INDIVIDUENBEZOGENE FAIRNESS METRIKEN

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

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

FAIRNESS DURCH AWARENESS

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

FAIRNESS DURCH UNAWARENESS

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

GRUPPENBEZOGENE FAIRNESS METRIKEN

Was ist mit qualifizierten, aber weniger selbstbewussten Gruppen? Es gäbe weniger Bewerbungen von ihnen, aber der Anteil der Qualifizierten könnte höher sein, und die würden abgelehnt ...

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

DEMOGRAPHIC PARITY

Fairness bedeutet, **dass beide Gruppen dieselbe Wahrscheinlichkeit haben, akzeptiert zu werden.**

TREATMENT EQUALITY

Fairness bedeutet, **dass sich Fehler des KI-Systems auf beide Gruppen in gleicher Weise auswirken.**

Deep dive auf der Rückseite!

FAIRNESS-METRIKEN SIND NICHT GENUG!

Fühlt sich das fair an? Jede der Metriken mag fair klingen, aber es ist nicht einfach, Fairness zu erreichen. Die Metriken liefern Informationen für Maßnahmen. Doch sie alle beruhen auf binären Unterscheidungen und widersprechen sich teilweise! Kritik an den Metriken findest du auf der Posterseite des nächsten Zines.
Was kann Techie sonst noch tun, um faire(re) KI zu entwickeln? Erforsche fairere Prozesse auf den Seiten des nächsten Zines!

3 WIE WIRD FAIRNESS GEMESSEN?

Unter Fairness versteht man die Gleichbehandlung verschiedener Gruppen. Doch wie lässt sich dieses fluide Konzept in KI-Kontexten umsetzen? Auf den folgenden Seiten erfahrt ihr mehr über Fairness-Metriken. Das sind Ansätze, um Fairness messbar zu machen.

SENSIBLE MERKMALE

Wenn es bei Fairness um die Gleichbehandlung verschiedener Gruppen geht – wer sind dann diese Gruppen? Bei der Definition davon werden **sensible Merkmale**, binäre Unterteilungen auf der Grundlage von Privilegien, hinzugezogen:

Privilegiert

z. B. Menschen, die als männlich, *weiß*, nicht-queer gelesen werden

z. B. Menschen, die als nicht-männlich, nicht-*weiß*, queer gelesen werden

Nicht-privilegiert

Mit Hilfe der Testdaten kann herausgefunden werden, **ob sich das KI-Modell für verschiedene Gruppen unterschiedlich verhält.** Dafür werden **Fairness-Metriken** verwendet. Davon gibt es viele verschiedene und jede einzelne ist eine spezifische Definition von Fairness. Auf den folgenden Seiten gibt es einige Beispiele.

FAIRNESS MESSEN?

Fairness-Metriken basieren auf dem Vergleich von Prognosen für Einzelpersonen oder Gruppen. Aber wie werden sie verglichen? Es gibt zwei Ansätze:

Individuenbezogene Fairness-Metriken …

… beruhen auf dem Vergleich der Vorhersagen von Einzelpersonen, z. B. die Metrik Fairness durch awareness oder unawareness (auf den Zine-Seiten). Hier wird z. B. die Vorhersage für ein Individuum mit den Ergebnissen für ein ähnliches oder anderes Individuum verglichen.

Gruppenbezogene Fairness-Metriken …

… beruhen auf dem Vergleich der Vorhersagen des KI-Modells mit den gelabelten Daten für verschiedene Gruppen. Siehe z. B. Demopraphic Parity oder Treatment Equality (auf den Zine-Seiten). Der Vergleich der Vorhersagen pro Gruppe basiert auf einer Tabelle, die einen Überblick darüber gibt, wo und für welche Gruppe das KI-System richtig vorhergesagt hat:

	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

Jede Fairness-Metrik ist eine bestimmte Definition von Fairness. Mathematisch ausgedrückt, könnte so eine Definition so aussehen so aussehen:

Das KI-System ist fair, wenn

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

Das kann auch so ausgedrückt werden:

Das Verhältnis zwischen den voraussichtlichen Einstellungen und der Gesamtzahl der Bewerbungen ist in allen Gruppen ähnlich. 40 % der Bewerber*innen sind weiblich – 40 % der eingestellten Mitarbeiter*innen sollten weiblich sein!
Diese Metrik definiert Fairness als die gleiche Wahrscheinlichkeit für beide Gruppen, eingestellt zu werden. Das nennt man demografische Parität – und diese Definition hast du bereits auf den Zine-Seiten kennengelernt!

NERD NOTIZEN:

Es gibt viele verschiedene Fairness-Metriken. Ihre große Anzahl zeigt, wie schwierig es ist, eine übergreifende und anwendbare Definition von Fairness zu finden – diese Unschärfe ist aber auch eine Stärke von Fairness.
Wir haben hier eine Metrik im Detail dargestellt, um zu erklären, wie die Metriken funktionieren. Wenn man andere Metriken genauer vergleicht, wird deutlich, dass sie sich manchmal widersprechen. Oder die Ergebnisse einer Metrik würden ein KI-System als fair beschreiben, während eine andere Metrik das nicht tun würde.
Die Fairness-Metriken sind eine Auswahl der in Mehrabi et al. (2022) vorgestellten Fairness-Metriken: A survey on bias and fairness in Machine Learning; und Verma & Rubin (2018): Fairness definitions explained.

IDEE · UMSETZUNG · I. Mastenbroek, I. Kühnlein, B. Lübbert DESIGN · I. Kühnlein, I. Mastenbroek, L. Wunderlich