Date: 30TH September 2020 Segment: Research

AISWITCH AI PRACTICE COOKBOOK: HOW TO DEBIAS ENTERPRISE AI SOLUTIONS- 15 TYPES OF DATA BIASES

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 should AI leaders ensure bias-free models and solutions?

A global 2020 survey shows that 48% tech leaders' vision in 3-5 years is to use emerging technologies like AI and automation, for competitive differentiation for their company. 41% want tech to drive business innovation with these technologies. These indicators clearly show that, in coming years, AI and automation are going to become key strategic initiatives across organizations, sectors and geographies.

While these technologies are extremely promising in terms of creating new business models, revenue opportunities, cost efficiencies, services and experiences, there are some very real challenges on the ground:

- In absence of a uniform governance framework to cover AI adoption in a holistic way, the strategies and technology implementations remain disconnected.
- The core factors of trust and bias handling are still not practiced by default, in most of the AI use-case implementations.
- AI-automation management & governance systems like AISWITCH^{IP} are starting to address these issues but their widespread adoption is going to take as much time as typical AI usecases have taken, to scale.

The explicit and intrinsic impact of data and algorithmic biases creeping into the decisions and actions by AI-automation systems, have long-term strategic risks and consequences.

Two possibilities present themselves:

- The first, is to use AI in identifying and reducing the effect of human prejudices.
- The second is to improve AI systems themselves, in terms of how they use data and how they are developed, deployed, and used, in order to prevent a fostering and continuation of human biases and preventing AI systems from developing biases themselves.

AI systems can reduce unpredictability in choices as they use only that data which helps to increase predictive accuracy. This is particularly useful in cases like the criminal justice system where judgements are powered by stereotypes towards Black and minority groups as those in power are overwhelmingly from dominant groups. Yet if that were the only side of the story, this would not be an ethical issue for AI at all.

Storyboard: The most (in)famous case of a biased AI application- COMPAS

Julia Angwin and others at ProPublica have shown how COMPAS, used to predict recidivism, incorrectly labelled African-Americans as “high-risk”, at nearly twice the rate it mis-labelled white defendants, while Latanya Sweeney’s research on racial differences in online ad-targeting showed that searches for African-American names are more likely to display the word “arrest” In another example, Google image search results for “CEO” were skewed towards men, which researchers said is an example of AI furthering gender biases in worldview, since people re likely to believe what they see.

What are the types of data biases?

In most cases, it is underlying data, rather than the algorithm that reinforces the stereotype. Data collected from sources like news articles, may, through natural language processing techniques, pick up gender- stereotype. The fact remains that AI can embed and replicate human biases.

In the following table, we explain 15 types of data-related biases, with explanations on why they occur.

Common types of data biases	Description
Measurement bias	How we choose and measure a specific feature, e.g. biases on minority groups (COMPAS project to predict recidivism/ re-crime risks)
Representation bias	Induced by choosing samples from a predefined population, e.g. ImageNet not capturing geographical diversity
Population bias	User characteristics in selected population being different from the original target population
Historical bias	Pre-existing social bias- can creep in despite careful feature selection & right sampling techniques.
Aggregation bias	Overgeneralization based on false conclusions and assumptions drawn on a subgroup, based on different subgroups, e.g. stereotyping in clinical cases
Evaluation bias	Due to inappropriate and disproportionate benchmarks and thresholds, e.g. face recognition systems biased on skin color and gender
Simpson’s paradox	Bias in heterogeneous data composed of subgroups with different behavioral trends
Sampling bias	Non-random, selective sampling-induced bias

Behavioral bias	Different user behaviors across platforms or datasets, may often lead to communication errors in chats.
Longitudinal study bias	If observational data are grouped as longitudinal, even if they are cross-section data.
Temporal, popularity and emergent bias	Time-variant bias based on popular topics, e.g. hashtags, most trending topics, cultural values, social bots, fake reviews and ratings, often generated by machines
Content bias	Semantic or syntactic biases in contents generated by users, e.g. usage of gender-specific pronouns and subsequent verb-forms, in certain languages.
Omitted variable bias	Important dimensions or variables are left out of the model
Causality bias	Due to confusion between correlation and causality
Algorithmic bias	Bias absent in data but added by the algorithms

Table 1: Most common types of data biases in AI solutions

AI leaders, architects and data engineers may consider the following aspects while handling data bias:

- It's not necessary to check all of these biases in all types of AI applications scenarios, while designing the training, test and validation data-sets to train and test the models, or while preparing the data pipelines for the AI-automation systems.
- Some of the biases are statistical and mathematical hence can be avoided by conscious efforts by analysts and model owners, with rigorous testing and comprehensive yet holistic data quality assessment factoring in debiasing and fairness-ensuring techniques.
- Observational biases and biases related to causality and algorithm selections can also be factored into the KRAs and KPIs of the data scientist/ engineering teams.
- Biases like content bias, linking bias etc. are relatively harder to detect and handle, given they may not be explicitly monitorable and may hide in piles of unstructured data/ corpus.
- Depending on the potential risks and maximum likelihood of certain types of biases creeping into the target models or solutions, a contextually prioritized set of biases must be checked for every AI solution.

