

[Back](#)[Home](#)[AI Guides](#)

ML Metrics

# Fairness Metrics In Machine Learning



Yaniv Zohar

Jul 25, 2023

14 mins read

During their extensive training and learning process, machine Learning engineers will inevitably encounter ML bias and fairness questions, which casts a shadow over the development and deployment process and their models. In a machine learning model, bias manifests as an unfair preference or prejudice towards a specific class, distorting learning and ultimately compromising the model's performance.

You can rely on the power of fairness metrics to eliminate bias in your models. This post will explore how these indicators can be used, showing how they help ensure consistent performance across all classes, regardless of any training data imbalances. Thankfully, a solution is within reach.

Let us first delve into the significance of finding the right balance while exploring solutions and the

Share article



On this page

[Understanding Fairness Bias in AI Models](#)[Examples of Fairness Bias](#)[Common Fairness Metrics in Machine Learning](#)[Using Fairness Metric Tools to Reduce Bias in Machine Learning](#)[General Practices to Keep Your Learning Fair](#)[Fairness-Aware Model Evaluation](#)[Conclusion](#)

consequent outcomes for Fairness Bias in AI Models.

## Understanding Fairness Bias in AI Models

Machine learning models must not exhibit unfair or discriminatory behavior, exemplified through biased predictions or discriminatory decision-making. This occurs when models unfairly favor or discriminate against certain groups or classes, resulting in unequal treatment and disparate outcomes.

Biases in fairness can stem from multiple origins, such as discriminatory training data, biased assumptions in model design, or the inclusion of biased features. The model can potentially replicate existing biases if the training data is skewed or reflects historical discrimination. Likewise, the model may display little tendencies if specific features are given more significance during training.

The consequences of fairness bias can be detrimental, as it can perpetuate and amplify existing societal inequalities or discrimination, which various compliances like HIPAA may not even detect. For example, in the context of hiring decisions, a biased model might favor candidates from a particular gender, race, or socioeconomic background, leading to unfair outcomes and exclusion of qualified individuals.

Addressing fairness bias is crucial to ensure ethical and equitable machine learning applications. Various approaches, such as fairness metrics, pre-processing techniques, algorithmic modifications, and post-processing interventions, can be employed to mitigate and monitor bias in machine learning models. These techniques aim to promote fairness, transparency, and accountability in the decision-making processes of AI systems.

## Examples of Fairness Bias

Artificial Intelligence (AI) models, touted for their ability to automate decision-making processes, have become integral in various domains. However, the unchecked integration of AI can lead to unintended

consequences, such as fairness bias. Fairness bias occurs when AI models exhibit discriminatory behavior or amplify existing societal inequalities.

Let us see some examples in real life where your Bias model can ultimately dictate or even destroy human lives if not rectified and trained properly:

1. **Sentencing Disparities:** AI models are increasingly used in the criminal justice system to aid in sentencing decisions. However, studies have shown that AI-powered algorithms can perpetuate racial biases. For instance, an AI model may unknowingly be trained on historical data that disproportionately criminalize certain racial groups. Consequently, the model may recommend harsher sentences for individuals from these groups, exacerbating existing disparities. This fairness bias can perpetuate systemic injustice, leading to detrimental consequences for affected communities.
2. **Biased Hiring Practices:** AI-powered tools often screen and shortlist job applicants in recruitment processes. However, if the training data used to develop these models reflects biased hiring patterns, the AI system can inadvertently discriminate against specific demographics. For example, if historically male-dominated industries predominantly feature male employees, the AI model may learn to favor male candidates, perpetuating gender biases. Such biased hiring practices hinder diversity and inclusivity, restricting equal opportunities for marginalized groups.
3. **Predatory Lending Algorithms:** In the financial sector, AI algorithms are employed to assess creditworthiness and determine loan approvals. However, if these models are built using partial

