Universidade de Brasília
Faculdade de Administração, Contabilidade, Economia e Gestão de Políticas Públicas - FACE
Departamento de Economia
Bacharelado em Economia

ELI JOSÉ BATISTA JÚNIOR

# Biases in Machine Learning
## (Vieses em Aprendizagem de Máquinas)

Brasília/DF
2024

**ELI JOSÉ BATISTA JÚNIOR**

# Biases in Machine Learning

**(Vieses em Aprendizagem de Máquinas)**

Trabalho de conclusão de curso apresentado ao Departamento de Economia da Universidade de Brasília como requisito parcial para a obtenção do título de bacharel em Ciências Econômicas.

Orientador: Prof. Dr. Daniel Oliveira Cajueiro

Brasília/DF
14 de Outubro de 2024

**ELI JOSÉ BATISTA JÚNIOR**

# Biases in Machine Learning

**(Vieses em Aprendizagem de Máquinas)**

Trabalho de conclusão de curso apresentado ao Departamento de Economia da Universidade de Brasília como requisito parcial para a obtenção do título de bacharel em Ciências Econômicas.

Este trabalho foi defendido e aprovado em 15/10/2024.

BANCA EXAMINADORA

---

Prof. Dr. Daniel Oliveira Cajueiro – UnB
Orientador

---

Prof. Dr. José Guilherme de Lara Resende – UnB
Avaliador

## AGRADECIMENTOS

Agradeço a Deus pela realização de mais um sonho, pelas experiências inesquecíveis e pelos desafios superados. Santificado seja Vosso nome e seja feita a Vossa vontade, sempre e em todo lugar.

Dedico esta vitória a minha mãe, SUELY MOURA ANDRADE BATISTA, quem me deu o dom da vida e me ensina a fazer da vida um dom a Deus, minha primeira professora e meu exemplo; a meu pai, ELI JOSÉ BATISTA, quem me ensinou e incentivou a continuamente crescer e a entregar sempre o meu melhor; aos meus irmãos, ÉLISSON ANDRADE BATISTA e SUENY ANDRADE BATISTA, meus melhores amigos, que me inspiram e acreditam em mim mais do que eu mesmo; ao meu irmão WELLINGTON MOURA ANDRADE (in memoriam), que sempre torceu de forma vibrante por minhas conquistas; e a minha linda e amada esposa, LAURA CRISTINA LOBATO DE OLIVINDO, sem a qual este trabalho nunca seria concluído, companheira e cúmplice de todas as aventuras, minha inspiração, suporte e apoio nos momentos mais desafiadores.

Agradeço aos professores de Economia da Universidade de Brasília, fontes inesgotáveis de conhecimento. Agradecimento especial aos professores Dr. Rodrigo Andrés de Souza Peñaloza e Dr. Michael Christian Lehmann, pelas aulas extraordinárias que modificaram profundamente minha visão econômica do mundo, conhecimentos estes que me empenharei em zelar e aplicar no exercício da Ciência Econômica. Agradecimento particular ao meu orientador, Prof. Dr. Daniel Oliveira Cajueiro que, no momento que mais precisei, me guiou e aconselhou.

Agradeço também aos amigos Gabriel Miranda Couto, Luísa Miranda Tauffer Padilha e Natasha de Souza Veloso, que Deus me deu a honra de conhecer durante os desafios do curso e que me ajudaram nas matérias em que mais tive dificuldades.

“EM TUDO DAI GRAÇAS.” (1 Tess 5, 18)

**Abstract**

We present a review of Machine Learning metrics, addressing the challenge of balancing bias mitigation with increasing model accuracy, focusing on the pursuit of fair ML models. We begin by defining bias, listing possible forms of occurrence, as well as the contexts in which they arise. We then analyze potential criteria that can serve as the basis for defining fairness, in order to establish possible metrics for implementation in algorithms. Each metric only partially represents fairness criteria, and the developer must choose which ones to prioritize depending on the context in which the ML tool will be used. Finally, we present a statistical analysis of the main metrics, aiming to ensure the conciliation between technical, ethical-philosophical, and legal criteria.
**Keywords**: Machine Learning, Bias, Fairness, Metrics.

# Contents



# List of Figures



# List of Tables

# 1 Introduction

Xiao (2022) defines Artificial Intelligence (AI) as a field of computer science aimed at enabling machines to perform intelligent activities, that is, capable of receiving and understanding information from the environment, performing logical and computational analysis, and then making decisions autonomously. McCarthy et al. (2007) defines Artificial Intelligence more broadly, constituting the science and engineering of making machines, especially computer programs, capable of achieving goals in the world, not necessarily as humans and animals do, but also through computational procedures that only machines can perform.

In our daily routine, we may not realize all the moments when AI has influenced our way of acting and making decisions, but there are numerous examples: customer service systems that simulate human language; personalized assistants like Amazon's Alexa, Microsoft's Cortana, Google Assistant, and iPhone's Siri; targeted advertising based on your daily activities; content offerings on social media aligned with your social circle; and suggestions for music and movies on streaming platforms.

Xiao (2022) states that AI tools can be organized into two subsets:

- **Machine Learning:** a set of mathematical algorithms capable of analyzing data in an automated way. It can be subdivided into:
  - *Supervised learning*: models are trained with labeled data. The algorithm constantly adjusts parameters to reduce the error between the calculated result and the desired result (contained in the database). The most well-known algorithms include support vector machines, naive Bayes, linear discriminant analysis, decision trees, Random forest, k-nearest neighbor algorithm, neural networks (multilayer perceptron), and similarity learning, regression algorithms (linear, logistic, and polynomial are the main ones);
  - *Unsupervised learning*: models are trained with unlabeled data. The algorithm divides the data into groups (clustering) or aims to find relationships in the database that describe the largest portion of the data. The most well-known algorithms include Hierarchical clustering, K-means clustering, K-NN, Principal component analysis, Singular-value decomposition, Independent component analysis;
  - *Semi-supervised learning*: uses labeled and unlabeled data, widely used in speech analysis, internet content classification, and protein sequence classification. The most well-known algorithms include self-training, generative methods, mixture models, and graph-based methods;
  - *Reinforcement learning*: algorithms learn through trial and error to perform tasks based on rewards, commonly used in robotics, video gaming, natural language processing, and navigation;
- **Deep Learning:** a special type of neural network with more than one hidden layer. It is considered a subset of machine learning and is divided into two types:
  - *Convolutional neural networks*: usually used in image analysis, recommendation systems, natural language processing, brain-computer interfaces, and financial time series;
  - *Recurrent neural networks*: specialized in sequence processing.

AI tools have increasingly gained space in various fields of knowledge and a wide range of applications, whether in the corporate world, in the provision of public services, or even by non-governmental organizations. Among the benefits are the growing speed of information

processing, the ability to identify patterns in unstructured data, and the possibility of greater accuracy in predictive models.

Applications of AI can be found, for example, in areas such as social identification in the criminal justice system, price prediction in financial markets, candidate selection in education systems, prioritization in disease treatment, identification of promising payers for loan approval, personnel recruitment, customized online advertising, monitoring of pedestrian and vehicle traffic, and the list continues to grow daily.

Although the expectation is that the gains will increase over time, the associated risks are still a little-explored topic, with solutions that are even less consensual among experts (Pagano et al., 2023). In addition to the intentional use of AI to produce outcomes that harm others, current AI technology must deal with undesirable or difficult-to-predict outcomes. The learning and feedback capacity of information has the potential to propagate and even amplify possible biases present in the database, in the algorithms' procedures, and in the information generated from user interaction (Mehrabi et al., 2021). Almost every sizable database contains sensitive information about people, institutions, or countries, which, if not used correctly, can lead to biased conclusions or perpetuate prejudiced behaviors already embedded in recorded data.

Pagano et al. (2023) presents a systematic review of bias and unfairness in machine learning models, listing the possible types of biases, tools for identifying injustice, bias mitigation techniques, the most commonly used databases for bias analysis, the main metrics used to analyze bias, and respective statistical approaches. The author presents one limitation regarding AI studies: the existence of few metrics that consider multiple classes of biases and sensitive information, as most existing studies present an approach focused on binary methods.

Carey and Wu (2023) presents an evaluation approach for the most commonly used metrics in machine learning studies, incorporating not only technical criteria but also ethical-philosophical and legal criteria. The author points out that the aim of his research is to fill a gap that allows AI practitioners to implement algorithms with designs that meet societal interests, while enabling social science experts to understand the reasoning behind the use of specific metrics.

The present work aims to list the main fair metrics in machine learning, as well as analyze them from the perspectives of motivation, practicality, and purpose. Thus, the Section 2 provides the foundation for the motivation to study fair metrics in machine learning, where the main concepts of bias (definition and types) are explained. The Section 3 explores the practicality and purpose of using metrics by defining technical, legal, and ethical-philosophical criteria that support measures aimed at correcting various types of bias. Also in section 3, the main metrics found in the machine learning literature are listed. The Section 4 presents results that allow machine learning practitioners to use the metrics safely, as they are aligned with technical, ethical-philosophical, and legal criteria. Finally, the Section 6 presents a summary of the findings, as well as proposals for future studies.

