Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Bachelorarbeit

Dominik Wachter

Diskriminierung durch KI - Ursachen und Prüfansätze am Beispiel Strafverfolgung

Fakultät Technik und Informatik
Studiendepartment Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Dominik Wachter

Diskriminierung durch KI - Ursachen und Prüfansätze am Beispiel Strafverfolgung

Bachelorarbeit eingereicht im Rahmen der Bachelorprüfung

im Studiengang Bachelor of Science Technische Informatik
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. Bettina Buth
Zweitgutachter: Prof. Dr. Kai von Luck

Eingereicht am: 24. Mai 2022

Dominik Wachter

Thema der Arbeit
Diskriminierung durch KI - Ursachen und Prüfansätze am Beispiel Strafverfolgung

Stichworte
KI, Ethik, Diskriminierung, Strafverfolgung, Explainable AI, XAI, Interpretierbarkeit

Kurzzusammenfassung
Diese Literaturarbeit setzt sich mit den negativen Auswirkungen von KI auseinander. Anhand von Anwendungsbeispielen in der Strafverfolgung, wie z.B. Facial Recognition Technology oder Predictive Policing, wird nach Belegen für Diskriminierung gesucht und deren Ursachen herausgearbeitet. Aus den anwendungsbezogenen Erkenntnissen werden anschließend allgemeine Aussagen zu den Ursachen von Diskriminierung abgeleitet. Ergänzend dazu wird ein Explainable AI Ansatz vorgestellt, der bei der Aufdeckung von Diskriminierung helfen kann.

Dominik Wachter

Title of the paper
Discrimination by AI - causes and testing methods in the field of law enforcement

Keywords
AI, ethics, discrimination, law enforcement, explainable ai, XAI

Abstract
This literature research is discussing the negative consequences of AI. On the basis of applications in the field of law enforcement, e.g. facial recognition technology or predictive policing, research is done for proving discrimination and its causes. These practice-related findings will be used to make genereal conclusions about the causes of discrimination. Additionally, there will be an explainable ai method introduced which can be used to test for discrimination.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

CNN Deep Convolutional Neuronal Network.

FAC Facial Attribute Classification.

FDA U.S. Food and Drug Administration.

FR Facial Recognition.

FRT Facial Recognition Technology.

GSM Global Surrogate Model.

KI Künstliche Intelligenz.

ML Machine Learning.

NIST National Institute for Standards and Technology.

XAI Explainable AI.

1 Einleitung

In der folgenden Einführung wird die Motivation für die Arbeit erläutert und die daraus abgeleiteten Forschungsfragen benannt. Anschließend erfolgt ein Überblick über den Aufbau der Arbeit.

1.1 Motivation

Der Einsatz von KI (Künstliche Intelligenz) ist in der heutigen Zeit kaum noch wegzudenken. Jede Google-Suchanfrage, Werbung, vorgeschlagene Youtube Videos, Spam Filter oder automatisierte Kundenbetreuung durch Chatbots werden durch KIs realisiert. Mittlerweile sind KIs in vielen Disziplinen dem Menschen überlegen, wie z.B. die KI OpenAI Five, die die Weltmeister OG im Spiel Dota geschlagen hat (OpenAI 2019).

KIs haben zudem einen immer weitreichenderen Einfluss auf die Gesellschaft durch Anwendungen wie z.B. automatisierte Bonitätsprüfungen, autonomes Fahren, Optimierung von Unternehmen durch automatisierte Bewerbungsprozesse oder KI-gestützte Strafverfolgung. In China werden schon jetzt automatisierte Gerichtsverfahren für einfache Delikte ausgesprochen, ohne dass die Fälle jemals von einem/einer Richter*in geprüft werden (Chen 2021). Gesichtserkennung wird zunehmend für die Strafverfolgung genutzt, wie Buzzfeed News (2021) anhand der Nutzungshäufigkeit von ClearviewAIs FRT (Facial Recognition Technology) herausfindet. Durch die einfache Handhabung der Software wird FRT zudem nicht nur in schwerwiegenden Fällen genutzt, sondern auch für einfache Straftaten, was u.A. von NBC News (2019) kritisiert wird. Da die Technologie nicht als Beweismaterial genutzt wird, bleibt die Nutzung häufig unerkannt. Zudem existieren wenig regulierende Gesetzte, die definieren, wann die Polizei FRT nutzen darf, welche Personen in der Datenbank zu finden sind, wie präzise die Systeme identifizieren sollen oder in welchen Fällen die Nutzung der Systeme publik gemacht werden muss (NBC News 2019). Diese Entwicklung soll als Motivation genutzt werden, sich näher mit derartigen Technologien auseinanderzusetzen und die negativen Folgen zu untersuchen.

KI bedeutet, automatisiert Entscheidungen zu treffen, mit variablem Einfluss durch den Menschen. Da die automatisierten Entscheidungen häufig für alle Teile der Gesellschaft gleichermaßen gelten, stellen sich einige Fragen:

1. Entscheiden KIs frei von Vorurteilen oder reproduzieren sie vorherrschende Diskriminierung?
2. Welche Ursachen hat Diskriminierung, bedingt durch die Entwicklung oder Anwendung von KI?
3. Kann Diskriminierung aufgedeckt werden und wenn ja, welche Methoden gibt es?

Dies sind zentrale Fragestellungen dieser Arbeit. Sie sollen anhand von praktischen Anwendungen in der Strafverfolgung untersucht werden. Unter Strafverfolgung werden alle Maßnahmen verstanden, die für die Aufdeckung und Verfolgung von Straftaten genutzt werden können. Der Bereich wurde ausgesucht, da er durch die große Entscheidungsgewalt besonders wichtig erscheint und Ungerechtigkeit vermutet wird. Die daraus hervorgehenden Erkenntnisse sollen anschließend für allgemeingültige Aussagen über die Ursachen von Diskriminierung durch KIs genutzt werden.

1.2 Übersicht über die Arbeit

In Kapitel 2 sollen die Grundlagen erläutert werden, die für das Verständnis der Arbeit notwendig sind.

In Kapitel 3 werden Anwendungsbeispiele von KI im Bereich der Strafverfolgung vorgestellt und dabei Ursachen von Diskriminierung herausgearbeitet. Von den gefundenen Ursachen werden allgemeine Aussagen über KI abgeleitet.

In Kapitel 4 wird die Analyse der Software „COMPAS“ zur Risikobewertung von Strafttäter*innen von Angwin et al. (2016) vorgestellt. Anschließend werden die Erkenntnisse der Analyse mit den Erkenntnissen aus Kapitel 3 verglichen. Zudem wird die Übertragbarkeit der Analyse von Angwin et al. (2016) auf weitere Anwendungen bewertet.

Im letzten Kapitel 5 werden die Erkenntnisse der Arbeit zusammengetragen und bewertet. Zudem wird auf weitere, an diese Arbeit anschließende, Fragestellungen eingegangen.

2 Einführung in KI und Ethik

In diesem Kapitel sollen die Grundlagen erläutert werden, die für das Verständnis der Arbeit notwendig sind. Es wird zunächst ein Überblick der verschiedenen Arten von KI gegeben und die Verwendung dieses Begriffes in der Arbeit definiert. Im Anschluss soll geklärt werden, was unter ethischer KI verstanden wird. Hierfür werden unterschiedliche Leitsätze herangezogen, u.A. von der EU Komission (2018). Anschließend soll darauf eingegangen werden, wie die Qualität und Fairness von KI gemessen werden kann und wie *Transparenz*, *Interpretierbarkeit* und *Erklärbarkeit* zu verstehen sind. Wie die Interpretierbarkeit von KI verbessert werden kann, wird im Kontext von XAI geklärt. Damit klar ist, wie der Begriff Diskriminierung im Kontext der Arbeit verwendet wird, muss dieser ebenfalls im Anschluss definiert werden. Abschließend sollen unterschiedliche Arten von Bias vorgestellt werden, die für eine differenziertere Untersuchung der Ursachen von Diskriminierung notwendig sind.

2.1 KI

KI steht für *Künstliche Intelligenz* und wird von Deutscher Bundestag (2020) wie folgt beschrieben:

„KI-Systeme sind von Menschen konzipierte, aus Hardware- und/oder Softwarekomponenten bestehende intelligente Systeme, die zum Ziel haben, komplexe Probleme und Aufgaben in Interaktion mit der und für die digitale oder physische Welt zu lösen.“

Daraus lässt sich ablesen, dass KI in erster Linie eine *intelligente* Technologie darstellt, welche von Menschen für Menschen entwickelt wird und Problemstellungen im Zusammenspiel mit der Umwelt bewältigt. Im Rahmen der Arbeit wird der Begriff „KI“ nach dieser abstrakten Beschreibung verwendet. Die im Folgenden genannte Einteilung von KI Systemen wurde ebenfalls von Deutscher Bundestag (2020) entnommen:

KI Systeme können in zwei Arten aufgeteilt werden:

- **Regelbasierte KI-Systeme**: Systeme, die vollständig durch algorithmische Regeln und maschinenlesbares Wissen von Menschen definiert werden
- **Lernende KI-Systeme**: Systeme, die initial durch den Menschen konfiguriert werden und anschließend anhand von Daten lernen, wie sie ein Problem lösen können

Im Rahmen der Arbeit werden nur lernende KI-Systeme betrachtet und die Begriffe im Folgenden synonym verwendet. Weitere Begriffe, die für lernende KI-Systeme verwendet werden, sind KI-Modelle, ML-Algorithmen (Machine Learning oder Maschinelles Lernen) oder ML-Modelle.

Zu Beginn erhält das ML-Modell durch den Menschen initiale Werte für eine Menge von Parametern. Die Werte werden während des Trainings angepasst. Während des Trainings werden dem Modell nacheinander Beispiele gezeigt und die Parameter angepasst.

ML kann eingeteilt werden in drei Arten des Lernens:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Beim Supervised Learning werden sogenannte *gelabelte* Daten verwendet. Labels, von Deutscher Bundestag (2020) Annotationen genannt, sind Informationen zu dem jeweiligen Datum. Für die FRT, die in Abschn. 3.2 näher behandelt wird, können das z.B. beschreibende Attribute, wie Haarfarbe und Augenfarbe der Person sein, die auf einem Bild zu sehen sind. Das Lernen nach diesem Verfahren ist besonders effizient, aber auch aufwendig, da die Labels häufig von Menschen vergeben werden müssen Deutscher Bundestag (2020).

In Abb. 2.1 sind die verschiedenen Lernverfahren (bzw. Lernstile), die Lernaufgabe, sowie Beispiele für konkrete Lernverfahren und ML-Modelle zu sehen. Das Supervised Learning wird demnach in zwei Lernaufgaben aufgeteilt: Klassifikation und Regression. Bei der Klassifikation werden Daten durch die KI klassifiziert. Als Beispiel sei hier die Risikobewertungssoftware „COMPAS“ genannt: Sie klassifiziert angeklagte Straftäter*innen in unterschiedliche Risikostufen. Der Ausgang der Klassifizierung ist somit eine Klasse und kann z.B. durch eine *logistische Regression* erfolgen. Sie wird im nächsten Abschn. 2.1.1 genauer beschrieben, da sie später in der Analyse von „COMPAS“ in Kapitel 4 verwendet wird. Wie in Abb. 2.2 zu sehen, ist die logistische Regression mit 63.5% das am Meisten genutzte ML-Verfahren der Befragten von Kaggle. Auch der Entscheidungsbaum mit 49.9% und eine Weiterentwicklung dessen, die

Lernstil	Lernaufgabe	Lernverfahren	Modell
Überwacht	Regression	Lineare Regression	Regressionsgerade
		Klassifikations- und Regressionsbaum-verfahren (CART)	Regressionsbaum
	Klassifikation	Logistische Regression	Trennlinie
		Iterative Dichotomizer (ID3)	Entscheidungsbaum
		Stützvektormaschine (SVM)	Hyperebene
		Bayessche Inferenz	Bayessche Modelle
Unüberwacht	Clustering	K-Means	Clustermittelpunkte
	Dimensionsreduktion	Kernel Principal Component Analysis (PCA)	Zusammengesetzte Merkmale
Bestärkend	Sequentielles Entscheiden	Q-Lernen	Strategien
Verschiedene	Verschiedene	Rückwärtspropagierung	Künstliche Neuronale Netze

Abbildung 2.1: Einteilung der KI Modelle (Döbel et al. 2018)

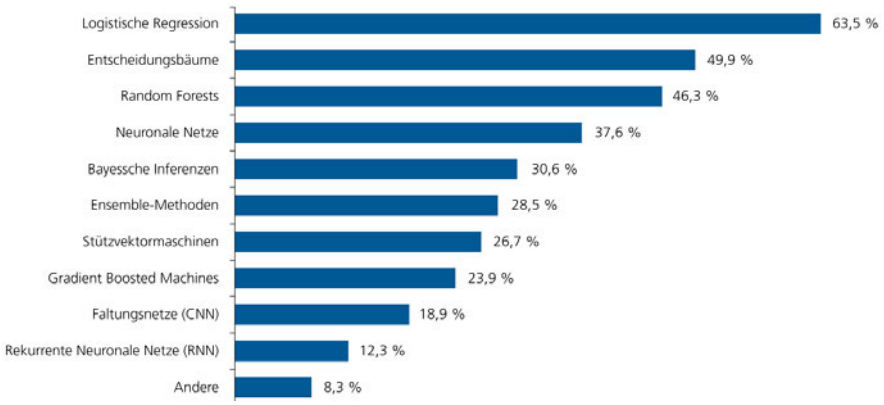


Abbildung 2.2: Verteilung der angewendeten KI Modelle von Kaggle (Döbel et al. 2018)

Random Forests, werden mit 46.3% für Klassifizierungsprobleme genutzt. So handelt es sich auch bei dem Großteil der Anwendungen in dieser Arbeit um Klassifizierungsprobleme.

