Tech Mahindra

AI Ethics and ML Fairness

WHITEPAPER

Connected World.
Connected Experiences.

AI has been bringing sweeping changes to how we interact, do our work, understand our world and pretty much at everything we do. At societal levels, AI can bring in sustainable & equitable economic growth and tackle global challenges that have proven difficult to tackle at scale till recently. At the same time AI has the potential to become the run-way technology that increases the societal imbalances, impact human rights of individuals & minorities negatively and erode finely balanced institutions we have built over our history.

In the Enterprise context, AI systems have the potential to improve productivity and open up new avenues of revenue by building insights from data that define the business and customer interaction patterns. On the flip side, perceived unfair treatment of customers or groups by the AI model can lead to trust issues, brand value erosion, talent retention & acquisition, missed business opportunities—all directly impacting the top line.

This is where AI Ethics & Fairness become important.

Ethics, Bias and Fairness

AI Ethics is a branch of Digital Ethics that includes Fairness. In addition to Fairness, AI Ethics is concerned with wide variety of subjects like privacy & surveillance, behavior manipulation, accountability of autonomous systems, robotic rights etc. In this article, we shall mainly focus on ML fairness.

To understand what is bias and fairness in ML model sense, let us look at a scenario. You are creating a model to create transcriptions from speech recordings. It works well for a set of users (Caucasian male), but performs poorly on a different set (female African American). In this simple scenario, it can be said that there is no implicit unfairness in the model, it is just a failure to produce consistent results—most likely due to bias introduced by the data on which the model was built. Next, let us consider a slightly extended context where the output created by the model is used to select right candidates for customer help desk positions. In this changed scenario, the ML model is said to discriminate (or have a fairness issue) against female African American candidates. In summary, bias is related to the skewness of the input data used in the modeling process. Fairness is a measure of if and how a ML model discriminates against group or individual resulting in material impact during the inference phase. Normally the discrimination can be correlated to certain attributes in the data (race and gender in our above scenario)—called sensitive attributes in some ML literature.

Let us look at ways to measure bias & ML fairness, how the bias gets introduced into the system and ways to increase ML fairness of the models.

Different types of bias enter into the data science lifecycle at various phases as depicted in the figure below.

