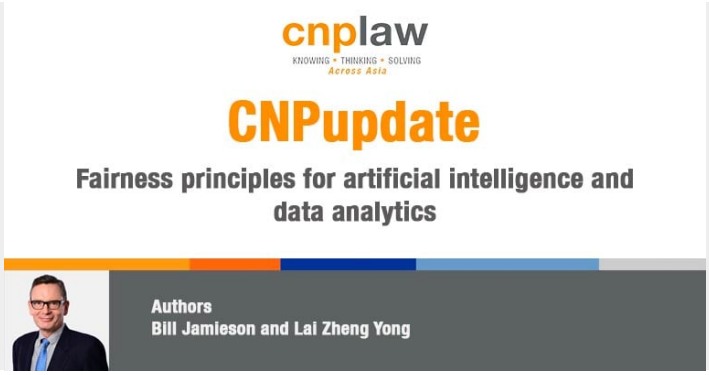
FAIRNESS PRINCIPLES FOR ARTIFICIAL INTELLIGENCE AND DATA ANALYTICS

Posted on March 26, 2021



Category: CNPupdates

Authors: Bill Jamieson and Lai Zheng Yong.

Following the conclusion of Phase One of the Veritas initiative, on 6 January 2021, the Veritas Consortium (the "**Consortium**") published two whitepapers detailing the Fairness, Ethics, Accountability and Transparency ("**FEAT**") fairness assessment methodology (the "**Methodology**") and its application in the two use cases. This article provides an update on Singapore's fairness framework for the adoption of artificial intelligence in finance.

Background

Artificial intelligence and data analytics ("**AIDA**") technology is increasingly employed for its ability to optimise decision-making processes. AIDA removes human-decision-making as a variable, and replaces it with a data-driven approach. The adoption of AIDA by Financial Services Institutions ("**FSI**") has been observed in areas involving internal-process automation and risk management, in the form of credit scoring and fraud detection.

In response to the plethora of risks associated with the adoption of AIDA in finance, regulators across the globe have developed their own guidelines to address what they identify as the major risk categories. In a research study of 36 guidelines on ethics and principles for artificial intelligence, the team at Berkman Klein Center found the theme of "fairness and non-discrimination" to be featured in all of the guidelines studied, the Monetary Authority of Singapore's ("**MAS**") FEAT principles being one of which.

Fairness of AIDA

The effectiveness of artificial intelligence is fundamentally predicated on the data it analyses. It follows that AIDA technology is limited by both latent biases within the data and the algorithmic perpetuation of the

same. To counter such risks, it is essential to identify the context of the data being utilised and have an understanding of how such data is relevant to the end-product.
Context is of particular significance as the abovementioned latent biases can impede the system's ability to process the data. Such latent biases can be observed from the following example:

> *"if one obtusely inputs white-collar professional labour data from the 1940s to the 1970s into an artificial intelligence system to predict what demographics of individuals would be the most successful applicants for white-collar professions, the suggestion would likely be white males of a certain age."*

Insofar as we accept that data may always contain some form of bias, extra precaution must be taken when handling the end-product, and appropriate adjustments made to mitigate such biases. Such adjustments are necessary to not only improve accuracy of the end-product, but to also incorporate a human assessment on the ethics, morality, and social acceptability of the end-product to the decision-making process.

Having highlighted concerns over the fairness of AIDA in decision-making, we assess the findings put forth by the Consortium in the following sections.

Fairness Principles
In Singapore, a set of principles has been published by MAS in relation to FSIs' use of AIDA. The Fairness Principles form the tenets of the Consortium's Methodology, and its application keeps the AIDA's decision-making process aligned with the overarching business and fairness objectives.
The four Fairness Principles are as follows:

> *F1 – Individuals or groups of individuals are not systematically disadvantaged through AIDA-driven decisions, unless these decisions can be justified*
> *F2 – Use of personal attributes as input factors for AIDA-driven decisions is justified*
> *F3 – Data and models used for AIDA-driven decisions are regularly reviewed and validated for accuracy and relevance, and to minimise unintentional bias*
> *F4 – AIDA-driven decisions are regularly reviewed so that models behave as designed and intended*

The Methodology
The Methodology consists of five steps:

> *(A) describe system objectives and context;*
>
> *(B) examine data and models for unintended bias;*
>
> *(C) measure disadvantage;*
>
> *(D) justify the use of personal attribute; and*
>
> *(E) examine system monitoring and review.*

Steps A, B, and C invite the assessor to establish both the business and fairness objectives of the system, which sets the benchmark against which the system's fairness and potential tradeoffs are measured against. In HSBC's simulated case study on the marketing of unsecured loans, consideration was given to the potential harms and benefits of having marketing intervention to the AIDA-selected individuals. Historically, foreign nationals have a lower rate of loan application approval. It was noted in the study that there is a potential harm of further disadvantaging foreign nationals where such historical data is utilised. By identifying latent bias at an early stage, FSIs are able to input mitigating mechanisms such as lifting the threshold for foreign nationals to mitigate the bias present within the data.

The concept of fairness must not be regarded as being blind to personal attributes. A gender- or racially-blind algorithm can widen any pre-existing disparity, and intervention may be necessary to promote fairness. It was observed in HSBC's study that a higher loan application rejection rate for foreign nationals would materialise if the nationality of the applicant was not taken into account by the system. Such inclusion of personal attribute was justifiable to ensure that the system meets its intended objectives set out in step A, and satisfies the Fairness Principles F1 and F2.

Lastly, the Methodology calls for an ongoing monitoring of the system, in accordance to Fairness Principles F3 and F4. HSBC hypothesised that such monitoring can be implemented by conducting an analysis before a campaign was initiated, to prevent a significant shift in the parameter of the system; monitoring the system's output during the campaign; and having the senior management team review the end-result of the campaign to ensure that the system meets the established objectives. In order to keep humans in the loop with the operations of the AIDA technology, it has been suggested that such accountability framework should be built upon existing infrastructure. In Singapore, this may be in the form of extending the scope and responsibilities of senior managers in FSIs under the MAS-issued Proposed Guidelines on Individual Accountability and Conduct (IAC Proposed Guidelines), to incorporate responsibility over the day-to-day operations of the AIDA technology.

Remarks

We note that the Methodology is principles-based and does not prescribe mandatory responsibilities or regulatory obligations that FSIs must comply with. It remains to be seen how the Fairness Principles will fare beyond simulated studies. Moving forward, Phase Two of the Veritas Initiative will focus on the development of the ethics, accountability and transparency assessment methodology.