developers.google.com

Machine Learning Glossary: Fairness | Google Developers

17-21 minuten

This page contains Fairness glossary terms. For all glossary terms, click here.

A

attribute

Synonym for **feature**. In fairness, attributes often refer to characteristics pertaining to individuals.

automation bias

When a human decision maker favors recommendations made by an automated decision-making system over information made without automation, even when the automated decision-making system makes errors.

B

bias (ethics/fairness)

1. Stereotyping, prejudice or favoritism towards some things, people, or groups over others. These biases can affect collection and interpretation of data, the design of a system, and how users interact with a system. Forms of this type of bias include:

- **automation bias**
- **confirmation bias**
- **experimenter's bias**
- **group attribution bias**
- **implicit bias**
- **in-group bias**
- **out-group homogeneity bias**

2. Systematic error introduced by a sampling or reporting procedure. Forms of this type of bias include:

- **coverage bias**
- **non-response bias**
- **participation bias**
- **reporting bias**
- **sampling bias**

- **selection bias**

Not to be confused with the bias term in machine learning models or **prediction bias**.

C

confirmation bias

The tendency to search for, interpret, favor, and recall information in a way that confirms one's preexisting beliefs or hypotheses. Machine learning developers may inadvertently collect or label data in ways that influence an outcome supporting their existing beliefs. Confirmation bias is a form of **implicit bias**.

Experimenter's bias is a form of confirmation bias in which an experimenter continues training models until a preexisting hypothesis is confirmed.

counterfactual fairness

A **fairness metric** that checks whether a classifier produces the same result for one individual as it does for another individual who is identical to the first, except with respect to one or more **sensitive attributes**. Evaluating a classifier for counterfactual fairness is one method for surfacing potential sources of bias in a model.

See "When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness" for a more detailed discussion of counterfactual fairness.

coverage bias

See **selection bias**.

D

demographic parity

A **fairness metric** that is satisfied if the results of a model's classification are not dependent on a given **sensitive attribute**.

For example, if both Lilliputians and Brobdingnagians apply to Glubbdubdrib University, demographic parity is achieved if the percentage of Lilliputians admitted is the same as the percentage of Brobdingnagians admitted, irrespective of whether one group is on average more qualified than the other.

Contrast with **equalized odds** and **equality of opportunity**, which permit classification results in aggregate to depend on sensitive attributes, but do not permit classification results for certain specified ground-truth labels to depend on sensitive attributes. See "Attacking discrimination with smarter machine learning" for a visualization exploring the tradeoffs when optimizing for demographic parity.

disparate impact

Making decisions about people that impact different population subgroups disproportionately. This usually refers to situations where an algorithmic decision-making process harms or benefits some subgroups more than others.

For example, suppose an algorithm that determines a Lilliputian's eligibility for a miniature-home loan is more likely to classify them as “ineligible” if their mailing address contains a certain postal code. If Big-Endian Lilliputians are more likely to have mailing addresses with this postal code than Little-Endian Lilliputians, then this algorithm may result in disparate impact.

Contrast with **disparate treatment**, which focuses on disparities that result when subgroup characteristics are explicit inputs to an algorithmic decision-making process.

disparate treatment

Factoring subjects' **sensitive attributes** into an algorithmic decision-making process such that different subgroups of people are treated differently.

For example, consider an algorithm that determines Lilliputians’ eligibility for a miniature-home loan based on the data they provide in their loan application. If the algorithm uses a Lilliputian’s affiliation as Big-Endian or Little-Endian as an input, it is enacting disparate treatment along that dimension.

Contrast with **disparate impact**, which focuses on disparities in the societal impacts of algorithmic decisions on subgroups, irrespective of whether those subgroups are inputs to the model.

E

equality of opportunity

A **fairness metric** that checks whether, for a preferred **label** (one that confers an advantage or benefit to a person) and a given **attribute**, a classifier predicts that preferred label equally well for all values of that attribute. In other words, equality of opportunity measures whether the people who should qualify for an opportunity are equally likely to do so regardless of their group membership.

For example, suppose Glubbdubdrib University admits both Lilliputians and Brobdingnagians to a rigorous mathematics program. Lilliputians’ secondary schools offer a robust curriculum of math classes, and the vast majority of students are qualified for the university program. Brobdingnagians’ secondary schools don’t offer math classes at all, and as a result, far fewer of their students are qualified. Equality of opportunity is satisfied for the preferred label of "admitted" with respect to nationality (Lilliputian or Brobdingnagian) if qualified students are equally likely to be admitted irrespective of whether they're a Lilliputian or a

Brobdingnagian.