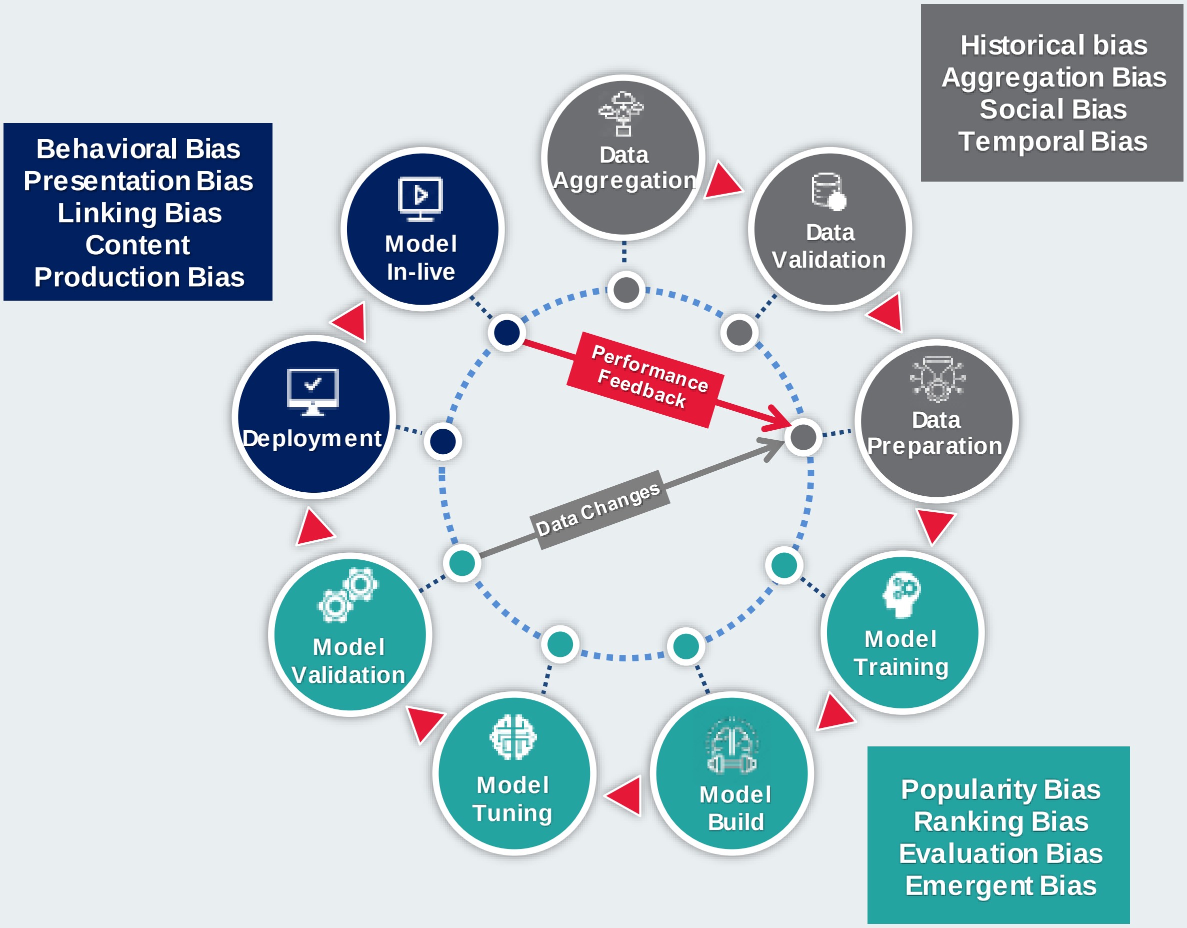

Fig 1: Bias types and introduction in different phases. Adapted from: "A Survey on Bias and Fairness in Machine Learning" - Mehrabi et al.

As we can see, behavioral import from human-data touch points like presentation, aggregation and interpretation during data pipeline, model build and live-model-performance-feedback phases contribute to the bias introduction. Historical bias is a pre-existing skewness in the data that reflects prevailing societal prejudices of the sampled population against a particular group. A user preference prediction model that was built on data reflecting user preferences five years back may display behavior that is out-of-tune with today's population—this is emergent bias. The way data gets arranged & presented to the user can introduce ranking bias: search results at the top tend to be the ones clicked more which increases their popularity score causing their ranking to go up pushing them up in the results display arrangement—kind of a self-reinforcing loop.

Discriminations due to fairness issues can be direct (correlates directly to the sensitive attributes like race, gender) or indirect (correlates to the sensitive attribute through other attributes like zip code as a proxy to race identity).

Different definitions of ML fairness exist based on the approach to correction. Please note that most of these concepts have been around for decades within the legal framework to handle discrimination in the real world.

Few of these definitions are at group level—like Demographic parity where acceptance/rejection rate of disadvantaged group should be within a limit as compared to the advantaged group. Then there is an individual fairness definition which looks to apply fine-grained corrections to ensure individuals that are equal on given metrics are treated the same. Another one is unawareness where the sensitive attributes are omitted completely during the prediction to ensure no discrimination.

Couple of caveats to be kept in mind in applying these corrections. First, some of these definitions are mutually exclusive and striving to achieve fairness fitting all definitions is not possible and so one needs to be clear on which measure is appropriate and set up the systems to measure through the lifecycle continuously. Secondly, there is a trade-off to be made between model accuracy and model fairness.

How to improve fairness in ML models?

Different techniques can be employed at different phases of the Data science lifecycle to increase ML model fairness.

