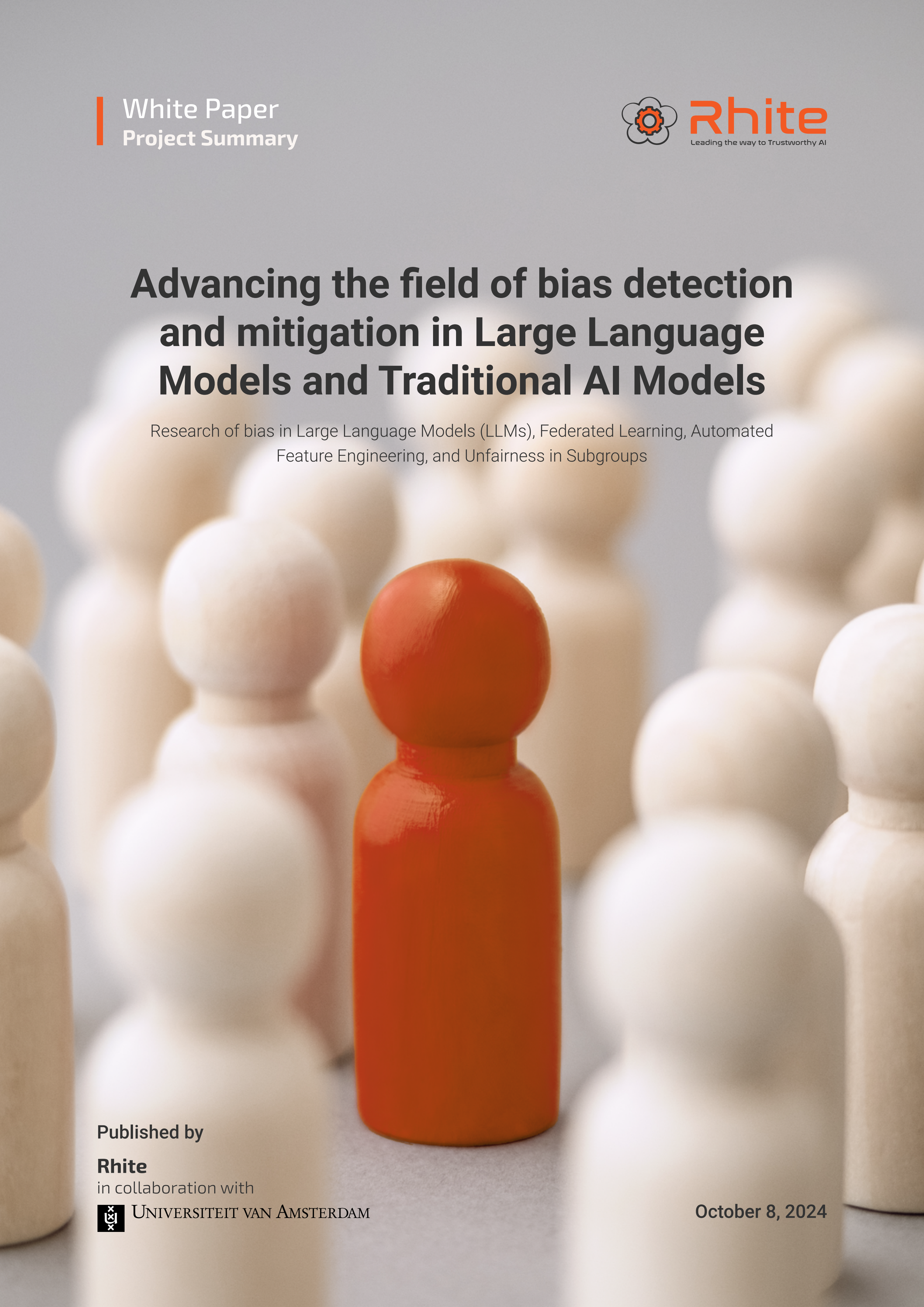
White Paper
Project Summary

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.

READ THE **FULL STORY:**

Check out the **extended white paper** to learn how we got to the results presented in this document.
See our previous **research** for an overview of bias throughout the lifecycle of AI systems.

