FEELS FAIR?

WIE WIRD EIN KI MODELL ENTWICKELT?

Lerne **Techie** (they/them) kennen, eine* Datenwissenschaftler*in, die KI für das Gemeinwohl nutzen möchte. Techie wird vom Unternehmen C.O.R.P. beauftragt, ihre Einstellungsprozesse zu automatisieren.
Dieser KI-Lebenszyklus zeigt die Schritte der KI-Entwicklung.

DIE HERAUSFORDERUNG VON C.O.R.P.:

Entwickle ein faires KI-Einstellungsmodell:

Erstelle ein binäres Klassifizierungsmodell[NN], das den Lebenslauf einer Person betrachtet und vorhersagt, ob sie eingestellt werden soll (1) oder nicht (0). Ein gutes Modell würde „einstellen" vorhersagen, wenn die Person zuvor eingestellt wurde.

Finde anhand des Lebenslaufs heraus, ob neue Bewerber*innen in die C.O.R.P.-Kultur passen.

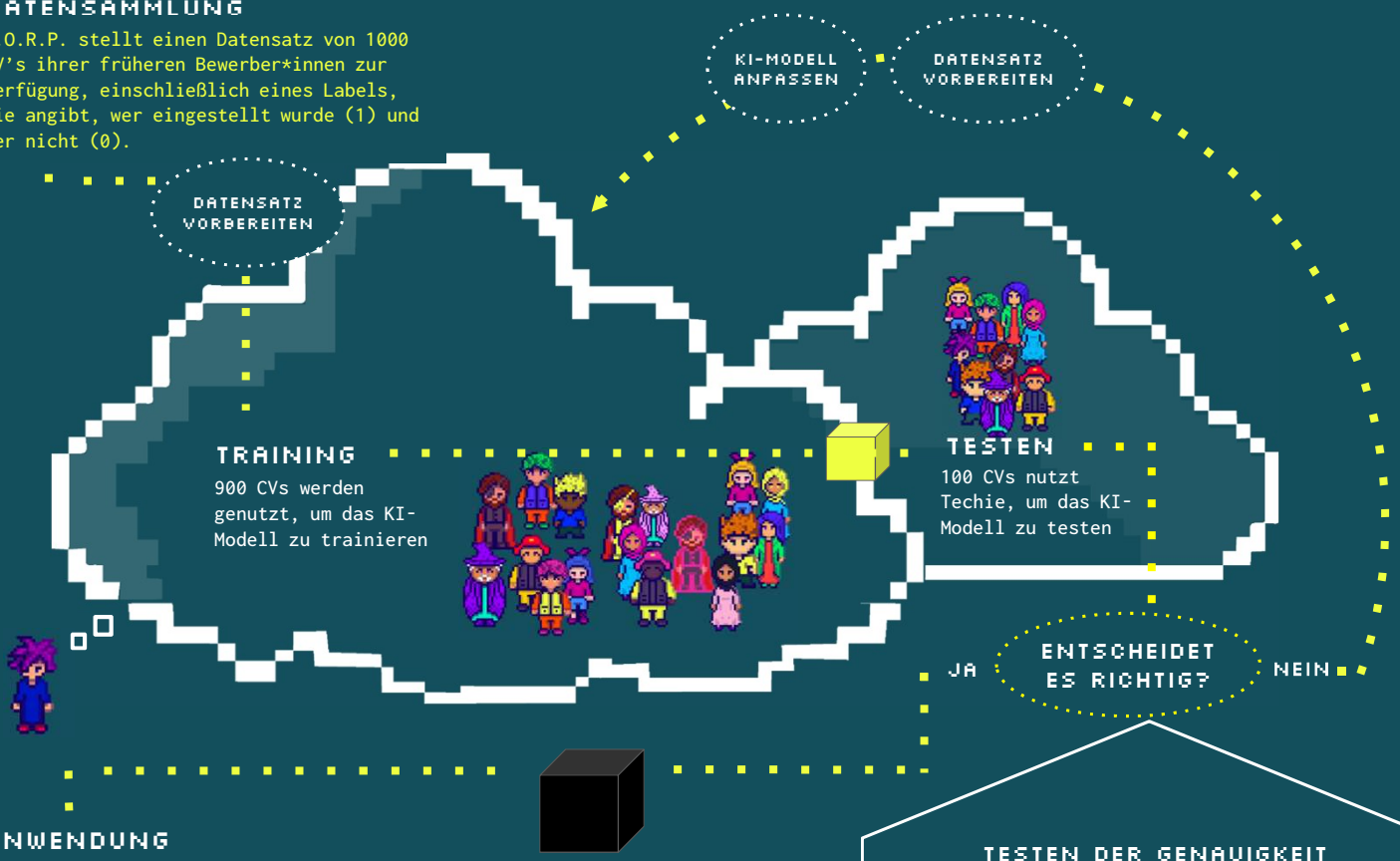
Hintergrundinformationen:

Die Mehrheit der derzeitigen Mitarbeiter*innen von C.O.R.P. ist weiß, männlich und verdient ein hohes Einkommen.

