Deutsche Stiftung für Recht und Informatik

Die Rechtmäßigkeit der Verarbeitung sensibler Daten nach Art. 10 Abs. 5 AI Act

Ein Durchbruch für das Debiasing von KI-Systemen?

Yakin Surjadi, LL.M. (Hong Kong)
Schürmann Rosenthal Dreyer Rechtsanwälte

Herbstakademie 2024

GLIEDERUNG

1. Bias und Debiasing
2. Rechtlicher Rahmen für das Debiasing von KI-Systemen
   - Anforderungen nach der DSGVO
   - Die neue Vorschrift des Art. 10 Abs. 5 AI Act
3. Fazit

BIAS UND DEBIASING

Overcoming Racial Bias In AI Systems And Startlingly Even In AI Self-Driving Cars

Racial bias in a medical algorithm favors white patients over sicker black patients

AI expert calls for end to UK use of 'racially biased' algorithms

AI Bias Could Put Women's Lives At Risk - A Challenge For Regulators

Gender bias in AI: building fairer algorithms

Bias in AI: A problem recognized but still unresolved

Amazon, Apple, Google, IBM, and Microsoft worse at transcribing black people's voices than white people's with AI voice recognition, study finds

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

When It Comes to Gorillas, Google Photos Remains Blind

Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

The Week in Tech: Algorithmic Bias Is Bad. Uncovering It Is Good.

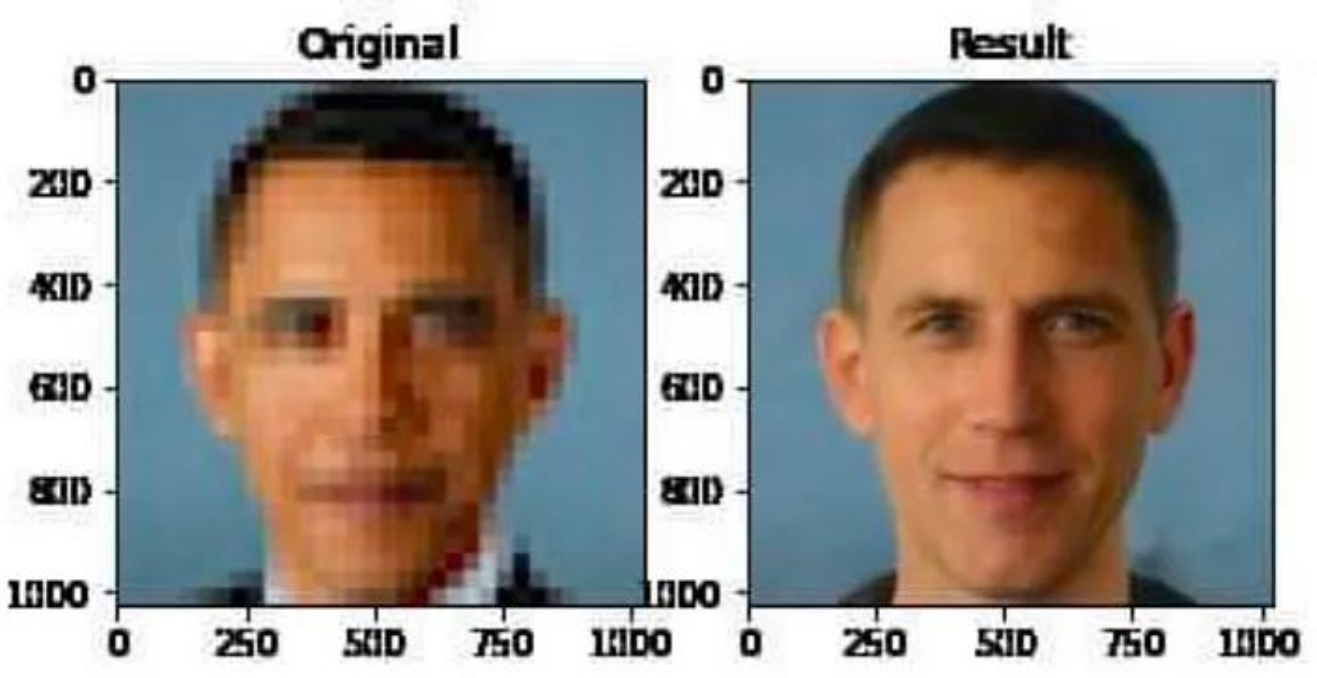
Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

Artificial Intelligence has a gender bias problem – just ask Siri

The Best Algorithms Struggle to Recognize Black Faces Equally

US government tests find even top-performing facial recognition systems misidentify blacks at rates five to 10 times higher than they do whites.