# 2 Bias

## 2.1 Definitions of Bias

Many authors understand that the terms bias and unfairness can be interchangeable (Booth et al., 2021). Di Noia et al. (2022) defines unfairness as the disparity between how the world should be and how it is, while they define bias as the disparity between how the world is and how the world is structured across various systems. The authors delve deeper into the concepts by describing unfairness as a subjective concept of social justice that can vary between countries, cultures, and even within institutions (such as schools, hospitals, companies, etc.), while bias refers to a systematic error that alters people's behavior or judgment about other individuals due to them having specific attributes that, when misused, may distinguish them in a harmful way.

Every system aiming to impact human behavior, positively or not, will encounter ethical, social, political, legal, and technical limits for its implementation; limits that may be surpassed when AI tools produce unexpected or undesirable results. In this sense, studying possible ways of identifying and mitigating biases becomes essential. Otherwise unfairness results may be unintentionally produced, potentially going unnoticed, since the complexity of AI models distances the user and developer from the results achieved (Jones et al., 2020). Such results, when undesirable, are a source of unfairness, which can have different understandings depending on the context, culture, and time of a given society.

There are various techniques for recognizing bias and unfairness in models and databases, each associated with a different metric, whose suitability to the practical case is a challenge in mitigating undesirable effects (Nielsen, 2020; Kordzadeh and Ghasemaghaei, 2021). For example, we may find a database where a certain social group is underrepresented due to poorly conducted data collection, so the model's accuracy will be compromised when dealing with the identifying or predicting attributes associated with such individuals. However, the underrepresentation may be merely a consequence of collective social behavior. In the first case (poor collection), the problem could be solved by adding more information about individuals from the underrepresented group to the database. In contrasto, in the second case (collective behavior), attempting to add more information could distort the sample or raise ethical and moral questions about the model's predicted results.

Suppose we have a database with information about loan applicants, where only 1/4 of the observations correspond to individuals with favorable characteristics for loan approval. A Machine Learning model with good accuracy will tend to approve 1/4 of the loans. The model cannot be considered biased because the loan is denied in about 3/4 of the cases. Now, suppose that after more careful analysis of the database, we find that the 3/4 of individuals classified as bad payers share the characteristic of being Black or poor. We still cannot categorically state that the model is biased, but we certainly cannot claim it is fair either. On the other hand, the database may be a good representation of the population distribution, leading us to suspect that the data merely reflect a real scenario, i.e., the data captured the consequences of population's prejudiced behavior over time. Situations like this raise questions about the database's quality and the sample's statistical representativeness. Finally, even if all of this could be ignored, the question of whether the results align with what is expected in a given context would remain.

This simple reasoning exercise about biases in artificial intelligence raises some questions:

- How to identify if the data are capturing some unethical behavior from the population?
- How to ensure that the data statistically represent the population?
- Should or can protected attributes (such as gender, race, social class, illnesses, address, kinship, etc.) be used in predicting results by AI?
- Is not using protected attributes enough to avoid undesirable outcomes?
- How to develop algorithms that identify the context in which the results are undesirable?

Paaßen et al. (2019) points out that many classification models inherently already carry human bias, so they tend to repeat and escalate the segregation of certain groups through a vicious cycle.

A well-known example, if not the most, is the so-called Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), software used in some states in the United States in criminal judicial processes, which aims to measure the probability of a person committing a crime again if granted parole. An investigation conducted by Angwin et al. (2016), published in 2016, found bias in the parole system against African Americans. Despite the enormous advantages of using artificial intelligence in processes involving large amounts of information, there is evidence that its use may not always be beneficial. In this line, Dressel and Farid (2018)

found that COMPAS software, when compared to the judgment of a non-expert human, did not achieve better results than a normal human. Additionally, despite COMPAS having more than 137 variables, only seven were used by people in their analyses. Finally, these authors also highlight that COMPAS performs worse when compared to a simple logistic regression in decision-making.

## 2.2 Types of Bias

Biases can originate from various sources, but in a way, we can affirm that everything begins and ends with human behavior (Mehrabi et al., 2021). People are social beings, and therefore internalize information as they interact with others or as they observe third-party interactions. Thus, the knowledge already learned and the context in which the individual is embedded form the basis for learning new concepts and prejudices.

In this way, biases may arise depending on the stage and manner in which human knowledge is transferred to the Artificial Intelligence tool:

- **in database**: during the collection and gathering of information in a database, several issues may arise, such as underrepresentation of individuals, direct and indirect social prejudices, absence of relevant information, biased information due to a lack of randomness, etc.;
- **in algorithm development**: the developer's worldview and knowledge limitations may lead them to encode procedures that become noticeable only after many iterations or on a large scale;
- **in the interaction between user and algorithm**: users may make decisions or react based on the AI tool's result. This user's action may serve as information to feed back into the database, and depending on the interaction between the algorithm and the user, biased information may emerge. For example, Twitter (now X) is known for presenting users with only content aligned with their preferences, producing real information bubbles and polarizing debates Kowald and Lex (2018).

Mehrabi et al. (2021), Suresh and Guttag (2019), and Olteanu et al. (2019) present the main sources of biases in the AI field, which are summarized in the following topics.

### 2.2.1 Data-to-algorithm bias

This section presents the first class of bias, which corresponds to the situation where the database used to train the model is biased, such that the algorithm will learn these biases and reflect them in its results, possibly even amplifying and perpetuating them (Mehrabi et al., 2021):

1. **Measurement Bias**: bias stemming from how the information was chosen, recorded, or measured in the database (Suresh and Guttag, 2019). For example, in the COMPAS project database, one of the variables corresponded to a proxy for "crime risk," which was measured by the number of family members' arrests. It turns out that marginalized groups are more frequently monitored by the police ( whether due to historical prejudice or having fewer resources for security investment), which strongly correlates with the number of arrests;
2. **Omitted Variable Bias**: occurs when important variables are not recorded or not considered by the algorithm (Clarke, 2005; Mustard, 2003; Riegg, 2008). For example, the emergence of a competing company may represent information that the model was not previously prepared to learn.

3. **Representation Bias**: relates to how the information is collected (Suresh and Guttag, 2019);

4. **Aggregation Bias**: corresponds to using group information as a proxy for individual information (Suresh and Guttag, 2019). For example, a model aimed at assisting diabetes treatment may have its accuracy decreased if it does not consider individual information within the various gender and ethnic groups (Suresh and Guttag, 2019).

   (a) *Simpson's Paradox*: heterogeneous data, when clustered or disaggregated into subgroups, may alter analysis results, sometimes even inverting them (Blyth, 1972);

   (b) *Modifiable Areal Unit Problem*: bias obtained by the algorithm when trained on different levels of spatial aggregation (Gehlke and Biehl, 1934)

5. **Sampling Bias**: Similar to representation bias, but here the sample presents a problem due to a lack of randomness in the subgroups of data (Mehrabi et al., 2021);

6. **Longitudinal Data Fallacy**: individuals followed in research over several years may exhibit biased behavior, so cross-sectional and longitudinal analyses may diverge significantly (Mehrabi et al., 2021). Barbosa et al. (2016) presents the case where an analysis on the length of comments made by individuals on Reddit apparently decreased over time. However, upon deepening the analysis, the author concluded that the length of each individual's comments had been growing over time, but the individuals in each time period varied considerably, causing the result in mentioned result in the analysis of the grouped data.

#### 2.2.2 Algorithm-to-User bias

The second type of bias results from the optimization and classification that models perform on the database, even if they accurately represent the population, through processes that accentuate segregations between groups or individuals (Pagano et al., 2023). Biases produced by algorithms can potentially modify user behaviors, generating behavioral biases in future databases (Mehrabi et al., 2021). In this class of biases, we have:

1. **Algorithmic Bias**: when the bias is produced by the design of the algorithm (Baeza-Yates, 2018), which can occur for various reasons, such as when the algorithm analyzes only part of the data, uses biased estimators, or through the combination of optimization, classification, regularization, and clustering procedures (Danks and London, 2017);

2. **User Interaction Bias**: resulting from the interface imposed on the user or from the behavioral bias caused by the user when interacting with the algorithm (Baeza-Yates, 2018):

   (a) *Presentation Bias*: a consequence of how information is presented to the user (Baeza-Yates, 2018) ;

   (b) *Ranking Bias*: the presentation of ranked information may induce the user to limit their analysis to the top-ranked options, attracting more clicks to those options and reinforcing their position among the top-ranked (Mehrabi et al., 2021).

