CENTRE of EXCELLENCE in FINANCIAL SERVICES

Directed Risk Research Programme

DIRECTED RESEARCH PROBLEM STATEMENT

Research Theme	Bias in Advanced Analytics	**Problem Nr.**	PS21007

Client Name	Lian van Oudheusden			**Client Org.**	FirstRand
Designation	Head: Model Risk Management				
E-mail	Lian.vanoudheusden@firstrand.co.za	**Tel (w)**	0112821213	**Mobile**	0834594227

1. **Project Title:** Determining Methodologies for Ongoing Monitoring for Model Bias

2. **Project Goal:** To define methodologies for ongoing monitoring of model inputs, outputs and decision outcomes for emerging bias

3. **Higher level description of problem:** As access to diverse sources of data increases and financial services organisations seek competitive advantage through innovation, application of advanced analytics tools such as artificial intelligence and machine learning is receiving increased focus and investment. South Africa's overall investment in artificial intelligence (AI) over the last decade is significant, with around $1.6-billion invested by 2019. These investments have seen businesses experimenting with a range of different technologies, including chatbots, robotic process automation and advanced analytics.

   These methods provide powerful tools for the enhancement of business value if used responsibly. Misuse of these tools, on the other hand, can result in severely detrimental outcomes. Examples of such detrimental outcomes have received widespread publicity, leading in part to potential distrust amongst both internal stakeholders and external stakeholders such as clients and regulators. The amount of money devoted to artificial intelligence is significant. But for all its reliance on notionally unbiased data, AI can end up very biased, because it is designed by people, and trained on data sets chosen and created by them. How do we recognise and guard against bias in artificial intelligence?

   Financial services organisations have recognized the incremental risks that use of these tools introduces and have responded by investing in research and development of risk management frameworks aimed at addressing such risks. Particular focus has been placed on developing

mechanisms for fairness and avoidance of unintended bias in analytics-based decision-making tools.

The nuance of human intelligence is invaluable when it comes to shaping AI, and it's that intelligence that's ultimately liable for anything generated by, or resulting from, placing faith in an artificial version of it. The biases outlined above exist in people long before they do in machines, so while it's imperative to weed them out of machines, it's also an excellent opportunity for reflection. If we examine our own human weaknesses and failings and try to address them, we're vastly less likely to introduce them into the AI systems we create, whether now or in the future

While significant research has been done on assessing data, model outputs and decision outcomes for bias prior to deployment, further work is required to establish robust and implementable methodologies for ongoing monitoring of solutions once in production.

4. **Project objectives:** To research methodologies for ongoing monitoring of models for fairness/unintended bias and to make recommendations regarding robust, implementable approaches for adoption within the financial services industry.

5. **Outputs required:** A research report detailing available methodologies for ongoing monitoring for bias and providing recommendations for suitable methodologies that could be adopted within the banking industry. The research report should include examples demonstrating the application of recommended methodologies and suitable thresholds, where relevant.

6. **Funding for project:** Project funding can be discussed on a case be case basis with input from the relevant academic institution and would be dependent on the level of research to be conducted.

7. **Strategic value to directed risk research:** This research will contribute to the development of suitable approaches for ongoing avoidance of unintended bias in artificial intelligence and machine learning models. Such monitoring approaches will protect clients from such unintended bias, contributing to enhanced trust in financial services.