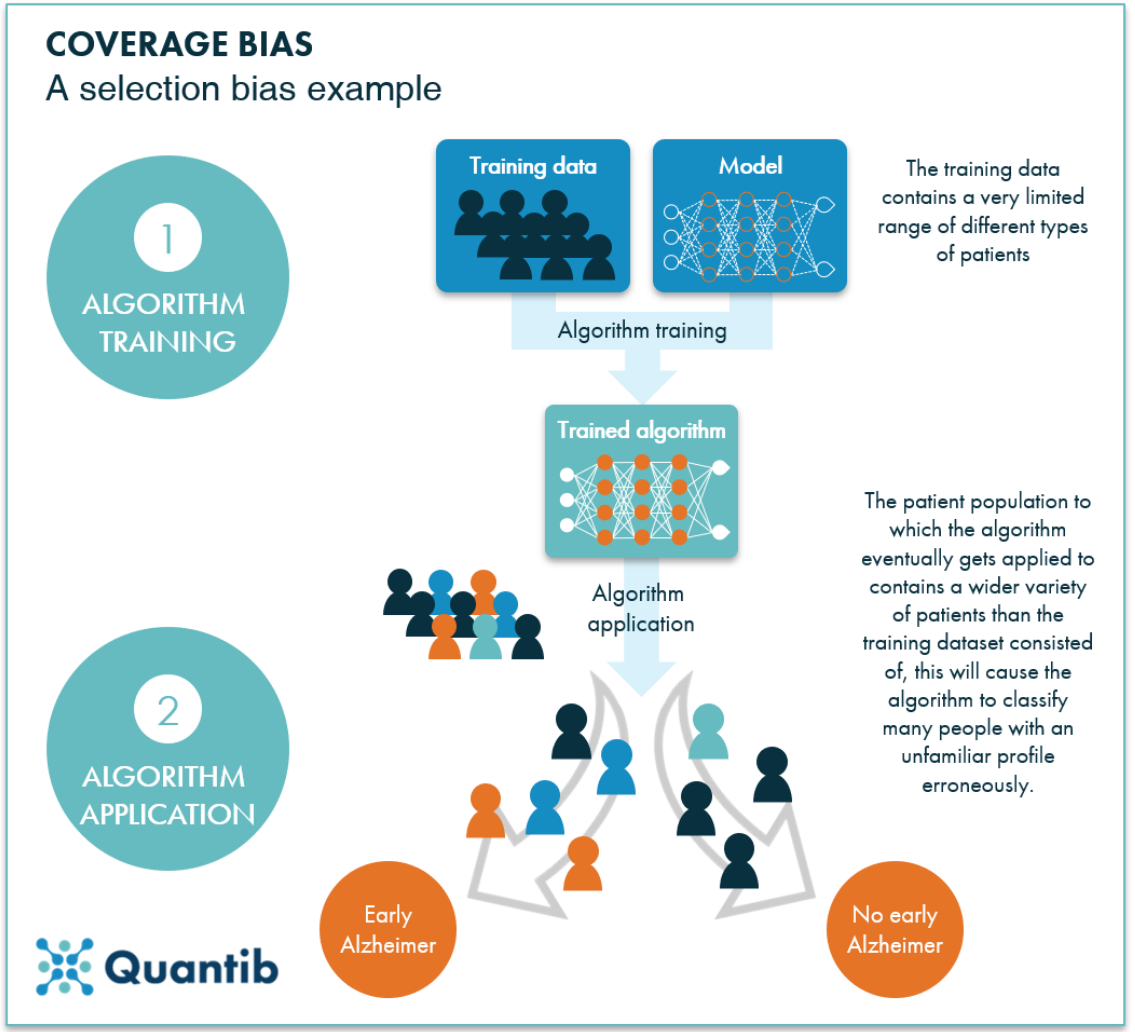
https://towardsdatascience.com



Bias in Bilderkennung

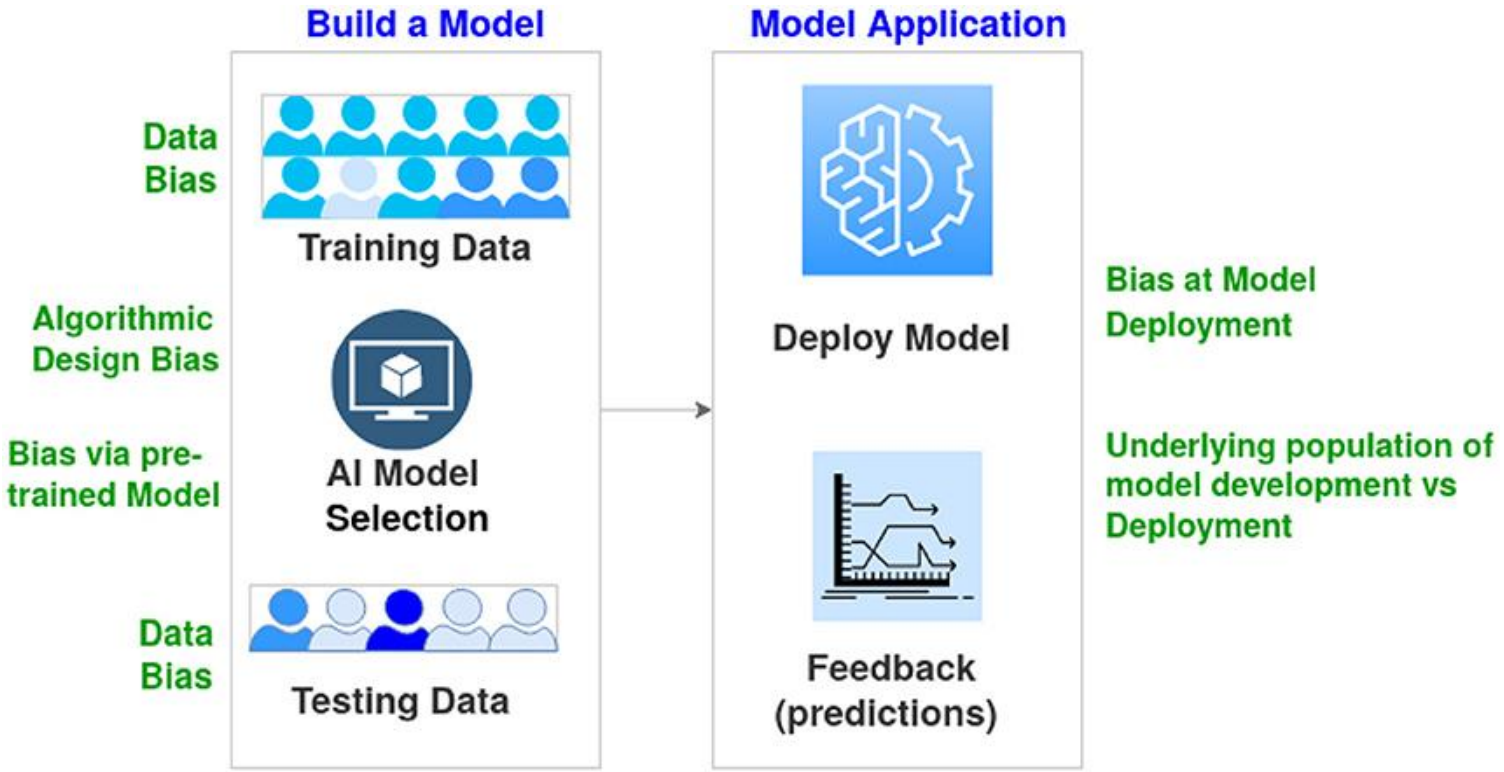
https://towardsdatascience.com/how-are-algorithms-biased-8449406aaa83

Bias in der Medizin



www.quantib.com/blog/understanding-the-role-of-ai-bias-in-healthcare

Auftreten von Bias in den Entwicklungsphasen einer KI



www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2022.1070493/full

Bias-Metriken und Debiasing-Methoden

- Um eine quantitative Indikation für Verzerrungen in Daten und Modellen zu ermöglichen, werden in der Wissenschaft verschiedene Bias- bzw. Fairness-Metriken vorgeschlagen
  - **Gruppenfairness:** statistische Größen in Bezug auf Personengruppen in Daten werden verglichen
  - **Individuellen Fairness:** basiert auf dem Vergleich und der Gleichbehandlung von Individuen
- Das Debiasing kann durch das Einwirken auf die **Trainingsdaten**, durch die **Änderung des Machine-Learning-Algorithmus** sowie durch die **nachträgliche Bearbeitung des Outputs** erfolgen

