(K)ein Widerspruch:

Diversity und vertrauenswürdige KI

Können vertrauenswürdige KI und Diversity in HR für Chancengleichheit sorgen?

Madeleine Bauer-Eder
CHRO IBM Österreich

Bias und Diskriminierung

Quelle:
Progress 2024

65%

der Geschäfts- und IT-Führungskräfte sind der Meinung, dass es in ihrem Unternehmen derzeit eine Datenverzerrung gibt.

78%

glauben, dass die Verzerrung von Daten mit der zunehmenden Nutzung von KI/ML ein größeres Problem werden wird.

13%

der Unternehmen befassen sich derzeit mit Bias.

Was sind Bias?

Menschen haben Vorurteile.
Maschinen haben Bias.
Beides kann diskrimieren.

Was sind klasische Beispiele für Voreingenommenheit?

Confirmation Bias (Bestätigungsfehler)

Selektive Suche nach Informationen, die bestehende Überzeugungen bestätigen

Stereotypen-Bias

Vorurteile über Gruppen, z. B. basierend auf Geschlecht oder Ethnizität, beeinflussen Wahrnehmung und Entscheidungen.

Ingroup-Bias

zeigt sich, wenn Menschen die Mitglieder ihrer eigenen Gruppe bevorzugen und andere benachteiligen.

Wie kommen Bias in unsere Daten?

Daten als Spiegel der Gesellschaft.

Welche Faktoren führen zu unausgewogenen oder voreingenommenen Datensätzen?

Historische Ungleichheiten

Daten spiegeln oft bestehende gesellschaftliche Vorurteile wider, z. B. in Jobdaten, die die Unterrepräsentation von Frauen in bestimmten Berufen zeigen.

Ungleichgewicht bei der Datenerfassung

Wenn bestimmte Gruppen unterrepräsentiert sind oder Daten ungleichmäßig gesammelt werden, entstehen Verzerrungen.

Menschliche Entscheidungen

Bias kann durch die Menschen entstehen, die die Daten auswählen, kennzeichnen oder interpretieren, da ihre Vorurteile und Annahmen in den Datensatz einfließen.

Bias und Diskriminierung

Verzerrte Darstellungen durch KI

Wie generative Modelle, die mit voreingenommenen Trainingsdaten trainierert warden, Stereotype verstärken.

CEO

Richter/in

Lehrer/in

Sozialarbeiter/in

Quelle:
Bloomberg 2023

IBM

Bias und Diskriminierung

Vorurteils-Echos aus den Daten

KI, die mit voreingenommenen Trainingsdaten trainierert wurde, kann Geschlechter- und Ethnien-Stereotype in Berufbildern verstärken.

Gutbezahlter Beruf

CEO

Richter/in

Niedrigbezahlter Beruf

Lehrer/in

Sozialarbeiter/in

Quelle:
Bloomberg 2023

Risikobereich Daten

Datenqualität
und Selektivität

Prinzipien zur Vermeidung von Bias

Eine IBM Antwort: https://www.ibm.com/topics/ai-bias

Eine IBM Antwort

Was tun?

Ethische Grundsätze

Klare ethische Werte und Richtlinien, um Fairness Inklusion und den Schutz individueller Rechte sicherzustellen.

Governance-Struktur

Richtlinien und Kontrollen für den Einsatz von KI. Ziel: Vorurteile minimieren, objektive Entscheidungen und Chancengleichheit fördern

Transparenz und Verantwortlichkeit

Ki muss nachvollziebar und erklärbar sein. Ernenne Verantwortliche und stelle sicher: Die Entscheidungen trifft immer der Mensch.

IBM‘s Säulen vertrauenswürdiger KI

Gerechtigkeit

Unparteilichkeit und Umgang mit Verzerrungen (Bias)

Sind privilegierte Gruppen im Vergleich zu anderen Gruppen systematisch im Vorteil?

Wird Bias verstärkt?

Robustheit

Effektive Bewältigung außergewöhnlicher Bedingungen

Wird die KI in verschiedenen Szenarien immer wieder das gleiche Ergebnis ausgeben?

Privatsphäre

Hohe Integrität der Daten und Einhaltung der Geschäfts-anforderungen

Wie stellen wir sicher, dass die Eigentümer die Kontrolle über Daten und Erkenntnisse behalten?

Erklärbarkeit

Einfach zu verstehende Ergebnisse/ Entscheidungen

Warum ist die KI zu einem bestimmten Ergebnis gekommen? Wann wäre es anders gewesen?

Transparenz

Offen für die Überprüfung von Fakten und Details

Können wir verstehen, warum und wie KI geschaffen wurde?

Eine IBM Antwort

Checklist

1. Wähle das richtige Datenmodell.
2. Trainiere mit den richtigen Daten
3. Wähle ein ausgewogenes Team
4. Führen Datenverarbeitung mit Bedacht durch
5. Richte eine kontinuierliche Überwachung ein

Was macht eine generative KI-Plattform vertrauenswürdig?

Wie wurde die KI trainiert?

- Die Trainingsdaten müssen umfangreich und umfassend sein, aber auch kuratiert warden.

