h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

ATHENE
National Research Center
for Applied Cybersecurity

Diskriminierende KI: Ursachen und Lösungsansätze

Christian Rathgeb

Ringvorlesung Cybersicherheit

16.11.2023

Überblick

Agenda

1. Algorithmische Fairness
2. Gesichtserkennung
3. Diskriminierung am Beispiel Gesichtserkennung
4. Ursachen für Diskriminierung
5. Lösungsansätze
6. Zusammenfassung

Algorithmische Entscheidungsfindung

Kopplung von Daten mit Algorithmen (KI, maschinelles Lernen) zum Zweck datengesteuerter automatisierter Entscheidungen, z. B. biometrische Systeme

Vorteile:

- Konsistenz
- Verarbeitungsgeschwindigkeit
- Rationalität und auf Logik basierende Entscheidungen
- etc.

Potentielle Nachteile:

- Genauigkeit
- Erklärbarkeit
- **Fairness**

Algorithmische Fairness

Diskriminierung

Benachteiligung von Menschen aufgrund eines unrechtmäßigen Merkmals, wie zum Beispiel [1]:

- Hautfarbe,
- Geschlechts,
- sexuellen Orientierung und/oder Identität,
- ihrer Religion,
- ihres Alters,
- der Staatsangehörigkeit,
- der Sprache,
- etc.


The Guardian

Facial recognition bias frustrates Black asylum applicants to US, advocates say


The New York Times

Many Facial-Recognition Systems Are Biased, Says U.S. Study

Algorithms falsely identified African-American and Asian faces 10 to 100 times more than Caucasian faces, researchers for the National Institute of Standards and Technology found.

Prominentes Beispiel:

- **Gesichtserkennung** (insbesondere Strafverfolgung)

[1] Amnesty International, Definition: Was ist Diskriminierung.

https://www.amnesty.ch/de/themen/diskriminierung/zahlen-fakten-und-hintergruende/was-ist-diskriminierung#

Fairness

Definition von Fairness:

„the quality of treating people equally or in a way that is right or reasonable" [2]

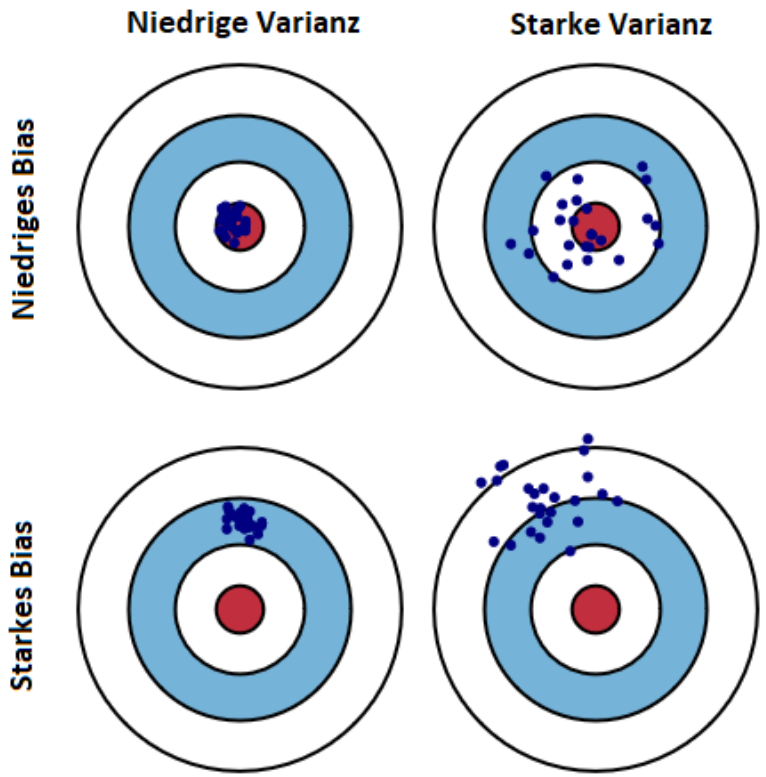
- Ethisches und soziales Konzept
- Beeinflusst durch kulturelle, historische, rechtliche, religiöse, und andere Faktoren
- Keine einzige Vorstellung oder Definition von Fairness in der Praxis
- Mindestens 30 verschiedene Definitionen von algorithmischer Fairness in der Literatur
- Hinsichtlich Diskriminierung meist Bezug auf Gruppen Fairness

[2] https://dictionary.cambridge.org/dictionary/english/fairness

Statistischer Bias

Systematischer Fehler der mathematisch gut definiert werden kann

- Trade-Off zwischen Varianz und Bias
- Vorhersagefehler können in zwei Hauptunterkomponenten zerlegt werden,
  - Fehler aufgrund von Bias und
  - Fehler aufgrund von Varianz
- **Herausforderung**: Bias in den Daten
- Kein statistisches Bias kann als notwendige (aber nicht hinreichende) Voraussetzung für Fairness gesehen werden



Gesichtserkennung

Biometrie

ISO-Definition des Begriffs Biometrie:

"Automatisierte Erkennung von Individuen anhand deren Verhalten und ihrer biologischen Charakteristika."