Neben der Klassifizierung ist eine weitere Lernaufgabe des Supervised Learnings die Regression. Der Ausgang ist hier ein absoluter Wert. Dies kann z.B. ein berechneter Aktienindex sein oder die Vorhersage der Temperatur durch eine Wetterprognose. Der Unterschied zwischen der Regression und Klassifizierung wird in Abschn. 2.1.1 noch tiefgreifender beschrieben.

Beim Supervised Learning wird dem Modell durch die Labels vorgegeben, was es lernen soll. Anders ist das beim Unsupervised Learning: Hier erkennen die ML-Modelle selbstständig Regeln und können z.B. Gruppen von ähnlichen Daten erkennen. Dies wird als Clustering bezeichnet. Zudem werden sie für die Dimensionsreduktion verwendet, bei der uninteressante

Informationen der Daten herausgefiltert werden. Bei dieser Art des Lernens werden keine Labels benötigt. Unsupervised Learning wird häufig in Kombination mit Supervised Learning genutzt (Deutscher Bundestag 2020) und auch Semi-Supervised Learning genannt.

Beim Reinforcement Learning (Bestärkendem Lernen) entscheidet das ML-Modell selbstständig, ohne dass dem Modell explizit gesagt wird, welche Entscheidung richtig ist. Es benötigt also, wie auch beim Unsupervised Learning, keine Daten mit der vorgegebenen Lösung. Stattdessen werden durch vorgegebene Belohnungs- und Bestrafungsmechanismen selbstständig Regeln erlernt. Ein Beispiel ist die KI „AlphaGO“ von Google: Sie ist die erste KI, die es schaffte, die besten Spieler des chinesischen Brettspiels „GO“ zu schlagen. Es galt lange Zeit als eines der schwierigsten Probleme der KI Geschichte (Li & Du 2018).

Häufig wird im Zusammenhang mit KI von Neuronalen Netzen gesprochen. Wie in Abb. 2.1 zu sehen, sind künstliche Neuronale Netze, auch tiefe Neuronale Netze oder Deep Learning genannt, keinem Lernstil zugeordnet, da sie überall verwendet werden können. Die Funktionsweise eines Feed-forward Networks ist in Abb. 2.3 schematisch dargestellt: Bei einem Feed-forward Netz werden Informationen in eine Richtung weitergegeben (in der Abb. von links nach rechts). Links in der Eingabeschicht befinden sich die Features der Daten. Features sind Eigenschaften, die in den Daten erkannt werden sollen. Für eine Berechnung der Schadensklasse einer Versicherung könnten Features bspw. „Alter“ und „Wohnort“ sein. Die Attribute werden mit Gewichten versehen und in der Zwischenschicht aufsummiert. In der Praxis können Neuronale Netze über sehr viele Zwischenschichten verfügen, sind in der Abbildung aber nur vereinfacht mit einer Schicht dargestellt. In der Ausgabeschicht wird dann das Ergebnis ausgegeben, in dem genannten Beispiel die „Schadensklasse“.

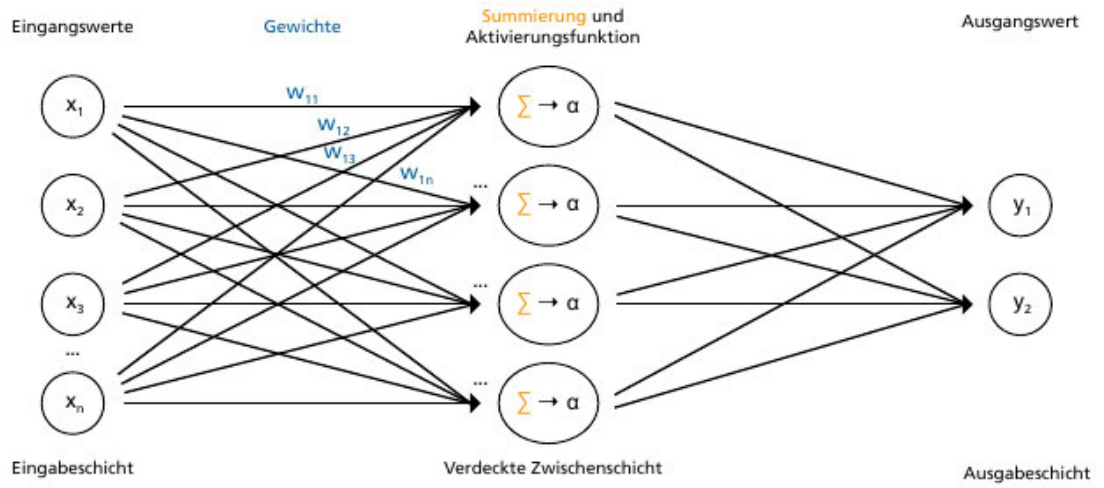


Abbildung 2.3: Funktionsweise eines Feed-forward Networks (Döbel et al. 2018)

Nachdem in diesem Abschnitt ein Überblick über die verschiedenen Arten des Lernens und ML-Modelle gegeben wurde, soll die logistische Regression im nächsten Abschnitt genauer erklärt werden.

2.1.1 Logistische Regression

In der in Kapitel 4.2 vorgestellten Analyse von Angwin et al. (2016) wird die *logistische Regression* als GSM (Global Surrogate Model, s. Abschn. 2.5) verwendet.

Für ein besseres Verständnis der Analyse soll sie hier genauer erklärt werden: Folgende Formel gilt für die lineare Regression mit n Features und den Regressionskoeffizienten α_n:

$$f(x_n) = \alpha_0 + \alpha_1 x_1 + ... + \alpha_n x_n \tag{2.1}$$

Die Regressionskoeffizienten α_i mit $i \in \{0, ..., n\}$, folgend auch einfach Koeffizienten oder Gewichte genannt, werden während des Trainings ermittelt. Der von der KI vorhergesagte Wert $f(x_n)$ ist also die Summe aller Features mit ihren Gewichten. An den Koeffizienten kann direkt abgelesen werden, wie stark das zugehörige Feature in die Vorhersage einfließt. Es ist ersichtlich, dass $f(x_n)$ auch Werte außerhalb von 0 und 1 annehmen kann und deshalb nicht für eine Klassifizierung geeignet ist, sondern nur für eine Regression (vgl. Abschn. 2.1).

Die logistische Regression wird folgendermaßen berechnet:

$$P(y(x_n) = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + ... + \beta_n x_n)}} \tag{2.2}$$

mit $P(y(x_n) = 1)$, der Wahrscheinlichkeit dafür, dass ein Ereignis $y(x_n)$ eintritt und den Koeffizienten β_i mit $i \in \{0, ..., n\}$. Im Gegensatz zur linearen Regression ist die Interpretation der Koeffizienten der logistischen Regression jedoch etwas komplexer. Für eine Interpretation müssen die Koeffizienten gem. Formel 2.2 in eine Wahrscheinlichkeit umgerechnet werden. Die Koeffizienten werden auch häufig als *Log Odds Ratio* angegeben. Diese Darstellungsform wird jedoch im Rahmen der Arbeit nicht verwendet.

Der Graph der Funktion 2.2 sieht folgendermaßen aus:

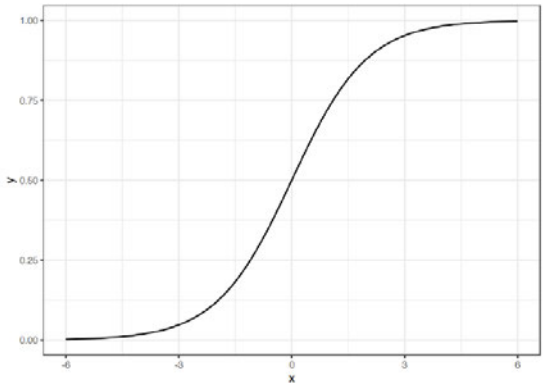


Abbildung 2.4: Logistische Regressionskurve (Molnar 2019)

Auf der Abszisse lassen sich die Wahrscheinlichkeiten $P(x_n)$ ablesen. Da es sich bei Formel 2.2 um eine Wahrscheinlichkeitsfunktion handelt, muss $P(x_n) \in [0, 1]$ sein. Dadurch können Aussagen getroffen werden wie: „Datenpunkt x ist zu 20% Klasse 1 bzw. 80% Klasse 2“. Hier wird nochmal deutlich, warum sich die logistische Regression für Klassifizierungen eignet.

Es wird hier nur die binäre Klassifizierung vorgestellt, also eine Differenzierung nach zwei Klassen. Die logistische Regression kann auch für mehr als zwei Klassen verwendet werden, wird im Rahmen der Arbeit jedoch nicht verwendet und deshalb auch nicht weiter berücksichtigt.

Nachdem in diesem Abschnitt eine Einführung in die Funktionsweise von KIs gegeben wurde, soll im Folgenden geklärt, was unter ethischer KI zu verstehen ist.

2.2 Ethische KI

In diesem Abschnitt soll betrachtet werden, wie eine KI gestaltet werden muss, um ethisch zu handeln. Dafür wurden verschiedene Leitlinien entwickelt, u.a. von der EU Kommission, der OECD oder von ACM. Diese sind so abstrakt formuliert, dass sie für unterschiedlichste KI-Anwendungen genutzt werden können.

Generell haben die Leitsätze nicht den Anspruch auf Vollständigkeit, sondern dienen als Anregung für einen Diskurs, der im Idealfall von der Planung der KI-Anwendung und über den gesamten Produktlebenszyklus bestehend, geführt wird. Dementsprechend sind die Leitsätze auch nicht rechtlich bindend. Zudem sind sie nicht als absolutes Regelwerk für ethisch korrektes Verhalten zu verstehen; es können sich auch Dilemmata aus den Leitsätzen bilden, die die

Verletzung eines Leitsatzes für die Realisierung eines anderen hervorruft. In der Strafverfolgung kommt man bspw. schnell zu der Frage, ob Sicherheit wichtiger ist, als individuelle Freiheits- und Datenschutzrechte, wie auch von der EU Komission (2018) angemerkt. Hier muss im konkreten Fall abgewogen werden, welcher Leitsatz Priorität hat und ob ein Kompromiss eingegangen werden muss. Als generelle Überlegung für solche Fälle hat die EU Komission (2018) formuliert, dass „[...] die Vorteile von KI-Systemen insgesamt die vorhersehbaren individuellen Risiken erheblich überwiegen“ sollen. Im Falle von Interessenkonflikten unterschiedlicher Gruppen fordert die ACM Code 2018 Task Force (2018) bspw., es solle im Zweifel die am meisten benachteiligte Gruppe besondere Priorität erhalten.

Die Leitsätze decken sehr viele unterschiedliche Aspekte ab, wie z.B. Fairness, Transparenz, Security und Datenschutz, Rechenschaftspflicht, Wissenstransfer in die Gesellschaft (Lehre), Urheberrecht, Privatsphäre, oder Naturschutz. Neben Anforderungen an das KI-System selbst werden z.B. auch Handlungsempfehlungen für den Entwicklungsprozess oder die internationale Zusammenarbeit ausgesprochen.