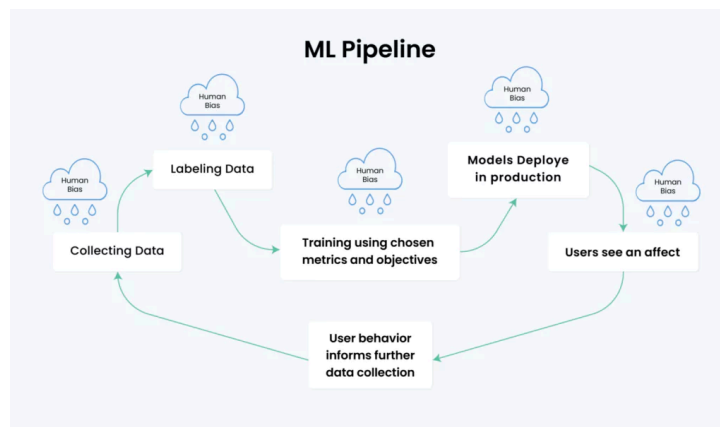
historical lending data, they can inadvertently discriminate against underprivileged communities. For instance, if the training data indicates that certain minority groups have been unfairly denied loans, the AI model may adopt this discriminatory behavior, perpetuating the cycle of financial exclusion. This fairness bias deepens socio-economic disparities and limits access to resources for marginalized communities.

4. **Healthcare Disparities:** AI models are increasingly utilized in healthcare settings to aid in diagnosing diseases and recommending treatments. These models can exhibit fairness bias if trained on partial healthcare data. If historically marginalized communities have faced inadequate healthcare access, an AI model trained on such data may perpetuate these disparities. This can lead to delayed or incorrect diagnosis and adverse health outcomes.

Now that we have a somewhat clear understanding of why dealing with Fairness Bias is essential for your model and your product and service's lifeline let us look at the various ways you can first detect and then deal with such biases.

## Common Fairness Metrics in Machine Learning

Data has proven itself to be the most valuable resource recently, from the billions of dollars earned by research institutes (Microsoft invested \$10 Billion in OpenAI before the launch of ChatGPT) to complete organizations coming down to their knees through unfair use of data. It should be poetic that an implosion of such a large-scale disparity in machine learning models through bias to be dealt with and rectified from the data itself.



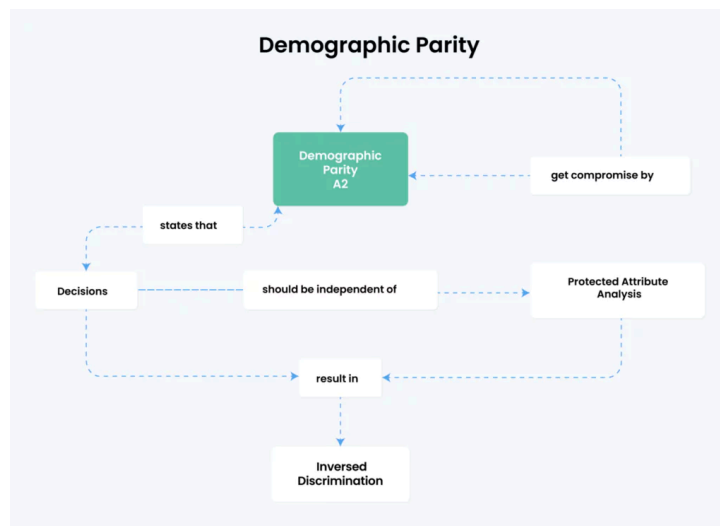
Metrics is a term thrown around a lot in machine learning, mainly in accuracy (F1 Score, Confusion Matrix, AUC, Loss Functions, etc.). Metrics help users understand the performance, efficiency, and eventual expectations from a machine learning model trained on intricate and confusing architectures, which otherwise may not be as intuitive for the average working professional and more so suited to PhD-level scholars.

The same can be observed and practiced when trying to find bias disparities in machine learning models, especially when even 1 percent of error due to a bias error can misclassify tens of thousands of potentially different classes. Let us see what some of the most used, discussed, and proven ways that you, as an organization or professional, identify the extent of bias in your model are.

