recruitment prediction using machine learning. 2020 5th International Conference on Computing, Communication and Security (ICCCS), pages 1–4, 2020.
- Jonas Rieskamp, Lennart Hofeditz, Milad Mirbabaie, and Stefan Stieglitz. Approaches to improve fairness when deploying AI-based algorithms in hiring – using a systematic literature review to guide future research. Proceedings of the Annual Hawaii International Conference on System Sciences, 2023.
- Joel Silas, Prajakta Udhan, Pranali Dahiphale, Vaibhav Parkale, and Poonam Lambhate. Automation of candidate hiring system using machine learning. International Journal of Innovative Science and Research Technology, 8, 2023.
- Andrew C. Wicks, Linnea P. Budd, Ryan A. Moorthi, Helet Botha, and Jenny Mead. Automated hiring at amazon. 2021.
- Paris Will, Dario Krpan, and Grace Lordan. People versus machines: Introducing the hire framework. Artificial Intelligence Review, 56(2):1071–1100, May 2022.
- Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, and et al. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: A model evaluation study. The Lancet Digital Health, 6(1), Jan 2024.
- Sijing Zhang, Ping Li, and Ziyan Cai. Are male candidates better than females? Debiasing bert resume retrieval system. 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Oct 2022.
- Jiaxu Zhao, Meng Fang, Shirui Pan, Wenpeng Yin, and Mykola Pechenizkiy. Gptbias: A comprehensive framework for evaluating bias in large language models, 2023.




## Unveiling the Mechanisms of Bias in LLMs by Eliciting Latent Knowledge


- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, pages 610–623, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445922. URL https://dl.acm.org/doi/ 10.1145/3442188.3445922.
- Paul Christiano, Ajeya Cotra, and Mark Xu. Eliciting Latent Knowledge. Technical report, Alignment Research Center, December 2021. URL https://docs.google.com/document/d/1WwsnJQstPq91_Yh-Ch2XRL8H_EpsnjrC1dwZXR37PC8/.
- Jesse Dodge, Maarten Sap, Ana Marasovi´c, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1286–1305, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.98. URL https://aclanthology.org/2021. emnlp-main.98.
- Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. Large Language Models in Education: Vision and Opportunities, November 2023. URL http://arxiv. org/abs/2311.13160. arXiv:2311.13160 [cs].
- Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. A Survey of Large Language Models for Healthcare: from Data, Technology, and Applications to Accountability and Ethics, October 2023. URL http://arxiv.org/abs/ 2310.05694. arXiv:2310.05694 [cs].
- Samuel Marks and Max Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets, December 2023. URL http://arxiv.org/abs/2310.06824. arXiv:2310.06824 [cs].
- Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. Societal Biases in Language Generation: Progress and Challenges. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4275–4293, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.330. URL https://aclanthology.org/2021.acl-long.330.
- Oskar van der Wal, Dominik Bachmann, Alina Leidinger, Leendert van Maanen, Willem Zuidema, and Katrin Schulz. Undesirable Biases in NLP: Addressing Challenges of Measurement. Journal of Artificial Intelligence Research, 79:1–40, January 2024. ISSN 1076-9757. doi: 10.1613/jair.1.15195. URL http://arxiv.org/abs/2211.13709. arXiv:2211.13709 [cs].
- Jingying Zeng, Richard Huang, Waleed Malik, Langxuan Yin, Bojan Babic, Danny Shacham, Xiao Yan, Jaewon Yang, and Qi He. Large Language Models for Social Networks: Applications, Challenges, and Solutions, January 2024. URL http://arxiv.org/ abs/2401.02575. arXiv:2401.02575 [cs].


## Assessing and Addressing Gender Bias in Large Language Models


- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language Models are Few-Shot Learners.
- Cabrera, J., Loyola, M. S., Magaña, I., & Rojas, R. (2023). Ethical dilemmas, mental health, artificial intelligence, and LLM-Based chatbots.
- Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018, October 11). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
- Eigner, E., & Händler, T. (2024, February 27). Determinants of LLM-assisted Decision-Making. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/
- Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2023, September 2). Bias and Fairness in Large Language Models: A survey.
- Workshop, B., Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., . . . Wolf, T. (2022, November 9). BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.
- Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., Chen, J., Chi, E., & Petrov, S. (2020). Measuring and reducing gendered correlations in pre-trained models.
- Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders.
- Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020, September 30). CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models.

