Date: November 2020 **Segment: Research**

AISWITCH AI PRACTICE LEADERS: TRENDSETTER SERVICE PROVIDERS IN FAIR & RESPONSIBLE AI

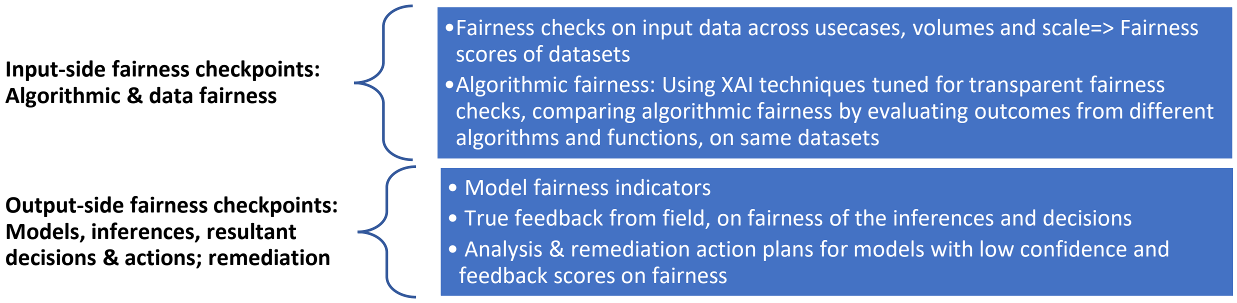
Who should read this: Enterprise AI CoE leaders, CDO, CIO, CEO (for strategic AI initiatives), AI Business User Leaders, AI Solution Architects, AI Solutions & Service Providers

Why must enterprise AI leaders ensure fairness in all AI usecases?

Discussing globally evolving ethical standards must also involve exploring emerging artificial intelligence technologies being developed and deployed by global organizations. Regulations need to consider what exactly needs to be regulated, hence any discussion around fair, ethical and legal standards of AI is incomplete without discussing globally evolving use-cases of AI.

This rapidly emerging use of AI by global organizations must include fairness checks as de facto practices, since the scale of these technologies ensures that they have the greatest impact. AI fairness and trust assurance initiatives and standards are emerging in the horizon, e.g. RAISE 2020 (Responsible AI for Social Empowerment), ISO/IEC TR 24028:2020 – to evaluate and ensure trustworthiness of the decisions and actions taken or handled by enterprise AI solutions. Gartner Hype Cycle on AI 2021, has mentioned AI governance to be a fast-emerging albeit nascent practice.

Trendsetting end-user companies, especially in the BFSI space, e.g. Capital One, have been working on explainable and fair AI practices for quite a few years now.



Input-side fairness checkpoints: Algorithmic & data fairness

- Fairness checks on input data across usecases, volumes and scale=> Fairness scores of datasets
- Algorithmic fairness: Using XAI techniques tuned for transparent fairness checks, comparing algorithmic fairness by evaluating outcomes from different algorithms and functions, on same datasets

Output-side fairness checkpoints: Models, inferences, resultant decisions & actions; remediation

- Model fairness indicators
- True feedback from field, on fairness of the inferences and decisions
- Analysis & remediation action plans for models with low confidence and feedback scores on fairness

In this Fair AI-focused landscape where both regulators and leading end-user client companies are focusing on including fairness in algorithms & models, a few technology and service providers have also started walking the talk along with their client teams. Such exemplary service providers with leading initiatives in the ethical and fair AI world include:

- IBM- AIFair30
- Google TensorFlow Fairness Indicators
- Microsoft Transparent and Interpretable AI
- Accenture Responsible AI
- Ernst and Young Trusted AI

- Deloitte Trustworthy AI Framework
- Wipro Holmes Ethica

IBM AIFair 360

IBM has made its AIFair 360 APIs openly accessible and easy-to-consume, for AI solution architects and builders, to democratize the practice of fairness in AI right from the start, of any AI application’s lifecycle in an enterprise:

- The techniques include algorithmic fairness check such as dynamic un-sampling of training data and adversarial debiasing - by reducing weights of sensitive attributes by design, i.e. reducing the adverse effects of discriminatory parameters like race or country of origin, that may affect fairness of output model decisions but should not be dropped completely as there may be high correlation or influences of these variables on the other parameters.
- The APIs also help AI builders to leverage techniques like prevention of disparity amplifications, semi-supervised variational auto-encoders and VAE-GANs etc.