Im Fokus sollen im Folgenden die Leitsätze der EU Komission (2018) stehen. Da in dieser Arbeit Diskriminierung und dessen Aufdeckung im Vordergrund steht, liegt das Hauptaugenmerk auf Textpassagen, die Informationen über Fairness und *Transparenz* enthalten. Transparenz wird zudem nochmal tiefgehender im Zusammenhang mit den Begriffen *Erklärbarkeit* und *Interpretierbarkeit* im Abschn. 2.4 erklärt.

Im Folgenden soll eine Idee darüber geschaffen werden, wie die Leitsätze formuliert sind und was ethische KI bedeutet. Dabei soll keine tiefgreifendere Auseinandersetzung oder gar Bewertung der Leitsätzen gemacht werden.

Ethisches Handeln ist laut EU Komission (2018), neben Rechtmäßigkeit (Einhaltung aller Gesetze) und Robustheit (Zuverlässigkeit und Sicherheit), eine von drei Eigenschaften, die eine vertrauenswürdige KI definieren. Dafür wurden aus der EU-Grundrechtecharta vier Grundsätze (ethische Imperative) abgeleitet, die durch KI-Akteure stets befolgt werden sollen:

1. Achtung der menschlichen Autonomie
2. Schadensverhütung
3. Fairness
4. Erklärbarkeit

Die Leitsätze haben gemein, dass der Mensch und seine Grundrechte an erster Stelle stehen, vor dem technischem und wirtschaftlichem Nutzen. Neben dem Schutz der Menschenwürde und

der „geistigen und körperlichen Unversehrtheit“ wird auch besonders der Schutz von Kindern, Menschen mit Behinderungen und Minderheiten betont: So sollen „schutzbedürftige Personen“ mit in den Entwicklungsprozess von KI-Systemen mit einbezogen werden (Schadensverhütung) und „[...] Personen und Gruppen vor unfairer Verzerrung, Diskriminierung und Stigmatisierung geschützt werden“ (Fairness). Weiterhin ist darauf zu achten, dass die Kontrolle von KI Systemen stets beim Menschen liegt und diesen „[...] nicht auf ungerechtfertigte Weise unterordnen, nötigen, täuschen, manipulieren, konditionieren oder in eine Gruppe drängen“ darf (Achtung der menschlichen Autonomie).

Ein weiterer, häufig betonter Punkt, ist die Transparenz von KI Systemen. So kann eine KI nur fair sein, wenn Entscheidungen nachvollziehbar sind, „[...] verantwortliche Stellen identifizierbar und der Entscheidungsfindungsprozess erklärbar [...]“ (Fairness) und die Fähigkeiten von KIs bekannt sind (Erklärbarkeit). Letzteres sei auch eine Grundvoraussetzung, um sich gegen „[...] Entscheidungen der KI-Systeme und der sie betreibenden Menschen [...]“ wehren zu können. Da Systeme aber häufig, manchmal sogar von Natur aus (Bsp. Neuronale Netze) intransparent sind, wird in Abschn. 4.2 eine Methode vorgestellt, um intransparente Systeme für den Menschen besser verständlich zu machen. Wie transparent ein System sein muss, sei allerdings „[...] sehr stark vom Kontext und der Tragweite der Konsequenzen eines fehlerhaften oder anderweitig unzutreffenden Ergebnisses [...]“ abhängig.

Da die ethischen Imperative sehr abstrakt sind, wurden daraus sieben Anforderungen abgeleitet, die etwas näher an der Praxis sind, wie in Abb. 2.5 zu sehen:

„Technische Robustheit und Sicherheit“ ist somit leichter auf KI-Anwendungen zu übertragen als „Schadensverhütung“. Sicherheit umfasst hier sowohl Security Maßnahmen, als auch Sicherheitsmaßnahmen zur Verhinderung von Schaden an Lebewesen oder Umwelt. Letzteres ist u.A. eng gekoppelt an „Vorrang menschlichen Handelns und menschliche Aufsicht“. Als Beispiel sollte eine Gesichtserkennung nicht selbstständig darüber urteilen, ob es sich bei Person X tatsächlich um die gefahndete Person handelt. Stattdessen sollte die Entscheidung in jedem Fall nochmal von einem Menschen überprüft werden. Eine Person sollte zudem immer über das Recht verfügen, „[...] nicht einer ausschließlich auf einer automatisierten Verarbeitung beruhenden Entscheidung unterworfen zu werden, die ihr gegenüber rechtliche Wirkung entfaltet oder sie in ähnlicher Weise erheblich beeinträchtigt“. Im Falle einer falschen Beurteilung durch eine KI ist in der Abbildung der Punkt „Rechenschaftspflicht“ zu sehen: Entscheidungen durch KI sollen begründet werden können. Wenn das Rückfallrisiko von Straftäter*innen bewertet wird, wie durch die Software „COMPAS“ (s. Kapitel 3), sollte begründet werden, warum jemand mit hohem Risiko bewertet wurde. Dies führt auch schnell zum Punkt „Vielfalt, Nichtdiskriminierung und Fairness“: Nur eine Begründung für die Bewertung mit hohem

Abbildung 2.5: Sieben Anforderungen abgeleitet von vier ethischen Leitsätzen (EU Komission 2018)

Risiko kann eine Diskriminierung nach der Ethnizität (oder anderen sensitiven Eigenschaften, s. Abschn. 2.6) ausschließen. Weiterhin wird hier beschrieben, dass die Datensätze, welche für KI-Systeme genutzt werden, „unbeabsichtigte historische Verzerrungen“, auch *Historical Bias* genannt (vgl. Abschn. 2.7), beinhalten können und somit „[...] Vorurteile und Marginalisierung potenziell verschärfen“ können. Dieser Bias sei nach Möglichkeit während der Datenerhebung zu beseitigen. In Kapitel 3 wird anhand von Anwendungsbeispielen gezeigt, welche Rolle Bias in Daten für die Diskriminierung spielt.

Die Verwobenheit der einzelnen Anforderungen aus Abb. 2.5 ist durch die Linien gekennzeichnet: Alle Anforderungen sind miteinander verbunden. Es soll einerseits signalisieren, dass die Anforderungen sich gegenseitig beeinflussen. Anderseits soll es auch zeigen, dass sie alle parallel bestehen. Alle Anforderungen sollen außerdem während des gesamten Lebenszyklus mitgedacht werden, weshalb in der Mitte die Anforderung „Kontinuierliche Bewertung und Berücksichtigung während des gesamten Lebenszyklus des KI-Systems“ steht.

Es lassen sich viele weitere Ethische Leitsätze finden, die im Kern denen der EU Komission (2018) ähneln. Hierzu gehören zum Beispiel die (OECD 2019). Ihre Leitsätze lauten:

1. Inklusives Wachstum, nachhaltige Entwicklung und Lebensqualität
2. Menschenzentrierte Werte und Fairness
3. Transparenz und Erklärbarkeit
4. Robustheit und Sicherheit
5. Rechenschaftspflicht

Sie sollen hier nicht tiefgehender erklärt werden, da sie denen der EU Komission (2018) stark ähneln. Auch hier steht eine menschenzentrierte, auf Fairness und Transparenz beruhende Nutzung von KI-Anwendungen im Fokus. Der menschliche Eingriff soll stets möglich sein und Entscheidungen nie rein maschinell getroffen werden. Dabei sollen besonders Werte wie „[...] Freiheit, Würde und Selbstbestimmung, Schutz der Privatsphäre und Datenschutz, Nichtdiskriminierung und Gleichbehandlung, Vielfalt, Fairness, soziale Gerechtigkeit und international anerkannte Arbeitsrechte“ Achtung bekommen.

Bei der Entwicklung und Nutzung von KI muss damit gerechnet werden, dass Schäden entstehen. Hier fordert ACM Code 2018 Task Force (2018) eine generelle Schadensminimierung. Schaden definiert ACM Code 2018 Task Force (2018) allgemein durch negative Konsequenzen, besonders wenn diese schwerwiegend und ungerecht sind. In Schadensfällen wären die Verantwortlichen dazu verpflichtet, im Rahmen ihrer Möglichkeiten, den Schaden rückgängig zu machen oder zu verringern. Schadensminimierung beginne zudem bereits damit, dass grundsätzlich der potentielle Schaden an Betroffenen durch Entscheidungen mitbedacht wird. Jede*r Akteur*in sei zudem dazu verpflichtet, Risiken, die potentiell zu Schaden führen könnten, zu melden.

Nachdem hier ein Überblick geschaffen wurde, was ethische KI bedeutet, sollen im Kapitel 3 einige Anwendungsbeispiele analysiert werden, in denen ethische Leitsätze nicht bzw. in Teilen nicht beachtet wurden und deshalb Diskriminierung verursachen. Was unter Diskriminierung genauer zu verstehen ist, wird in Abschn. 2.6 definiert.

2.3 Qualitäts- und Fairnessmaß

Im Abschnitt 2.2 wurde beschrieben, welche Anforderungen an eine ethische KI gestellt werden. Dabei taucht in allen Leitsätzen auch die Forderung nach *Fairness* auf. Um Diskriminierung aufzudecken, wäre es offensichtlich hilfreich, wenn es eine Metrik für Fairness geben würde. Dies wird durch das so genannte *Fairnessmaß* realisiert.

Das Fairnessmaß ist abhängig von der Qualität der KI Entscheidungen. Die Metrik für die Qualität nennt sich *Qualitätsmaß*. Deutscher Bundestag (2020) nennt als Beispiel für ein Qualitätsmaß die *Korrektheit* eines KI-Systems. Sie kann gemessen werden, indem die Vorhersagen der KI mit der Realität abgeglichen werden: Wenn die KI bpsw. vorhersagt, dass eine Person einen Kredit zurückzahlt, war die Entscheidung „korrekt“, wenn die Person den Kredit tatsächlich irgendwann zurück gezahlt hat.

Ein Beispiel für die Messung der Korrektheit ist die *Konkordanz*. Sie soll an folgendem Beispiel erklärt werden:

	Wahrscheinlichkeit f. Rückfall	tatsächlich rückfällig geworden
Person 1	Niedrig	nein
Person 2	Hoch	ja
Person 3	Hoch	nein

Tabelle 2.1: Beispiel für die Vorhersage der Rückfälligkeit von Straftäter*innen

In der Tabelle 2.1 sind exemplarisch drei Fälle dargestellt. Die erste Spalte gibt an, wie hoch das Risiko von einer KI bewertet wurde, dass eine Person wiederholt straffällig wird. Die zweite Spalte beinhaltet, ob die Person tatsächlich rückfällig geworden ist. Aus den drei Personen können nun drei Paare gebildet werden: (P1,P2), (P1,P3), (P2,P3). Nun werden die Paare ausgewählt, in denen eine Person rückfällig geworden ist, und eine nicht. Es bleiben die Paare (P1,P2) (P2,P3) übrig. Ein Paar gilt nun als konkordant, wenn die Wahrscheinlichkeit für die eine Person höher ist, als für die andere und dies tatsächlich eingetroffen ist. Konkordant ist somit (P1,P2). Nicht konkordant ist folglich (P2,P3). Die Konkordanz ergibt sich dann aus

$$Konkordanz = \frac{\textit{Anzahl konkordanter Paare}}{\textit{Gesamtanzahl Paare}} \tag{2.3}$$

Daraus ergibt sich eine Konkordanz von $1/2 = 0.5$ und kann so interpretiert werden, dass 50% der Vorhersagen korrekt sind. Die Metrik wird u.A. von Angwin et al. (2016) verwendet, um die Korrektheit der Software „COMPAS“ zu bewerten. Dabei handelt es sich, wie in dem Beispiel, um eine Software zur Ermittlung des Risikos von Straftäter*innen, rückfällig zu werden.