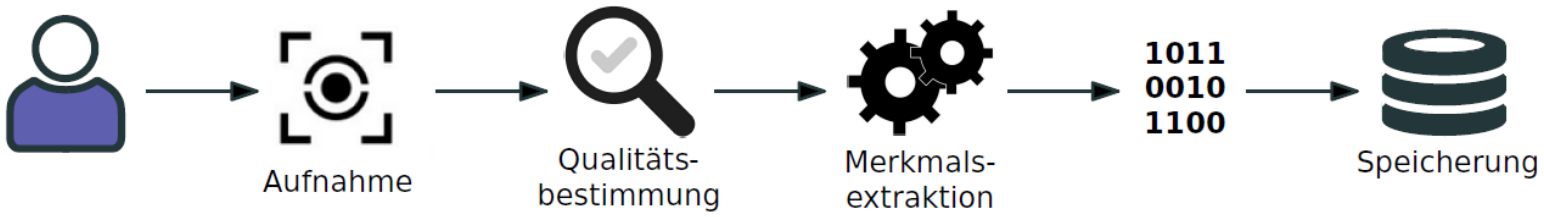
Es gibt verschiedene Anwendungsfelder und Ziele der Biometrie :

- **Benutzerfreundlichkeit**, z. B. Entsperren von Smartphones
- **Zugangskontrolle**, z. B. automatisierte Grenzkontrolle
- **Forensik**, z. B. Identifikation von Personen

Gesichtserkennung

Funktionsweise

Generische Funktionsweise einer biometrischen Registrierung (Enrolment):



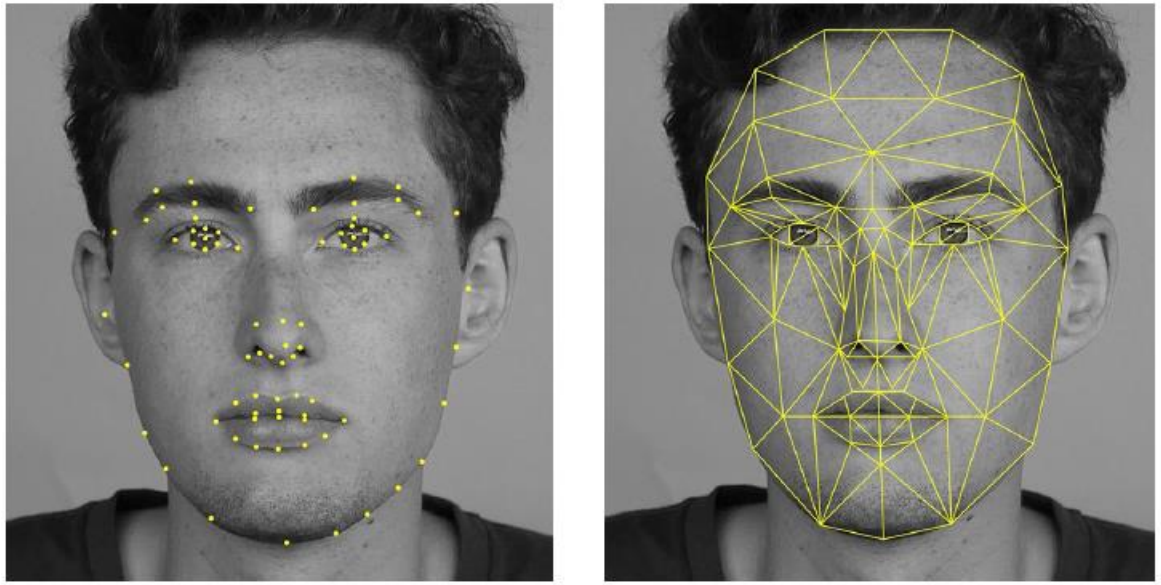
1. **Aufnahme**: präsentierte biometrische Charakteristik (Gesicht) wird durch einen (biometrischen) Sensor aufgenommen
2. **Qualitätsbestimmung**: Prüfung der Eignung der aufgenommenen biometrischen Daten zum Zweck der Wiedererkennung
3. **Merkmalsextraktion**: Extrahierung eines Merkmalsvektors
4. **Speicherung**: Speicherung des Merkmalsvektor, z. B. auf Smartphone

Bei der Registrierung gespeicherte Merkmalsvektoren bezeichnet man als **Referenzen**.