Google TensorFlow Fairness Indicators- TensorFlow Data Validation (TFDV), TensorFlow Model Analysis (TFMA)

TensorFlow, an open source library for machine learning and numerical computation designed to enhance ease of access to training models, acquiring data and refined future predictions.

- It integrates responsible AI practices within machine learning workflows, through a shared combination of tools and resources, to provide customized model building for customers.
- TensorFlow’s responsible AI practices revolve around a human-centric approach built on accountability, fairness and transparency for all.
- Tensorflow Fairness Indicators include fairness metrics for binary as well as multiclass classifiers. They seem to work well for checking fairness in large-scale datasets and models, across any size of use case.
- These indicators enable AI solutioning teams to evaluate the distribution of datasets for fairness, and to evaluate model performance- sliced across defined groups of users. These indicators build confidence about the model results at multiple thresholds and dive deep into individual slices showing low confidence of fairness, to explore root causes and opportunities for improvement. [https://github.com/tensorflow/fairness-indicators]

Microsoft’s Transparent and Interpretable ML

Microsoft’s transparent and interpretable machine learning capitalizes on the idea that there is generally a compromise between accuracy and interpretability of ML.

- It uses general additive models to deliver results that are as accurate as complex random forecasts, but as interpretable as linear regression models.
- These models are applied to diverse fields from healthcare data, where diabetes, pneumonia and hospital readmission risks are calculated, to recidivism scores and credit rating.

Accenture’s Responsible AI

Accenture has been working on Responsible AI for a long time now and therefore has a mature practice of Responsible and Fair AI embedded in its critical AI solutions and capabilities. Accenture’s responsible AI practices include:

- security against biased data and algorithms to enhance justifiability of automated decisions.
- focusses on transparency and accountability to deliver practices consistent with organizational aims, user expectations and social norms.
- Accenture’s Applied Intelligence developed a fairness check toolkit that aims to ensure equal opportunity and unbiased classifications, reducing the influence of discriminatory parameters as much as possible, while the trade-off between model accuracy and algorithmic fairness being left in the hands of highly capable and informed solution/ model builders and users.

Ernst and Young’s Trusted AI platform

E&Y has created a robust trust-based AI framework, building on the idea that trust in AI systems is an end in itself, and not a secondary consideration.

- It uses analytical models to compute risk factors and provide guidance to AI design teams to monitor, evaluate and quantify AI risks and provide insights to organisations.
- Its schematic and assessment tools monitors risk profiles and delivers a score of the risk involved, to aid informed decision making.

Deloitte’s Trustworthy AI framework

Deloitte’s recently unveiled trustworthy AI framework aims to provide guidance to organisations on responsible, ethical and accountable use of artificial intelligence.

- Recognizing that dealing with emerging AI security issues is tough, Deloitte seeks to provide a common platform for these security concerns to be addressed.
- It incorporates six key checks and balances that companies must abide by. These involve: fairness and impartial use, transparency and explainability; responsibility and accountability; cybersecurity checks to prevent physical and digital harms, data privacy and reliability monitor.

Wipro ETHICA

Wipro’s framework for ethical use of AI is called ETHICA, which stands for Explainability, Transparency, Human-first, Interpretability, Common sense, and Auditability.

- It aims to reduce biases prevalent in current usage of AI.
- It promotes the usage of explainability and transparency in all critical AI solutions, using XAI algorithms like deconvolution to LRP and LIME, to explain the classification features and parameters be it in image or text classification usecases.
- Wipro Holmes ETHICA prioritizes a customer-first approach, integrating it into an overall organizational mission of trust and transparency

For further information on techniques and systems: admin@aiswitch.org