3. **Popularity Bias**: options that are more popularly known tend to receive greater exposure, even without representing better quality (Mehrabi et al., 2021). Popularity metrics can be manipulated, for example, by social bots (Ciampaglia et al., 2018);

4. **Emergent Bias**: a consequence of changes in population characteristics and collective cultural values over time, as well as the complexity of interactions between users and algorithms (Friedman and Nissenbaum, 1996).

5. **Evaluation Bias**: caused by the use of improper references during the evaluation stage of the algorithm (Suresh and Guttag, 2019). .

#### 2.2.3 User-to-data bias

The third type of bias concerns how the user chooses and uses the model. The user may not be aware of how the tool processes the data, so their actions translate into a biased interaction, thus compromising the results (Pagano et al., 2023). Many machine learning applications use a set of user-generated data models as their data source, meaning that the user's behavior and the way the algorithm influences them may introduce biases into the data (Mehrabi et al., 2021). In this class, we have:

1. **Historical Bias**: even if the sample data is a statistical population representation, it may contain biases resulting from prejudices and socio-technical issues (Suresh and Guttag, 2019). An example pointed out by Suresh and Guttag (2019) relates to the generation of images of female CEOs, whose results are biased due to the fact that only about 5% of *Fortune 500 CEOs* were women in 2018;

2. **Population Bias**: refers to the lack of representativeness of the algorithm when applied to a population demographically different from the one it was trained on (Olteanu et al., 2019). An example presented by Mehrabi et al. (2021) is the difference in preferences regarding the use of social media, where women tend to prefer platforms like Facebook and Instagram, while men are more active on forums like Reddit and Twitter (now X).

3. **Self-selection Bias**: occurs when the use of the algorithm or the insertion of information into the data reflects the user's self-selection process, as seen in electoral polls, where supporters tend to engage more in opinion polls about the candidates they support (Mehrabi et al., 2021).

4. **Social Bias**: occurs when an individual's judgment is influenced by the actions of others (Baeza-Yates, 2018).

5. **Behavioral Bias**: occurs when users' behavior diverges across platforms, contexts, or different datasets (Olteanu et al., 2019); Miller et al. (2021) show that differences in emojis across platforms lead users to different behaviors on those platforms, sometimes causing communication errors.

6. **Temporal Bias**: Olteanu et al. (2019) explain that this refers to the change in behavior over time, such as when people use a certain hashtag to draw attention to a topic, but as the conversation progresses, the hashtag falls out of use.

7. **Content Production Bias**: arises when an individual's context involves structural, lexical, semantic, and syntactic differences in the content (Olteanu et al., 2019). For example, the difference in the meaning of certain linguistic expressions between people of different ages and genders (Nguyen et al., 2013).

## 3 Fairness

### 3.1 Definition of Fairness

Systems that use Machine Learning may present biases or unfair results for various reasons, for example, to the social bias present in the database used during the training stage, to social bias embedded (intentionally or not) in the algorithm's design, or even due to biases emerging from the interaction between the user and the algorithm over time. One way to address this issue is

to try to measure the bias or fairness of the results produced by Machine Learning. However, while some metrics may indicate the existence of bias in a certain model, others may suggest that the results are fair (Pagano et al., 2023).

Carey and Wu (2023) argue that many studies have achieved good results in explaining the mathematical and algorithmic aspects of machine learning tools, but without delving into the philosophical and legal details that are fundamental to adapting the technique to the social system (which he calls the sociotechnical system). On the other hand, several studies have detailed the philosophical and legal foundations in proposing fair metrics for machine learning but disconnected from the mathematical and algorithmic aspects, making their technical implementation in the social system impossible. In their research, Carey and Wu (2023) aim to bridge this gap, allowing machine learning practitioners to understand all these aspects, presenting how fair ML metrics work while linking them to the foundation of social sciences. The following subsections present a summary of the criteria outlined by Carey in his literature review.

### 3.1.1 Technical Criteria (Mathematical-Statistical)

Barocas and Selbst (2016) emphasized that most proposed fairness criteria concern the properties of the joint distributions of the protected attribute $S$ ($S = 0$ non-marginalized, $S = 1$ marginalized) regarding a marginalized group (e.g., gender or race), the actual attribute $Y$, and the classification resulting from the model $\hat{Y}$.

**Independence:** occurs when the protected attribute $S$ has no relation to the classification outcome $\hat{Y}$. In mathematical terms: $\hat{Y} \perp S$. For the binary case, Barocas and Selbst (2016) present two possible formulations:

$$\text{Exact: } P[\hat{Y} = 1|S = 0] = P[\hat{Y} = 1|S = 1]$$

$$\text{Relaxed: } \frac{P[\hat{Y} = 1|S = 0]}{P[\hat{Y} = 1|S = 1]} \geq 1 - \epsilon$$

.

The metrics that meet this criterion do not consider any relationship with the variable that stores the actual value $Y$, which limits the model's accuracy and creates incentives for using lazy metrics (Carey and Wu, 2023). For example, equal random selection among students from marginalized and non-marginalized groups satisfies the independence criterion without considering individual characteristics, such as effort and knowledge.

**Separation:** This criterion considers that, in many scenarios, it is desirable for the characteristic $S$ that defines the marginalized group to be correlated with the actual value of the target variable $Y$. In other words, conditional on the values of $Y$, it is expected that the classification $\hat{Y}$ be independent of the characteristic $S$: $\hat{Y} \perp S|Y$ (Barocas and Selbst, 2016). Mathematically, this is equivalent to stating that all groups should have the same true positive and false positive rates (Carey and Wu, 2023):

$$\text{TP: } P[\hat{Y} = 1|Y = 1 \cap S = 0] = P[\hat{Y} = 1|Y = 1 \cap S = 1]$$

$$\text{FP: } P[\hat{Y} = 1|Y = 0 \cap S = 0] = P[\hat{Y} = 0|Y = 1 \cap S = 1]$$

**Sufficiency:** This criterion is met when, for the purpose of predicting $Y$, the value of $S$ does not need to be used if $\hat{Y}$ is provided (Barocas and Selbst, 2016). For example, in the case of student admissions, if students' GPA and SAT scores are sufficient in relation to race, then the admissions committee does not need to actively consider race when making their decision (Carey and Wu, 2023). Formally, $Y \perp S|\hat{Y}$ (i.e., $Y$ is conditionally independent of $S$, given $\hat{Y}$). In binary classification, this is equivalent to requiring parity of positive and negative values between groups $\hat{y} \in \hat{Y} = \{0, 1\}$ (Carey and Wu, 2023):

$$P[Y = 1|\hat{Y} = \hat{y} \cap S = 0] = P[Y = 1|\hat{Y} = \hat{y} \cap S = 1].$$

Metrics based solely on statistical grounds are not sufficient to guarantee fairness for individuals or marginalized groups Corbett-Davies et al. (2023). On the contrary, such metrics only provide reasonable guarantees for the "average" members of marginalized groups and are often conflicting with each other (Carey and Wu, 2023). For example, among the metrics presented in section 3.2, it is impossible to simultaneously satisfy false-positive rates, false-negative rates, and positive predictive value across marginalized groups (Carey and Wu, 2023). This observation was exposed by Barocas and Selbst (2016), demonstrating that the fairness criteria of representational independence, separation, and sufficiency constitute statistical limits that cannot be implemented simultaneously.

#### 3.1.2 Ethical-Philosophical Criteria

Carey and Wu (2023) presents the notion of Equality of Opportunity (EOP) in three ways: formal EOP, substantive EOP, and luck-egalitarian EOP. EOP explores the idea of equality of competition among all members of society. The author warns that the concept of formal EOP is important for understanding the other forms but does not constitute a good foundation for constructing fair metrics.

**Formal EOP:** Any desirable position must be within reach of all, and selection should be based on relevant individual qualifications, so that the most qualified achieves the result (Carey and Wu, 2023). In machine learning, this criterion has been implemented as fairness through blindness or fairness through unawareness (Kusner et al., 2018).

While formal EOP has the benefit of excluding protected information in the classification process, its application does not imply the correction of arbitrary privileges (Carey and Wu, 2023). Individuals belonging to marginalized groups are disproportionately impacted by the correlation between the protected attribute and challenges such as poverty, racism, bullying, and discrimination. In other words, the impacts of these challenges represent a tax on individuals' cognitive abilities (Berk et al., 2017).

As Carey and Wu (2023) states, the problem of not accounting for arbitrary privileges can be divided into two components:

- **before** the classification, arbitrary privileges allow some individuals to better develop the qualifications that will be relevant for selection;
- **after** the classification, the winners will be in even more advantageous positions to compete in future selections.

Formal EOP introduces a "snowball" component where winners win faster, but losers lose faster (Khan et al., 2021). From a symmetrical perspective, formal EOP amplifies both advantages and disadvantages, which is commonly known as *discrimination laundering* (Green, 2016).

