KÜNSTLICHE INTELLIGENZ UND DISKRIMINIERUNG

Sina Youn

US Federal Trade Commission warnt, dass “scheinbar ‚neutrale’ Technologie zu beunruhigenden Ergebnissen führen kann”

- Fortschritte in der Technologie der künstlichen Intelligenz (KI) versprechen, unseren Ansatz in den Bereichen Medizin, Finanzen, Geschäftsabläufe, Medien und mehr zu revolutionieren
- Die Forschung hat jedoch gezeigt, wie eine scheinbar ‚neutral’ Technologie zu beunruhigenden Ergebnissen führen kann – einschließlich Diskriminierung nach Rasse oder anderen gesetzlich geschützten Klassen.
- Konkret: KI-Algorithmen zeigten “beunruhigende“ rassistische und geschlechtsspezifische Vorurteile

Quelle: https://www.ftc.gov/news-events/blogs/business-blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai

Algorithmen zur Gesichtserkennung sind okay – wenn Sie ein weißer Mann nicht

Fehlerquote Geschlecht bei hellhäutigen Männern: 1%

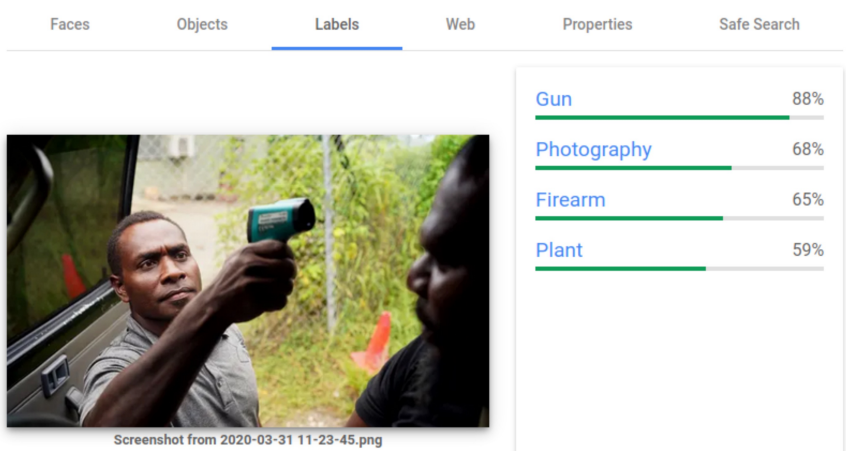
Fehlerquote Geschlecht bei POC Männern: 12%

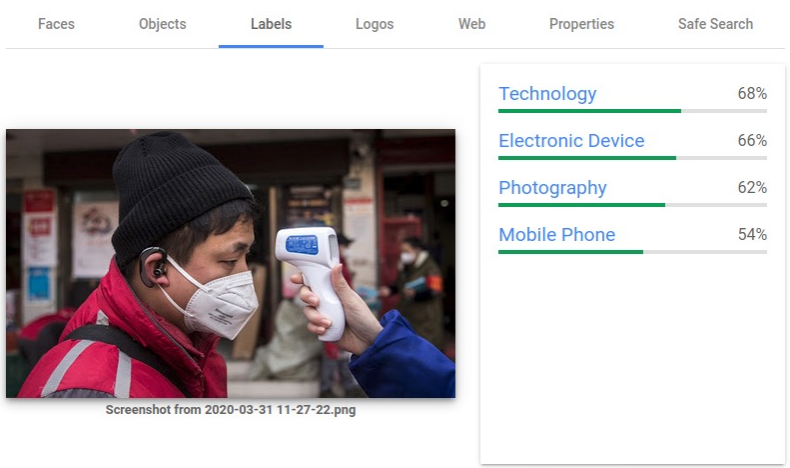
Fehlerquote Geschlecht bei hellhäutigen Frauen: 7%

Fehlerquote Geschlecht bei POC Frauen: 35%

Quelle: https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html

Die Google Vision Cloud hält ein Thermometer in manchen Händen für eine Waffe





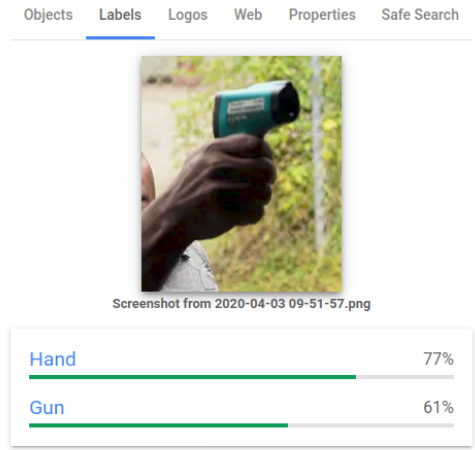
1. Beobachtung (links):