Gesichtserkennung

Schlüsselfunktion Merkmalsextraktion

Ziel ist es **diskriminative** und **robuste** Merkmale zu extrahieren, z. B. Landmarken bei Gesichtserkennung



Stand der Technik: auf maschinellen Lernverfahren basierte Methoden, insbesondere tiefe neuronale Netze, die mit großen Datenbanken trainiert werden (Deep Face Recognition)

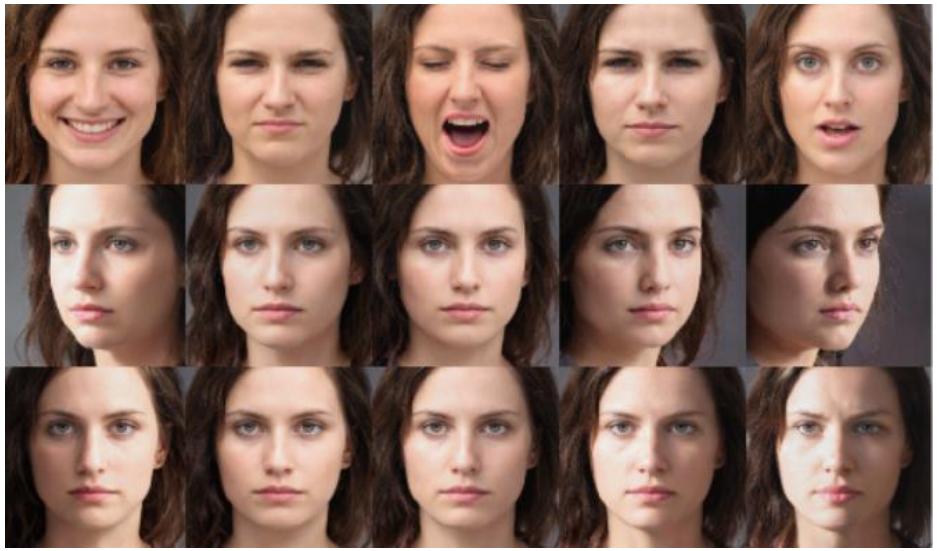
Gesichtserkennung

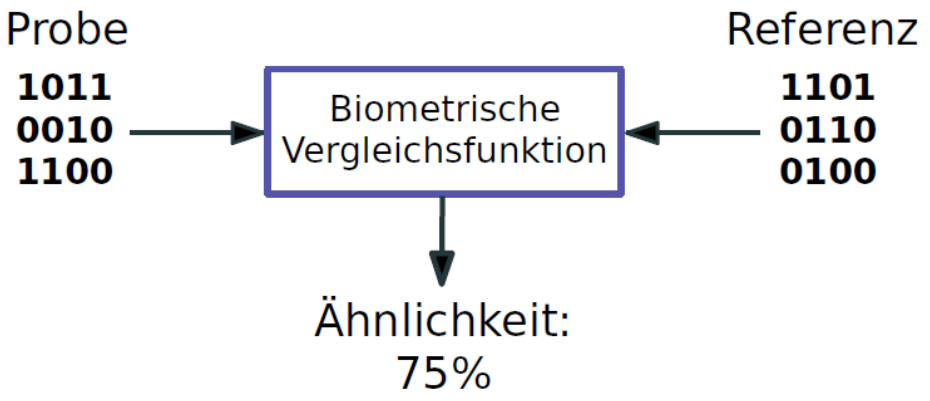
Biometrische Varianz und Vergleiche

Die Robustheit der Merkmalsextraktion ist wichtig, da biometrische Messungen eine gewisse **Intra-Klassen Varianz** aufweisen.

Bei einem **Authentisierungsversuch** wird das Gesicht erneut aufgenommen und analog zur Registrierung verarbeitet (biometrische Probe)

Ziel eines biometrischen Vergleichsfunktion ist es einen **Ähnlichkeitswert** zwischen einer biometrischen **Referenz und einer Probe** zu bestimmen





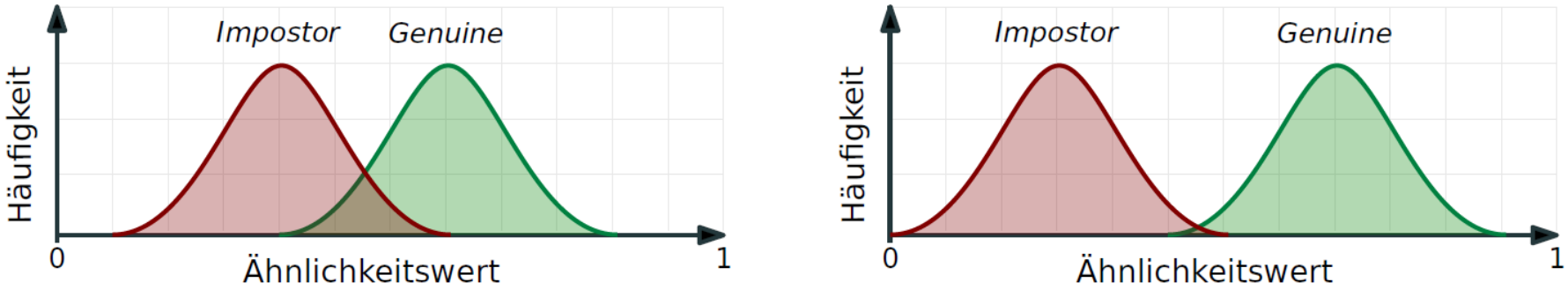
Gesichtserkennung

Verteilung von Vergleichswerten

Zwei Arten von Vergleichen zw. biometrischen Referenzen und Proben als Test der biometrischen Erkennungsleistung:

- **Genuine Vergleich**: selbe Person
- **Impostor Vergleich**: verschiedene Personen

Beispiele für Verteilungen von Genuine und Impostor Ähnlichkeitswerten:



Schwellwert bestimmt ob eine Probe und eine Referenz übereinstimmen (Match)

Arten von Fehler

In biometrischen Systemen können zwei (algorithmische) Arten von Fehler auftreten:

- **False Reject**: ein Genuine Vergleich wird fälschlicher Weise abgelehnt
- **False Accept**: ein Imposter Vergleich wird fälschlicher Weise akzeptiert

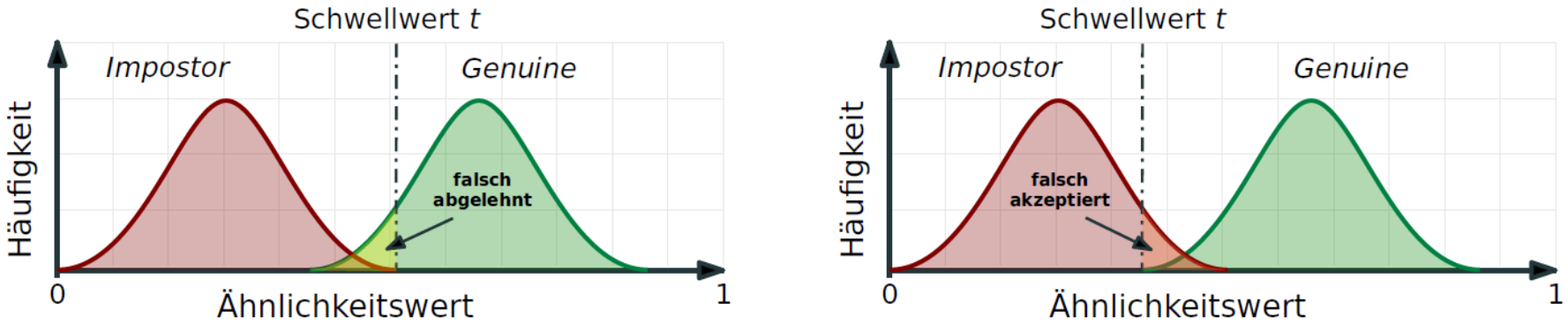
Solche Fehler treten mit gewisser Wahrscheinlichkeit auf (abhängig von dem verwendeten Schwellwert)

Beispiele:



Erkennungsleistung

Ein Schwellwert *t* wird gesetzt um zwischen Genuine und Impostor Vergleichen zu unterscheiden.



Dadurch ergeben sich zwei Fehlerraten:

- **False Reject Rate**: *FRR(t)* = Anteil von Genuine Ähnlichkeinstwerten kleiner *t*
- **False Accept Rate**: *FAR(t)* = Anteil von Impostor Ähnlichkeinstwerten größer *t*

FRR ist eine Maß für die Benutzbarkeit und FAR ein Maß für die Sicherheit

Gesichtserkennung

Betriebsarten

In der Biometrie existieren zwei Operationsmodi:

Verifikation:

- Es liegt eine Identitätsbehauptung vor - Vergleich einer Probe mit einer Referenz (1:1 Vergleich)
- **Beispiele**: automatisierte Grenzkontrolle, Entsperren eines Smartphones

Identifikation:

- Es liegt keine Identitätsbehauptung vor - Vergleich einer Probe mit meherern Referenzen (1:*N* Vergleich)
- **Beispiel**: Abgleich eines Gesichtsbildes im Zuge einer forensischen Untersuchung

Diskriminierung & Gesichtserkennung

Erkennungsleistung

Fokus liegt in der Regel auf zwei demographischen Eigenschaften:

- Geschlecht
- Hautfarbe (Ethnie)

In einem fairen System sollten die Wahrscheinlichkeit für algorithmische Fehler für verschiedene demographische Gruppen gleich sein

Konsequenzen falscher Entscheidungen:

Fehler	Verifikation	Identifikation
False Rejection	Unannehmlichkeit	Entgangene Spur
False Acceptance	Sicherheitsrisiko	Falsche Spur