## Bridging Fairness and Privacy: Bias Assessment in Federated Learning


- Annie Abay, Yi Zhou, Nathalie Baracaldo, Shashank Rajamoni, Ebube Chuba, and Heiko Ludwig. Mitigating bias in federated learning. arXiv preprint arXiv:2012.02447, 2020.
- Annie Abay, Ebube Chuba, Yi Zhou, Nathalie Baracaldo, and Heiko Ludwig. Addressing unique fairness obstacles within federated learning. AAAI RDAI-2021, 2021.
- Hongyan Chang and Reza Shokri. Bias propagation in federated learning. arXiv preprint arXiv:2309.02160, 2023.
- Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A Salman Avestimehr. Fairfed: Enabling group fairness in federated learning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, pages 7494–7502, 2023.
- Peter Kairouz, H Brendan McMahan, Brendan Avent, Aur´elien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. Foundations and trends® in machine learning, 14(1–2):1–210, 2021.
- Daniel Yue Zhang, Ziyi Kou, and Dong Wang. Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In 2020 IEEE International Conference on Big Data (Big Data), pages 1051–1060. IEEE, 2020.

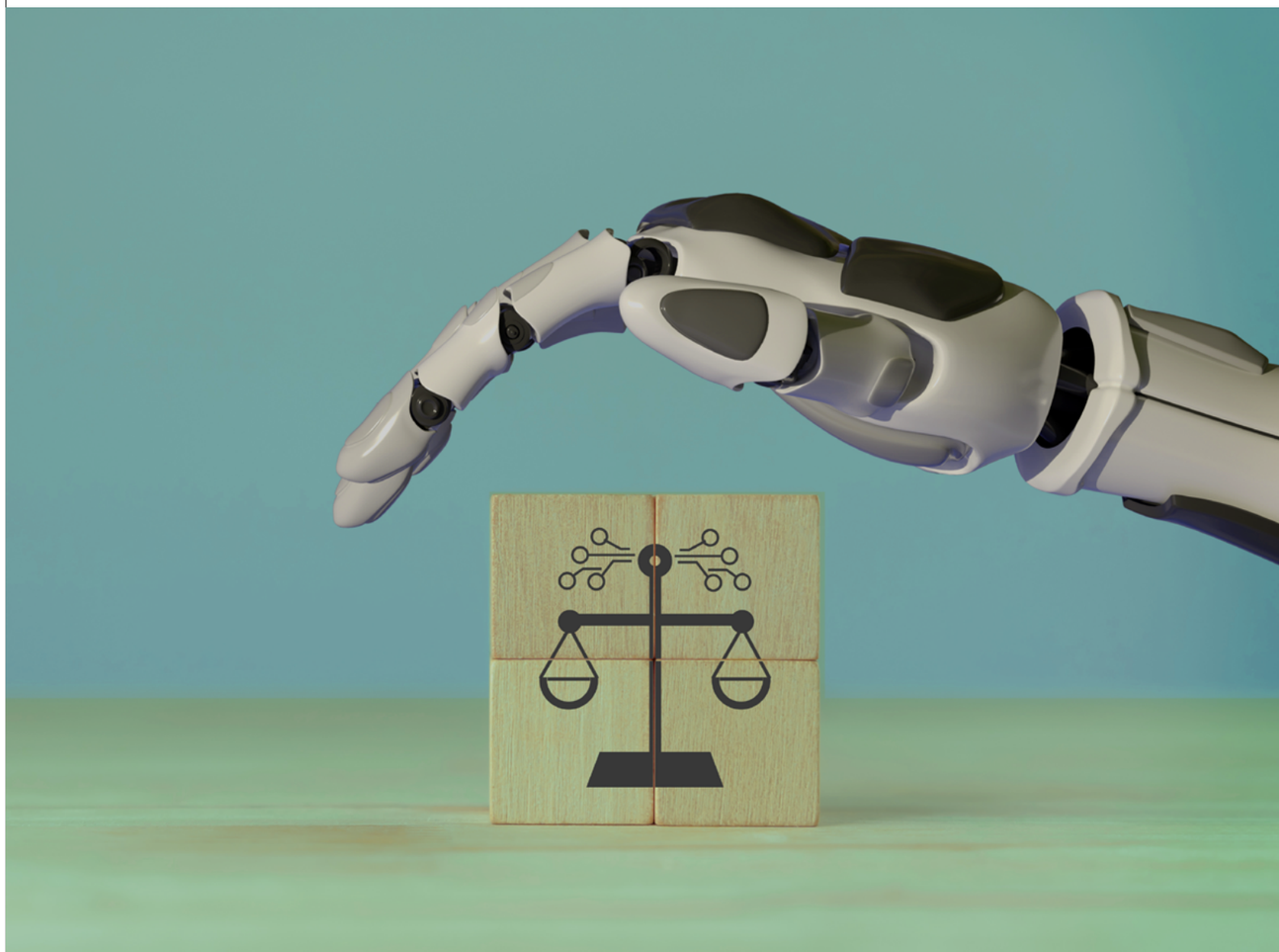

## Causal Fairness Analysis with Automated Feature Engineering


- Silvia Chiappa and Thomas P. S. Gillam. Path-specific counterfactual fairness. stat.ML, arxiv(arXiv:1802.08139), 2018.
- Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, 2017.
- Sainyam Galhotra, Karthikeyan Shanmugam, Prasanna Sattigeri, and Kush R. Varshney. Causal feature selection for algorithmic fairness. arxiv, 2022.
- Lauren Kirchner Jeff Larson, Surya Mattu and Julia Angwin. How we analyzed the compas recidivism algorithm. ProPublica, 2016.
- Udayan Khurana, Deepak Turaga, Horst Samulowitz, and Srinivasan Parthasrathy. Cognito: Automated feature engineering for supervised learning. 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), pages 1304–1307, 2016. Conference Name: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW) ISBN: 9781509059102 Place: Barcelona, Spain Publisher: IEEE.
- Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.
- Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. ACM, 54(6):115:1–115:35, 2021.
- Tiago P. Pagano, Rafael B. Loureiro, Fernanda V. N. Lisboa, Rodrigo M. Peixoto, Guilherme A. S. Guimaraes, Gustavo O. R. Cruz, Maira M. Araujo, Lucas L. Santos, Marco A. S. Cruz, Ewerton L. S. Oliveira, Ingrid Winkler, and Erick G. S. Nascimento. Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. Big Data and Cognitive Computing, 7(1), 2023.
- J. Pearl, M. Glymour, and N.P. Jewell. Causal Inference in Statistics: A Primer. Wiley, 2016.
- Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 50(302):157–175, July 1900.
- Drago Plecko and Elias Bareinboim. A causal framework for decomposing spurious variations. SIGMOD, arxiv(arXiv:2306.05071), 2023.
- Drago Plecko and Elias Bareinboim. Reconciling predictive and statistical parity: A causal approach. SIGMOD, arxiv(arXiv:2306.05059), 2023.
- Drago Plecko and Nicolai Meinshausen. Fair data adaptation with quantile preservation. J. Mach. Learn. Res., 21(1), jan 2020.
- Drago Plecko and Elias Bareinboim. Causal fairness analysis: A causal toolkit for fair machine learning. FNT in Machine Learning, 17(3):304–589, 2024.
- Ricardo Salazar, Felix Neutatz, and Ziawasch Abedjan. Automated feature engineering for algorithmic fairness. VLDB, 14(9):1694–1702, 2021.
- Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In Proceedings of the 2019 International Conference on Management of Data, pages 793–810. ACM, 2019.
- Yishai Shimoni, Ehud Karavani, Sivan Ravid, Peter Bak, Tan Hung Ng, Sharon Hensley Alford, Denise Meade, and Yaara Goldschmidt. An evaluation toolkit to guide model selection and cohort definition in causal inference. stat.ML, arxiv(arXiv:1906.00442), 2019.
- Jin Tian and Judea Pearl. A general identification condition for causal effects. In Eighteenth National Conference on Artificial Intelligence, page 567–573, USA, 2002. American Association for Artificial Intelligence.