Ein weiteres Beispiel für ein Qualitätsmaß ist die Bewertung nach Klassifizierungskriterien. Sie sind in folgender Abbildung dargestellt:

Event	Condition	Resulting notion ($\mathbb{P}\{\text{event} \mid \text{condition}\}$)
$\hat{Y} = 1$	$Y = 1$	True positive rate, recall
$\hat{Y} = 0$	$Y = 1$	False negative rate
$\hat{Y} = 1$	$Y = 0$	False positive rate
$\hat{Y} = 0$	$Y = 0$	True negative rate

Abbildung 2.6: Klassifizierungskriterien (Barocas et al. 2019)

mit der Wahrscheinlich $\mathbb{P}\{\gamma = \hat{\gamma}\}$ für die korrekte Vorhersage einer binären Klasse. Am Beispiel von „COMPAS“ entspricht das Event der Bewertung der Risikoklasse, entweder mit „Niedrig“ oder „Hoch“. Die Condition ist die Information darüber, ob ein Angeklagter tatsächlich rückfällig geworden ist. *True Positive* wäre dementsprechend die Bewertung eines Angeklagten mit der Risikoklasse „Hoch“ und ein tatsächlicher Rückfall. *True Negative* umgekehrt die Bewertung mit der Risikoklasse „Niedrig“ und kein Rückfall. *False Negative* entspräche dem Fall, dass die Risikostufe „Niedrig“ vergeben wurde, der Angeklagte aber dennoch rückfällig geworden ist. *False Positive* dementsprechend die Bewertung mit „Hoch“ und keine Rückfälligkeit. Für die Ermittlung der Raten, die häufig in Prozent angegeben werden, müssen bpsw. alle false-positive Fälle addiert werden und durch die Gesamtanzahl der Bewertungen geteilt werden. Wenn also von 100 Bewertungen eine false-negative ist, ergibt die *false-negative Rate* 1%.

Die Korrektheit ist ein Beispiel für ein Qualitätsmaß. Anhand dessen kann eine Metrik abgeleitet werden, die das Fairnessmaß genannt wird. Deutscher Bundestag (2020) erklärt das korrespondierende Fairnessmaß zu einem Qualitätsmaß folgendermaßen: Die Qualität (in diesem Fall die Korrektheit) solle für alle Teilgruppen, getrennt nach ihren sensitiven Eigenschaften (s. Abschn. 2.6), ungefähr gleich groß sein. Um zum ersten Beispiel zurückzukehren: Wenn 80% der Vorhersagen über die Kreditwürdigkeit korrekt sind, dann sollte dies auch für die Gruppe „Männlich“ und „Weiblich“ gelten. Für das konstruierte Beispiel von „COMPAS“ könnte das bedeuten, dass für die Gruppe der „Schwarzen“ und „Weißen“ Angeklagten jeweils eine Konkordanz von 50% vorliegen soll.

Für die Messung von Fairness muss sich allerdings darauf geeinigt werden, wie Fairness definiert werden soll. Dass das nicht immer eindeutig ist, wird in Kapitel 4 am Beispiel von „COMPAS“ weiter erläutert.

2.4 Transparenz, Interpretierbarkeit und Erklärbarkeit

In den Leitsätzen aus Abschn. 2.2 werden immer wieder *Transparenz*, *Interpretierbarkeit* und *Erklärbarkeit* als Voraussetzung für eine ethische und diskriminierungsfreie KI genannt. Die Begriffe sollen deshalb im Folgenden noch einmal näher betrachtet werden:

Transparenz bedeutet im Zusammenhang mit KI die Offenlegung von Informationen, die nötig sind, um Entscheidungen durch eine KI nachzuvollziehen. Hierzu gehöre laut Krafft & Zweig (2019) z.B. das Lernverfahren, Feature Auswahl, die Qualität der Entscheidungen (s. Abschn. 2.3) und die Datengrundlage, also auch Informationen zur Datenerhebung und Datenaufbereitung. Diese Informationen müssten „[...] präzise, leicht zugänglich und verständlich, sowie in klarer und einfacher Sprache vermittelt werden“. Der Fokus liegt hier also auf der Bereitstellung von Informationen zur verwendeten Technologie.

Nach der Taxonomie der Erklärbarkeit von Lipton (2018) wurde von Waltl (2019) die Abb. 2.7 abgeleitet. Transparenz umfasst demnach drei Ebenen: Das gesamte Model (Simulierbarkeit), einzelne Komponenten wie die Features (Dekomposition) und den Trainingsalgorithmus (Algorithmische Transparenz). Waltl (2019) begründet das folgendermaßen: Die bloße Betrachtung von Entscheidungen durch KI (Simulierbarkeit) sei für die Aufdeckung von Diskriminierung oft unzureichend. Diese sei häufig bereits in den Daten manifestiert (vgl. Abschn. 2.7), oder durch die Auswahl der Features bedingt (Dekomposition), wie auch in Kapitel 3 näher erklärt.

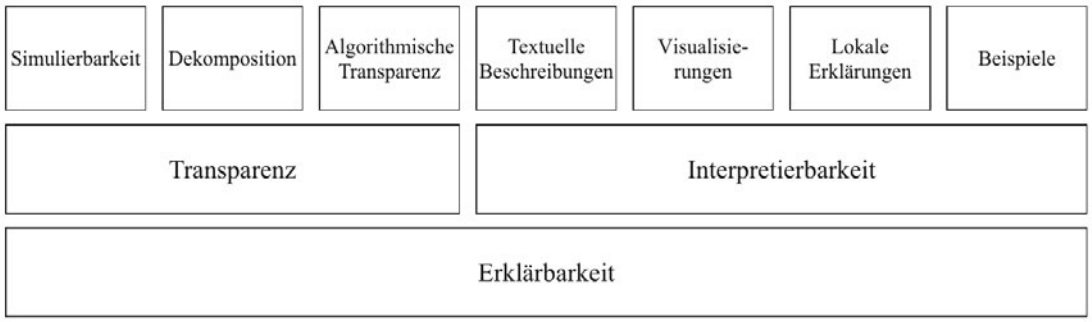


Abbildung 2.7: Taxonomie der Erklärbarkeit nach Lipton (Waltl 2019)

Zu sehen ist, dass Transparenz erst zusammen mit der Interpretierbarkeit (auch Nachvollziehbarkeit genannt) die Erklärbarkeit ergibt. Die Offenlegung alleine reicht also noch nicht, um eine KI als Ganzes zu verstehen. Dazu gehört auch, sie überprüfen zu können. Hier können beschreibende Maßnahmen wie textuelle Beschreibungen oder Visualisierungen notwendig werden. Auch Beispiele für Eingabe- und Ausgabedaten können hier das Verständnis unter-

stützen. Typischerweise seien hierfür Systemschnittstellen nötig, um das Verhalten der KI mit eigenen Eingaben validieren zu können (Deutscher Bundestag 2020).

Es gibt ML-Algorithmen, die besser interpretierbar sind, als andere. Beispiele für gut interpretierbare Algorithmen sind logistische Regressionen, regelbasierte Systeme oder Entscheidungsbäume. Die logistische Regression wird deshalb später in Kapitel 4 noch eine Rolle spielen. Wie in Abbildung 2.8 zu sehen, haben aber vor allem die hoch performanten Ansätze wie Deep Learning den Nachteil, dass sie nur schwer interpretierbar sind und wie eine „Black Box“ zu verstehen sind. Leistungsfähigkeit und Verständlichkeit von ML-Modellen scheinen in einem inversen Verhältnis zueinander zu stehen (Gunning & Aha 2019).

Hier setzt *XAI (Explainable AI)* an (vgl. Abschn. 2.5). Es handelt sich um Methoden, die darauf abzielen, das Verhalten von KI für den Menschen nachvollziehbar zu machen. Sie sollen im nächsten Abschn. 2.5 weiter erklärt werden.

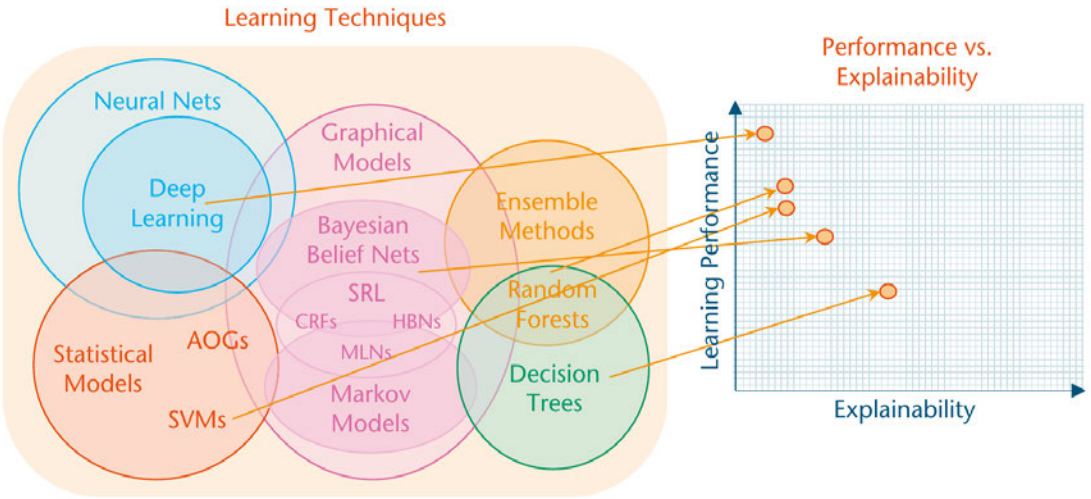


Abbildung 2.8: Einfluss der Lernmethode auf die Erklärbarkeit (Gunning & Aha 2019)

2.5 XAI

In folgendem Abschnitt soll geklärt werden, was unter *XAI* zu verstehen ist. Im Abschn. 2.4 wurde bereits der Begriff der Interpretierbarkeit eingeführt und aufgezeigt, dass einige Modelle besser interpretierbar sind (Bsp. Entscheidungsbaum) als andere (Bsp. Neuronales Netz). Leistungsfähige Lernalgorithmen, wie die aktuell weit verbreiteten neuronalen Netze, haben den Nachteil, nicht gut interpretierbar zu sein. Dieser Zusammenhang wird ausführlich

von (Gunning & Aha 2019) erklärt. In den Leitsätzen aus Abschn. 2.2 wird Interpretierbarkeit ausdrücklich als Voraussetzung für ethische und damit diskriminierungsfreie KI gefordert.

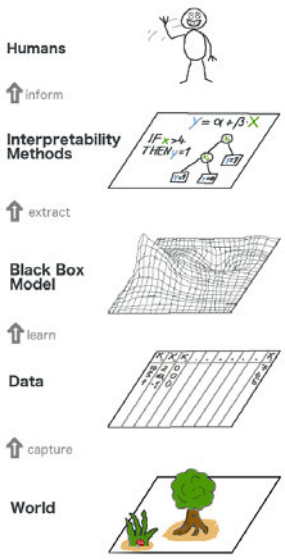


Abbildung 2.9: Abstraktionsebenen des Machine Learning (Molnar 2019)

Molnar (2019) spricht im Zusammenhang mit XAI von fünf Abstraktionsebenen (s. Abb. 2.9): Auf der untersten Ebene befindet sich die Umwelt, z.B. Personen, die durch eine Gesichtserkennung erkannt werden sollen. Die Umwelt wird auf der zweiten Ebene durch Daten beschrieben, welche in der nächst höheren Ebene von einem KI Modell (in der Abb. Black Box Model genannt) für das Training verwendet werden. Von dem KI Modell werden dabei Informationen extrahiert. Diese werden durch die nächste Ebene, den Interpretationsmethoden bzw. XAI für den Menschen verständlich gemacht. Laut Gunning & Aha (2019) sind XAIs „[...] AI systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future“.

XAI umfasst verschiedene Ansätze, um die Interpretierbarkeit von KI zu verbessern. Hierzu zählen z.B. modell-agnostische Methoden, die unabhängig vom zugrunde liegenden KI-Modell angewendet werden können. Sie versuchen, das generelle Verhalten des KI Modells zu erklären. Dies ist einerseits vorteilhaft für die Entwickler, da sie nicht in der Modellauswahl eingeschränkt werden. Auch nach grundlegenden Veränderungen am Modell oder sogar einem Wechsel auf einen anderen Algorithmus können die modell-agnostischen Methoden weiter verwendet werden. Anderseits können sie z.B. auch dann verwendet werden, wenn Unternehmen ihre Software nicht offenlegen wollen. In Abschnitt 4.2 wird eine modell-agnostische Methode

vorgestellt, um die Risk Assessment Software „COMPAS“ auf Diskriminierung zu analysieren. Es handelt sich um das „Global Surroagte Model“ (GSM), bei dem ein Ersatzmodell trainiert wird, welches über eine gute Interpretierbarkeit verfügt.

Weitere Methoden sind u.a. der Partial Dependence Plot, bei dem der Einfluss einzelner Features auf das Gesamtergebnis grafisch dargestellt wird. So könnte z.B. das Feature „Alter“ sequentiell von 0 auf 80 erhöht werden, während andere Features gleich bleiben. Für jedes Alter wird dann die Ausgabe des Modells, z.B. Kosten einer Versicherung, in einem Diagramm aufgetragen und so der Einfluss des Alters auf die Versicherungskosten visualisiert. Die Methode hat den Vorteil, dass sie relativ leicht umzusetzen und auch zu verstehen ist. Ein großer Nachteil ist jedoch, dass die Features unabhängig voneinander sein müssen (Molnar 2019).