**Substantive (Rawls') EOP:** This criterion addresses the problem of discrimination laundering, as it requires that all individuals have the same opportunities to develop qualifications (Carey and Wu, 2023). However, the assumption that talents are equally distributed is often not verified in real life. Rawls' EOP has been implemented in fair machine learning in the sense that a candidate should be selected based only on characteristics that are not related to unequal opportunities for qualification prior to selection Khan et al. (2021).

**Luck-Egalitarian EOP:** This criterion requires that an individual's selection should be based solely on their choices and not their circumstances (Segall, 2013), but making this separation is not trivial and is, in fact, very difficult to implement (Carey and Wu, 2023). Another issue is the simultaneous influence between choices and circumstances (Khan

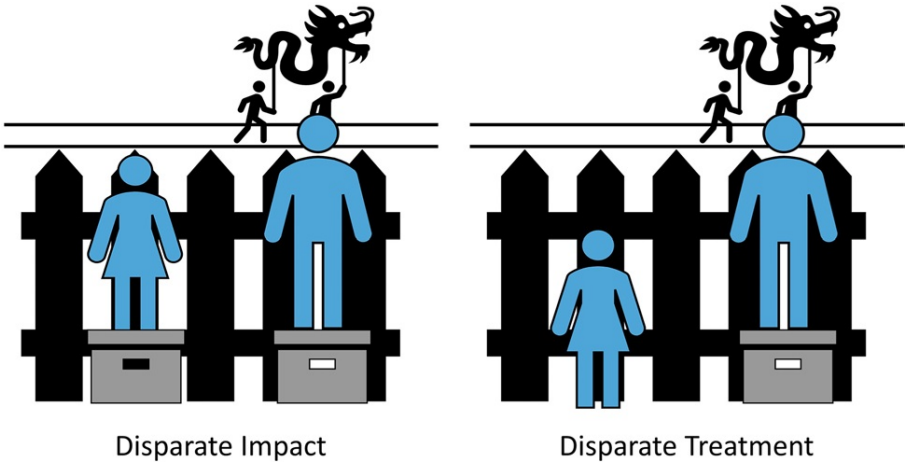


Figure 1: Difference beetwen DI and DT. Fonte: Carey and Wu (2023)
Graphical depiction of disparate impact vs. disparate treatment. For the disparate impact (left) side, while both people were given boxes to stand on to see over the fence, the person on the left is disproportionately impacted by being shorter. For disparate treatment, the person on the left if directly being discriminated against by not being given a box to stand on. (Carey and Wu, 2023)

et al., 2021); for example, wealthy students may choose to dedicate more time to studies because they have better conditions (Carey and Wu, 2023).

### 3.1.3 Legal Criteria

Carey and Wu (2023) points out some legal principles commonly used in machine learning, due to legal requirements contained in the United States anti-discrimination legislation. Among them, the anti-classification and anti-subordination criteria are the principles that motivated the creation of such legislation (Xiang and Raji, 2019; Balkin and Siegel, 2003).

**Disparate Impact (DI):** This refers to avoiding the undesirable consequence of using neutral policies, where individuals from marginalized groups are more affected than others (Pessach and Shmueli, 2020). In fair machine learning, this occurs when, even unintentionally, the number of positive outcomes differs drastically between different groups (Lipton et al., 2019).

**Disparate Treatment (DT):** This refers to avoiding negative consequences arising from intentionally different treatment between individuals from marginalized and non-marginalized groups. Carey and Wu (2023) provided Figure 1 in their work, which clearly summarizes the difference between DI and DT.

A common approach to implementing DT is the exclusion of protected attributes from the input arguments of algorithms (Carey and Wu, 2023). This approach is also known as equal treatment, as equal individuals are treated equally, regardless of which group they belong to (Corbett-Davies et al., 2023).

**Anti-classification:** This principle imposes that the U.S. government cannot classify people using protected attributes (Carey and Wu, 2023). Any metric that meets the independence criterion also satisfies anti-classification, as $\hat{Y} \perp S$.

**Anti-subordination:** This refers to the ideal that laws should address existing social hierarchies, in the sense that it is inappropriate for any group to be subordinated to others. Thus, the law must aim to empower marginalized groups, even if at the expense of non-marginalized groups (Colker, 1986).

### 3.2 Fairness metrics

Machine Learning models have been improved with bias and fairness measurement methodologies (Paaßen et al., 2019). However, when it comes to long-term decisions, models have produced unsatisfactory results despite the vast number of metrics already developed, as stated by Pagano et al. (2023). The authors argue that some measurement metrics are insufficient because they either only consider the impact on the individual, or only on the group, or because the metrics are incapable of predicting the model's behavior over time.

Most existing metrics use the concepts of *true positive (TP)*, *false positive (FP)*, *true negative (TN)*, and *false negative (FN)*. These values are obtained by counting the number of errors and correct predictions the model is capable of producing. In supervised machine learning models, this involves comparing the predicted or classified results with the labeled data in the database. In unsupervised machine learning models, this involves comparing the results with some reward criterion embedded in the algorithm.

Let $Y$ be the variable that represents the actual value of an observation. For example, if a banking institution intends to use an AI tool to select good payers, good payers are those with a positive classification, represented by $Y = 1$. Bad payers are those with a negative classification, represented by $Y = 0$. Now, let $\hat{Y}$ be the result obtained by the ML model. What the models do is attempt to classify or predict to which group of payers individuals belong, so that a positive result corresponds to $\hat{Y} = 1$ and a negative result to $\hat{Y} = 0$.

Thus, the measure **TP** corresponds to the number of observations where the model provides a positive result that coincides with the actual value ($Y = 1 \cap \hat{Y} = 1$); **FP** corresponds to the number of observations where the model provides a positive result, but it diverges from the actual value ($Y = 0 \cap \hat{Y} = 1$). **TN** corresponds to the number of observations where the model provides a negative result that coincides with the actual value ($Y = 0 \cap \hat{Y} = 0$); **FN** corresponds to the number of observations where the model provides a negative result, but it diverges from the actual value ($Y = 1 \cap \hat{Y} = 0$).

These measures are sufficient to understand bias and fairness metrics for models with binary outcomes. To understand metrics for problems with multiple classes, the concepts of TP, FP, TN, and FN need to be explained more thoroughly. In these cases, the values are assigned and counted relative to each class (see fig. 2). Let's use an example to better explain. Continuing with the hypothetical scenario of loan approval, suppose now we intend to classify individuals into three types of possible payers: low risk (class 1), medium risk (class 2), and high risk (class 3). Let's now restrict our attention to group 1, the low-risk group, i.e., those individuals who have this classification originally in the database. When using a machine learning model in the training stage, an individual will either be classified in group 1 or in the other groups. In this case, if the individual is classified correctly, we have a *TP* result; regarding class 1, the other classes correspond to negative values, so if the individual is classified in group 2 or 3, the model will present a negative result that does not match reality, i.e., it is a case of *FN*; regarding class 1, individuals from classes 2 and 3, when classified in class 1, receive a positive result from the model, but this does not match reality, i.e., it is a case of *FP*; the other cases (individuals from group 2 who are classified in group 3 and vice versa) are not related to class 1, so they correspond to *TN*.

Pagano et al. (2023), among all the metrics used for evaluating fairness and bias, highlights the most frequently mentioned in the works analyzed in their literature review: Equalized Odds (EO), Equality of Opportunity (EOP), Demographic Parity (DP), Individual Differential Fairness. In addition to these, Pagano et al. (2023) and Carey and Wu (2023) present other metrics, specified below, whose references can be found in Table 1:

1. **Positive Predctive Value (PPV) or precision**:

$$PPV = \frac{TP}{TP + FP}$$

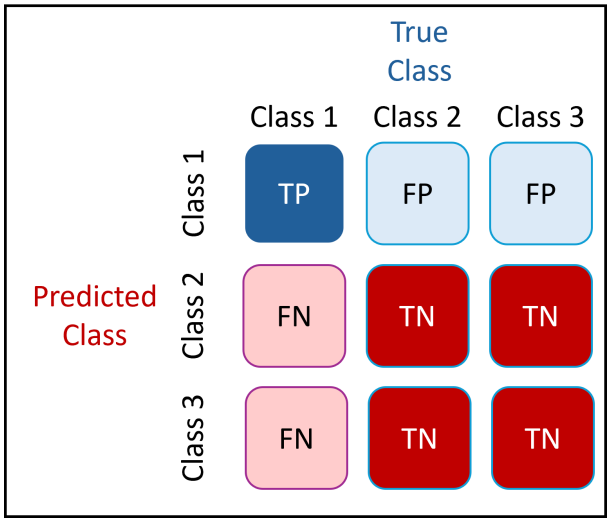


Figure 2: Matrix Multiclass

2. **Negative Predictive Value (NPV)**:

$$NPV = \frac{TN}{TN + FN}$$

3. **Predictive Value Parity (PVP)** - also called Conditional Use Accuracy, it is satisfied when both PPV and NPV are similar in both marginalized and non-marginalized groups. In notation, for $y \in 0, 1$: $y \in \{0, 1\}$:

$$P\big[Y = y|\hat{Y} = y \cap S = 0\big] = P\big[Y = y|\hat{Y} = y \cap S = 1\big]$$

4. **Predctive Parity (PP)** - also known as the Outcome Test, this metric requires that PVP be equal for both marginalized and non-marginalized groups. In notation, for $y \in \{0, 1\}$:

$$P\big[Y = y|\hat{Y} = 1 \cap S = 0\big] = P\big[Y = y|\hat{Y} = 1 \cap S = 1\big]$$

   Predictive Parity corresponds to PVP when $\hat{Y} = 1$.