Bias und Halluzinationen

- Wie kann die Plattform Verzerrungen erkennen und korrigieren?
- Werden Bias und Halluzinationen zuverlässig erkannt und minimiert?

Ist sie transparent?

- Offene vs. Blackbox
- Wie lassen sich Modell und die von ihm generierten Antworten prüfen und erklären?
- Verfolgt das Modell Drift und Verzerrungen? Und wie geht es mit ihnen um?

Wird die Einhaltung von Vorschriften unterstützt?

- Wie werden die Basismodelle und ihre Verwendung mit dem Datenschutz und den staatlichen Vorschriften in Einklang gebracht?

Ist die KI sicher?

- wer hat die Kontrolle über das Modell, die Eingabe- und Ausgabedaten?
- Können Sie sicherstellen, dass keine vertraulichen Informationen weitergegeben werden?
- Wie wird es überwacht?
- Welche Sicherheitsvorkehrungen gibt es?

Kann es angepasst werden?

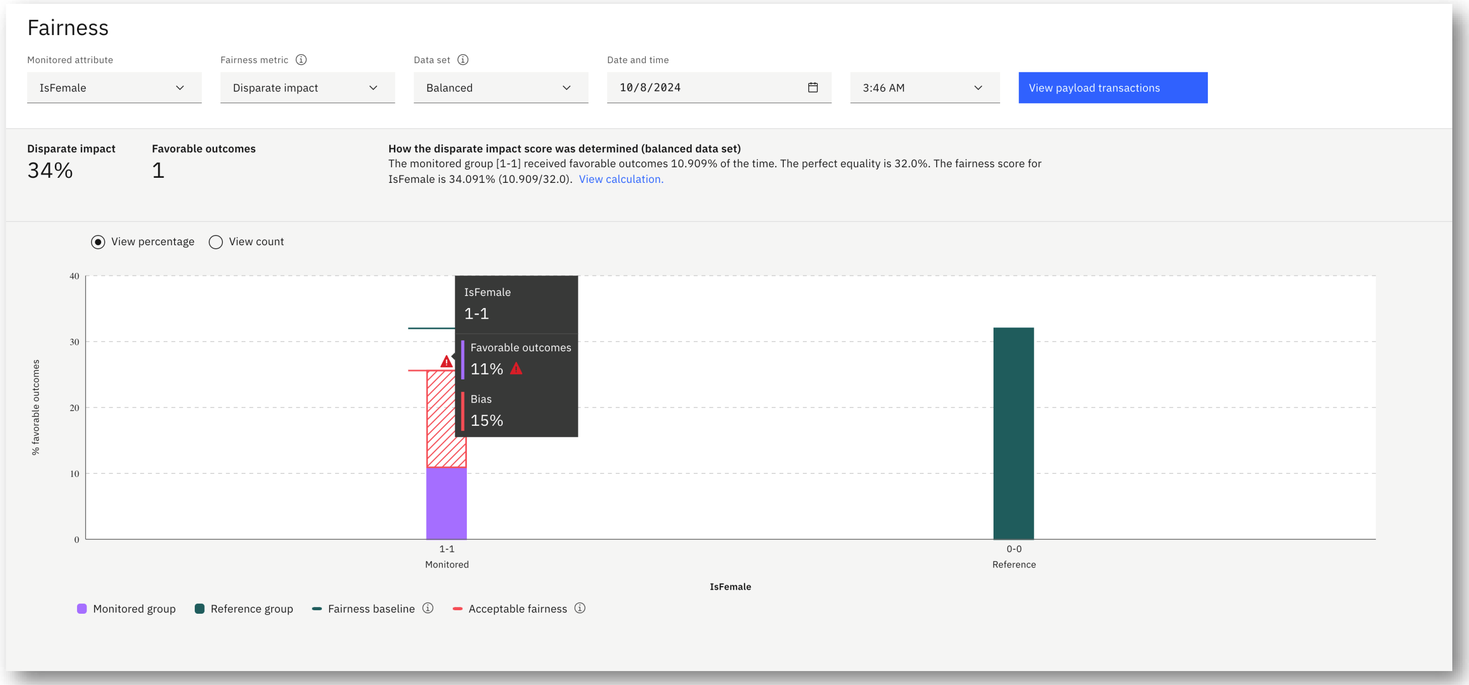
- Kann das Modell mit Ihren Daten fein abgestimmt werden?
- Kann es für bestimmte Anwendungsfälle angepasst warden?
- Wie wird es mit anderen Anwendungen integriert?

watsonx.
governance

IBM

watsonx. governance

Erkennen und Behebung von Verzerrungen (Bias)



- **Verzerrungen Erkennen:**
  Erkennung von Verzerrungen während der Laufzeit

- **Risikomanagement:**
  Einstellen von Warnmeldungen im Dashboard

- **Behebung von Verzerrungen**
  Bereitstellung von Metriken und Daten zur Unterstützung von Data Scientists bei der Behebung von Verzerrungen

- **‘Orten’ der Verzerrung**
  Zeigt die exakte Eingabe an, ab der die Verzerrung stattfand

Fragen?

Madeleine Bauer-Eder

Thomas Jirku