*Note: Some metrics may contradict each other, but each organization must decide which metric meets its requirement.*

## Demographic Parity

Demographic parity is a fairness metric in AI that aims to ensure equal proportions of positive outcomes across different demographic groups. It assesses whether the model treats other groups fairly, regardless of their protected attributes. Identifying disparities and biases can be identified by comparing results. Achieving demographic parity helps mitigate discrimination and promote equal opportunities.



Simply put, protective and sensitive information like Gender, Race, and Religion shouldn't be the driving factor of your machine learning model; instead, fields relevant to the task, such as qualifications, education, etc.

The formula for Demographic Parity is:

$$P(A=0)=P(A=1)$$

This essentially means that the probability of having favorable outcomes across different groups should be the same.

## Equal Opportunity

Equal opportunity in AI promotes fairness and equal access to opportunities, regardless of demographics. It aims to eliminate biases and discrimination, ensuring decisions are based on qualifications and merits rather than protected attributes. AI systems should be unbiased and fair and address historical disadvantages. Steps include collecting square data and making appropriate decisions based on objective criteria.

## Proxy Attribute

One of the simplest ways to make sure unnecessary bias is not generated while training the data substitutes or *proxy* can be made for sensitive attributes, i.e., instead of sensitive characteristics like

Gender, Race, or Ethnicity, fields can be replaced by a value that is much less harsh or controversial, like their area code or their city.

## Statistical Parity Difference

Statistical Parity Difference (SPD), also referred to as disparate impact, serves as a standard fairness metric in assessing and quantifying the discrepancies in outcomes among various demographic groups within machine learning models. It enables the measurement of whether there exists an equal likelihood, across different groups, of experiencing positive effects such as loan approvals or job offers.



The Statistical Parity Difference is calculated by finding the discrepancy between the average probability of a positive outcome for the privileged group (typically the majority or non-protected group) and that of the disadvantaged group (often the minority or protected group).

The formula for Statistical Parity Difference is:

$$SPD = P(\text{outcome} \mid \text{privileged group}) - P(\text{outcome} \mid \text{disadvantaged group})$$

*Note: While this blog primarily focuses on key fairness metrics in machine learning, please note there are numerous other metrics, making the field vast and continuously evolving.*

## Using Fairness Metric Tools to Reduce Bias in Machine Learning

At the core implementing these algorithms can be a tough job, given the nature of complex advanced statistics that go into it; however, because of the prevalence of fairness in AI, with the ever evolving field, you may want to check your model for fairness on simple commands. Let us see some of the tools that you can use to make sure that your data and model are fair and are having the perfect balance of bias and variance.

### **Google's Model Card Toolkit**

Google has developed the Model Card Toolkit to simplify and automate the creation of Model Cards, which are informative documents that offer insights and transparency into the performance of machine learning models. By integrating this toolkit into your ML pipeline, you can easily share important model information, metrics, and metadata with researchers, developers, and other stakeholders. If your ML pipeline uses TensorFlow Extended (TFX) or ML Metadata, you can take advantage of the automated model card generation offered by the toolkit.

### **IBM's AI Fairness 360**

AI Fairness 360 (AIF360) is a Python toolkit developed by IBM specifically designed to promote algorithmic fairness. It facilitates the integration of fairness research algorithms into industrial settings and provides a common framework for fairness researchers to collaborate, evaluate, and share their algorithms. This open-source toolkit can be extended and applied throughout the AI application lifecycle to examine, report, and mitigate discrimination and bias in machine learning models.

### **Microsoft's Fairlearn**

Microsoft's Fairlearn is a freely available toolkit aimed at evaluating and enhancing the fairness of AI systems. The toolkit comprises two main elements: an interactive visualization dashboard and algorithms for mitigating unfairness. These components are specifically designed to help users comprehend the trade-offs that may exist between fairness and the performance of the model. By leveraging Fairlearn,



users can gain insights into fairness considerations and work towards improving the equity of their AI systems.

## General Practices to Keep Your Learning Fair

However, many tools or algorithms are introduced to the pipeline, and the efforts for sustainable and fair machine learning development, deployment, and monitoring should start at your own core of operations.