For example, let's say 100 Lilliputians and 100 Brobdingnagians apply to Glubbdubdrib University, and admissions decisions are made as follows:

Table 1. Lilliputian applicants (90% are qualified)

	Qualified	**Unqualified**
Admitted	45	3
Rejected	45	7
Total	90	10
Percentage of qualified students admitted: 45/90 = 50% Percentage of unqualified students rejected: 7/10 = 70% Total percentage of Lilliputian students admitted: (45+3)/100 = 48%		

Table 2. Brobdingnagian applicants (10% are qualified):

	Qualified	**Unqualified**
Admitted	5	9
Rejected	5	81
Total	10	90
Percentage of qualified students admitted: 5/10 = 50% Percentage of unqualified students rejected: 81/90 = 90% Total percentage of Brobdingnagian students admitted: (5+9)/100 = 14%		

The preceding examples satisfy equality of opportunity for acceptance of qualified students because qualified Lilliputians and Brobdingnagians both have a 50% chance of being admitted.

See "Equality of Opportunity in Supervised Learning" for a more detailed discussion of equality of opportunity. Also see "Attacking discrimination with smarter machine learning" for a visualization exploring the tradeoffs when optimizing for equality of opportunity.

equalized odds

A **fairness metric** that checks if, for any particular label and attribute, a classifier predicts that label equally well for all values of that attribute.

For example, suppose Glubbdubdrib University admits both Lilliputians and Brobdingnagians to a rigorous mathematics program. Lilliputians' secondary schools offer a robust curriculum of math classes, and the vast majority of students are qualified for the university program. Brobdingnagians' secondary schools don’t offer math classes at all, and as a result, far fewer of their students are qualified. Equalized odds is satisfied provided that no matter whether an applicant is a Lilliputian or a Brobdingnagian, if they are qualified, they are equally as likely to get admitted to the program, and if they are not qualified, they are equally as likely to get rejected.

Let’s say 100 Lilliputians and 100 Brobdingnagians apply to Glubbdubdrib University, and admissions decisions are made as follows:

Table 3. Lilliputian applicants (90% are qualified)

	Qualified	Unqualified
Admitted	45	2
Rejected	45	8
Total	90	10
Percentage of qualified students admitted: 45/90 = 50% Percentage of unqualified students rejected: 8/10 = 80% Total percentage of Lilliputian students admitted: (45+2)/100 = 47%		

Table 4. Brobdingnagian applicants (10% are qualified):

	Qualified	Unqualified
Admitted	5	18
Rejected	5	72
Total	10	90
Percentage of qualified students admitted: 5/10 = 50% Percentage of unqualified students rejected: 72/90 = 80% Total percentage of Brobdingnagian students admitted: (5+18)/100 = 23%		

Equalized odds is satisfied because qualified Lilliputian and Brobdingnagian students both have a 50% chance of being admitted, and unqualified Lilliputian and Brobdingnagian have an 80% chance of being rejected.

Equalized odds is formally defined in "Equality of Opportunity in Supervised Learning" as follows: "predictor $\hat{Y}$ satisfies equalized odds with respect to protected attribute A and outcome Y if $\hat{Y}$ and A are independent, conditional on Y."

experimenter's bias

See **confirmation bias**.

F

fairness constraint

Applying a constraint to an algorithm to ensure one or more definitions of fairness are satisfied. Examples of fairness constraints include:

- **Post-processing** your model's output.

- Altering the **loss function** to incorporate a penalty for violating a **fairness metric**.
- Directly adding a mathematical constraint to an optimization problem.

fairness metric

A mathematical definition of “fairness” that is measurable. Some commonly used fairness metrics include:

- **equalized odds**
- **predictive parity**
- **counterfactual fairness**
- **demographic parity**

Many fairness metrics are mutually exclusive; see **incompatibility of fairness metrics**.

G

group attribution bias

Assuming that what is true for an individual is also true for everyone in that group. The effects of group attribution bias can be exacerbated if a **convenience sampling** is used for data collection. In a non-representative sample, attributions may be made that do not reflect reality.

See also **out-group homogeneity bias** and **in-group bias**.

I

implicit bias

Automatically making an association or assumption based on one’s mental models and memories. Implicit bias can affect the following:

- How data is collected and classified.
- How machine learning systems are designed and developed.

For example, when building a classifier to identify wedding photos, an engineer may use the presence of a white dress in a photo as a feature. However, white dresses have been customary only during certain eras and in certain cultures.

See also **confirmation bias**.

incompatibility of fairness metrics

The idea that some notions of fairness are mutually incompatible and cannot be satisfied simultaneously. As a result, there is no single universal **metric** for quantifying fairness that can be applied to all ML problems.

While this may seem discouraging, incompatibility of fairness metrics doesn’t imply that fairness efforts are fruitless. Instead, it suggests that fairness must be defined contextually for a given ML problem, with the goal of preventing harms specific to its use cases.

See "On the (im)possibility of fairness" for a more detailed discussion of this topic.

individual fairness

A fairness metric that checks whether similar individuals are classified similarly. For example, Brobdingnagian Academy might want to satisfy individual fairness by ensuring that two students with identical grades and standardized test scores are equally likely to gain admission.