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

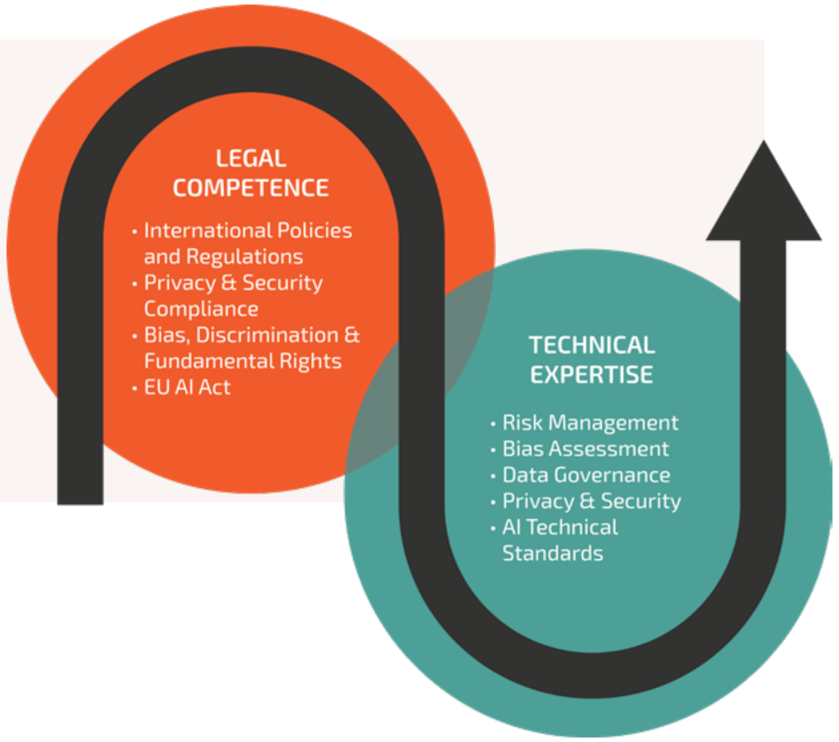
- Legal and technical consultancy on AI;
- Guidance to comply with the requirements of the EU AI Act;
- Auditing of algorithms and AI systems;
- Privacy, security, safety and fundamental rights Impact assessments of AI solutions;
- Bespoke trainings on AI Risk Management;
- Implementation of Responsible 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 #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 #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 #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 #3
TOPIC
Gender Bias in LLMs
Comparing in-processing and post-processing techniques to assess gender bias.

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)**

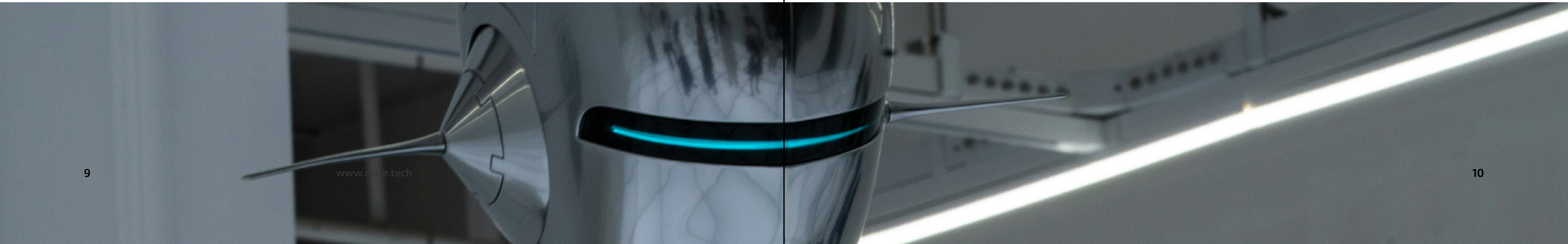
Models: **Structural Causal Model (SCM) and Standard Fairness Model (SFM)**
Dataset: **COMPAS**

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."

