datacamp DataRobot

Data Literacy for Responsible AI

As Artificial intelligence (AI) matures as a technology, 70% of customers expect organizations to use it responsibly, ethically, and transparently. What are the risks with deploying AI systems and how can organizations better mitigate them? *

*Capgemini, AI and the ethical conundrum

Algorithmic Bias

In general, algorithmic bias refers to the observation that an algorithm is treating groups in your data differently. Those groups, when we are concerned about societal bias, are identified by what are referred to as protected or sensitive characteristics—like

Race · Gender · Age · Pregnancy · Veteran Status

This is especially risky in highly sensitive industries like Healthcare or Finance.

Examples of Algorithmic Bias in Action

1. Recidivism rate prediction algorithm discriminating based on skin color
2. Language models producing harmful stereotypes associated with certain groups
3. Hiring algorithms disproportionately favoring men
4. Healthcare algorithm exhibiting racial bias

[1] "How We Analyzed the Compas Recidivism Algorithm, ProPublica", Accessed April 2, 2021.
[2] Abid, Farooqi, and Zhou. "Per sistent Anti-Muslim Bias in Large Language Models". arXiv Preprint, 2021.
[3] "Amazon scraps secret AI recruiting tool that showed bias ....", Accessed April 6, 2020"
[4] "Dissecting racial bias in an algorithm used to manage the ...." Accessed April 6, 2020.

The Two Main Categories of Algorithmic Bias

While research and academia propose that there are over 70 metrics that can measure bias, they can be grouped into two categories.

Fairness by Representation

Fairness by representation focuses directly on what outcomes the model predicts to evaluate if there are different likelihoods of receiving the more favorable outcome by each group.

Example: Hiring algorithms disproportionately favoring men.

Fairness by Error

In fairness by error, the quality of model performance and accuracy is compared across groups; are some groups disproportionately affected by certain kinds of error?

Example: Healthcare algorithms disproportionately denying access to treatment for one group.

Where Bias Comes from

Below are some ways bias can manifest from data:

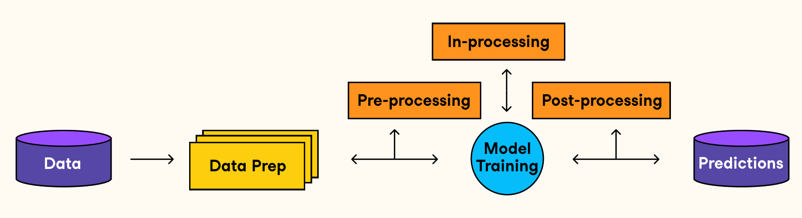
1. **Skewed dataset:** Lack of representation in the data can affect an AI's ability to learn from diverse sets of examples, which can result in biased model performance.
2. **Tainted examples:** Unreliable labels or historical bias in the data have a direct impact on AI's discriminatory behavior.
3. **Limited features:** Feature collection for certain groups may not be informative or reliable, which can occur under bad data collection practices.
4. **Sample size:** Small datasets limit the ability of the AI's effective learning process and can result in bias.
5. **Proxy features:** Features can indirectly leak information about the protected attributes, even in cases when that protected feature has been removed. Zipcodes, sports activities, and university attended can be used by the model to indirectly infer race or gender.

Limiting AI Risk

Mitigation Techniques

Mitigation strategies can occur at different points of the machine learning pipeline: pre-processing, in-processing, and post-processing.

For more details, make sure to download the white paper



AI Governance Framework

Organizations can draw on risk management to govern the risk posed by deployed AI systems. DataRobot's AI Risk Framework classifies AI systems by risk size, and prescribes a human role in governance based on the risk size.

	Type I - Low	Type II - Medium	Type III - High
Risk Size	Loss <$100K	$100k < Loss < $1M, or Injury to Human Livelihood	Loss > $1M, or Death
Example	Probability of an ad-click	Probability of Mortgage Default	Medical Imaging Machine Vision
Human Role in Governance	Construction, Maintenance & Monitoring	Type I & Risk Assessment & Mitigation	Type I, II & Final Augmented Decision Outcome "Human over the loop"

DataRobot's AI Risk Framework. Thresholds need to be adjusted according to organizational definitions.

Most importantly, AI Governance requires multi-stakeholder engagement and the sign-off of personas stemming from various backgrounds and levels of technical proficiency.

Data Literacy for Responsible AI

Data literacy can be defined as the ability to critically understand data science and AI applications, distinguish between various data roles, communicate insights from data, and make data-driven decisions. More importantly, it promotes a two-way conversation between subject matter experts and AI experts that allows non-technical stakeholders to inject their domain expertise into the problem setup, scoping, and implementation of AI projects.

What is data literacy in the context of responsible AI?

1. Understanding the data science and machine learning workflow
2. Understanding the distinction between various data roles
3. Understanding how data is collected and flows through an organization
4. Grasping the distinction between various types of AI systems and their explainability
5. Learning the different evaluation metrics for machine learning models

datacamp DataRobot

Learn More

Do you want to know more about how data literacy fuels responsible AI? Get the white paper!

Get the White Paper →