5. **False Discovery Rate (FDR)**:

$$FDR = \frac{FP}{TP + FP}$$

6. **False Omission Rate (FOR)**:

$$FOR = \frac{FN}{TN + FN}$$

7. **True Positive Rate (TPR) or Recall Score**:

$$TPR = \frac{TP}{TP + FN}$$

8. **False Positive Rate (FPR)**:

$$FPR = \frac{FP}{FP + TN}$$

9. **False-Positive Error Rate Balance (FPERB)** - also known as Predictive Equality, it requires that the FPR (False Positive Rate) be similar across different groups. In notation, for $y \in \{0, 1\}$:

$$P\big[\hat{Y} = \hat{y} | Y = 0 \cap S = 0\big] = P\big[\hat{Y} = \hat{y} | \hat{Y} = 0 \cap S = 1\big]$$

10. **False Negative Rate (FNR)**:

$$FNR = \frac{FN}{TP + FN}$$

11. **False-Negative Error Rate Balance (FNERB):** - also known as Equal Opportunity, it is the opposite of FPERB. Here, it requires that the FNR (False Negative Rate) be similar across groups. In notation, for $y \in \{0, 1\}$:

$$P\big[\hat{Y} = \hat{y} | Y = 1 \cap S = 0\big] = P\big[\hat{Y} = \hat{y} | \hat{Y} = 1 \cap S = 1\big]$$

12. **True Negative Rate (TNR)**:

$$TNR = \frac{TN}{FP + TN}$$

13. **Treatment Equality (TE)** - it requires that the ratio of FN (False Negatives) and FP (False Positives) be equal across groups:

$$\frac{FN_{S=0}}{FP_{S=0}} = \frac{FN_{S=1}}{FP_{S=1}}$$

14. **Equalized Odds (EO)** - also called Conditional Procedure Accuracy Equality and Disparate Mistreatment, it aims to ensure that (i) the probability of an individual in the positive class receiving a positive estimate in the protected group is the same as in the non-protected group; and (ii) the probability of an individual in the negative class erroneously receiving a positive estimate in the protected group is the same as in the non-protected group:

$$EO = \frac{1}{2} \cdot \Big( \Big| \frac{FP_p}{FP_p + TN_p} - \frac{FP_u}{FP_u + TN_u} \Big| + \Big| \frac{TP_p}{TP_p + FN_p} - \frac{TP_u}{TP_u + FN_u} \Big| \Big)$$

In notation, for $y \in \{0, 1\}$:

$$P\big[\hat{Y} = 1 | Y = y \cap S = 0\big] = P\big[\hat{Y} = 1 | Y = y \cap S = 1\big]$$

15. **Equality of Opportunity (EOP)** - It aims to ensure that the probability of an individual in the positive class receiving a positive result is the same in both the protected group and the non-protected group:

$$EOP = \frac{TP_p}{TP_p + FN_p} - \frac{TP_u}{TP_u + FN_u}$$

16. **Demographic Parity (DP)** - also known as Statistical Parity, Statistical Fairness, Equal Acceptance Rate, or Benchmarking. It represents the probability of being classified positively for both marginalized and non-marginalized groups:

$$DP = \frac{TP + FP}{N}$$

In notation, it is expected to obtain:

$$P\big[\hat{Y} = 1 | S = 0\big] = P\big[\hat{Y} = 1 | S = 1\big]$$

17. **Conditional Statistical Parity (CSP)** - it constitutes a relaxation of Statistical Parity, where the results may be related to other sets of attributes (whether correlated or not with the actual values of $Y$). This condition is satisfied when both marginalized and non-marginalized groups have equal probabilities of being classified positively, after controlling for a set of factors, say $L$:

$$P\big[\hat{Y} = 1 | L = \ell \cap S = 0\big] = P\big[\hat{Y} = 1 | L = \ell \cap s = 1\big]$$

In some cases, CSP has the ability to bypass Simpson's paradox, since the conditional information added to the statistic can mitigate the aggregation of the original classification labels (Carey and Wu, 2023).

18. **Disparate Impact (DI)** - compares the proportion of individuals who receive a positive outcome between groups with protected and non-protected characteristics:

$$DI = \frac{\frac{TP_p + FP_p}{N_p}}{\frac{TP_u + FP_u}{N_u}}$$

19. **K-Nearest Neighbors Consistency (KNNC)** - an individual fairness metric that aims to measure the sensitivity of an attribute between similar individuals:

$$KNNC = 1 - \frac{1}{n} \sum_{i=1}^{n} \left| \hat{y}_i - \frac{1}{k} \sum_{j \in N_k(x_i)} \hat{y}_j \right|$$

20. **Absolute Balanced Accuracy Difference (ABAD)** - represents the accuracy difference between the protected and non-protected groups:

$$ABAD = \frac{1}{2} \cdot \Big| (TPR_p + TNR_p) - (TPR_u + TNR_u) \Big|$$

21. **Absolute Average Odds Difference (AAOD)** - represents the difference between TPR (True Positive Rate) and FPR (False Positive Rate) between two protected groups:

$$AAOD = \frac{1}{2} \cdot \Big| \big(FPR_p + FPR_p\big) - \big(TPR_u + TPR_u\big) \Big|$$

22. **Absolute Equal Opportunity Rate Difference (AEORD)** - the difference between Recall Scores between the protected and non-protected groups:

$$AEORD = \big| TPR_p - TPR_u \big|$$

23. **Statistical Parity Difference (SPD)** - the difference between the Statistical Parity of the protected group and the non-protected group:

$$SPD = \frac{TP_p + FP_p}{N_p} - \frac{TP_u + FP_u}{N_u}$$

24. **F1-score** - it constitutes a weighted average between precision and recall:

$$F1 - score = 2 \cdot \frac{recall \cdot precision}{recall + precision}$$

25. **Overall Accuracy Equality or Accuracy** - represents the number of correct classifications by the model:

$$accuracy = \frac{TN + TP}{TN + TP + FN + FP}$$

In notation, for $y, \hat{y} \in \{0, 1\}$:

$$P\big[\hat{Y} = y | Y = y \cap S = 0\big] = P\big[\hat{Y} = y | Y = y \cap S = 1\big]$$

26. **Number of Positive Instances (NIP):**

$$NIP = TP + FP$$

27. **Number of Negative Instances (NIN):**

$$NIN = TN + FN$$

28. **Base Rate (BR):**

$$BR = \frac{NIP}{N}$$

29. **Test fairness (TF)** - also known as Calibration, Equal Calibration, and Matching Conditional Frequencies:

$$P[Y = 1 | \mathcal{P} = p \cap S = 0] = P[Y = 1 | \mathcal{P} = p \cap S = 1]$$

30. **Well Calibration (WC):**

$$P[Y = 1 | \mathcal{P} = p \cap S = 0] = P[Y = 1 | \mathcal{P} = p \cap S = 1] = p$$

31. **Balance for the Positive Class (BPC):**

$$\mathbb{E}[\mathcal{P} | Y = 1 \cap S = 0] = \mathbb{E}[\mathcal{P} | Y = 1 \cap S = 1]$$

32. **Balance for the Negative Class (BNC):**

$$\mathbb{E}[\mathcal{P} | Y = 0 \cap S = 0] = \mathbb{E}[\mathcal{P} | Y = 0 \cap S = 1]$$

# 4 Statistical analysis

In this section, we will present a summary of the results developed by Carey and Wu (2023), concerning the adequacy of fair machine learning metrics to the ethical-philosophical, legal, and technical criteria presented in subsections 3.1.1, 3.1.2 e 3.1.3. However, we will leave aside the formal proofs, which we recommend reading directly from the authors' article.