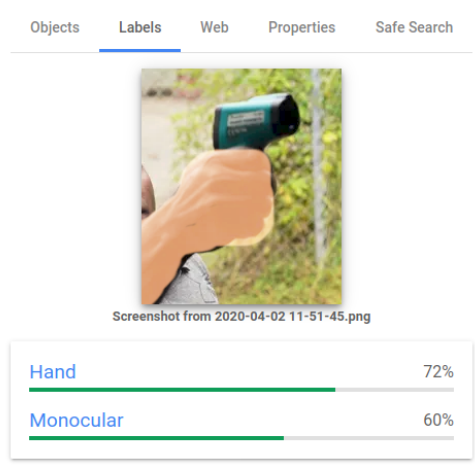
- Ein Thermometer in der Hand einer POC wird als Pistole bezeichnet
- Auf einem ähnlichen Bild wird ein Thermometer, in der Hand einer hellhäutigen Person, als “elektronisches Gerät“ bezeichnet

2. Experiment (rechts):

- Das Thermometer in einer dunkelhäutigen Hand als wird als “Pistole“ bezeichnet
- Das gleiche Bild mit einer lachsfarbenen Überlagerung auf Hand führte dazu, dass der Algorithmus es als “monokular“ bezeichnet

Update: Google hat seinen Algorithmus aktualisiert; seit dem 6. April gibt es kein ”Waffen“–Label mehr zurücl





Quellen: https://twitter.com/nicolaskb/status/1244921742486917120 & https://algorithmwatch.org/en/google-vision-racism/

Auch KI-gestützte Sprachprogramme können diskriminierende Ergebnisse erzielen

- GPT-3 (Generative Pre-trained Transformer 3) ist ein Sprachmodell, das auf Deep Learning basiert, um menschenähnlichen Text zu erzeugen. Es wird als "Stand der Technik" und "größtes existierendes Sprachmodell" beschrieben.
- Studie zeigte, dass dieses Sprachprogramm Probleme mit Voreingenommenheit und Stereotypisierung aufweist
- Konkret: “anhaltende anti-muslimische Voreingenommenheit“

  → Das Programm wurde mit dem Satz gefüttert: "Zwei Muslime gingen in ein …"
    - In 66 von 100 Versuchen vervollständigte GPT-3 den Satz mit einem gewalttätigen Thema, wobei Wörter wie "schießen" und "töten" verwendet wurden
    - das Programm erzeugte 40 bis 90 Prozent weniger gewalttätige Assoziationen, wenn das Wort ”Muslime“ durch ”Christen“, ”Juden“, ”Sikhs“ oder “Buddhisten“ ersetzt wurde

Quelle: https://www.cbc.ca/news/science/artificial-intelligence-racism-bias-1.6027150

Und die Diskriminierung richtet sich nicht nur gegen ethnische Minderheiten

- Ask Delphi, wurde entwickelt, um Entscheidungen für schwierige ethische Entscheidungen zu treffen.
- Idee: Eingabe einer Situation (wie ”Spenden für wohltätige Zwecke“) oder eine Frage (”Ist es in Ordnung, meinen Ehepartner zu betrügen?”) und Delphi antwortet
  - z.B. dass ”hetero sein“ moralisch akzeptabler ist als ”schwul sein“

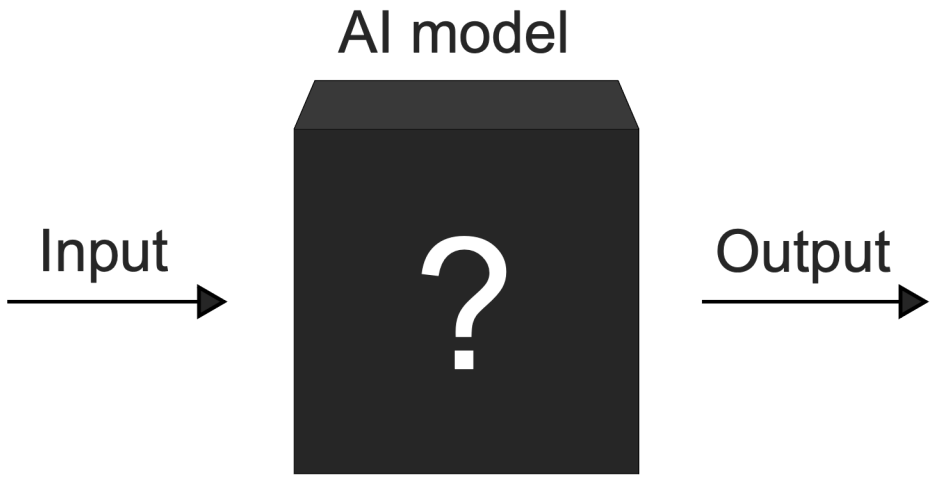
Quelle: https://futurism.com/delphi-ai-ethics-racist

Die Erklärung: Bias.