Neben modell-agnostischen gibt es auch modell-spezifische Methoden. Sie können nur für ein spezielles Modell verwendet werden. Hierzu gehört z.B. die „Feature Visualization“, durch die gelernte Features von Neuronalen Netzen visualisiert werden können.

Weiterhin gibt es z.B. die so genannten Example-based Methoden. Sie versuchen anhand von konkreten Instanzen des Datensatzes Erklärungen über das Modell zu liefern. Sie werden eher bei Daten angewendet, die eine Struktur aufweisen, wie z.B. Bilder oder Texte, und weniger für tabellarische Daten, in denen kein für den Menschen erkennbarer Zusammenhang zwischen den (häufig 100-1000) unterschiedlichen Parametern existiert. Ziel bei den Example-based Methoden ist nämlich die Aussage zu treffen: „Situation A hat X ausgelöst, Situation B ist ähnlich wie A, also wird Situation B vermutlich ebenfalls X auslösen“. Example-based und modell-spezifische Methoden sollen hier nur als weitere Arten von Ansätzen vorgestellt, in dieser Arbeit aber nicht weiter behandelt werden. Sie werden ausführlich von Molnar (2019) erklärt.

2.6 Diskriminierung

In diesem Abschnitt soll ein kurzer Überblick geschaffen werden, wie *Diskriminierung* im Kontext dieser Arbeit verstanden wird.

Diskriminierung bedeutet zunächst eine „[...] benachteiligende, ungerechtfertigte Ungleichbehandlung [...]“ (Orwat 2019) auf Grund von folgenden geschützten Merkmalen, welche der „Charta der Grundrechte der Europäischen Union“ (EU 2000) entnommen wurde:

- Geschlecht
- Rasse
- ethnische oder soziale Herkunft
- genetische Merkmale
- Sprache
- Religion oder Weltanschauung
- politische oder sonstige Anschauung
- Zugehörigkeit zu einer nationalen Minderheit
- Vermögen
- Geburt
- Behinderung
- Alter
- sexuelle Ausrichtung

Ungerechtfertigt bedeutet dabei, dass kein sachlicher Grund für die Ungleichbehandlung existiert (Orwat 2019). Im Rahmen dieser Arbeit wird der Begriff „Diskriminierung“ immer in Bezug auf ungerechtfertigte Diskriminierung genutzt. In der Literatur, speziell im Bereich der KI, wird Diskriminierung häufig synonym verwendet für *Differenzierung*. In diesen Fällen wird dann der Begriff „Differenzierung“ verwendet.

Diskriminierung kann in unmittelbare und mittelbare Diskriminierung eingeteilt werden: Unmittelbare Diskriminierung würde dann vorliegen, wenn ein direkter Bezug zu oben genannten Merkmalen hergestellt werden kann. Mittelbare Diskriminierung hingegen entsteht

durch scheinbar neutrale „[...] Vorgaben und Verfahrensweisen, die im Effekt gleichwohl zu Benachteiligungen bestimmter Personenkategorien und sozialer Gruppen führen“ (Scherr 2016).

Diskriminierende Handlungen folgen dabei nicht zwingend aus diskriminierenden Einstellungen, sondern können auch bspw. „[...] aus rationalen ökonomischen Kalkülen vorurteilsfreier Akteure resultieren“ (Scherr 2016). Dies kann auch auf diskriminierende Handlungen technischer Natur übertragen werden: Wenn z.B. Gesichtserkennungssoftware hauptsächlich an Weißen Menschen getestet wird und daraus eine geringere Korrektheit für Schwarze Menschen resultiert (vgl. Abschn. 3.2.2), ist dies zwar diskriminierend, muss jedoch keine diskriminierende Einstellung als Ursache besitzen.

Eine Mögliche Ursache für mittelbare Diskriminierung sei laut Scherr (2016) die so genannte *statistische Diskriminierung*. Sie wird abgegrenzt zur *präferenzbedinten Diskriminierung*, die eine Diskriminierung auf der Basis von persönlichen Vorurteilen darstellt. Bei der statistischen Diskriminierung entsteht die Diskriminierung über Ersatzmerkmale, die für die Betrachtung einer Person/Personengruppe hinzugezogen wird, weil die eigentlich interessanten Hauptmerkmale schwer messbar sind. Als Beispiel sei das Hinzuziehen von dem Ersatzmerkmal „Geschlecht“ genannt, weil das Hauptmerkmal „Produktivität“ eines Arbeitnehmers schwer messbar ist und eine individuelle Betrachtung nötig wäre (Beispiel von Scherr (2016) übernommen). Die individuelle Betrachtung wird aber häufig aus Kostengründen nicht gemacht. Ersatzmerkmale werden auch *Proxies* genannt (Orwat 2019).

Ein Hauptmerkmal wäre z.B. die Kriminalität. Sie ist schwer zu messen, da Straftaten auch unentdeckt bleiben können. Als Proxy für die „Kriminalität“ könnte man bspw. die Anzahl an Festnahmen nutzen. Die Anzahl der Festnahmen ist aber abhängig von der Polizeipräsenz. Diese Problematik wird in Kapitel 3 nochmal aufgegriffen.

Neben der statistischen Diskriminierung existiert laut Scherr (2016) die institutionelle Diskriminierung, organisationelle Diskriminierung und gesellschaftsstrukturelle Diskriminierung. Diese sollen im Rahmen dieser Arbeit jedoch nicht weiter behandelt werden, da im Zusammenhang mit KI vor allem die statistische Diskriminierung auftritt (Orwat 2019).

2.7 Bias

Der Begriff *Bias*, der in folgender Arbeit synonym mit dem Wort *Verzerrung* genutzt wird, findet sich in vielen wissenschaftlichen Disziplinen. In der Sozialforschung ist ein Bias die Verzerrung eines Ergebnisses durch fehlerhafte Untersuchungsannahmen bzw. –methoden. Auch eine KI kann einen Bias beinhalten und dadurch zu Diskriminierung führen, wie das klassische Beispiel

der von Amazon entwickelten KI zur Auswahl von Bewerber*innen zeigt: Die Trainingsdaten der KI basierten auf den historischen Daten der bisherigen Mitarbeiter*innen des Konzerns. Diese waren allerdings überwiegend männlich, sodass die KI lernte, weibliche Bewerberinnen auszuschließen (Saka 2020).

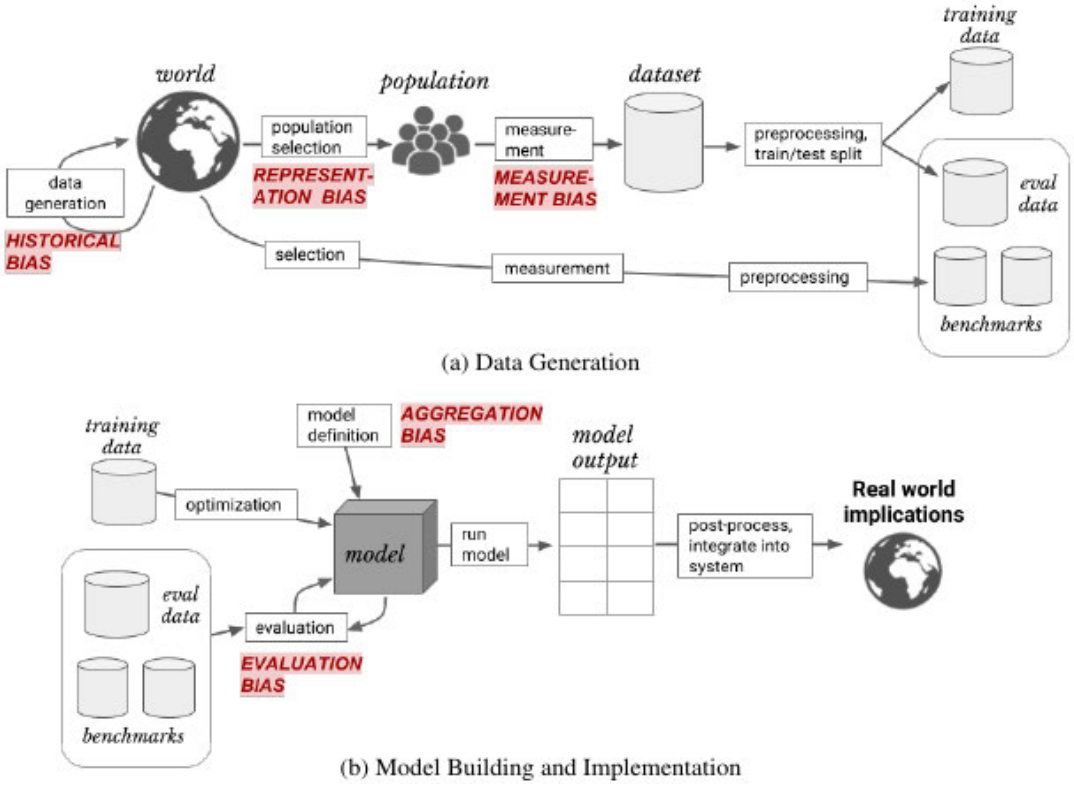


Abbildung 2.10: Unterschiedliche Arten von Bias (Suresh & Guttag 2019)

Suresh & Guttag (2019) beschreiben in ihrem Framework fünf verschiedene Typen von Bias:

- Historical Bias
- Representation Bias
- Measurement Bias
- Aggregation Bias
- Evaluation Bias

Die unterschiedlichen Typen werden im Folgenden anhand der Erklärungen von Suresh & Guttag (2019) zusammengefasst und nicht nochmal gesondert zitiert: Bias kann in unterschiedlichen Phasen des KI Modells entstehen und in zwei Phasen „Datengenerierung“ und „Modellerstellung und -implementierung“ eingeteilt werden. Beide Phasen können wiederum in einzelne Schritte zerlegt werden, wie in Abb. 2.10 zu sehen. Der *Historical Bias* tritt z.B. schon bei der Datengenerierung auf. So das Beispiel von Amazon, die überproportionale Einstellung von Männern ist historisch bedingt und wird in den Daten manifestiert.

Ein *Representation Bias* entsteht durch eine unzureichende Abbildung der Bevölkerung. In Abschn. 3.2.2 wird am Beispiel der Gesichtserkennung dargestellt, was dies für Auswirkungen haben kann.

Der *Measurement Bias* entsteht bei der *Messung* der Daten. Sie entstehen u.A. durch die Wahl von Proxies (vgl. Abschn. 2.6), also stellvertretende Parameter, die den ursprünglichen Parameter aber nur unzureichend abbilden können. Dieser Punkt wird im Kapitel 3 bei den „Predictive Policing“ Anwendungen wieder aufgegriffen. Weiterhin wird der *Measurement Bias* durch unterschiedliche Qualität und Granularität der Daten je Bevölkerungsgruppe und die Feature-Auswahl beeinflusst.

Die bereits genannten Formen des Bias werden der Phase „Datengenerierung“ zugeordnet, während die folgenden Arten der Phase „Modellerstellung und -implementierung“ zugeordnet werden: Ein *Aggregation Bias* entsteht, wenn ein allgemeingültiges Modell (one-size-fit-all Modell) auf Personengruppen angewendet wird, die nicht in dieses Modell passen, da z.B. genetische Unterschiede zwischen Ethnizitäten oder Geschlechtern herrschen.

Der *Evaluation Bias* entsteht, wenn das KI-Modell anhand von Benchmarkdatensätzen trainiert wird, welche keine adäquaten Repräsentationen für die spätere Anwendung bieten. So wird z.B. die Gesichtserkennung u.A. mit Fotos von Prominenten aus dem Internet trainiert. Die Qualität der Fotos übersteigt jedoch die derjenigen Bilder, die später für die Anwendung genutzt werden, wie in Abschn. 3.2 erklärt.

Die unterschiedlichen Typen von Bias ermöglichen eine differenziertere Betrachtung der Ursachen von Diskriminierung. Im Kapitel 3 wird anhand von Anwendungsbeispielen untersucht, wo welche Art von Bias auftritt (sofern dies möglich ist) und welche Folgen das für Entscheidungen durch KI haben kann.