The following variables will be used: $S = 1$ to denote individuals who have the protected attribute or belong to the marginalized group; $S = 0$ for individuals who do not have the protected attribute and do not belong to the marginalized group; $\hat{Y}$ corresponds to the predicted value or classification label by the model, and $Y$ is the actual value or true classification label.

| Statistical Metrics | References |
|---|---|
| PPV or precision | Verma and Rubin (2018) |
| NPV | Verma and Rubin (2018) |
| PVP | Berk et al. (2021) |
| PP | Chouldechova (2016) |
| FDR | Verma and Rubin (2018) |
| FOR | Verma and Rubin (2018) |
| TPR or recall score | Verma and Rubin (2018), Quadrianto and Sharmanska (2017), Georgopoulos et al. (2021) |
| FPR | Verma and Rubin (2018), Chiappa and Isaac (2019), Yang et al. (2020) Adel et al. (2019), Quadrianto and Sharmanska (2017), Grari et al. (2019), Jain et al. (2019) |
| FPERB | Chouldechova (2016) |
| FNR | Verma and Rubin (2018), Chiappa and Isaac (2019), Yang et al. (2020), Adel et al. (2019), Grari et al. (2019), Jain et al. (2019) |
| FNERB | Chouldechova (2016) |
| TNR | Verma and Rubin (2018) |
| TE | Berk et al. (2021) |
| EO | Mehrabi et al. (2021), Georgopoulos et al. (2021), Radovanovic et al. (2020), Du et al. (2021), Reddy et al. (2021), Hardt et al. (2016) |
| EOP | Mehrabi et al. (2021), Li et al. (2022), Amend and Spurlock (2021) Radovanovic et al. (2020), Reddy et al. (2021) |
| DP | Mehrabi et al. (2021), Jalal et al. (2021), Li et al. (2022), Amend and Spurlock (2021), Du et al. (2021), Reddy et al. (2021), Dwork et al. (2012), Kusner et al. (2018) |
| CSP | Corbett-Davies et al. (2023) |
| DI | Bryant et al. (2019) |
| KNNC | Bryant et al. (2019) |
| ABAD | Jang et al. (2021) |
| AAOD | Jang et al. (2021) |
| AEORD | Jang et al. (2021) |
| SPD | Bryant et al. (2019), Jang et al. (2021) |
| accuracy | Das et al. (2019), Amend and Spurlock (2021), Reddy et al. (2021), Berk et al. (2021) |
| precision | Das et al. (2019) |
| recall | Das et al. (2019) |
| F1-score | Das et al. (2019) |
| NIP | Bryant et al. (2019) |
| NIN | Bryant et al. (2019) |
| BR | Bryant et al. (2019) |
| TF | Chouldechova (2016) |
| WC | Kleinberg et al. (2016) |
| BPC | Kleinberg et al. (2016) |
| BNC | Kleinberg et al. (2016) |

Table 1: Statistical Metrics References. Fonte: Pagano et al. (2023); Carey and Wu (2023)

**Definition 1** (Rawls' EOP and luck-egalitarian EOP for supervised learning (Heidari et al., 2018))**.** A predictive model $h$ satisfies Rawlsian/luck-egalitarian EOP if for every $s \in S = \{0, 1\}$ and $y, \hat{y} \in Y, \hat{Y} = \{0, 1\}$ the following conditions are satisfied:

$$\text{Rawls': } F^h\big(U \leq u | S = 0 \cap Y = y\big) = F^h\big(U \leq u | S = 1 \cap Y = y\big)$$

$$\text{LE: } F^h\big(U \leq u | S = 0 \cap \hat{Y} = \hat{y}\big) = F^h\big(U \leq u | S = 1 \cap \hat{Y} = \hat{y}\big)$$

where $F^h\big(U \leq u\big)$ corresponds to the distribution of the utility $U$ of individuals under the model $h$, that is, it is the difference between the individual's actual effort $A$ and their circumstances $D$, $U = A - D$.

Carey and Wu (2023) organizes fairness metrics into three groups:

1. **Predicted Outcomes:** This group contains the simplest metrics, using predicted values, in which marginalized and non-marginalized classes have the same probability of receiving a positive prediction from the model (Verma and Rubin, 2018):
   (a) *Statistical Parity*: $P\big[\hat{Y} = 1 | S = 0\big] = P\big[\hat{Y} = 1 | s = 1\big]$
   (b) *Conditional Statistical Parity:* $P\big[\hat{Y} = 1 | L = \ell \cap S = 0\big] = P\big[\hat{Y} = 1 | L = \ell \cap s = 1\big]$
2. **Predicted and actual outcomes:** now the metrics are related to the actual values.
   (a) *Predictive Value Parity or COnditional Use Accuracy:* $P\big[Y = y | \hat{Y} = y \cap S = 0\big] = P\big[Y = y | \hat{Y} = y \cap S = 1\big]$
   (b) *Predctive Parity (PP)*: $P\big[Y = y | \hat{Y} = 1 \cap S = 0\big] = P\big[Y = y | \hat{Y} = 1 \cap S = 1\big]$
   (c) *Equalized Odds:* $P\big[\hat{Y} = 1 | Y = y \cap S = 0\big] = P\big[\hat{Y} = 1 | Y = y \cap S = 1\big]$
   (d) *False-positive error rate balance:* $P\big[\hat{Y} = \hat{y} | Y = 0 \cap S = 0\big] = P\big[\hat{Y} = \hat{y} | \hat{Y} = 0 \cap S = 1\big]$
   (e) *False-negative error rate balance:* $P\big[\hat{Y} = \hat{y} | Y = 1 \cap S = 0\big] = P\big[\hat{Y} = \hat{y} | \hat{Y} = 1 \cap S = 1\big]$
   (f) *Overall Accuracy Equality:* $P\big[\hat{Y} = y | Y = y \cap S = 0\big] = P\big[\hat{Y} = y | Y = y \cap S = 1\big]$
   (g) *Treatment Equality (TE):*
   $$\frac{FN_{S=0}}{FP_{S=0}} = \frac{FN_{S=1}}{FP_{S=1}}$$
3. **Predicted probabilities and actual outcomes**:
   (a) *Test fairness:* $P[Y = 1 | \mathcal{P} = p \cap S = 0] = P[Y = 1 | \mathcal{P} = p \cap S = 1]$
   (b) *Well calibration:* $P[Y = 1 | \mathcal{P} = p \cap S = 0] = P[Y = 1 | \mathcal{P} = p \cap S = 1] = p$
   (c) *Balance for the positive class:* $\mathbb{E}[\mathcal{P} | Y = 1 \cap S = 0] = \mathbb{E}[\mathcal{P} | Y = 1 \cap S = 1]$
   (d) *Balance for the negative class:* $\mathbb{E}[\mathcal{P} | Y = 0 \cap S = 0] = \mathbb{E}[\mathcal{P} | Y = 0 \cap S = 1]$

**Proposition 1** (Statistical parity as Rawls' EOP (Heidari et al., 2018))**.** Consider the binary classification task where $Y, \hat{Y} = \{0, 1\}$. Suppose that $U = A - D$, $A = \hat{Y}$, and $D = Y = 1$ (i.e., the utility based on the effort of all individuals is assumed to be the same). Then, the conditions for Rawls' EOP to be satisfied are equivalent to Statistical Parity when $\hat{Y} = 1$.

**Proposition 2** (Conditional use accuracy as luck-egalitarian EOP (Heidari et al., 2018))**.** Consider the binary classification task where $Y, \hat{Y} = 0, 1$. Suppose that $U = A - D$, $A = \hat{Y}$, and $D = Y$ (here, we do not need to have $Y = 1$). Then, the conditions that satisfy luck-egalitarian EOP are equivalent to those that satisfy PVP.

**Proposition 3** (Equalized Odds as Rawls' EOP (Heidari et al., 2018))**.** Consider the binary classification task where $Y, \hat{Y} = \{0, 1\}$. Suppose that $U = A - D$, $A = \hat{Y}$, and $D = Y$. Then, the conditions that satisfy Rawls' EOP are equivalent to those that satisfy EO (Equalized Odds).

**Proposition 4** (Overall accuracy equality as Rawls' EOP (Heidari et al., 2018))**.** Consider the binary classification task where $Y, \hat{Y} = \{0, 1\}$. Suppose that $U = A - D$, $A = (\hat{Y} - Y)^2$, and $D = 0$. Then, the conditions that satisfy Rawls' EOP are equivalent to those that satisfy Overall Accuracy Equality.

**Proposition 5** (Test Fairness as luck-egalitarian EOP (Carey and Wu, 2023))**.** Test fairness satisfies luck-egalitarian EOP.

**Remark.** Predictive Parity corresponds to PVP when $\hat{y} = 1$, so it falls into the same philosophical category as PVP, which is luck-egalitarian EOP.

**Remark.** False-positive error rate balance can be seen as a relaxed version of EO (Equalized Odds), and thus it satisfies Rawls' EOP.

**Remark.** Treatment Equality can be seen as the ratio between FPERB (False Positive Error Rate Balance) and FNERB (False Negative Error Rate Balance), and thus it satisfies Rawls' EOP.

## 5 Mitigation techniques

Given the diversity of existing biases, the scientific community has advanced in the development of fair methods for mitigating machine learning, with attempts depending on the respective domains of AI . (Mehrabi et al., 2021).

Pagano et al. (2023); Meher and Panda (2021); Carey and Wu (2023); Agarwal and Mishra (2021) present a categorization of bias elimination methods into three groups: *pre-processing*, *in-processing*, and *post-processing*.