Diskriminierung & Gesichtserkennung

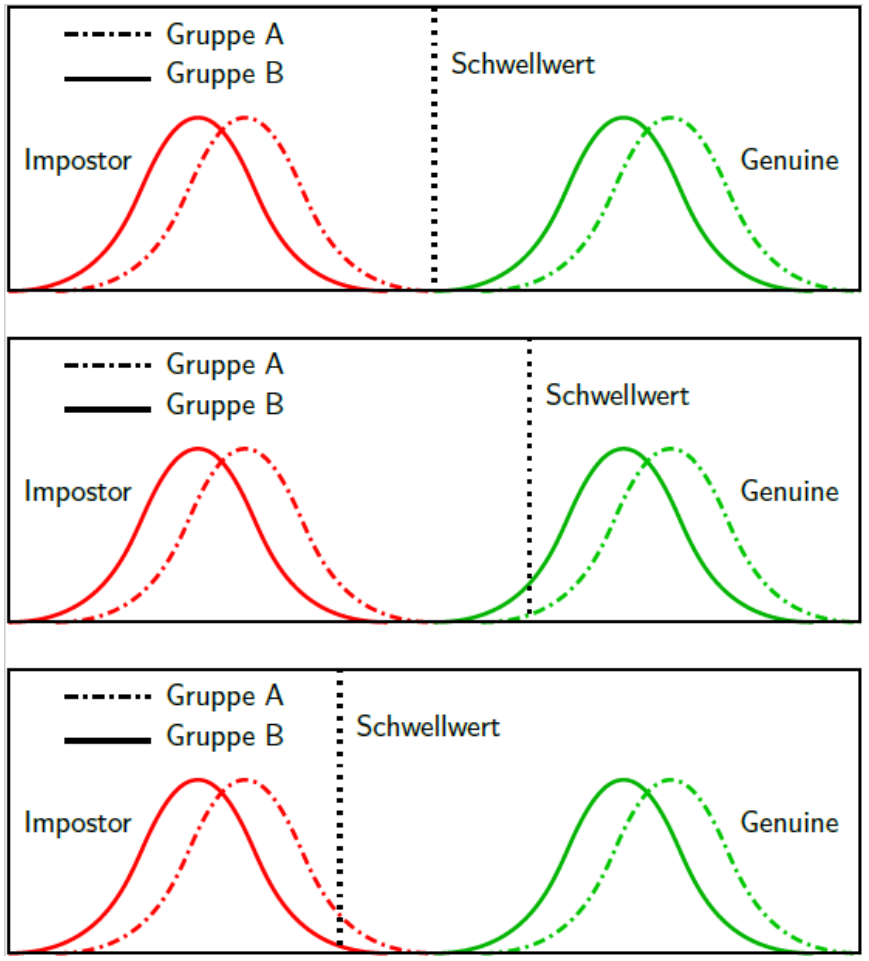
Demographisch unterschiedliche Performanz und Ergebnisse

Demographisch unterschiedliche Performanz beschreibt statistisch unterschiedliche Ergebnisse (Verteilung von Vergleichswerten)

Demographisch unterschiedliche Ergebnisse treten auf, wenn bei einem Schwellwert die demographisch unterschiedliche Performanz zu Unterschieden in den Fehlerraten (FRR, FAR) führen

Demographische Fairness (bzw. **Diskriminierung**) bezieht sich auf die Konsequenzen unterschiedlicher Entscheidungen

Es gibt verschiedene Metriken um Fairness zu quantifizieren [3]



[3] ISO/IEC 19795-10 „Quantifying biometric system performance variation accross demographic groups“