Maintaining fairness in a machine learning algorithm is a complex and ongoing process. While the specific approach may vary depending on the context and the specific fairness concerns, here are some general guidelines to help promote fairness:

- 1. Define fairness criteria:** Clearly define the fairness objectives and criteria that align with the ethical considerations and legal requirements relevant to your application. Consider both individual fairness (treating similar individuals similarly) and group fairness (avoiding disparate impacts on different demographic groups).
- 2. Collect diverse and representative data:** Ensure that the training data used to develop the algorithm is diverse and representative of the population it will be applied to. Biases and disparities in the training data can lead to biased model outcomes. Regularly review and update the training data to account for changes and maintain fairness.
- 3. Evaluate and mitigate biases:** Use fairness metrics to assess the algorithm's performance across different demographic groups. Identify and understand any biases or disparities that

exist. Employ techniques such as algorithmic adjustments, pre-processing, or post-processing to mitigate biases and promote fairness.

4. **Transparent and explainable models:** Strive to develop models that are transparent and explainable. This helps in understanding how the algorithm arrives at its decisions and facilitates the identification and mitigation of any unfair biases. Consider using interpretable models or techniques that provide insights into the decision-making process.
5. **Regular monitoring and updates:** Continuously monitor the algorithm's performance for fairness after deployment. Implement processes for feedback collection, regular model updates, and ongoing evaluation to address emerging fairness concerns and adapt to changes in the application context or user population.
6. **Engage diverse stakeholders:** Involve diverse stakeholders, including experts in ethics, fairness, and domain knowledge, as well as individuals from the affected demographic groups. Their input and feedback can provide valuable insights and help uncover biases or blind spots that may have been overlooked.
7. **Document and communicate fairness efforts:** Document the steps taken to promote fairness, including the data collection process, fairness evaluation results, and any algorithmic adjustments made. Transparently communicate the limitations, trade-offs, and potential biases associated with the algorithm to build trust and accountability with users and stakeholders.

It is important to note that fairness is a multidimensional concept, and trade-offs may exist between different fairness objectives or fairness and performance. Balancing these trade-offs requires careful consideration and a nuanced approach specific to the application domain and context.

By following these general guidelines and adopting a proactive and iterative approach, organizations can work towards developing and maintaining fair machine learning algorithms that promote equal treatment and minimize discriminatory biases.

## **Fairness-Aware Model Evaluation**

In evaluating machine learning models, fairness is crucial to ensure performance and equity. Relying solely on performance metrics such as accuracy or precision may neglect potential biases and unfair practices towards specific demographic groups. To address this limitation, fairness-aware evaluation incorporates fairness metrics alongside performance metrics, aiming to achieve accuracy and fairness simultaneously.

### **Importance of cross-validation in fairness evaluation**

Cross-validation plays a vital role in evaluating fairness by assessing how well a model's fairness performance can be generalized. It achieves this through the division of the dataset into multiple subsets and training the model on different combinations of these subsets. This process ensures a more robust evaluation, enhancing our understanding of the model's overall fairness.



Fig: K-Folds Cross Validation

Techniques for cross-validation encompass various approaches, including k-fold cross-validation, stratified cross-validation, and leave-one-out cross-validation. These diverse methods ensure that the evaluation of fairness is not confined to a particular subset of data but rather captures the model's overall fairness performance.

Cross-validation plays a crucial role in assessing fairness. It helps determine if the valid concerns extend consistently across different data subsets or if they are specific to specific training data. Additionally, it aids in pinpointing potential sources of bias and comprehending the model's performance variations across diverse contexts.

## Analyzing trade-offs between Fairness and Performance

In the realm of designing and optimizing machine learning models, it becomes crucial to grasp the trade-offs between fairness and performance. Here lies a delicate balance — enhancing fairness may lead to a decline in performance, and vice versa. Hence, striking an equilibrium between these two objectives assumes paramount importance.

When organizations analyze trade-offs, they evaluate how different algorithmic adjustments, feature selections, or model architectures impact both fairness and performance. By systematically exploring various configurations, they can identify strategies that optimize fairness while maintaining acceptable levels of performance.