**Pre-processing** techniques aim to eliminate biases through the manipulation of the dataset d'Alessandro et al. (2017), which can be performed through various treatment methods: filtering biased data, standardizing missing values, changing the dataset's proportions after replacing NaN values, removing sensitive data, removing proxy variables, generating synthetic data, reweighting the dataset, fair representation learning, and relabeling (Calders et al., 2009; Kearns and Roth, 2019; He et al., 2008; **?**; Feldman et al., 2015; Yang et al., 2020; Bryant et al., 2019; Zemel et al., 2013; Paviglianiti and Pasero, 2020; Pedreshi et al., 2008; Jang et al., 2021; Georgopoulos et al., 2021).

**In-processing** techniques aim to adapt the algorithm during model training, either by incorporating fairness constraints and criteria or by implementing changes in the objective function (d'Alessandro et al., 2017; Berk et al., 2017): reweighting, neutralizing the sensitive attribute, fairness constraints, adversarial debiasing, fairness regularization, Bayesian fairness, and fair ensemble learning. (Ashokan and Haas, 2021; Neda et al., 2021; Jain et al., 2019; Radovanovic et al., 2020; Reddy et al., 2021; Du et al., 2021);

**Post-processing** techniques involve a validation step, using data not present during the model's training (d'Alessandro et al., 2017). This is very common in the analysis of algorithms known as "black boxes," which, because their mathematical complexity, are difficult for practitioners to understand well (Loyola-González, 2019). In this group, we have the following mitigation techniques: threshold adjustment, transformation, and reject option.classification (Das et al., 2019; Hardt et al., 2016; Baumann et al., 2023; Wei et al., 2020; Kamiran et al., 2018).

# 6 Conclusion

In this research, we address a recurring problem for Machine Learning practitioners, both implementers and users, concerning the association between metrics and biases aimed at achieving fairness objectives. Well-developed algorithms can perform excellent prediction and classification tasks, but their use in an inappropriate context can have severe impacts on people's lives. This concern motivated the study of the possible types of biases existing in the production chain of ML tools.

Given the variety of bias sources, some beyond the control of the developer or user, ML practitioners must choose, based on the real-world context, which fairness objectives to achieve, knowing that (so far) there is no way to satisfy all aspects of fairness without compromising model accuracy.

The statistical analysis of metrics aims to provide a safe ground for developers to understand the ethical and legal fairness criteria that their algorithms are producing, while also allowing users and social scientists to comprehend the impossibility of implementing certain metrics due to the limitations imposed by context.

Finally, during the review of metrics found in the literature, we noted the absence of methodologically organized works containing a broad economic analysis of Machine Learning tools. Some studies analyze economic contexts, but they are restricted to specific problems. Therefore, it is suggested that this work be expanded by incorporating fairness criteria through the lens of economic science.

## References


T. Adel, I. Valera, Z. Ghahramani, and A. Weller. One-network adversarial fairness. Proceedings of the AAAI Conference on Artificial Intelligence, 33:2412–2420, 07 2019. doi: 10.1609/aaai.v33i01.33012412.

S. Agarwal and S. Mishra. Responsible AI Implementing Ethical and Unbiased Algorithms. London: Springer, 2021.

J. J. Amend and S. Spurlock. Improving machine learning fairness with sampling and adversarial learning. Journal of Computing Sciences in Colleges, 36(5):14–23, 2021.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. ProPublica, 23:77–91, 2016.

A. Ashokan and C. Haas. Fairness metrics and bias mitigation strategies for rating predictions. Information Processing Management, 58:102646, 09 2021. doi: 10.1016/j.ipm.2021.102646.

R. Baeza-Yates. Bias on the web. Communications of the ACM, 61:54–61, 05 2018. doi: 10.1145/3209581.

J. Balkin and R. Siegel. The american civil rights tradition: Anticlassification or antisubordination. Issues in Legal Scholarship, 2, 01 2003. doi: 10.2202/1539-8323.1039.

S. Barbosa, D. Cosley, A. Sharma, and R. M. Cesar. Averaging gone wrong: Using time-aware analyses to better understand behavior. In Proceedings of the 25th International Conference on World Wide Web, volume 8 of WWW '16, page 829–841. International World Wide Web Conferences Steering Committee, Apr. 2016. doi: 10.1145/2872427.2883083. URL http://dx.doi.org/10.1145/2872427.2883083.

S. Barocas and A. D. Selbst. Big data's disparate impact. Calif. L. Rev., 104:671, 2016.

J. Baumann, A. Hannák, and C. Heitz. Fair machine learning through post-processing: The case of predictive parity. In EWAF, 2023.

R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art, 2017. URL https://arxiv.org/abs/1703.09207.

R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in Criminal Justice Risk Assessments: The State of the Art. Sociological Methods & Research, 50(1):3–44, February 2021. doi: 10.1177/0049124118782533. URL https://ideas.repec.org/a/sae/somere/v50y2021i1p3-44.html.

C. R. Blyth. On simpson's paradox and the sure-thing principle. Journal of the American Statistical Association, 67(338):364–366, 1972. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2284382.

B. M. Booth, L. Hickman, S. K. Subburaj, L. Tay, S. E. Woo, and S. K. D'Mello. Integrating psychometrics and computing perspectives on bias and fairness in affective computing: A case study of automated video interviews. IEEE Signal Processing Magazine, 38(6):84–95, Nov. 2021. ISSN 1558-0792. doi: 10.1109/msp.2021.3106615. URL http://dx.doi.org/10.1109/MSP.2021.3106615.

R. Bryant, C. Cintas, I. Wambugu, A. Kinai, and K. Weldemariam. Analyzing bias in sensitive personal information used to train financial models. 11 2019. doi: 10.48550/arXiv.1911.03623.

T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In 2009 IEEE international conference on data mining workshops, pages 13–18. IEEE, 2009.

A. N. Carey and X. Wu. The statistical fairness field guide: perspectives from social and formal sciences. AI and Ethics, 3(1):1–23, 2023.

S. Chiappa and W. S. Isaac. A Causal Bayesian Networks Viewpoint on Fairness, page 3–20. Springer International Publishing, 2019. ISBN 9783030167448. doi: 10.1007/978-3-030-16744-8_1. URL http://dx.doi.org/10.1007/978-3-030-16744-8_1.

A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, 2016. URL https://arxiv.org/abs/1610.07524.

G. L. Ciampaglia, A. Nematzadeh, F. Menczer, and A. Flammini. How algorithmic popularity bias hinders or promotes quality. Scientific Reports, 8(1), Oct. 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-34203-2. URL http://dx.doi.org/10.1038/s41598-018-34203-2.

K. A. Clarke. The phantom menace: Omitted variable bias in econometric research. Conflict Management and Peace Science, 22(4):341–352, 2005. ISSN 07388942, 15499219. URL http://www.jstor.org/stable/26273559.

R. Colker. Anti-subordination above all : Sex , race , and equal protection. 1986. URL https://api.semanticscholar.org/CorpusID:6605248.

S. Corbett-Davies, J. D. Gaebler, H. Nilforoshan, R. Shroff, and S. Goel. The measure and mismeasure of fairness. The Journal of Machine Learning Research, 24(1):14730–14846, 2023.

B. d'Alessandro, C. O'Neil, and T. LaGatta. Conscientious classification: A data scientist's guide to discrimination-aware classification. Big Data, 5:120–134, 06 2017. doi: 10.1089/big.2016.0048.

D. Danks and A. London. Algorithmic bias in autonomous systems. pages 4691–4697, 08 2017. doi: 10.24963/ijcai.2017/654.

A. Das, S. Anjum, and D. Gurari. Dataset bias: A case study for visual question answering. Proceedings of the Association for Information Science and Technology, 56:58–67, 10 2019. doi: 10.1002/pra2.7.

T. Di Noia, N. Tintarev, P. Fatourou, and M. Schedl. Recommender systems under european ai regulations. Communications of the ACM, 65:69–73, 04 2022. doi: 10.1145/3512728.

J. Dressel and H. Farid. The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4:eaao5580, 01 2018. doi: 10.1126/sciadv.aao5580.

M. Du, S. Mukherjee, G. Wang, R. Tang, A. Hassan, and X. Hu. Fairness via representation neutralization, 06 2021.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference, pages 214–226, 2012.

M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pages 259–268, 2015.

B. Friedman and H. Nissenbaum. Bias in computer systems. ACM Trans. Inf. Syst., 14(3): 330–347, July 1996. ISSN 1046-8188. doi: 10.1145/230538.230561. URL https://doi.org/10.1145/230538.230561.

C. E. Gehlke and K. Biehl. Certain effects of grouping upon the size of the correlation coefficient in census tract material. Journal of the American Statistical Association, 29(185A):169–170, 1934. doi: 10.1080/01621459.1934.10506247. URL https://doi.org/10.1080/01621459.1934.10506247.