## Profile-Based Subgroup Discovery (PSD) for Fairness Analysis


- Sacha E. Buijs. Clustering algorithms and concept descriptors in constructing conceptual spaces. Bachelor's thesis, University of Amsterdam, Faculty of Science, Science Park 900, 1098 XH Amsterdam, 2023. Supervisor: Dr. G. Sileno.
- Maarten Buyl and Tijl De Bie. Inherent limitations of ai fairness. Communications of the ACM, 67(2):48–55, 2024.
- CJ Carmona and David Elizondo. Subgroup discovery: Real-world applications. Technical report, Techincal Report, 2011.
- Aidan Cooper, Orla Doyle, and Alison Bourke. Supervised clustering for subgroup discovery: An application to covid-19 symptomatology. In Joint European conference on machine learning and knowledge discovery in databases, pages 408–422. Springer, 2021.
- Edwin S Dalmaijer, Camilla L Nord, and Duncan E Astle. Statistical power for cluster analysis. BMC bioinformatics, 23(1):205, 2022.
- Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In International conference on machine learning, pages 2564– 2572. PMLR, 2018.
- Kenji Kobayashi and Yuri Nakao. One-vs.-one mitigation of intersectional bias: A general method to extend fairness-aware binary classification. arXiv preprint arXiv:2010.13494, 2020.
- Mieke Wilms, Giovanni Sileno, and Hinda Haned. Pebam: A profile-based evaluation method for bias assessment on mixed datasets. In German Conference on Artificial Intelligence (K¨unstliche Intelligenz), pages 209–223. Springer, 2022.



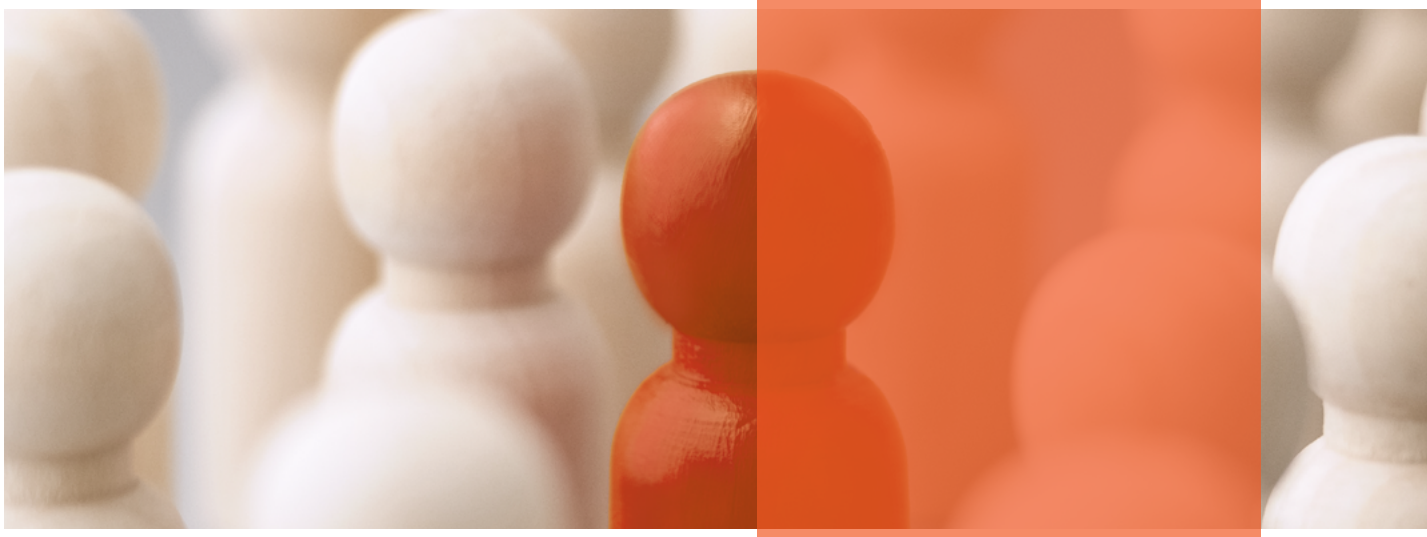



WHITE PAPER

# Advancing the field of bias detection and mitigation in Large Language Models and Traditional AI Models

Research of bias in Large Language Models (LLMs), Federated Learning, Automated Feature Engineering, and Unfairness in Subgroups

**Leading the way to Trustworthy AI**



**Visit our website**

www.rhite.tech

Rhite

Leading the way to Trustworthy AI