- During the data modeling phase, take care to handle the protected classes under the regulatory & legal framework within which the ML model is to perform. For example, things like race, gender, age, genetic information etc. are ‘protected’ by prevailing legal framework applicable to certain areas in the US.
- Techniques like Reweighting, Disparate Impact Removal, Learning Fair Representation etc. can help identify and weed out some of the bias in data.
- At model build phase, bias removal is usually handled as part of the build process itself. One of the most straight forward approaches would be to have a fairness penalty integrated into the loss function. Another is to use a Generative Adversarial Network (GAN) approach to reduce the bias
- Choice of right data sampling strategy based on expected discrimination issues in the ML model and proper fairness metrics should be done at the testing phase to weed out fairness related issues before ML model is taken live.
- In the customer acceptance testing phase, customers or proxies should be given the tools to query the ML model using AI Explainability frameworks like IBM AIX360. This will give them a feel of inner working of the model and help identify any potential fairness issues.
- A live model has to be continuously assessed for fairness as part of the overall performance monitoring. Cohort-level sampling for individual predictions should be done, assessed for ML model fairness and the case would have to be sent for manual intervention on what looks like unfair/extreme cases. Manual corrections will have to be incorporated into the ML model in subsequent training cycles. Care must be taken to ensure new bias does not seep into the model during model retraining cycle.

Preparing for Fair and Ethical AI

Ethical AI and ML fairness are very new and still evolving areas. Given the importance of Ethics in the AI world, now-a-days companies are creating an ethics officer role which is tasked with ensuring compliance at all levels - people, process and technology.

- At people level, a higher level of awareness of the AI ethical practices, knowledge on possibility of injecting bias through data science lifecycle and potential business impact of bias & fairness are necessary. Establishing courses to train data engineers, data scientists, ML modelers and operations personnel is essential
- Process level practices include establishing an end-to-end Data Science Lifecycle definition and establishing proper governance. This should encompass information architecture that brings in the necessary bias related checks on the input data and AI fairness algorithmic design principles. Data and delivery process should include clear guidelines for diversity requirements in data, choosing fairness measures, coverage of model user profiles, identification of advantaged & disadvantaged groups and data sampling to check bias & fairness. Establishment of end to end traceability and accountability is an essential foundation to audit, trace, identify and fix issues as they occur. Operational practices would include process to flag issues related to bias and steps for manual intervention.
- Technology practices would include choice of right set of tools, technology governance principles and technology best practices based guidelines at every level.

Tech Mahindra and Ethical AI

Tech Mahindra has an evolved AI Center of Excellence framework that weaves industry best practices principles of AI Ethics through the three legs of service delivery – people, process and technology. As part of the "Responsible AI" drive, Tech Mahindra has instituted the following actions.

- All AI engineers are required to take an AI Engineer Oath like Canadian Iron ring oath for Engineers.
- An Ethics Lead is constituted as part of our AI Business Unit.
- Mandated involvement of empathy experts in the relevant industry verticals that will play a key part in all the AI/ML projects.
- All AI projects delivered by Tech Mahindra will be checked for undesirable behaviors of intelligent agents.

GAiA is Tech Mahindra's industry leading end-to-end AI/ML lifecycle management platform. GAiA is built upon Acumos™, a platform and open source framework that makes it easy to build, share, and deploy AI apps. GAiA contains an AI/ML fairness framework that enables Tech Mahindra's Enterprise customers to quickly adopt the necessary best practices and rollout well-balanced ML models.

Author

Mukund Kannan
AI Competency, Tech Mahindra

Tech Mahindra

www.youtube.com/user/techmahindra09
www.facebook.com/techmahindra
www.twitter.com/tech_mahindra
www.linkedin.com/company/tech-mahindra
www.techmahindra.com
Data&AnalyticsCommunications@techmahindra.com

Disclaimer. Brand names, logos and trademarks used herein remain the property of their respective owners.