3 KI in der Strafverfolgung

In diesem Kapitel sollen Belege von Diskriminierung und deren Ursache anhand von Anwendungsbeispielen aus der Strafverfolgung gefunden werden. Dafür soll zunächst ein Überblick geschaffen werden, welche Anwendungen existieren. Im Anschluss wird die FRT näher behandelt, da sie besonders umstritten ist. Zum Schluss sollen aus den Erkenntnissen der untersuchten Anwendungen allgemeine Schlüsse zu den Ursachen von Diskriminierung durch KI gemacht werden.

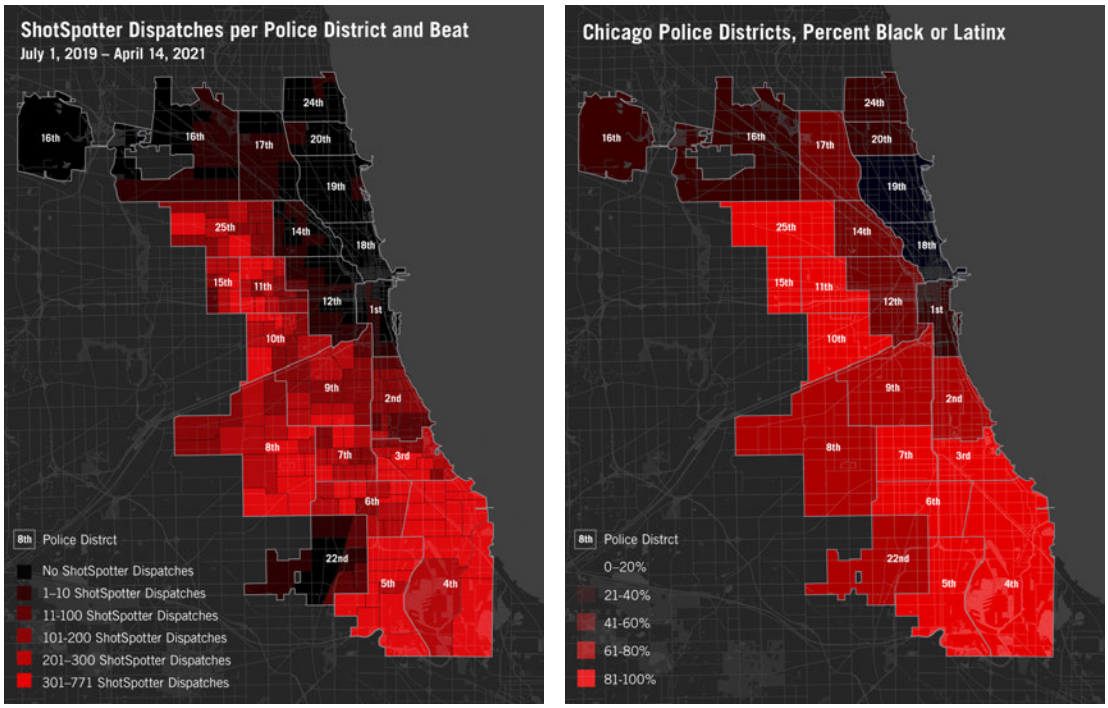
3.1 Überblick

KIs werden vielseitig in der Strafverfolgung genutzt. Vorläufig ist das Ziel der Anwendungen, die Gesellschaft sicherer zu machen, indem sie zur Aufklärung von Straftaten beitragen oder diese präventiv vermeiden sollen. Zu den klassischen Anwendungsgebieten gehört z.B. FRT. Sie wird u.A. verwendet, um gefahndete Personen automatisiert auf Fotos und Videos zu identifizieren und im Abschn. 3.2 näher behandelt.

Neben FRT gibt es viele weitere Anwendungen. Hierzu zählt z.B. die KI gestützte Identifizierung und Lokalisierung von Schüssen durch Waffen. Als Beispiel sei die Firma „ShotSpotter“ genannt, die u.A. in Chicago zwischen 20 und 25 Mikrofonen pro Quadratmeile aufstellen, um potentielle Schüsse aufzuzeichnen. Die Software kann bestimmen, um welchen Waffentyp es sich handelt und über Triangulation der Geräuschpegel bestimmen, wo der Schuss gefallen ist. Diese Informationen werden auch vor Gericht verwendet (Stanley 2021).

MacArthur Justice Center (2021) finden in ihrer Studie heraus, dass „ShotSpotter“ fast ausschließlich dort eingesetzt wird, wo der Anteil an Schwarzen und Hispanics sehr hoch ist. Dies ist in Abb. 3.1 zu sehen: Links ist die Anzahl der „ShotSpotter“ Installationen je District zu sehen. Rechts ist der prozentuale Anteil von Schwarzen und Hispanics zu sehen. Districts mit 100 oder mehr „ShotSpotter“ Installationen sind fast ausschließlich dort, wo der Anteil an Schwarzen und Hispanics mehr als 60% beträgt.

Laut MacArthur Justice Center (2021) würde dies eine ohnehin erhöhte Polizeipräsenz in den Districts weiter verstärken. Dies hätte auch eine Verzerrung der Kriminalstatistik zur Folge, denn mehr Polizeipräsenz erhöhe alleine durch die Präsenz die Wahrscheinlichkeit, Straftaten



(a) Installationen von „ShotSpotter“ aufgeteilt nach Districts

(b) Anteil Schwarze und Hispanics

Abbildung 3.1: Vergleich der Installationen von „ShotSpotter“ und Verteilung von Schwarzen und Hispanics in Chicago (MacArthur Justice Center 2021)

zu entdecken. Die erhöhte Kriminalitätsrate habe wiederum eine Erhöhung der Polizeipräsenz zur Folge.

Eine Ursache von Diskriminierung scheint also der ungleiche Einsatz der Software zu sein. Schwarze und Hispanics werden durch „ShotSpotter“ stärker überwacht und erfahren dadurch eine höhere Polizeipräsenz. Der Wohnort fungiert hier als Proxy für die Ethnizität, wodurch eine statistische Diskriminierung entsteht (vgl. Verma (2019)).

MacArthur Justice Center (2021) finden weiterhin heraus: In Abb. 3.2 sind die falschen Alarme je District zu sehen. Jeder Balken zeigt den Anteil der falschen Alarme, verursacht durch 9-1-1 Notrufe (grau) und durch „ShotSpotter“ (rot). Die falschen Alarme von „ShotSpotter“ entsprechen False-Positives (vgl. Abschn. 2.3). Es ist deutlich zu sehen, dass die durch „ShotSpotter“ verursachten falschen Alarme teilweise um ein vielfaches höher sind, als durch herkömmliche 9-1-1 Notrufe. Dies erhöhe weiter die Polizeipräsenz, da auch ein falscher Alarm zu einem Polizeieinsatz führt. Stanley (2021) schreibt dazu, dass die Software deshalb in vielen Städten nicht mehr genutzt wird, weil sie für den praktischen Einsatz nicht genau genug sei.

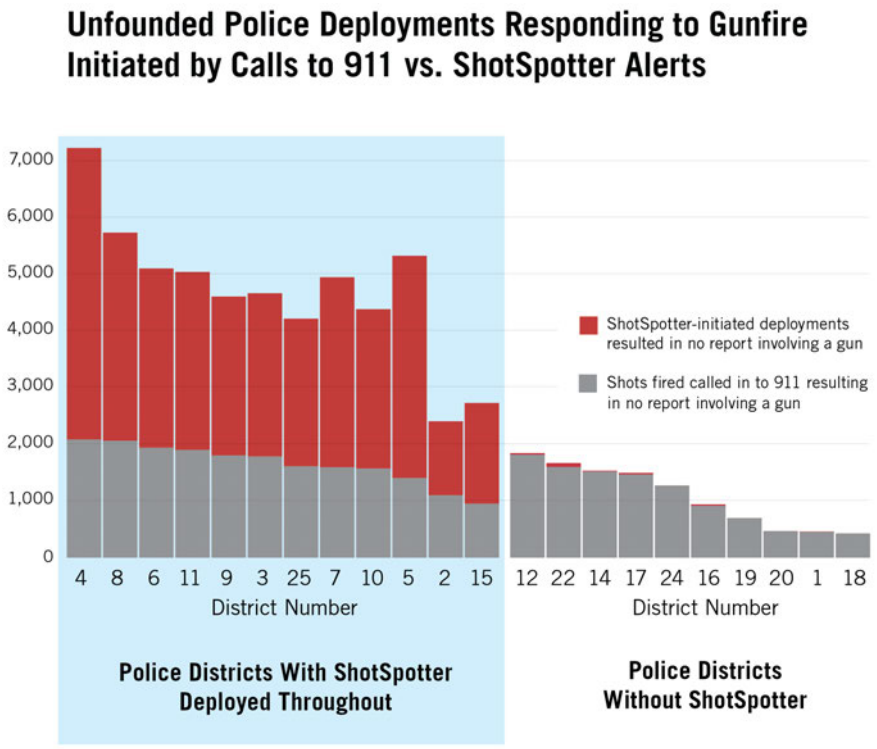


Abbildung 3.2: Vergleich falscher Alarm durch 9-1-1 Notrufe und „ShotSpotter“ (MacArthur Justice Center 2021)

Demnach ist die Ungenauigkeit der KI, also falsche Alarme bzw. False-Positives, eine weitere Ursache von Diskriminierung. Zunächst folgt daraus eine erhöhte Polizeipräsenz. Durch den selektierten Einsatz trifft das zudem nur die Districts, in denen „ShotSpotter“ verwendet wird und somit vor allem Schwarze und Hispanics.

Eine weitere Anwendung findet KI in der Forensik. Marciano & Sweder (2016) entwickelten z.B. eine Methode, bei der Forensiker durch KI unterstützt werden, um vermischte DNA Spuren von unterschiedlichen Personen zu trennen und zu identifizieren. Bei dieser Anwendungen konnten keine Ursachen von Diskriminierung identifiziert werden.

Wie zu Beginn erwähnt, wird KI auch präventiv für vorausschauende Polizeiarbeit, „Predictive Policing“ genannt, genutzt. Hierzu gehört bspw. die Software „Geolitica“, die eher bekannt ist unter dem alten Namen „PredPol“ der gleichnamigen Firma PredPol Inc. Andere Anbieter mit ähnlichem Produkt sind „CompStat“, welches in New York verwendet wird, oder „HunchLab“ in Philadelphia. Diese Softwares leiten anhand von historischen und aktuellen Daten ab, in welcher Gegend ein erhöhtes Risiko für folgende Straftaten herrscht. Sie haben ihren Ursprung

in der Vorhersage von Erdbeben (Verma 2019). Zu den Daten, auf denen die Vorhersagen basieren, gehören „[...] Kriminalfälle aus der Vergangenheit, aber auch soziodemografische Daten, Bonität, Wetterprognosen, Verkehrsdaten, zum Teil auch aktuelle Informationen aus sozialen Netzwerken“ (Peteranderl 2016). In Gegenden mit hohem Risiko kann dann z.B. verstärkt Polizei patrouillieren.

Die Straftäter*innen werden erst festgenommen, wenn die Tat bereits begangen wurde. Softwares wie „Geolitica“ machen keine individuellen Vorhersagen, sondern geografische, wodurch man sie als fair betrachten könnte. Verma (2019) kritisiert jedoch, ähnlich wie bei „ShotSpotter“ von MacArthur Justice Center (2021) beschrieben, dass die Polizeipräsenz selbst neue Daten generiere, weil aufgenommene Straftaten wieder als Eingabedaten für die Software genutzt werden.

Es wird deutlich, dass Anwendungen wie PredPol oder „ShotSpotter“ Diskriminierung verstärken können. Eine, womöglich diskriminierende, verstärkte Polizeipräsenz, kann sich durch eine Software wie PredPol selbst verstärken.

Weitere Anwendung findet KI im Risk-Assessment. Hier sei die Software „COMPAS“ genannt, die das Rückfälligkeitsrisiko von Straftäter*innen bewerten soll. Sie wurde von Angwin et al. (2016), auch bekannt unter dem non-profit-newsroom ProPublica, mithilfe eines modellagnostischen Ansatzes auf Diskriminierung untersucht. Die Software und Vorgehensweise der Analyse wird in Kapitel 4 genauer vorgestellt.

