FEELS FAIR?

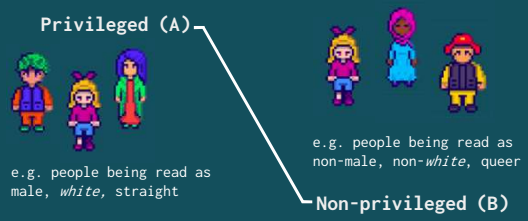
MEASURE FAIRNESS?

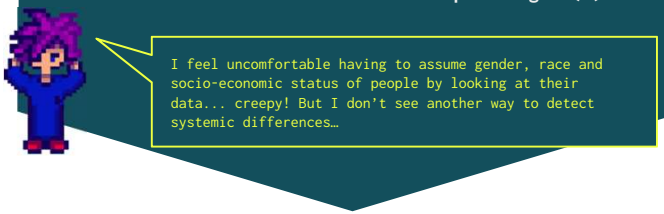
Biases and unfairness are everywhere - how can Techie ensure that their AI model is fair? Fairness is described as treating different groups equally. However, this can be interpreted differently.

Fairness is measured using fairness metrics that rely on binary partitioning. To measure fairness, Techie must decide which groups to compare. Techie does this by defining sensitive attributes and dividing the test data into two groups: privileged and non-privileged. Thus, Techie can determine if the AI model behaves differently for different groups.

SENSITIVE ATTRIBUTES

for predefined characteristics such as gender, race, and sexual orientation.





For each attribute, Techie creates a table for the privileged (A) and the non-privileged (B) groups. It shows, where the model predicted correctly and where not. By comparing these numbers, it becomes clear if the model is more accurate for privileged or non-privileged groups.

	AI predicted hire	AI predicted reject
Ground Truth said hire	TRUE POSITIVE for Group A AND Group B	FALSE NEGATIVE for Group A AND Group B
Ground Truth said reject	FALSE POSITIVE for Group A AND Group B	TRUE NEGATIVE for Group A AND Group B

There are many different fairness metrics - which one is the best? Here are some examples:

Fairness Through Awareness

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

That can only compare (similar) individuals, 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, complete unawareness is impossible. Is unawareness even fairness?

Demographic Parity

Is one group more likely to be hired?

The model is fair, if

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

Fairness means that both groups have the same likelihood to be accepted. **40% female applicants – 40% of the hires should be female**"

Applicants from a non-privileged group that are more competent but less confident might not apply so often. This smaller group of applicants would then contain more qualified candidates - and they would be dismissed!

Treatment Equality

Are errors impacting one group more?

The model is fair, if

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

The ratio of the falsely rejected to the falsely hired individuals is roughly the same. Fairness means that e**rrors of the AI system impact both groups similarly.**

This still relies on an unchecked Ground Truth. Even if the error rates are the same, the hiring rates could be very different - and unfair!

All fairness metrics are based on binary categories and complicate accounting for the nuances of life. There are **two lines of criticism against fairness metrics:**

Accuracy focused approach:
Mitigating against bias makes models **less accurate.**

Accuracy Fairness

Reducing existing bias means changing data sets. These changes could lead to more fairness, but also to less accuracy. There needs to be a process of weighing up which criteria to aim for in AI models.

Intersectional approach:
Using metrics to measure fairness is **too one-dimensional.**

Fairness evades metrics

Class, gender or race are not binary, they intersect and overlap. Fairness – treating different groups equally – is not always the right way. When it comes to justice, approaches such as equity – providing equal opportunities to participate – are crucial. More important than metrics is recognition: Do the people affected have a say in the process?

Not convinced!

Fairness means that we **shift power and share processes.** Quantifying fairness is neither possible nor helpful.

Aren't flawed ways to measure fairness better than none? At least we have a discussion!

What do you think and feel is fair? Which one of the approaches convinced you?

Vote with a sticker!

INDIVIDUAL FAIRNESS METRICS

GROUP FAIRNESS METRICS

CRITICISM