RECHTLICHER RAHMEN FÜR DAS DEBIASING

Anforderungen nach der DSGVO

- Nach Art. 9 Abs. 1 DSGVO ist die Verarbeitung sensibler Daten (z.B. Gesundheitsdaten, Daten der sexuellen Orientierung) grundsätzlich verboten
- In Bezug auf ein Debiasing von KI-Systemen mithilfe sensibler Daten sind vor allem zwei Ausnahmetatbestände nach Art. 9 Abs. 2 DSGVO relevant:
  - die **Einwilligung** der Betroffenen nach Art. 9 Abs. 2 lit. a DSGVO
  - sowie **Zwecke der Gesundheitsvorsorge** nach Art. 9 Abs. 2 lit. h DSGVO
- **(P)** Freiwilligkeit der Einwilligung nach Art. 7 Abs. 4 DSGVO, z.B. wenn ein Unternehmen für das Debiasing personenbezogene Daten seiner Mitarbeitenden verarbeiten möchte

Art. 10 Abs. 5 AI Act

*Soweit dies **unbedingt erforderlich** ist, um die Aufdeckung und Korrektur von Verzerrungen im Zusammenhang mit den **KI-Systemen mit hohem Risiko** gemäß Absatz 2 Buchstaben f und g dieses Artikels zu gewährleisten, können die **Anbieter** solcher Systeme ausnahmsweise besondere Kategorien personenbezogener Daten verarbeiten, sofern angemessene Garantien für die Grundrechte und Grundfreiheiten natürlicher Personen bestehen. Zusätzlich zu den Bestimmungen der Verordnungen (EU) 2016/679 und (EU) 2018/1725 und der Richtlinie (EU) 2016/680 müssen alle folgenden Bedingungen erfüllt sein, damit eine solche Verarbeitung stattfinden kann:*

*(a) Die Aufdeckung und Korrektur von Verzerrungen kann nicht wirksam durch die Verarbeitung anderer Daten, einschließlich **synthetischer oder anonymisierter Daten**, erfolgen;*

*(b) die besonderen Kategorien personenbezogener Daten unterliegen technischen Beschränkungen der Weiterverwendung der personenbezogenen Daten und dem Stand der Technik entsprechenden Sicherheits- und Datenschutzmaßnahmen, einschließlich **Pseudonymisierung**;*

*(c) die besonderen Kategorien personenbezogener Daten Gegenstand von Maßnahmen sind, die gewährleisten, dass die verarbeiteten personenbezogenen Daten **gesichert und geschützt** sind und **geeigneten Garantien** unterliegen, einschließlich strenger Kontrollen und Dokumentation des Zugangs, um Missbrauch zu vermeiden und sicherzustellen, dass nur befugte Personen mit angemessenen Vertraulichkeitsverpflichtungen Zugang zu diesen personenbezogenen Daten haben;*

*(d) die besonderen Kategorien personenbezogener Daten dürfen nicht **an andere Parteien übermittelt, weitergegeben oder anderweitig zugänglich** gemacht werden;*

*(e) die besonderen Kategorien personenbezogener Daten werden **gelöscht, sobald die Verzerrung behoben** ist oder die Aufbewahrungsfrist der personenbezogenen Daten abgelaufen ist, je nachdem, was zuerst eintritt;*

*(f) die Aufzeichnungen von Verarbeitungstätigkeiten gemäß den Verordnungen (EU) 2016/679 und (EU) 2018/1725 sowie der Richtlinie (EU) 2016/680 die Gründe enthalten, warum die Verarbeitung besonderer Kategorien personenbezogener Daten **unbedingt erforderlich** war, um Verzerrungen aufzudecken und zu korrigieren, und warum dieses Ziel nicht durch die Verarbeitung anderer Daten erreicht werden konnte.*

Beschränkung auf Anbieter von Hochrisiko-KI-Systemen

- Anbieter nach Art. 3 Nr. 3 AI Act:

  *„Anbieter eine natürliche oder juristische Person, Behörde, Einrichtung oder sonstige Stelle, die ein KI-System oder ein KI-Modell mit allgemeinem Verwendungszweck **entwickelt oder entwickeln lässt** und es unter ihrem eigenen Namen oder ihrer Handelsmarke **in Verkehr bringt** oder das KI-System unter ihrem eigenen Namen oder ihrer Handelsmarke **in Betrieb nimmt**, sei es entgeltlich oder unentgeltlich.“*