Strategies for optimizing both fairness and performance encompass several approaches. These include utilizing fairness-aware algorithms, incorporating fairness constraints during model training, or employing pre-processing or post-processing techniques to mitigate biases.

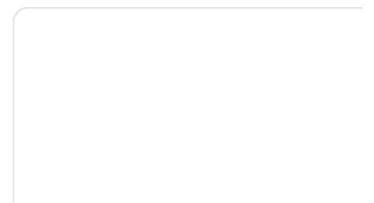
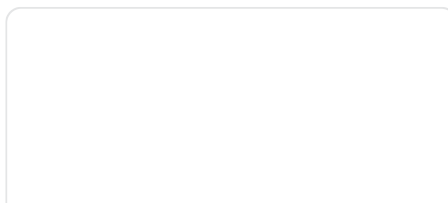
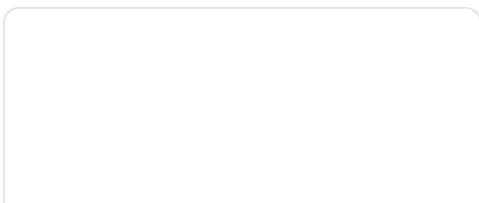
## Conclusion

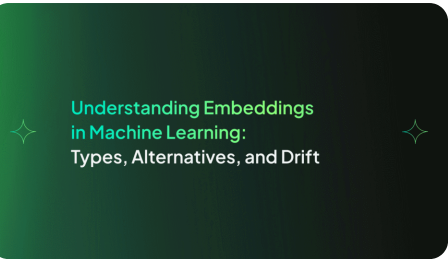
In conclusion, ensuring fairness in machine learning models goes beyond mere compliance with policies and regulations. It represents a critical ethical responsibility. The presence of bias in these models can perpetuate societal inequalities and result in unjust outcomes for specific demographic groups. However, organizations have the opportunity to mitigate this bias and enhance model fairness by harnessing proven algorithms and mathematical formulations. By doing so, they can ensure that all users and beneficiaries receive the best possible version of these models.

Fairness in model development is a continuous process that necessitates regular updates, transparency, and accountability. Organizational vigilance is key to monitoring, evaluating, and refining models based on new data and feedback. By doing so, they ensure the preservation of fairness while promptly addressing any emerging biases.

By placing a strong emphasis on fairness in machine learning models, organizations fulfill not only compliance requirements but also make valuable contributions towards creating a more inclusive and just society.

## Related articles





Understanding Embeddings  
in Machine Learning:  
Types, Alternatives, and Drift

ML Metrics

## Understanding Embeddings in Machine Learning: Types, Alternatives, and Drift

Introduction Machine learning algorithms, specifically in NLP, LLM, and computer vision models, often deal with...

11 mins  
read

[Read Now](#)



Recall: A Key Metric for Evaluating  
Model Performance

ML Metrics

## Recall: A Key Metric for Evaluating Model Performance

Measuring the performance of ML models is crucial, and the ML evaluation metric – Recall...

7 mins  
read

[Read Now](#)



Understanding Binary  
Cross-Entropy and Log Loss  
for Effective Model Monitoring

ML Metrics

## Understanding Binary Cross Entropy and Log Loss for Effective Model Monitoring

Introduction Accurately evaluating model performance is essential understanding how well your ML model is...

7 mins  
read

[Read Now](#)

# Observability without compromising

Store infinitely, search instantly and  
cut through the noise across every  
stack, region, and cloud



Platform capabilities

Remote, index-free querying

Infinite retention

DataPrime engine

In-stream analysis & alerting

Cross-stack dashboards

Cost optimization tool

Platform

APM

RUM

Infrastructure monitoring

Log analytics

AI discovery

AI observability

AI guardrails

AI security & compliance

SIEM

CSPM

MDR

Developers

Docs

Integrations

Coralogix Academy

Resources

E-books and whitepapers

Guides  
AI Guides

Coralogix blog

Case Studies

Ab

Ca

Ev  
we

Ne

Lo

Privacy Policy

Terms & Conditions

Cookies Policy

FAQs Coralogix AI Tools

Trust Center

Website Terms

© Coralogix Inc. 2025