BIAS IM KI KONTEXT

Biases können KI-Modelle in verschiedenen Entwicklungsschritten auf unterschiedliche Weisen beeinflussen. Oftmals wird entweder auf eine verzerrte Datengrundlage oder das Training des Modells verwiesen. KI-Modelle übernehmen dann Voreingenommenheiten, was diskriminierende Effekte haben kann.

Zwei Beispiele für Biases werden auf den folgenden Seiten erklärt, noch mehr gibt es auf der Posterseite dieses Zines!

*Wir verwenden das engl. Wort „Bias", da der Diskurs auch in Deutschland um das Wort "Bias" geführt wird.

BIAS

Der Begriff **Bias*** (dt. Voreingenommenheit, Verzerrung) beschreibt eine **systematische Abweichung**. Verschiedene Forschungsgebiete haben jeweils eigene Verständnisse davon, was genau "Bias" bezeichnet. **Diese verschiedenen Konzepte von Bias können zu Missverständnissen und falschen Vorstellungen dessen, wie mit ihnen umzugehen ist, führen.**

Im Kontext von KI wurde Bias zunächst als ein Begriff verwendet, der auf technisch bedingte Abweichungen fokussierte. Die Auseinandersetzung mit gesellschaftlich bedingten Biases im Kontext von KI ist neuer.

1 WO KOMMT UNFAIRNESS HER?

Nach der Implementierung der Einstellungs-KI bemerkt Techie, dass das KI-Modell diskriminierende Entscheidungen trifft. Also beschließt Techie, der Sache nachzugehen. Wie gelangen Ungerechtigkeiten in den Einstellungsalgorithmus?

Vorurteile gelangen an vielen Stellen in in KI-Modelle – und es gibt keinen alleinigen Ansatz, um sie alle zu bekämpfen.

Trotzdem möchte Techie wissen, wie diskriminierend deren KI-Modell tatsächlich ist. Wie kann Techie den Bias des Modells messen?
Mehr dazu in den nächsten Zines!

EXAMPLE BIASES

ABMESSUNGS-BIAS

Die im Modell verwendeten Merkmale und Bezeichnungen sind **ungenaue Darstellungen dessen, was uns wirklich interessiert**. Sie verzerren die Sicht des KI- Modells auf die Realität.

```
Wie gut kann ein kurzes
PDF das Leben einer Person
zusammenfassen? Bestimmte
wichtige Qualitäten wie
Freundlichkeit oder
relevante Netzwerk- und
Branchenverbindungen
finden sich vielleicht
nicht im Lebenslauf.
```


ALGORITHMISCHER BIAS

Modelle führen zu Biases in dem KI-System, indem das Modell systematisch bestimmte Faktoren gegenüber anderen bevorzugt.

```
Während des Trainings lernt
das KI-Modell, dass die
wichtigsten Informationen
im Abschnitt "Karriere" zu
finden sind. Deshalb
konzentriert es sich auf
diesen Abschnitt. Wenn
ein*e Bewerber*in für die
Stelle förderliche Hobbys
hat, würde die KI dies
nicht erkennen.
```


WIE KÖNNEN WIR MIT BIASES UMGEHEN?

Biases sickern also an vielen Stellen in die Daten, KI-Modelle und die Art und Weise, wie wir mit ihnen umgehen, ein. Ein Ansatz für den Umgang mit Datensätzen mit Bias ist das so genannte **"debiasing"**. Dieser Begriff wird häufig in Policy-Dokumenten verwendet und als Mittel zur Lösung von diskriminierenden KI-Anwendungen dargestellt*. Debiasing beschreibt die Idee, eine Anwendung mit Bias mithilfe von Fairness-Metriken und technischen Methoden zu behandeln, um einen "unverzerrten" Datensatz zu erhalten.

Das Versprechen von einer technischen Lösung klingt verlockend. Es ist aber kein guter Weg, um gesellschaftliche Probleme zu lösen. Verwendete Wörter, wie debiasing, können sogar irreführend sein. Es impliziert, dass Biases, und damit Diskriminierung, technisch reduziert werden könnten. Doch Ungerechtigkeit bleibt bestehen, solange die gesellschaftlichen Hintergründe und Annahmen, die Ungerechtigkeiten zugrunde liegen, nicht hinterfragt werden.

