Lawrence Alaso Krukrubo

AI Safety Researcher — Lecturer — Explainable AI (XAI) Specialist

Wolverhampton, UK — +44 7553 726128 — L.A.Krukrubo@wlv.ac.uk

LinkedIn — GitHub — Medium — ORCID

PROFESSIONAL SUMMARY

Academic Researcher and Data Scientist specializing in **AI Safety, Causal Fairness, and Explainable AI (XAI)**. Currently pursuing a PhD focused on mitigating algorithmic bias amplification in LLM-enhanced systems. Proven expertise in fine-tuning Large Language Models (LLMs) for alignment and deploying defense-in-depth strategies for AI safety. Combining a distinction-level academic background with practical industry experience, I aim to bridge the gap between theoretical fairness frameworks and robust, deployable machine learning systems.

RESEARCH & TECHNICAL PROJECTS

PhD Research: Causal Fairness in LLM-Enhanced Recommender Systems
University of Wolverhampton — Current

- **Focus:** Investigating how LLMs causally influence distributional and representational harms in multi-sided platforms.
- **Methodology:** Moving beyond correlation-based fairness metrics to **Causal Structural Modeling**. Analyzing "rich-get-richer" feedback loops where LLM latent representations amplify pre-existing biases.
- **Mitigation:** Developing LLM-driven post-processing re-ranking modules to enforce multi-stakeholder fairness constraints (balancing Utility vs. Fairness).

Visiting Member
London Initiative for Safe AI (LISA) — 2024 – Present

- Engaging with the AI safety research community to collaborate on robust evaluation frameworks and alignment methodologies.

AI Alignment & Safety Engineering
Bluedot Impact — 2024

- **Defense-in-Depth (Technical AI Safety Course):** Currently optimizing safety techniques for high-stakes AI deployment. Designing a multi-layered security protocol to prevent adversarial attacks and model misuse (Fundable Project).
- **Bias Mitigation in SMOL Models (AI Alignment Course):** Successfully fine-tuned Hugging Face open-source SMOL models to identify and mitigate gender bias. Utilized Supervised Fine-Tuning (SFT) and evaluation benchmarks to reduce stereotypical output generation.

MSc Dissertation: Improving Credit Approval Decisions with Explainable ML
University of Wolverhampton — 2023

- **Implementation:** Deployed **IBM's Contrastive Explanation Method (CEM)** to improve transparency in "Black-box" Deep Neural Networks (DNNs) for credit scoring.
- **Innovation:** Decomposed predictions into **Pertinent Negatives (PN)** and **Pertinent Positives (PP)** to provide actionable "counterfactual" explanations for rejected loan applicants.

EDUCATION

PhD in Computer Science (AI Safety & Fairness)
University of Wolverhampton — Sept 2025 – Present

- *Roadmap:* Causal modeling of LLM bias, empirical validation of feedback loops, and design of fairness-aware re-ranking algorithms.

Technical AI Safety Curriculum
Bluedot Impact — 2024

- Advanced curriculum covering mechanistic interpretability, scalable oversight, and governance.

MSc Artificial Intelligence (Distinction)
University of Wolverhampton — 2023

BSc Banking and Finance
Rivers State University

TECHNICAL SKILLS

- **AI Safety & XAI:** Causal Inference, Counterfactual Analysis, IBM AIX360 (CEM, LIME, SHAP), Fairness Metrics (EUR/RUR), Adversarial Robustness, RLHF.
- **Machine Learning & LLMs:** PyTorch, TensorFlow/Keras, Hugging Face Transformers, Fine-tuning (PEFT/LoRA), LangChain, Scikit-learn.
- **Data Science & Analytics:** Python, R (Tidyverse, ggplot2), SQL, SAS (Visual Analytics/CAS), BigQuery, Apache Spark.
- **Visualization:** Matplotlib, Seaborn, Tableau, Power BI, Plotly.

ACADEMIC & PROFESSIONAL EXPERIENCE

Lecturer: Faculty of Science and Engineering
The University of Wolverhampton — Jan 2024 – Present

- Delivering advanced curriculum on **Statistics in R**, **Data Science**, and **SAS Cloud Analytic Services**.
- Supervising postgraduate research projects, specifically guiding students in ML interpretability and methodologies.
- Integrating modern industry standards (e.g., Tidyverse workflows) into academic modules to enhance employability.

Visiting Lecturer: School of Mathematics & Computer Science
The University of Wolverhampton — Nov 2023 – Jan 2024

- Taught postgraduate modules on Data Visualization and Statistics.
- Developed coursework focused on rigorous statistical testing and visual storytelling using R and SAS Data Studio.

Data Boot Camp Instructor
Black Girls In Tech, London — Aug 2023 – Nov 2023

- Designed and delivered a comprehensive Data Science curriculum (Python, SQL, Scikit-learn).
- Mentored students through end-to-end ML projects, focusing on Data Wrangling and Model Deployment.

Data Analyst Session Lead
Udacity (Remote) — May 2021 – Jul 2023

- Led a team of mentors supporting students in the **Data Analyst** and **AI Programming** Nanodegrees.
- Performed root cause analysis on student churn, implementing interventions that increased completion rates from 53% to 78%.

Data Specialist
Tech Layer, Lagos — Jun 2019 – Mar 2021

- Engineered semi-automated ML pipelines using **AutoML** and **LIME** to create transparent predictive models.
- *Achievement:* Recipient of the IBM Explainable Data Science Badge (2020).

Early Career History (2011–2019): Held various Sales Management and Business Development leadership roles within the Telecommunications sector (Globacom, Airtel), demonstrating extensive stakeholder management and strategic planning experience.

PUBLICATIONS & CERTIFICATIONS

- **Certifications:** IBM Advanced Data Science Professional, DeepLearning.AI Convolutional Neural Networks, Udacity AI Nanodegree.
- **Writing:** Active contributor to Medium (Artificial Intelligence & Data Science) and Technical writer on XAI concepts.