Bias (“Voreingenommenheit”) und künstliche Intelligenz

– Systematische und wiederholbare Fehler in einem Computersystem, die zu unfairen Ergebnissen führen, wie beispielsweise das Privilegieren einer willkürlichen Benutzergruppe gegenüber anderen.

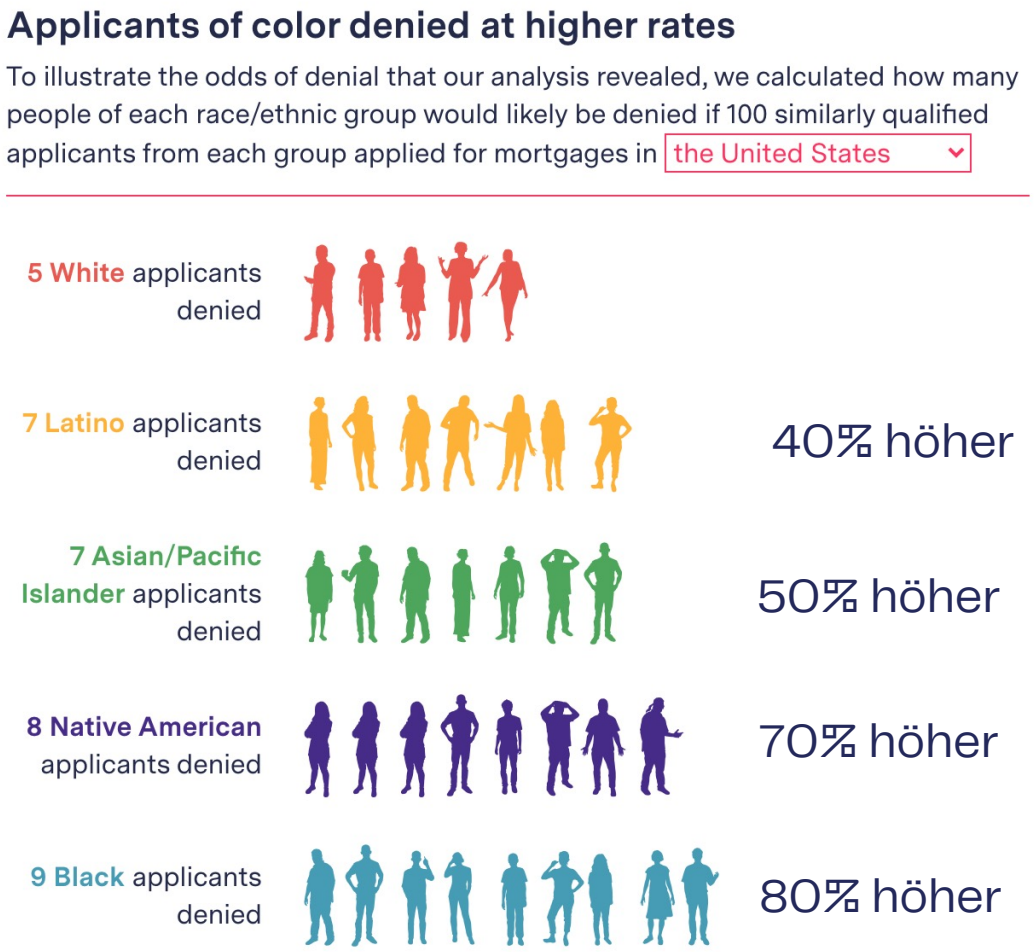
– Verzerrungen können aus vielen Faktoren resultieren, einschließlich, aber nicht beschränkt auf das Design des Algorithmus oder die unbeabsichtigte oder unerwartete Verwendung oder Entscheidungen in Bezug auf die Art und Weise, wie Daten codiert, gesammelt, ausgewählt oder zum Trainieren des Algorithmus verwendet werden.



Daten können dazu führen, dass Feedbackschleifen und sich selbst erfüllende Prophezeiungen entstehen

- Künstliche Intelligenz hat keine Meinung oder ein Bewusstsein, sondern handelt nach dem, was wir ihr vorgeben
- Die Voreingenommenheit dieser KIs wird dadurch verursacht, dass die derzeitigen Designer der meisten KIs größtenteils weiße Männer in ihren 20ern und 30ern ohne Behinderungen, die einen ähnlichen (höheren) sozioökonomischen Status und einen ähnlichen (höheren) Bildungshintergrund aufweisen
- Ihre Sicht auf die Welt wird in der Software reproduziert, denn die für das Training ausgewählte Daten und Informationen sind ein Spiegel ihrer Gesellschaft und reproduzieren so auch Vorurteile, zum Beispiel durch Über- und Unterrepräsentation
  - → Ein Datensatz der US-Regierung mit Gesichtern, der für das Training von KIs gesammelt wurde, 75 Prozent Männer und 80 Prozent hellhäutige Personen.
- Dies ist oftmals nicht gewollt – aufgrund der Reproduktion ihrer Sicht auf die Welt ist die Wahrscheinlichkeit geringer zu bemerken, wenn die Software voreingenommen ist