PROBLEMDEFINITION

DATENSAMMLUNG

C.O.R.P. stellt einen Datensatz von 1000 CV's ihrer früheren Bewerber*innen zur Verfügung, einschließlich eines Labels, die angibt, wer eingestellt wurde (1) und wer nicht (0).



ANWENDUNG

C.O.R.P. verwendet jetzt den Algorithmus um seine neuen Mitarbeiter*innen einzustellen.

EVALUATION IN ECHTER WELT

C.O.R.P. bewertet, wie effizient und nützlich der Algorithmus für das eigene Unternehmen ist. Aber ist der Algorithmus auch fair?

IST DEINE KI FAIR?

TESTEN DER GENAUIGKEIT

Techie testet die Leistung des KI-Modells im Hinblick auf Genauigkeit. Dafür vergleicht Techie die Vorhersagen mit früheren Entscheidungen von C.O.R.P.(Ground Truth[NN]):

	KI prognostiziert: einstellen	KI prognostiziert: ablehnen
Label von C.O.R.P.: einstellen	WAHR POSITIV	FALSCH NEGATIV
Label von C.O.R.P.: ablehnen	FALSCH POSITIV	WAHR NEGATIV

Techie berechnet Genauigkeitsmetriken, die Aufschluss über die Rate der „richtigen" Vorhersagen geben. Wenn diese Rate nicht zufriedenstellend[NN] ist, verfeinert Techie das Modell und trainiert es erneut.

$$\text{Präzision} = \frac{\text{Wahr positive Vorhersagen}}{\text{Alle positiven Vorhersagen}} \qquad \text{Genauigkeit} = \frac{\text{Alle wahren Vorhersagen}}{\text{Alle Vorhersagen}}$$

NERD NOTIZEN (NN)

Ein **binäres Modell** ist eine mathematische Darstellung eines Systems oder eines Prozesses, bei dem das Ergebnis entweder 0 oder 1 ist. Wir haben uns für ein einfaches **binäres Klassifizierungsmodell** entschieden (Einstellung vs. Nicht-Einstellung). Es bietet die einfachste Einführung in die Messung von Fairness. In der Realität wird eine Vielzahl von KI-Systemen darauf trainiert, komplexere Aufgaben auf der Grundlage einer Eingabeaufforderung auszuführen. Zum Beispiel bei der Generierung von Bildern/Videos/Texten/Audios. Obwohl diese generativen Algorithmen nicht für binäre Klassifizierungen trainiert werden, werden sie anhand derselben binären Fairness-Metriken evaluiert.

Wir verwenden zwei verschiedenfarbige Boxen für die Darstellung unseres KI-Modells, um einen wichtigen Phasenunterschied hervorzuheben. Beim Training verwenden wir gelabelte Daten, eine Situation, in der wir die "**Ground Truth**" kennen (im Fall von Techie: wurde diese Person eingestellt/nicht eingestellt). Mit diesem Wissen lässt sich die Genauigkeit des Modells testen. Wir setzen das Modell ein, wenn die Genauigkeit **zufriedenstellend** ist, d. h. über einem bestimmten Schwellenwert liegt. Nach dem Einsatz des Modells haben wir keinen Zugang zu solchen früheren Entscheidungen und verlassen uns stattdessen auf die angenommene Genauigkeit des Modells. Dieser Übergang macht das KI-Modell zu einer **Blackbox**, deren Entscheidungsprozesse nicht direkt beobachtbar oder mit der „Ground Truth" vergleichbar sind.

Ein Projekt des Alexander von Humboldt Institute für Internet und Gesellschaft innerhalb des AI Society Lab.

IDEE · REALISATION · Irma Mastenbroek, Irina Kühnlein, Birte Lübbert

DESIGN · Irma Mastenbroek, Larissa Wunderlich, Irina Kühnlein