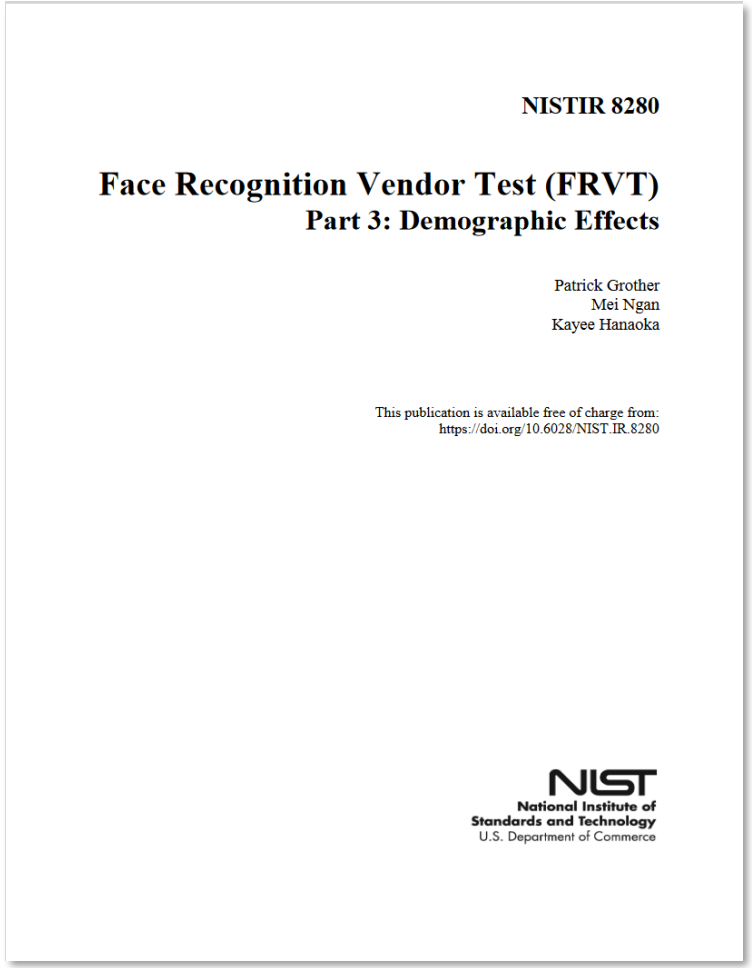
Diskriminierung & Gesichtserkennung

Wissenschaftliche Untersuchungen

Es gibt viele (oft zitierte) **fehlerhafte Studien** zum Thema Fairness und Gesichtserkennung

Erster relevante größere Untersuchung durch NIST (National Institute of Standards and Technology) basierend auf Einreisedaten der USA [4]

- Subjekte aus 24 Länder in 7 verschiedenen globalen Regionen
- Insgesamt 18,27 Millionen Bilder von 8,49 Millionen Menschen
- 189 meist kommerzielle Algorithmen von 99 Entwicklern


NISTIR 8280

Face Recognition Vendor Test (FRVT)
Part 3: Demographic Effects

Patrick Grother
Mei Ngan
Kayee Hanaoka

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8280

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

[4] Grother et al. „Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects“, 2019.

Diskriminierung & Gesichtserkennung

Wissenschaftliche Ergebnisse

Wichtigste Erkenntnisse der NIST Evaluierung:

- **Ethnie/Hautfarbe**:
  - Höhere Fehlerraten für afrikanische und asiatische Subjekte
  - Geringste Fehlerraten für europäische Subjekte
  - **ABER**: Umgekehrter Effekt für in China entwickelte Gesichtserkennungsalgorithmen
- **Geschlecht**: FAR bei Frauen ist höher als bei Männern für alle Algorithmen (Effekt geringer als bei Ethnie)
- **Alter**: erhöhte FAR bei älteren Menschen und Kindern (am geringsten bei Erwachsenen mittleren Alters)

Je genauer der Algorithmus desto geringer die demographisch unterschiedliche Performanz!

Ursachen für Diskriminierung

Qualität der Testdaten

Qualität der Testdaten ist unterschiedlich für demographische Gruppen, z. B. wegen regional unterschiedlichen Aufnahmestandards

Qualitativ schlechtere Daten führen zu **schlechterer Performanz** die als Diskriminierung fehlinterpretiert werden kann

Einflussfaktoren für Gesichtsbilder:

- Beleuchtung
- Kompression
- Auflösung

Performanz wird durch **andere Faktoren** beeinflusst, z. B. Gesichtsbehaarung [5]

[5] Bhatta et al. „The Gender Gap in Face Recognition Accuracy Is a Hairy Problem", 2022

Ursachen für Diskriminierung

Bias in Trainingsdaten

Demographische Gruppen die in den Trainingsdaten **unterrepräsentiert** sind werden benachteiligt

Häufiges Problem: Neuronale Netze sind „Daten-hungrig“ und Algorithmen werden oft mit Bildern aus dem Web trainiert (meist Bilder von berühmten Personen)

Auch: Dezidierte Entwicklung eines Algorithmus für eine demographische Gruppe:

- Regionaler Markt, z. B. Asian
- Spezielle Applikation, z. B. Grenzkontrolle

Ursachen für Diskriminierung

Bias in den Testdaten

Watchlist Inbalance Effect: Diskriminierung kann bei der biometrischen Identifikation (1:N) auftreten, wenn die Datenbank in der gesucht wird eine unausgewogene demographische Verteilung aufweist

Verwechslungen (False Accept) bzw. hohe Ähnlichkeiten treten meist **innerhalb einer demographischen Gruppe** auf!

A priori Wahrscheinlichkeit für eine **falsche Identifikation** ist dadurch für Subjekt der demographischen Gruppe am höchsten die in der Datenbank am meisten vertreten ist [6]

[6] Drozdowski et al. „The Watchlist Imbalance Effect in Biometric Face Identification", 2021.