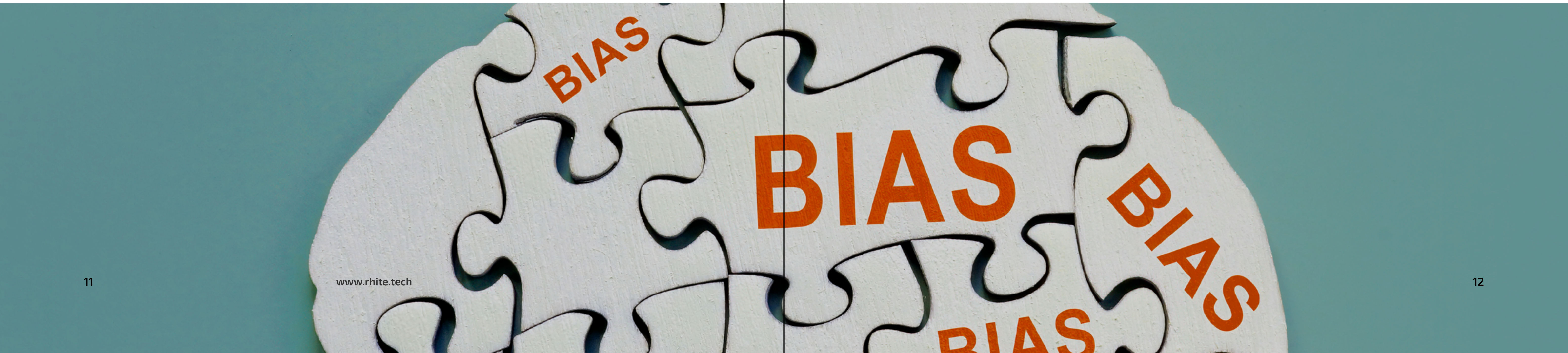
Conclusions

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.

GitHub Library

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



WANT TO KNOW MORE?

KEEP READING!

You've just reached the end of the project overview.
In the extended version of the white paper, we take a closer look at the research that shaped the work presented here. That 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.

Find the extended white paper here!

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 research, please contact:
Head of AI Research at Rhite: **Isabel Barberá -** email: **info@rhite.tech**

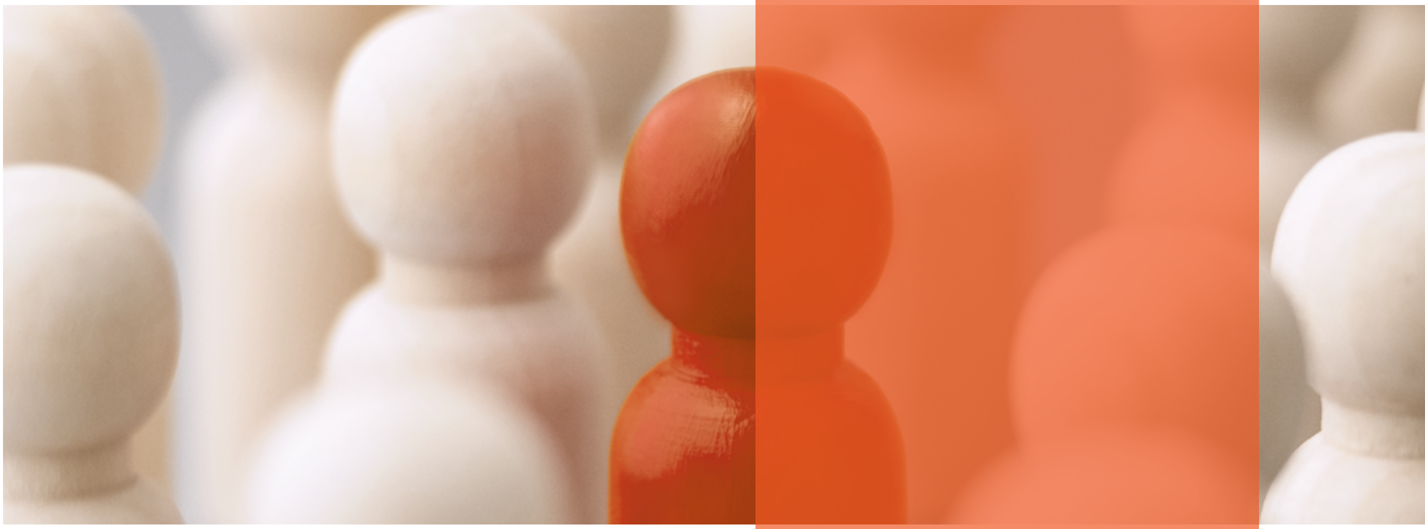
We welcome your feedback and look forward to engaging with you on the important topic of bias detection and mitigation in AI systems.

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