Ein letztes Beispiel, welches häufig im Kontext der Strafverfolgung auftaucht, ist die Software „Gotham“ der Firma Palantir. Es handelt sich dabei um ein (u.A. von der CIA u. Peter Thiel finanziertes) Datenaufbereitungs- und analyse Tool, welches für Geheimdienste, Militär, Polizei und Antiterroreinheiten entwickelt wurde. „Gotham“ setzt sich als Ziel, Daten jeglicher Art aus verschiedensten Quellen so aufzubereiten, dass sie ohne tiefgehende Kenntnisse über Datenanlyse verwendet werden können. Welche Daten die Software verwendet und was mit den Daten, die z.B. bei der Nutzung anfallen, passiert, ist nicht genau bekannt. Es soll z.B. möglich sein, soziale Netzwerke wie Facebook einzubinden und so auf private Daten und Freundschaftsnetzwerke zuzugreifen (Brühl 2018).

„Gotham“ ist international sehr umstritten, Haskins (2020) nennt sie „[...] one of the most controversial and powerful law enforcement tools in the world“. Sie wurde u.A. von Europol genutzt (Johannson 2020) und wird aktuell auch von deutschen Polizeibehörden getestet (Rosenbach & Sarovic 2021).

Im Zusammenhang mit „Gotham“ konnten keine Belege für Diskriminierung gefunden werden. Dennoch soll die Software als weiteres Anwendungsbeispiel für KI Anwendungen in der Strafverfolgung vorgestellt werden.

Nachdem ein Einblick in die Anwendungen der KI in der Strafverfolgung gegeben wurde, soll im Folgenden die FRT genauer vorgestellt werden. Innerhalb dieses Themenbereichs sollen weitere Belege für Diskriminierung und deren Ursachen herausgearbeitet werden, um die Forschungsfragen dieser Arbeit beantworten zu können. Zum Schluss werden die Erkenntnisse zusammengetragen und generelle Aussagen über die Ursachen von Diskriminierung, auch außerhalb der Strafverfolgung, abgeleitet.

3.2 Facial Recognition Technology (FRT)

FRT (Facial Recognition Technology), auf deutsch biometrische Gesichtserkennung, gehört zu den biometrischen Identifizierungsverfahren. Wie bei einem Fingerabdruck werden Merkmale einer Person genutzt, um sie zu identifizieren. Da das Gesicht im Gegensatz zum Fingerabdruck durch Kameras einfach digitalisiert werden kann, kann die Identifizierung voll automatisiert stattfinden. Dadurch kann sie mit anderen Technologien verbunden werden, wie z.B. Social Media oder Videoüberwachung (Marcus Smith 2021).

FRT wird zunehmend von der Polizei zur Strafverfolgung genutzt. Das zeigen z.B. Untersuchungen der Nutzungshäufigkeit von Clearview AIs FRT in den U.S.Buzzfeed News (2021). Zu den Anwendungen von FRT gehören laut United States. Governement Accountability Office (2020) außerdem:

- Zugangskontrollen z.B. Einlasskontrollen für Gebäude
- Erkennung von Straftäter*innen, wie z.B. Betrüger*innen im Casino
- Hotel-checkin
- Abwicklung von Zahlungsprozessen
- Tracking der Arbeitszeit oder Anwesenheit bei Vorlesungen
- Verifizierung der Identität für Wahlen

Die Begrifflichkeiten werden in Medien und Wissenschaft nicht eindeutig verwendet. So umfasst FRT neben der Identifizierung und Verifizierung von Personen auch häufig *Facial Analysis* oder *FAC (Facial Attribute Classification)*. Diese Techniken werden genutzt, um Gesichter auf bestimmte Eigenschaften zu analysieren. Sie sind teilweise sehr umstritten, da sie neben Geschlecht oder Herkunft (Buolamwini & Gebru 2018) auch sexuelle Orientierung (Wang & Kosinski 2018), Risiko von Straffälligkeit (Wu & Zhang 2016) oder die Fähigkeit, ein professioneller Pokerspieler zu sein (Faceception 2022) anhand von Bildern analysieren.

FRT ist eine der sich am schnellsten entwickelnden Methoden der biometrischen Identifizierung (Marcus Smith 2021). Auch das wirtschaftliche Interesse steigt enorm, so stiegen die Patentanmeldungen in den USA von 631 (2015) auf 1497 (2019). Der Umsatz betrug zwischen 2016-2019 etwa 3-5 Mrd. USD und wird für 2022-2024 auf 7-10 Mrd. USD geschätzt (United States. Governement Accountability Office 2020).

Dieser Anstieg ist auf die Weiterentwicklung von Deep Convolutional Neural Networks (CNNs), einer speziellen Form von Neuronalen Netzen für Bilder, zurückzuführen und die allgemein höhere Verfügbarkeit von Datensätzen (A. Bansal et al. 2021). Letzteres wird u.a. durch Social Media begünstigt und sorgte im Jahr 2020 für großes Aufsehen. Die Firma „Clearview AI“ lud automatisiert und ohne Erlaubnis 2.8 Milliarden Fotos von Instagram, Facebook, Youtube, Twitter und LinkedIn herunter und nutzte die Fotos als Datenbasis für ihre FRT Software (Marks 2021).

In den folgenden Abschnitten wird zunächst erklärt, wie FRT technisch funktioniert. Dabei wird unterschieden zwischen der Identifzierung/Verifizierung, die im Folgenden FR genannt wird, und der FAC. Anschließend sollen Probleme der Technologie genannt und einige Beispiele vorgestellt werden, in denen FRT zu diskriminierenden Entscheidungen geführt hat.

3.2.1 Funktionsweise FRT

Im Folgenden soll näher auf die Funktionsweise der FRT eingegangen werden, damit die im Anschluss beschriebenen Probleme der Technologien verstanden werden können. Es wird unterschieden zwischen der FR, die hauptsächlich für die Identifizierung/Verifizierung von Gesichtern genutzt wird und die FAC, die die Zuordnung von Attributen zu den Gesichtern als Ziel hat.

3.2.1.1 Facial Recognition (FR)

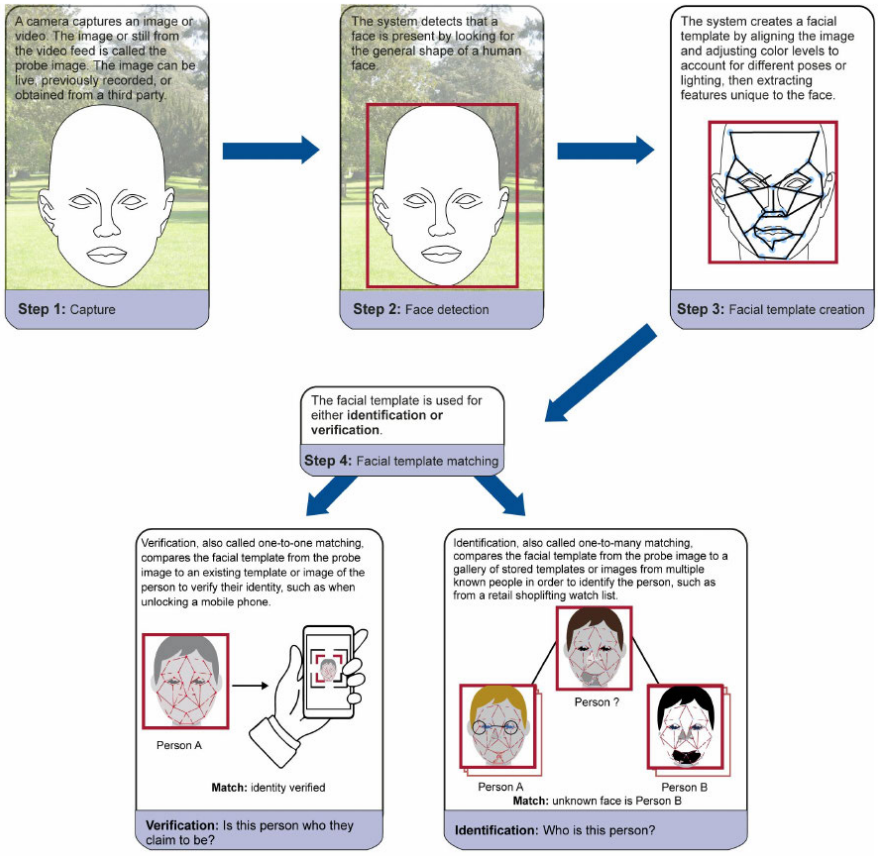


Abbildung 3.3: Facial Recognition Funktionsweise (United States. Governement Accountability Office 2020)

In Abb. 3.3 ist der Ablauf einer FR skizziert. Ein Bild der Person muss in digitaler Form (Foto oder Video) vorliegen. Zunächst muss das Gesicht erkannt (Face Detection) und ausgeschnitten werden. Dadurch werden die für die FR uninteressanten Hintergrundinformationen entfernt. Das ausgeschnittene Gesicht kann dann weiterverarbeitet werden und durch *Feature Extraction* (nicht in Abb. zu sehen) weiter auf wesentliche geometrische Merkmale des Gesichts reduziert werden. Das geschieht heutzutage meistens mit CNNs. Es entsteht das *Facial Template*, welches die geometrischen Eigenschaften des Gesichts in Form von Punkten repräsentiert und keine redundante oder irrelevante Information mehr enthält. In Abb. 3.4 ist ein Beispiel zu sehen, wie die Punkte im Gesicht markiert sein können.

Für das Training der CNNs wird vorzugsweise ein Datensatz mit mehreren Fotos der selben Person aus unterschiedlichen Winkeln genutzt. Dadurch lernt das CNN ein Facial Template zu erstellen, auch wenn das Gesicht der gesuchten Person nicht frontal aufgenommen wurde. In der Praxis ermöglicht dies z.B. eine Echtzeitidentifizierung mithilfe von Videoaufnahmen (Yang 2009). Das Facial Template kann anschließend genutzt werden um die FR durchzuführen. Wie oben beschrieben, wird dabei zwischen Identifizierung und Verifzierung unterschieden. Bei der Identifizierung wird das Bild einer Person, bspw. aufgenommen durch eine Überwachungskamera, mit einer Datenbank abgeglichen, um dadurch die Identität der gesuchten Person heraus zu finden. Bei der Verifizierung wird die Gleichheit zweier Bilder geprüft, bspw. um zu überprüfen, ob die Person, die versucht das Handy zu entsperren, auch wirklich der/die Besitzer*in ist.

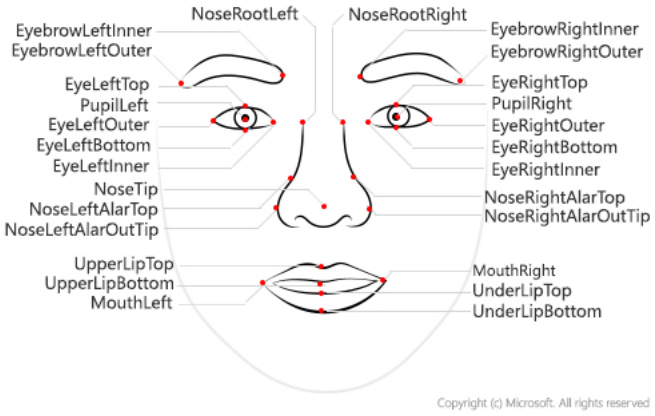


Abbildung 3.4: Face Landmarks (Microsoft 2021)

3.2.1.2 Face Attribute Classification (FAC)

Im vorherigen Abschn. 3.2.1.1 wurde beschrieben, wie FR durch die Bestimmung von geometrischen Eigenschaften funktioniert. Das Gesicht kann jedoch, neben rein geometrischen Parametern, auch auf andere Attribute untersucht werden und nennt sich FAC.

Auch bei der FAC muss zunächst eine Face Detection durchgeführt werden, um das Gesicht von uninteressanten Hintergrundinformationen zu trennen. Die anschließende Feature Extraction hingegen extrahiert beschreibende visuelle Eigenschaften wie Hautfarbe, Haarfarbe, das Tragen einer Brille, Bart etc. Ein Beispiel hierfür ist in Abb. 3.5 zu sehen. Es werden jeweils zwei Bilder miteinander verglichen: Im linken Bild handelt es sich um die selbe Person, im Rechten nicht. Unten sind die Attribute zu sehen, deren Ausprägung von der FAC bewertet wurde. Jedem Attribut wird eine Ausprägung zwischen -1 und 1 zugeordnet, je nachdem wie

stark das Attribut vorhanden ist. Auch hierfür werden häufig CNNs genutzt (Thom & Hand 2021).