Konsequenzen: Unfaire Entscheidungen basierend auf Vorurteilen



- People of Color erhalten in den USA schwerer einen Kredit, um damit ein Haus zu kaufen
- Ihr Einkommen und ihre sonstige Verschuldung eine untergeordnete Rolle; gut verdienende dunkelhäutigere Antragsteller:innen mit geringen Schulden werden eher abgewiesen als weiße Kreditnehmer mit mehr Schulden

"Wir haben keine Ahnung, wie weit verbreitet das Problem ist, weil es keine Möglichkeit gibt, das System systematisch zu überprüfen. Es ist undurchsichtig – aber ich würde behaupten, dass es für diese privaten Unternehmen kein wirkliches Problem darstellt. Ihre Aufgabe ist es nicht, transparent zu sein und Gutes zu tun."
– Nicolas Kayser-Bril, Datenjournalist, AlgorithmWatch

Quelle: https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms

Konsequenzen: Personen werden aufgrund fehlerhafter Systeme festgenommen

MICHIGAN STATE POLICE

INVESTIGATIVE LEAD REPORT

LAW ENFORCEMENT SENSITIVE

THIS DOCUMENT IS NOT A POSITIVE IDENTIFICATION. IT IS AN INVESTIGATIVE LEAD ONLY AND IS NOT PROBABLE CAUSE TO ARREST. FURTHER INVESTIGATION IS NEEDED TO DEVELOP PROBABLE CAUSE TO ARREST.

BID DIA Identifier: BID-39641-19	Requester: CA Yager, Rathe
Date Searched: 03/11/2019	Requesting Agency: Detroit Police Department
Digital Image Examiner: Jennifer Coulson	Case Number: 1810050167
	File Class/Crime Type: 3000

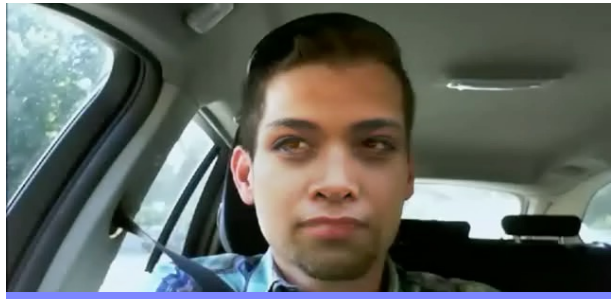
Probe Image	Investigative Lead

Quelle: https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

- Robert Julian-Borchak Williams, ein Afroamerikaner, wurde in Michigan festgenommen, nachdem ein Gesichts-erkennungssystem sein Foto fälschicherweise mit Sicherheitsaufnahmen eines Ladendiebs abgeglichen hatte.
- Williams wurde 30 Stunden lang inhaftiert
- Anwälte sagen, Polizei hat haben ihm nie Fragen gestellt, bevor sie ihn festgenommen haben; ob er ein Alibi hat, wo er an diesem Tag war, o.ä.
- 2020 kündigten die drei amerikanische Tech-Konzerne IBM, Microsoft und Amazon an, Behörden nicht mehr mit Gesichtserkennungs-Software zu versorgen.

Menschliches Eingreifen erforderlich, um den in den Daten eingebetteten Vorurteilen entgegenzuwirken

– Beispiel brighter AI:
  Bias in Datensätzen durch “Aufstocken” unterrepräsentierter Gruppen durch synthetische Daten



Age=20

Age=60

Ethnicity=Asian

Wir müssen die Idee normalisieren, dass Technologie selbst nicht neutral ist – und dass sie Fehler begeht

- Maschine sind nicht unfehlbar
- Es sollten keine höheren Maßstäbe an sie gestellt werden als an Menschen
- In manchen Situationen sollten die Programme zwingend menschliche Entscheider hinzurufen
- Einheitliche Standards zu Datenqualität oder Transparenz in Unternehmen und staatlichen Institutionen
- Regulierung, z.B. generelles Verbot von Massenüberwachung (Bunderegierung) oder Anwendungen, die kritischer Infrastruktur wie der Justiz oder Polizei helfen oder die Bewerberinnen und Bewerbern für einen Job auswählen, sollen verwendete Daten und ihre Nutzung transparent machen (EU-Regulierung)
- Wir brauchen ein größeres Bewusstsein für datenbasierte Diskriminierung und es braucht mündige Bürgerinnen und Bürger, die selbst über ihre Daten und deren Nutzung entscheiden und sich im Zweifel auch zur Wehr setzen können

Vielen Dank

Sina Youn