How to make AI solutions mostly bias-free?

Training AI models on relatively bias-free data is a question that makes us reflect upon:

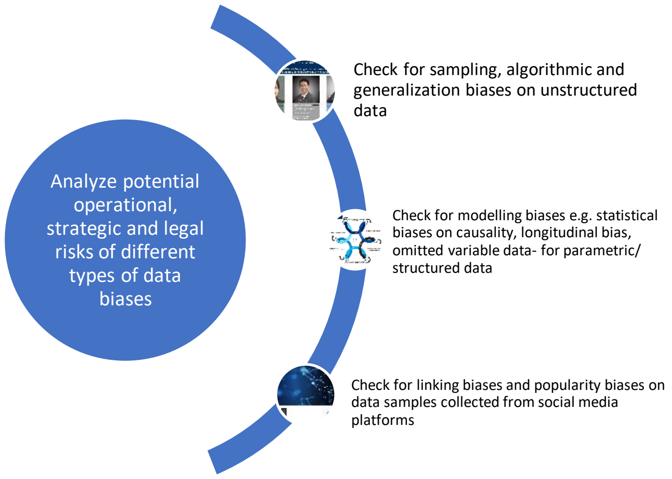
- Should AI show the world as it is, which is, manifestly unequal?

- Should it depict an ideal state of justice and equality that we are far from achieving?

One approach could be to do intentional oversampling, to add more data points to improve performance for minority groups. It has also been suggested that we use explainability techniques to explain why a judgement was arrived at and if the decision reflects biases.

Ultimately, given that AI depends on statistical frameworks and data sheets, it cannot analyze the socio-cultural and emotional origins of that data, ensuring that for unprejudiced AI to exist, humans and machines have to work alongside one another. The Aequitas (means ‘Justice’ in Latin) solution for AI bias-checking, and MIT-IBM Watson AI Lab’s efforts on recent advances in AI and computational cognitive modeling, have taken debiasing techniques forward. These initiatives include contractual approaches to ethics, descriptions of principles that people use in decision-making, to build machines that apply certain human values and principles in decision-making. IBM’s AI Fairness 360 toolkit is an opensource code-set that’s available for developers to check for common data biases and remediate/ alleviate them.

Here is a simple, iterative approach to make AI solutions relatively free of common and frequently occurring data biases:



Analyze potential operational, strategic and legal risks of different types of data biases

Check for sampling, algorithmic and generalization biases on unstructured data

Check for modelling biases e.g. statistical biases on causality, longitudinal bias, omitted variable data- for parametric/ structured data

Check for linking biases and popularity biases on data samples collected from social media platforms

As AI systems point out inconsistencies in human decision-making, learning where these deficiencies lie can lead to us becoming more egalitarian. Hence, it is imperative for all AI leaders and teams- in business, functions or technology domains, to be aware and responsible about using bias-mitigated data to design, build and train models for implementation enterprise AI usecases and solutions.

Action items next Monday Morning

Key actions to ensure bias-mitigated AI	Key actors
Make it mandatory for all enterprise AI applications to ensure usage of bias-mitigated datasets to train the models.	AI CoE leaders, organizational AI governance councils/ steering committee

Create awareness among enterprise data science teams and AI practitioners communities about the risks associated with usage of biased data	AI CoE leaders, solution architects, business leaders
Prioritize the types of data bias identification tests that must be conducted, especially for strategic AI initiatives/ usecases	AI CoE leaders, business leaders
Create and maintain a system of records and feedback mechanism to learn from past mistakes arising out of usage of biased data and resultant risks of various AI solutions deployed and running in the enterprise. Ensure that all relevant business and AI tech teams learn from these past mistakes and do not repeat them.	AI CoE leaders, solution architects, business leaders, AI solution owners

All Human perception is clouded by innate biases of the target, the perceiver and the context to be perceived. While completely objective judgement is a paradox, areas like finance, HR, criminal justice and healthcare do require an unbiased approach. The question then arises, if humans are fallible, what should one expect from Artificial Intelligence.

Then, almost 100% of current AI usecases in-prod are data-heavy, e.g. in BFSI- the E-KYC, AML, fraud detection, false claims in insurance, loan approval, credit scoring, algorithmic trading, vendor evaluations and trade negotiations etc. Now, if the petabytes of training datasets- structured, semi or unstructured- labelled or unlabeled images or text corpuses- have different types of explicit and implicit human biases captured in them, the output models will obviously take biased decisions and actions. These will have legal, regulatory repercussions as well as long-term damages on brand image and CSR. Therefore, AI practitioners must focus on data hygiene as a top priority.

Consequently, in the process of teaching machines to be unbiased, we may just become unbiased ourselves, or at the very least, take a step towards a more just, equitable, truly bias-free data-driven society.

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