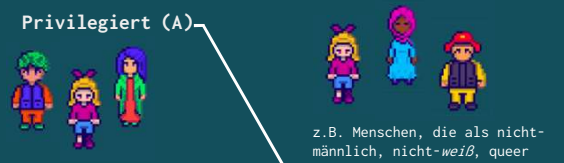
FEELS FAIR?

FAIRNESS MESSEN?

Biases und Unfairness sind überall - wie kann Techie sicherstellen, dass deren KI-Modell fair ist? Fairness wird als Gleichbehandlung verschiedener Gruppen beschrieben. Das kann jedoch unterschiedlich interpretiert werden.

Fairness wird mit Fairness-Metriken gemessen, die auf einer binären Einteilung beruhen. Um Fairness zu messen, muss Techie entscheiden, welche Gruppen verglichen werden sollen. Techie tut dies, indem sensible Merkmale (sensitive attributes) definiert und die Testdaten in zwei Gruppen aufgeteilt werden: privilegiert und nicht privilegiert. So kann Techie feststellen, ob sich das KI-Modell für verschiedene Gruppen unterschiedlich verhält.

Sensible Merkmale
für vordefinierte Merkmale wie Geschlecht, Ethnie und sexuelle Orientierung.



Ich fühle mich unwohl, Geschlecht, Ethnie und sozio-ökonomischen Status von Menschen auf Basis ihrer Daten anzunehmen... gruselig! Aber ich sehe keine andere Möglichkeit, um systemische Ungerechtigkeit zu erkennen...

Fairness durch Awareness

Fairness bedeutet, dass **ähnliche Personen unabhängig von ihrer Gruppe ähnlich eingeordnet werden.**

Damit können nur ähnliche Individuen verglichen werden, nicht alle Individuen. Wer darf überhaupt entscheiden, was Ähnlichkeit bedeutet?

Fairness durch Unawarness

Fairness bedeutet, dass **sensible Merkmale bei der Entscheidungsfindung nicht explizit berücksichtigt werden.**

Oft gibt es Möglichkeiten sensible Merkmale vorherzusagen, völlige Unkenntnis ist unmöglich. Sind Entscheidungen basierend auf Unkenntnis überhaupt fair?

INDIVIDUENBEZOGENE FAIRNESS METRIKEN

Für jedes Attribut erstellt Techie eine Tabelle für die privilegierten (A) und die nicht-privilegierten (B) Gruppen. Sie zeigt, wo das Modell korrekt prognostiziert hat und wo nicht. Durch den Vergleich dieser Zahlen wird deutlich, ob das Modell für privilegierte oder nicht-privilegierte Gruppen genauer ist.

	KI prognostiziert: einstellen	KI prognostiziert: ablehnen
Label von C.O.R.P.: einstellen	WAHR POSITIV für Gruppe A UND Gruppe B	FALSCH NEGATIV für Gruppe A UND Gruppe B
Label von C.O.R.P.: ablehnen	FALSCH POSITIV für Gruppe A UND Gruppe B	WAHR NEGATIV für Gruppe A UND Gruppe B

Es gibt viele verschiedene Fairness Metriken - welche ist die beste? Hier sind einige Beispiele:

Demographic Parity

Hat eine Gruppe eine höhere Wahrscheinlichkeit, eingestellt zu werden?

Das KI-Modell ist fair, wenn:

$$\text{Gruppe A: } \frac{\text{Alle Positiven}}{\text{Alle}} \approx \text{Gruppe B: } \frac{\text{Alle Positiven}}{\text{Alle}}$$

Fairness bedeutet, dass beide Gruppen dieselbe Wahrscheinlichkeit haben, akzeptiert zu werden: **40% der Bewerber*innen sind weiblich – 40 % der neu Eingestellten sollten weiblich sein!**

Bewerber*innen aus einer nicht privilegierten Gruppe, die kompetent sind, aber weniger Selbstvertrauen haben, würden sich seltener bewerben. Diese Gruppe Bewerber*innen würde im Schnitt qualifiziertere Kandidat*innen enthalten - und diese würden abgelehnt!

Treatment Equality

Wirken sich Fehler stärker auf eine Gruppe aus?

Das KI-Modell ist fair, wenn:

$$\text{Gruppe A: } \frac{\text{Falsch Negativ}}{\text{Falsch Positiv}} \approx \text{Gruppe B: } \frac{\text{Falsch Negativ}}{\text{Falsch Positiv}}$$

Das Verhältnis zwischen den fälschlicherweise abgelehnten und den fälschlicherweise eingestellten Personen ist in etwa gleich groß. Fairness bedeutet, dass sich **Fehler des KI-Modells auf beide Gruppen in ähnlichem Maße auswirken.**

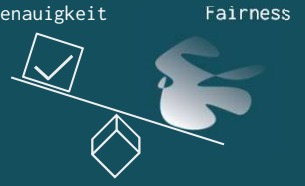
Dies beruht immer noch auf einer ungeprüften "Ground Truth". Selbst wenn die Fehlerquoten gleich sind, könnten die Einstellungsquoten sehr unterschiedlich - und ungerecht - sein!

GRUPPENBEZOGENE FAIRNESS METRIKEN

Fairness-Metriken basieren auf binären Kategorien und erschweren die Berücksichtigung der Nuancen verschiedener Lebensrealitäten. Es gibt **zwei Ansätze für Kritik an ihnen:**

Genauigkeitsbezogener Ansatz:

Der Versuch, Bias zu verringern, macht KI-Modelle **weniger genau.**



Bias zu reduzieren, bedeutet, Datensätze zu verändern. Diese Veränderungen könnten zu mehr Fairness, aber auch zu geringerer Genauigkeit führen. Es muss einen Abwägungsprozess geben, welche Kriterien angestrebt werden.

Intersektionaler Ansatz:

Die Verwendung von Metriken zur Messung von Fairness ist **zu eindimensional.**

Fairness entzieht sich Metriken

Klasse, Geschlecht oder Ethnie sind nicht binär, sie überschneiden und überlappen sich. Fairness – die Gleichbehandlung verschiedener Gruppen – ist nicht immer der richtige Ansatz. Wenn es um Gerechtigkeit geht, dann muss Gerechtigkeit im Sinne gleicher Zugangsmöglichkeiten (Equity) eine Rolle spielen. Wichtiger als Metriken ist Anerkennung: Haben die betroffenen Menschen ein Mitspracherecht in den Prozessen?

KRITIK

Nicht überzeugt!

Fairness bedeutet, dass wir **Machtverhältnisse verschieben und Prozesse teilen.** Sie zu quantifizieren ist weder möglich noch hilfreich.

Sind mangelhafte Ansätze zur Messung von Fairness nicht besser als gar keine? Zumindest haben wir eine Diskussion!

Was denkst du ist fair? Was empfindest du als fair? Welcher der Ansätze hat dich überzeugt?

Stimme mit einem Sticker ab!