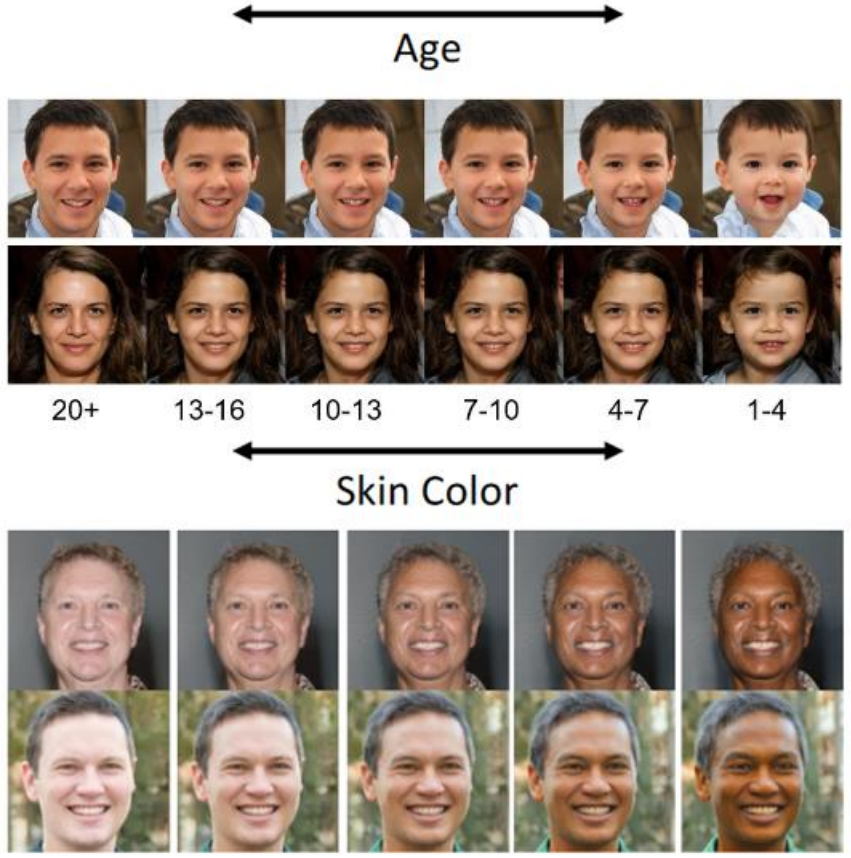
Lösungsansätze

Qualität der Testdaten

Kontrollierte Erstellung von Testdaten:

- Aufnahme in **kontrollierter Umgebung** (Kosten, Zeit)
- Verwendung (semi-)**synthetischer Daten** [7] [8]

Synthetische Gesichtsbilder können mittels **generativer Netzwerke** auf kontrollierte Weise erstellt werden!



[7] Balakrishnan et al. „Towards Causal Benchmarking of Bias in Face Analysis Algorithms", 2020

[8] Falkenberg et al. „Child Face Recognition at Scale: Synthetic DataGeneration and Performance Benchmark", 2023

Lösungsansätze

Bias in den Trainingsdaten

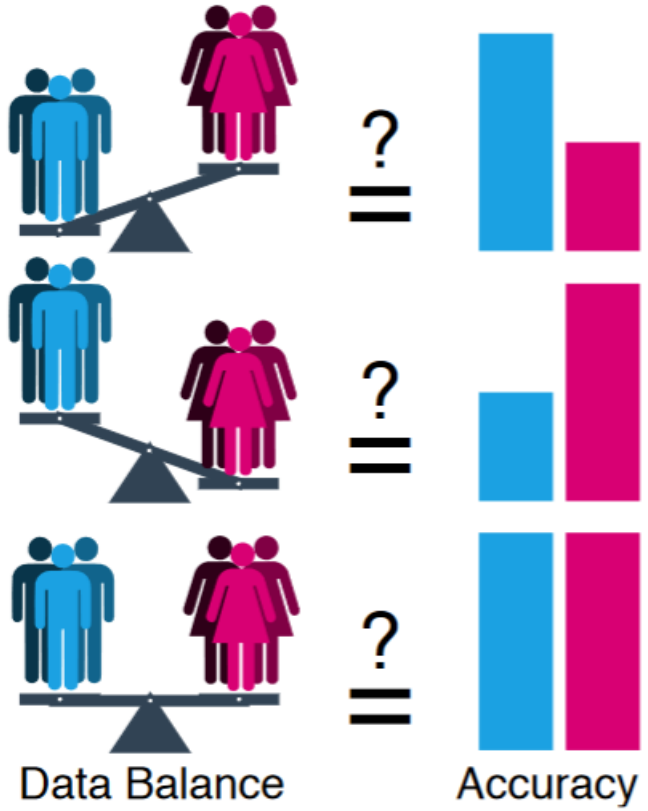
Viele Bemühungen zur Erstellung „**fairer Trainingsdaten**“

Gleichgewicht der Gruppen in den Trainingsdaten führt nicht zwingend zu Gleichgewicht in der Erkennungsleistung

Diskriminierung in der Testgenauigkeit wird meist **nicht** durch Gleichgewicht der Gruppen in den Trainingsdaten minimiert

Minimierung der Diskriminierung in der Erkennungsleistung kann **negative Auswirkung auf die Erkennungsleistung** aller Gruppen haben

[7] Albiero et al. „How Does Gender Balance In Training Data Affect Face Recognition Accuracy?”, 2020