*Der Bericht von Balayn & Güres (2021): Beyond Debiasing. Regulating AI and its inequalities geht tiefer darauf ein.

FEELS FAIR?

BIASES, DIE IN BEZUG AUF KI EINE ROLLE SPIELEN:

ABMESSUNGS-BIAS

Die im Modell verwendeten Merkmale und Bezeichnungen sind ungenaue Darstellungen dessen, was uns wirklich interessiert.

Wie gut kann ein kurzes PDF das Leben eines Menschen zusammenfassen?

LABEL-BIAS

Die Label geben nicht genau die Kategorien wieder, für die sie gedacht sind.

Binäre Label geben unsere komplexe Realität falsch wieder, indem sie z. B. den Eindruck erwecken, dass diejenigen, die nicht eingestellt werden, wertlos seien.

EVALUATIONS-BIAS

Das Modell wird anhand von Kriterien bewertet, die seine reale Verwendung in der Welt nicht genau widerspiegeln.

Die Bewertung wird von C.O.R.P. durchgeführt. Ihnen geht es um Genauigkeit und Effizienz - nicht um Fairness!

FEEDBACK-BIAS

Die Ergebnisse des Modells werden zu neuen Trainingsdaten, die die anfänglichen Biases tradieren und so verstärken.

Aufgrund der Überrepräsentation *weißer* Männer im KI Tool werden weiter mehr *weiße* Männer eingestellt und mehr Frauen of Colour abgelehnt werden.

HISTORISCHER BIAS

Tradierte gesellschaftliche Vorurteile spiegeln sich in den Trainingsdaten wider.

Aufgrund unserer von Sexismus und Rassismus geprägten Geschichte werden Vorurteile wie „*weiße* Männer sind erfolgreicher in Führungspositionen“ von KI-Modellen reproduziert.

DARSTELLUNGS-BIAS

Die verwendeten Daten spiegeln die Vielfalt und Komplexität der Welt nicht wider.

Wenn C.O.R.P.’s Datensatz vor allem *weiße* männliche Bewerber gut bewertet, wird das KI-System nicht viel über Schwarze Frauen* lernen. Wenn sich in Zukunft eine Schwarze Frau* bei C.O.R.P. bewerben würde, hätte sie kaum eine Chance - auch wenn ihre Fähigkeiten denen der Eingestellten gleichen.

ALGORITHMISCHER BIAS

Modelle bringen Bias in das System ein.

Während des Trainings lernt das KI-System, dass die interessantesten Informationen in der Rubrik „Karriere“ zu finden sind.

ANWENDUNGS-BIAS

Tritt auf, wenn sich der Kontext, in dem das Modell verwendet wird, von der Trainingsumgebung unterscheidet.

Unser Trainingsdatensatz erstreckt sich über die letzten 20 Jahre. Merkmale, auf die wir bei neuen Bewerber*innen Wert legen, sind bei den alten Bewerber*innen möglicherweise nicht vorhanden.

BIASES SIND UNFAIR

NERD NOTIZEN:

Diese Auswahl von Biases basiert auf van Giffen et al. (2022): Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods.

Einige der Biases sind auch unter anderen Namen bekannt, Einzelheiten dazu gibt’s im zitierten Artikel. „Debiasing“ als proklamierte Lösung für Datensätze mit Bias findet sich meist in Policy-Dokumenten, wie eine Untersuchung von Balayn & Gürses (2021) zeigt. Sie heben hervor, dass dort in der Regel nicht spezifiziert wird, wie limitiert debiasing in Bezug auf das Erreichen fairer KI-Modelle ist.

Debiasing basiert auf zwei Schritten: Schritt 1 sind Fairness-Metriken: Das sind Regeln zur Messung von Fairness. Schritt 2 sind technischen Methoden: Das sind Werkzeuge zur Anpassung im KI-Lebenszyklus auf der Grundlage des ermittelten Fairnessniveaus. Diese technischen Ansätze befassen sich jedoch nicht mit der großen Vielfalt von Biases und deren Ursachen.

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