- Die Vorschrift des Art. 10 Abs. 5 AI Act umfasst damit **keine Betreiber** (Art. 3 Nr. 4 AI Act) und keine KI-Systeme außerhalb des Hochrisiko-Bereichs
- Aber: Grds. auch außerhalb des Hochrisiko-Bereichs Bedürfnis für Debiasing von KI-Systemen

Zur Aufdeckung und Korrektur von Verzerrungen „unbedingt erforderlich“

- Prüfungsmaßstab im AI Act nicht weiter präzisiert, aber **hohe inhaltliche Hürde** („unbedingt“)
- Parallele zur **Erforderlichkeitsprüfung nach DSGVO**, hier aber tlw. andere Schutzrichtung des AI Act
- Überlegungen im Hinblick auf verwendete Bias-/Fairness-Metriken müssen umfassend **dokumentiert** werden (s. Art. 10 Abs. 5 S. 2 lit. f AI Act) -> insbesondere dahingehend, warum andere Daten als sensible Daten nach Art. 9 Abs. 1 DSGVO nach nicht ausreichend waren

Umfangreicher Pflichtenkatalog nach Art. 10 Abs. 5 S. 2 lit. a bis f AI Act

(a) *Die Aufdeckung und Korrektur von Verzerrungen kann nicht wirksam durch die Verarbeitung anderer Daten, einschließlich **synthetischer oder anonymisierter Daten**, erfolgen;*

(b) *die besonderen Kategorien personenbezogener Daten unterliegen technischen Beschränkungen der Weiterverwendung der personenbezogenen Daten und dem Stand der Technik entsprechenden Sicherheits- und Datenschutzmaßnahmen, einschließlich **Pseudonymisierung**;*

(c) *die besonderen Kategorien personenbezogener Daten Gegenstand von Maßnahmen sind, die gewährleisten, dass die verarbeiteten personenbezogenen Daten **gesichert und geschützt** sind und **geeigneten Garantien** unterliegen, einschließlich strenger Kontrollen und Dokumentation des Zugangs, um Missbrauch zu vermeiden und sicherzustellen, dass nur befugte Personen mit angemessenen Vertraulichkeitsverpflichtungen Zugang zu diesen personenbezogenen Daten haben;*

(d) *die besonderen Kategorien personenbezogener Daten dürfen nicht **an andere Parteien übermittelt, weitergegeben oder anderweitig zugänglich** gemacht werden;*

(e) *die besonderen Kategorien personenbezogener Daten werden **gelöscht, sobald die Verzerrung behoben** ist oder die Aufbewahrungsfrist der personenbezogenen Daten abgelaufen ist, je nachdem, was zuerst eintritt;*

(f) *die Aufzeichnungen von Verarbeitungstätigkeiten gemäß den Verordnungen (EU) 2016/679 und (EU) 2018/1725 sowie der Richtlinie (EU) 2016/680 die Gründe enthalten, warum die Verarbeitung besonderer Kategorien personenbezogener Daten **unbedingt erforderlich** war, um Verzerrungen aufzudecken und zu korrigieren, und warum dieses Ziel nicht durch die Verarbeitung anderer Daten erreicht werden konnte.*

FAZIT

Ausdifferenzierte Regelung – aber wohl kein umfassender „Durchbruch“

- Regelung zum Debiasing von KI-Systemen grundsätzlich zu begrüßen
- Wohl nur **begrenzte praktische Bedeutung** der Vorschrift
  - **Anwendungsbereich** beschränkt auf Anbieter von Hochrisiko-KI-Systemen
  - **„Unbedingt erforderlich“** stellt hohe Hürde dar und verlangt großen Begründungs- bzw. Dokumentationsaufwand
  - Pflichtenkatalog in Art. 10 Abs. 5 S. 2 lit. a bis f. AI Act ist **komplex** und seine Umsetzung wird für die Anbieter teilweise einen **hohen Aufwand** bedeuten
- Insgesamt kann daher wohl in Frage gestellt werden, ob die neue Vorschrift des Art. 10 Abs. 5 AI Act tatsächlich einen umfassenden Durchbruch für das Debiasing von KI-Systemen darstellt