Lösungsansätze

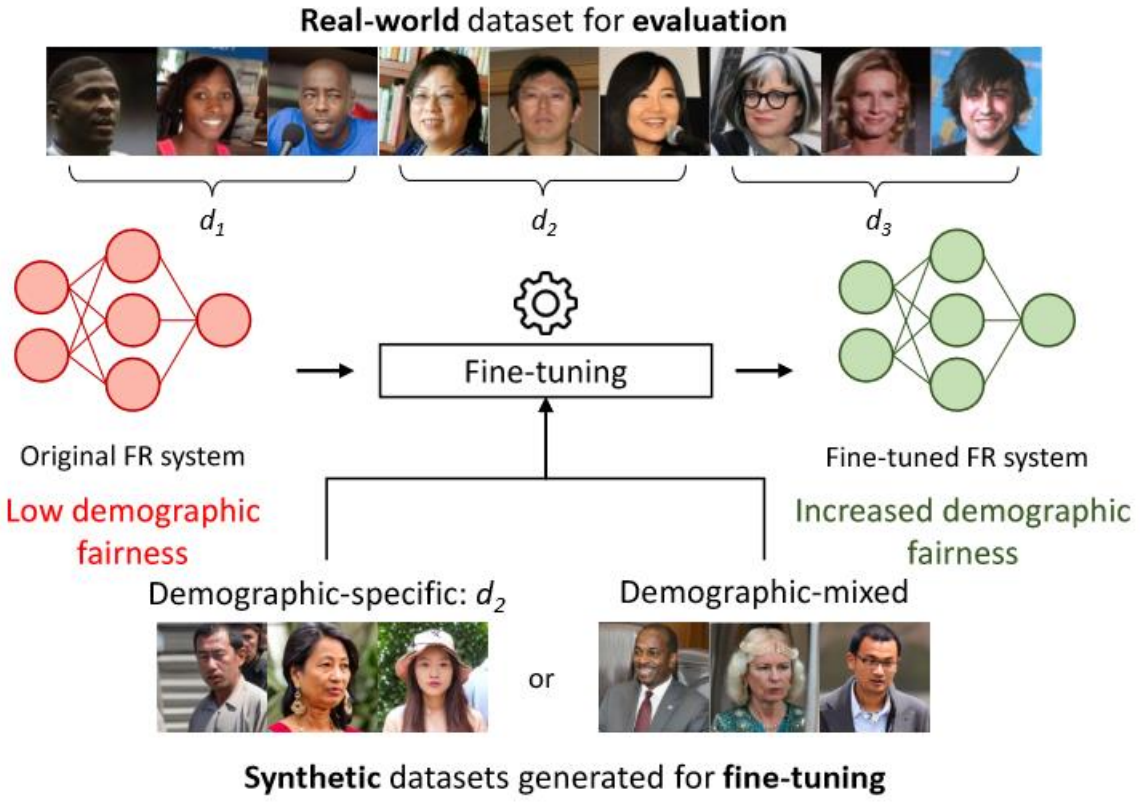
Bias in den Trainingsdaten

In vielen Fällen sind Trainingsdaten die zu einem faireren System führen **praktisch schwer zu beschaffen**

Hinzufügen von **realistische synthetische Daten** zum Training kann helfen Diskriminierung zu reduzieren

Bereits trainierte Netzwerke können **angepasst** werden (Fine-Tuning)

[7] Melzi et al. „Synthetic Data for the Mitigation of Demographic Biases in Face Recognition", 2023



Lösungsansätze

Bias in den Testdaten

Offenes Forschungsproblem: Bias in den Testdaten ist **schwer zu beheben** ohne weitere Diskriminierung einzuführen (z. B. unterschiedliche Schwellwerte für demographische Gruppen)

Verbesserung der Erkennungsleistung kann diese Art der Diskriminierung minimieren

Faire Ausgangsituation für Identifikation schaffen: Hinzufügen von (synthetischen) Daten, sodass die Anzahl aller demographischer Gruppen mit gleichem Anteil vertreten sind; wichtige Faktoren:

- Qualität der hinzugefügten Daten
- Möglicher Bias aus Verifikation

Zusammenfassung

Wichtigste Erkenntnisse

KI Systeme sind anfällig dafür **Vorurteile zu reproduzieren** wenn diese sich in den Trainingsdaten wiederspiegeln

Die Quantifizierung möglicher Diskriminierung erfordert geeignete **Metriken und Testdaten** (Standards)

Minimierung von Diskriminierung erfordert oft **komplexe Strategien** (z. B. führen faire Trainingsdaten nicht zwangsläufig zu fairen Algorithmen)

Durch die **Verbesserung von KI-Systemen** hat in der Regel einen **positiven Effekt bzgl. möglicher Diskriminierung**

Viele **Forschungsaktivitäten** um die Fairness von KI-Systemen zu verbessern

Ende

Vielen Dank für die Aufmerksamkeit!