3 HOW TO MEASURE FAIRNESS?

Fairness means treating different groups equally. But how can we implement this fluid concept in AI contexts? On the following pages, you'll learn about fairness metrics – approaches to make fairness measurable to act upon.

SENSITIVE ATTRIBUTES

Who are the "different groups" that should be treated equally? The definition of those groups relies on **sensitive attributes, binary divisions based on privilege**:

Privileged

e.g. people read as male, *white*, non-queer

e.g. people read as non-male, non-*white*, queer

Non-privileged

The test data can be used to **see if the AI model behaves differently for different groups**. To do so, **fairness metrics** are used. Many different fairness metrics exist and each one is a specific definition of fairness. On the following pages, there are some examples.

USING FAIRNESS METRICS IS NOT ENOUGH

Feels fair? Each of the metrics may sound fair at first, but achieving fairness is not so easy. The metrics are meant to provide information for action. Measuring fairness is not enough. Moreover, they are all based on binary divisions, contradict each other and are not uncontested! Read about the metrics' critique on the poster page of the next zine.
What else can Techie do to build fair(er) AI? Explore fairer processes on the next zine pages!

Deep dive on the back side!

INDIVIDUAL FAIRNESS METRICS

FAIRNESS THROUGH AWARENESS

Fairness means that **similar people receive similar classifications regardless of their group.**

That means only similar individuals can be compared, not all individuals. Who even gets to decide what similarity means?

FAIRNESS THROUGH UNAWARENESS

Fairness means that **sensitive attributes are not explicitly used** in the decision-making process.

Often, there are ways to predict sensitive attributes, so complete unawareness is impossible. Is unawareness even fair?

GROUP FAIRNESS METRICS

DEMOGRAPHIC PARITY

Fairness means that **both groups have the same likelihood of being accepted.**

What about qualified but less confident groups? There would be less applications from them, but the ratio of qualified ones could be higher, and they would be dismissed...

TREATMENT EQUALITY

Fairness means that **errors by the AI system impact both groups similarly.**

This still relies on an unchecked ground truth. Even if the errors are the same, the hiring rates could be different – and unfair!

MEASURE FAIRNESS?

Fairness metrics work by comparing predictions for individuals or groups. But how are they compared? There are two approaches:

Individual fairness metrics

...rely on comparing the predictions of individuals, for example the metrics fairness through awareness or unawareness (on the zine's pages). Here, the prediction for an individual is compared to the outcomes for another similar or different individual.

Group fairness metrics

...rely on comparing the predictions of the AI model with the actual labeled data for different groups, for example demographic parity or treatment equality (on the zine's pages). Comparing predictions per group is based on a confusion matrix - a table that gives an overview where and for which group the AI predicted correctly:

	AI predicted hire	AI predicted reject
Label by C.O.R.P.: hire	**TRUE POSITIVE** for Group A AND Group B	**FALSE NEGATIVE** for Group A AND Group B
Label by C.O.R.P.: reject	**FALSE POSITIVE** for Group A AND Group B	**TRUE NEGATIVE** for Group A AND Group B

Every fairness metric is a certain definition of fairness. Expressed mathematically, they could look like this:

The model is fair, if

$$\text{Group A: } \frac{\text{All Positives}}{\text{All}} \approx \frac{\text{All Positives}}{\text{All}} \text{ :Group B}$$

This translates into:

The ratio of predicted hires to the total number of applications is similar across groups. **40% female applicants – 40% of the hires should be female!**
It defines fairness as the same likelihood for both groups to be hired. It is called demographic parity - and you have already seen this definition on the zine's pages!

IDEA · REALISATION · I. Mastenbroek, I. Kühnlein, B. Lübbert DESIGN · I. Kühnlein, I. Mastenbroek, L. Wunderlich

NERD NOTES:

Many different fairness metrics exist. Their high number shows how hard finding one specific and overarching, applicable definition of fairness is – and this fuzziness is a strength of fairness as a concept.

We have shown one metric in detail here to explain how the metrics work. When comparing other metrics in more detail, it becomes clear that they also sometimes contradict each other or that one metric's results would describe an AI system as fair while another metric would not.

The fairness metrics are a selection of the fairness metrics presented in Mehrabi et al. (2022): A survey on bias and fairness in Machine Learning; Verma & Rubin (2018): Fairness definitions explained.