Note that individual fairness relies entirely on how you define "similarity" (in this case, grades and test scores), and you can run the risk of introducing new fairness problems if your similarity metric misses important information (such as the rigor of a student’s curriculum).

See "Fairness Through Awareness" for a more detailed discussion of individual fairness.

in-group bias

Showing partiality to one's own group or own characteristics. If testers or raters consist of the machine learning developer's friends, family, or colleagues, then in-group bias may invalidate product testing or the dataset.

In-group bias is a form of **group attribution bias**. See also **out-group homogeneity bias**.

N

non-response bias

See **selection bias**.

O

out-group homogeneity bias

The tendency to see out-group members as more alike than in-group members when comparing attitudes, values, personality traits, and other characteristics. **In-group** refers to people you interact with regularly; **out-group** refers to people you do not interact with regularly. If you create a dataset by asking people to provide attributes about out-groups, those attributes may be less nuanced and more stereotyped than attributes that participants list for people in their in-group.

For example, Lilliputians might describe the houses of other Lilliputians in great detail, citing small differences in architectural styles, windows, doors, and sizes. However, the same Lilliputians might simply declare that Brobdingnagians all live in identical houses.

Out-group homogeneity bias is a form of **group attribution bias**.

See also **in-group bias**.

P

participation bias

Synonym for non-response bias. See **selection bias**.

post-processing

Processing the output of a model *after* the model has been run. Post-processing can be used to enforce fairness constraints without modifying models themselves.

For example, one might apply post-processing to a binary classifier by setting a classification threshold such that **equality of opportunity** is maintained for some attribute by checking that the **true positive rate** is the same for all values of that attribute.

predictive parity

A **fairness metric** that checks whether, for a given classifier, the **precision** rates are equivalent for subgroups under consideration.

For example, a model that predicts college acceptance would satisfy predictive parity for nationality if its precision rate is the same for Lilliputians and Brobdingnagians.

Predictive parity is sometime also called *predictive rate parity*.

See "Fairness Definitions Explained" (section 3.2.1) for a more detailed discussion of predictive parity.

predictive rate parity

Another name for **predictive parity**.

preprocessing

Processing data before it's used to train a model. Preprocessing could be as simple as removing words from an English text corpus that don't occur in the English dictionary, or could be as complex as re-expressing data points in a way that eliminates as many attributes that are correlated with **sensitive attributes** as possible. Preprocessing can help satisfy **fairness constraints**.

proxy (sensitive attributes)

An attribute used as a stand-in for a **sensitive attribute**. For example, an individual's postal code might be used as a proxy for their income, race, or ethnicity.

R

reporting bias

The fact that the frequency with which people write about actions, outcomes, or properties is not a reflection of their real-world frequencies or the degree to which a property is characteristic of a class of individuals. Reporting bias can influence the composition of data that machine learning systems learn from.

For example, in books, the word *laughed* is more prevalent than *breathed*. A machine learning model that estimates the relative frequency of laughing and breathing from a book corpus would probably determine that laughing is more common than breathing.

S

sampling bias

See **selection bias**.

selection bias

Errors in conclusions drawn from sampled data due to a selection process that generates systematic differences between samples observed in the data and those not observed. The following forms of selection bias exist:

- **coverage bias**: The population represented in the dataset does not match the population that the machine learning model is making predictions about.
- **sampling bias**: Data is not collected randomly from the target group.
- **non-response bias** (also called **participation bias**): Users from certain groups opt-out of surveys at different rates than users from other groups.

For example, suppose you are creating a machine learning model that predicts people's enjoyment of a movie. To collect training data, you hand out a survey to everyone in the front row of a theater showing the movie. Offhand, this may sound like a reasonable way to gather a dataset; however, this form of data collection may introduce the following forms of selection bias:

- coverage bias: By sampling from a population who chose to see the movie, your model's predictions may not generalize to people who did not already express that level of interest in the movie.
- sampling bias: Rather than randomly sampling from the intended population (all the people at the movie), you sampled only the people in the front row. It is possible that the people sitting in the front row were more interested in the movie than those in other rows.
- non-response bias: In general, people with strong opinions tend to respond to optional surveys more frequently than people with mild opinions. Since the movie survey is optional, the responses are

more likely to form a bimodal distribution than a normal (bell-shaped) distribution.

sensitive attribute

A human attribute that may be given special consideration for legal, ethical, social, or personal reasons.

U

unawareness (to a sensitive attribute)

A situation in which **sensitive attributes** are present, but not included in the training data. Because sensitive attributes are often correlated with other attributes of one's data, a model trained with unawareness about a sensitive attribute could still have **disparate impact** with respect to that attribute, or violate other **fairness constraints**.