M. Georgopoulos, J. Oldfield, M. A. Nicolaou, Y. Panagakis, and M. Pantic. Mitigating demographic bias in facial datasets with style-based multi-attribute transfer. Int. J. Comput. Vision, 129(7):2288–2307, July 2021. ISSN 0920-5691. doi: 10.1007/s11263-021-01448-w. URL https://doi.org/10.1007/s11263-021-01448-w.

V. Grari, B. Ruf, S. Lamprier, and M. Detyniecki. Fair adversarial gradient tree boosting, 2019. URL https://arxiv.org/abs/1911.05369.

T. Green. Discrimination Laundering: The Rise of Organizational Innocence and the Crisis of Equal Opportunity Law. 01 2016. ISBN 9781107142008. doi: 10.1017/9781316494158.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. Advances in neural information processing systems, 29, 2016.

H. He, Y. Bai, E. A. Garcia, and S. Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), pages 1322–1328. Ieee, 2008.

H. Heidari, M. Loi, K. P. Gummadi, and A. Krause. A moral framework for understanding of fair ml through economic models of equality of opportunity, 2018. URL https://arxiv.org/abs/1809.03400.

B. Jain, M. Huber, L. Fegaras, and R. Elmasri. Singular race models: addressing bias and accuracy in predicting prisoner recidivism. pages 599–607, 06 2019. ISBN 978-1-4503-6232-0. doi: 10.1145/3316782.3322787.

A. Jalal, S. Karmalkar, J. Hoffmann, A. Dimakis, and E. Price. Fairness for image generation with uncertain sensitive attributes. In International Conference on Machine Learning, pages 4721–4732. PMLR, 2021.

T. Jang, F. Zheng, and X. Wang. Constructing a fair classifier with generated fair data. Proceedings of the AAAI Conference on Artificial Intelligence, 35:7908–7916, 05 2021. doi: 10.1609/aaai.v35i9.16965.

D. Jones, C. Snider, A. Nassehi, J. Yon, and B. Hicks. Characterising the digital twin: A systematic literature review. CIRP Journal of Manufacturing Science and Technology, 29: 36–52, 2020. ISSN 1755-5817. doi: https://doi.org/10.1016/j.cirpj.2020.02.002. URL https://www.sciencedirect.com/science/article/pii/S1755581720300110.

F. Kamiran, S. Mansha, A. Karim, and X. Zhang. Exploiting reject option in classification for social discrimination control. Information Sciences, 425:18–33, 2018.

M. Kearns and A. Roth. The ethical algorithm: The science of socially aware algorithm design. Oxford University Press, 2019.

F. A. Khan, E. Manis, and J. Stoyanovich. Fairness as equality of opportunity: Normative guidance from political philosophy, 2021. URL https://arxiv.org/abs/2106.08259.

J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807, 2016.

N. Kordzadeh and M. Ghasemaghaei. Algorithmic bias: review, synthesis, and future research directions. European Journal of Information Systems, 31:1–22, 06 2021. doi: 10.1080/0960085X.2021.1927212.

D. Kowald and E. Lex. Studying confirmation bias in hashtag usage on twitter, 2018. URL https://arxiv.org/abs/1809.03203.

M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva. Counterfactual fairness, 2018. URL https://arxiv.org/abs/1703.06856.

S. Li, J. Yu, X. Du, Y. Lu, and R. Qiu. Fair outlier detection based on adversarial representation learning. Symmetry, 14:347, 02 2022. doi: 10.3390/sym14020347.

Z. C. Lipton, A. Chouldechova, and J. McAuley. Does mitigating ml's impact disparity require treatment disparity?, 2019. URL https://arxiv.org/abs/1711.07076.

O. Loyola-González. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. IEEE Access, 7:154096–154113, 2019. doi: 10.1109/ACCESS.2019.2949286.

J. McCarthy et al. What is artificial intelligence. 2007.

S. K. Meher and G. Panda. Deep learning in astronomy: a tutorial perspective. The European Physical Journal Special Topics, 230(10):2285–2317, 2021.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6):1–35, 2021.

H. Miller, J. Thebault-Spieker, S. Chang, I. Johnson, L. Terveen, and B. Hecht. "blissfully happy" or "ready tofight": Varying interpretations of emoji. Proceedings of the International AAAI Conference on Web and Social Media, 10(1):259–268, Aug. 2021. doi: 10.1609/icwsm.v10i1.14757. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14757.

D. B. Mustard. Reexamining criminal behavior: The importance of omitted variable bias. The Review of Economics and Statistics, 85(1):205–211, 2003. ISSN 00346535, 15309142. URL http://www.jstor.org/stable/3211634.

B. Neda, Y. Zeng, and S. Gago-Masague. Using machine learning in admissions: Reducing human and algorithmic bias in the selection process. pages 1323–1323, 03 2021. doi: 10.1145/3408877.3439664.

D.-P. Nguyen, R. Gravel, R. Trieschnigg, and T. Meder. How old do you think i am?": A study of language and age in twitter. In Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, ICWSM 2013, pages 439–448, United States, July 2013. AAAI. ISBN 978-1-57735-610-3. URL http://www.icwsm.org/2013/. eemcs-eprint-23604 ; 7th International AAAI Conference on Weblogs and Social Media, ICWSM 2013, ICWSM ; Conference date: 08-07-2013 Through 10-07-2013.

A. Nielsen. Practical Fairness: Achieving Fair and Secure Data Models. O'Reilly Media, Incorporated, 2020. ISBN 9781492075738. URL https://books.google.com.br/books?id=palfzQEACAAJ.

A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. Frontiers in big data, 2:13, 2019.

B. Paaßen, A. Bunge, C. Hainke, L. Sindelar, and M. Vogelsang. Dynamic fairness - breaking vicious cycles in automatic decision making, 2019. URL https://arxiv.org/abs/1902.00375.

T. P. Pagano, R. B. Loureiro, F. V. Lisboa, R. M. Peixoto, G. A. Guimarães, G. O. Cruz, M. M. Araujo, L. L. Santos, M. A. Cruz, E. L. Oliveira, et al. Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. Big data and cognitive computing, 7(1):15, 2023.

A. Paviglianiti and E. Pasero. Vital-ecg: a de-bias algorithm embedded in a gender-immune device. In 2020 IEEE International Workshop on Metrology for Industry 4.0 IoT, pages 314–318, 2020. doi: 10.1109/MetroInd4.0IoT48571.2020.9138291.

D. Pedreshi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 560–568, 2008.

D. Pessach and E. Shmueli. Algorithmic fairness, 2020. URL https://arxiv.org/abs/2001.09784.

N. Quadrianto and V. Sharmanska. Recycling privileged learning and distribution matching for fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.

S. Radovanovic, A. Petrovic, B. Delibašić, and M. Suknovic. Enforcing fairness in logistic regression algorithm. pages 1–7, 08 2020. doi: 10.1109/INISTA49547.2020.9194676.

C. Reddy, D. Sharma, S. Mehri, A. Romero Soriano, S. Shabanian, and S. Honari. Benchmarking bias mitigation algorithms in representation learning through fairness metrics. In J. Vanschoren and S. Yeung, editors, Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/2723d092b63885e0d7c260cc007e8b9d-Paper-round1.pdf.

S. Riegg. Causal inference and omitted variable bias in financial aid research: Assessing solutions. The Review of Higher Education, 31:329–354, 03 2008. doi: 10.1353/rhe.2008.0010.

S. Segall. Equality and Opportunity. Oxford University Press, Oxford, GB, 2013.

H. Suresh and J. V. Guttag. A framework for understanding unintended consequences of machine learning. arXiv preprint arXiv:1901.10002, 2(8):73, 2019.

S. Verma and J. Rubin. Fairness definitions explained. In Proceedings of the International Workshop on Software Fairness, FairWare '18, page 1–7, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450357463. doi: 10.1145/3194770.3194776. URL https://doi.org/10.1145/3194770.3194776.

D. Wei, K. N. Ramamurthy, and F. P. Calmon. Optimized score transformation for fair classification. Proceedings of Machine Learning Research, 108, 2020.

A. Xiang and I. D. Raji. On the legal compatibility of fairness definitions, 2019. URL https://arxiv.org/abs/1912.00761.

P. Xiao. Artificial Intelligence Programming with Python: From Zero to Hero. Wiley, 2022. ISBN 9781119820963. URL https://books.google.com.br/books?id=bKBgEAAAQBAJ.

K. Yang, B. Huang, J. Stoyanovich, and S. Schelter. Fairness-aware instrumentation of preprocessing pipelines for machine learning. Workshop on Human-In-the-Loop Data Analytics (HILDA'20), 2020. doi: 10.1145/3398730.3399194. URL https://par.nsf.gov/biblio/10182459.

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In International conference on machine learning, pages 325–333. PMLR, 2013.