Prinzipiell können mit der FAC die gleichen Anwendungen realisiert werden, wie mit der FR, die nur geometrische Eigenschaften nutzt. Schon 2009 zeigte Vaquero et al. (2009), wie FAC für Überwachungszwecke genutzt und das System um Fragestellungen wie „Zeige mir alle Menschen mit Bart, rotem T-Shirt und Sonnenbrille, die letzten Samstag das Gebäude X betreten haben“ erweitert werden kann.

Viele gängige FRT Softwares verfügen auch über Funktionen der FAC. Im nächsten Abschnitt wird z.B. die Identifizierung des Geschlechts näher betrachtet.

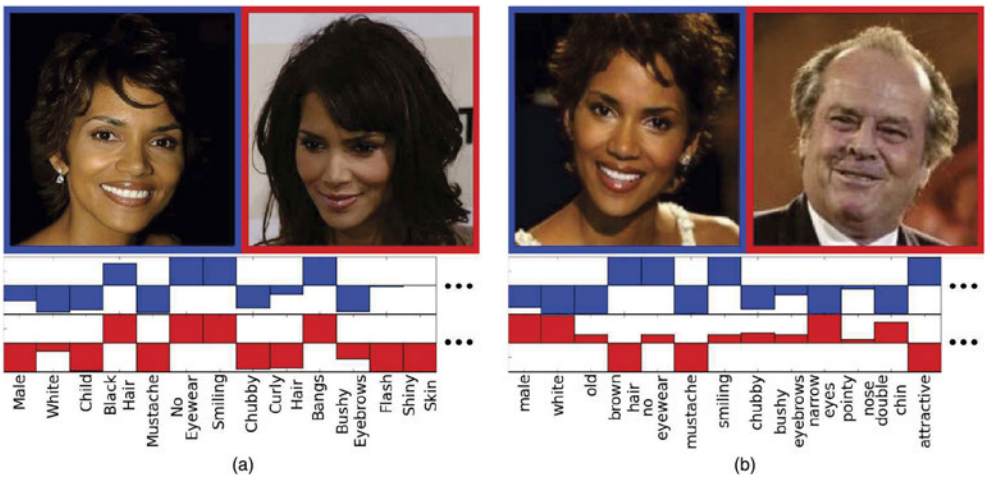


Abbildung 3.5: Beschreibende visuelle Attribute; (a) zwei Fotos der selben Person, (b) zwei Fotos unterschiedlicher Personen (Kumar et al. 2011)

3.2.2 Probleme der Technologien

Wie im vorigen Abschnitt beschrieben, werden für die FRT häufig CNNs verwendet. Ein großes Problem von CNNs ist, dass sie sehr viele Daten benötigen. In aktuellen Anwendungen werden teilweise Datensätze mit 500 Mio. Bildern von 10 Mio. Personen genutzt. Über die meisten großen Datenbanken verfügen Privatunternehmen und sind für die Öffentlichkeit nicht zugänglich. Grundsätzlich gibt es zwar genug Fotos auf Google oder anderen Suchmaschinen. Das Problem ist, dass die Bilder gelabelt sein müssen, da FRT i.d.R. supervised Methoden nutzt (s. Abschn. 2.1). Semi-supervised Methoden sind auf diesem Gebiet bisher noch nicht ausreichend erprobt (A. Bansal et al. 2021).

Neben gelabelten Daten ist auch die Qualität der Datensätze erheblich für das Ergebnis. So sollten sie nach Möglichkeit die gesamte Bevölkerung abbilden, mit allen Hauttypen und

gleicher Aufteilung nach Geschlechtern. Ansonsten werden die Teile der Bevölkerung, welche verstärkt in den Daten auftauchen, besser erkannt, als andere. Buolamwini & Gebru (2018) untersuchten hierfür FRT von Microsoft, IBM und Face++ auf die Fähigkeit, nach Geschlechtern zu klassifizieren. Hierfür erstellten sie eine Datenbank mit Fotos von Politiker*innen, da diese öffentlich zugänglich sind. Diese Datenbank soll alle Hauttypen und Geschlechter gleichermaßen abbilden. Für Untersuchungen dieser Art existieren eigentlich Benchmark Datensätze der NIST (National Institute for Standards and Technology), einer Bundesbehörde der U.S., welche für Standardisierungsprozesse zuständig ist. Diese seien laut Buolamwini & Gebru (2018) jedoch ebenfalls nicht ausreichend divers.

Mit der neu erstellten Datenbank von Buolamwini & Gebru (2018) wurde eine Statistik erhoben, wie gut die eben genannten Herstellerfirmen das Geschlecht klassifizieren. Dabei fanden sie heraus, dass alle Herstellerfirmen Frauen (zwischen 8.1% u. 20.6%) schlechter als Männer klassifizierten und Menschen mit dunklem Hauttyp (zwischen 11.8% u. 19.2%) schlechter klassifizierten als mit hellerem Hauttyp. Dies ist vor allem auf den Bias der Trainingsdatensätze zurückzuführen: Zum Beispiel besteht die weit verbreitete Datenbank „LFW“, welche Fotos von Prominenten beinhaltet, zu 77.5% aus Männern und 83.5% aus Weißen (Han & Jain 2014). Diese Art von Bias wird Representation Bias genannt.

Auch das NIST kommt laut United States. Governement Accountability Office (2020) zu ähnliche Beobachtungen. Sie kommen zu dem Schluss, dass zwar die Fehlerraten seit 2013 (bis 2020) um das 20 fache gesunken sind, dafür jedoch Unterschiede in der Genauigkeit für unterschiedliche Bevölkerungsgruppen zugenommen haben. Dies liegt neben dem Representation Bias auch daran, dass Trainingsdatenbanken häufig nur Fotos mit sehr guter Bildqualität beinhalten und Personen nur zu einem bestimmten Zeitpunkt zeigen. Bilder von der selben Person können aber sehr unterschiedlich aussehen, u.A. durch Unterschiede in der Beleuchtung, variierendem Gesichtsausdruck, durch das Tragen von Brillen, unterschiedliche Bildqualität oder Alterung.

Die Trainingsdaten verfügen somit über einen Evaluation Bias, da die Trainingsdaten nicht der tatsächlichen Anwendung entsprechen. Das Aussehen kann stark variieren und auch die Bildqualität, wie z.B. von einer Überwachungskamera, kann sehr unterschiedlich sein.

Wie auch schon bei „ShotSpotter“ in Abschn. 3.1 herausgefunden wurde, führt auch die Ungenauigkeit der Gesichtserkennung zu einer Diskriminierung gegen bestimmte Gruppen. Dies ist zurückzuführen auf den Representation Bias und den Evaluation Bias.

Der Bias im Trainingsdatensatz kann neben nicht vorhandener Daten auch durch falsches Labeling entstehen. Insbesondere bei der FAC wird das Labeling der Daten häufig durch Crowdsourcing Plattformen wie „Amazon Mechanical Turk“ realisiert. Auf der Plattform können so

genannte Mikrotasks von Menschen gegen eine geringe Summe Geld ausgeführt werden. Für FAC könnte das bedeuten: Man erhält eine geringe Menge an Bildern von Personen und muss ihnen Attribute wie Haarfarbe, Brillenträger, Lächeln, Alter, Attraktivität o.Ä. zuordnen. Vor allem bei den beiden letzten Attributen „Attraktivität“ und „Alter“ wird ersichtlich, dass es sich um Attribute handeln kann, die nur subjektiv bewertet werden können. Dadurch entsteht so genanntes Rauschen der Daten, was wiederum zu Bias führen kann (Thom & Hand 2021). Dieser Bias entspricht einem Measurement Bias.

Es entsteht auch die Frage, ob die Label an sich schon einen Bias aufweisen oder sogar diskriminierend sind. So beinhaltet die frei verfügbare Datenbank „Facetracer“ von Kumar et al. (2008), deren Labels häufig als Referenz für weitere Datenbanken genutzt wird (Thom & Hand 2021), folgende Labels (aus Textdatei zugehörig zur Datenbank entnommen):

```
# FaceTracer Dataset v1.0 - attributes.txt -
http://www.cs.columbia.edu/CAVE/databases/facetracer/
# Format: attribute label1 label2 ...
gender male female
race asian white black
age baby child youth middle_aged senior
hair_color blond not_blond
eye_wear none eyeglasses sunglasses
mustache true false
expression smiling not_smiling
blurry true false
lighting harsh flash
environment outdoor indoor
```

Ein anderes Beispiel ist die für die FAC häufig verwendete Datenbank „CelebA“ (Liu et al. 2015). Sie verfügt über folgende Labels (kopiert aus der Datenbank):

```
5_o_Clock_Shadow Arched_Eyebrows Attractive Bags_Under_Eyes
Bald Bangs Big_Lips Big_Nose Black_Hair Blond_Hair Blurry
Brown_Hair Bushy_Eyebrows Chubby Double_Chin Eyeglasses
Goatee Gray_Hair Heavy_Makeup High_Cheekbones Male
Mouth_Slightly_Open Mustache Narrow_Eyes No_Beard Oval_Face
Pale_Skin Pointy_Nose Receding_Hairline Rosy_Cheeks
Sideburns Smiling Straight_Hair Wavy_Hair Wearing_Earrings
Wearing_Hat Wearing_Lipstick Wearing_Necklace
Wearing_Necktie Young
```

Die Beispiele zeigen, dass auch Klassifizierungen nach Attributen gemacht werden, die zu den geschützten Merkmalen gehören (s. Abschn. 2.6), wie z.B. „Geschlecht“ oder „Ethnizität“. Das muss nicht zwingend eine Diskriminierung nach sich ziehen. Allerdings gibt es bspw. beim Attribut Geschlecht nur die Auswahlmöglichkeit der Labels „Männlich“ oder „Weiblich“. Non-Binary oder genderqueere Identitäten werden dabei völlig außer Acht gelassen. Auch die Einteilung nach „Asiatisch“, „Schwarz“ und „Weiß“ muss hinterfragt werden: So handelt es sich bei „Schwarz“ und „Weiß“ um gesellschaftlich konstruierte Gruppen, „Asiatisch“ hingegen betrifft die Herkunft.

Labels und damit die Features der KI können somit eine weitere Ursache von Diskriminierung sein, wenn Sie nicht über alle Ausprägungen verfügen und dadurch Personen in Gruppen drängen, zu denen sie nicht zugehören.

Abschließend soll ein Fall geschildert werden, welcher aufzeigen soll, welche Auswirkungen Fehlentscheidungen von FRT haben können:

2019 wurde laut (CNN Business 2021) der Schwarze Nijeer Parks aus New Jersey fälschlicherweise von FRT identifiziert. Er wurde beschuldigt für u.A. illegalen Waffenbesitz, Ladendiebstahl und soll versucht haben, einen Polizisten mit dem Auto anzufahren. Für die Erkennung nutzte das NYPD das Foto einer Fake-ID, welche am Tatort gefunden wurde. Nach elf Tagen Untersuchungshaft wurde Parks freigelassen, da er unschuldig war. Er befand sich zur Tatzeit nachweislich an einem 30 Meilen weit entfernten Ort. Von dem Einsatz von FRT zur Identifizierung soll er erst erfahren haben, nachdem er aus der Haft entlassen wurde.

Viele weitere Fälle und Auseinandersetzungen mit der durch FRT verursachten Diskriminierung werden u.A. durch die Initiative „Ban the Scan“ von Amnesty International (2021) dokumentiert.

3.3 Schlussfolgerung

Im Folgenden sollen die Ursachen von Diskriminierung, die in diesem Kapitel erfasst wurden, zusammengetragen und daraus allgemeine Aussagen für KI abgeleitet werden:

3.3.1 Korrektheit von KI Systemen und Bias

Die Untersuchung von MacArthur Justice Center (2021) hat gezeigt, dass die Anzahl an falschen Alarmen (False-Positives), verursacht durch „ShotSpotter“, um ein vielfaches höher sei, als durch herkömmliche 9-1-1 Notrufe (vgl. Abb. 3.2). Welche Folgen das hat, ist in Abschn. 3.1 beschrieben. In Kombination mit der in Abschn. 3.3.4 beschriebenen ungleichen Anwendung der Software, trifft die Ungenauigkeit der Software vorrangig Schwarze und Hispanics.

Wie in Abschn. 3.2.2 beschrieben, hat auch die Ungenauigkeit von FRT eine Diskriminierung gegen Schwarze zur Folge. Dies ist laut Buolamwini & Gebru (2018) zurückzuführen auf den Representation Bias der Trainingsdaten, die vorrangig Weiße Männer beinhalten und demnach weniger korrekt für Schwarze und Frauen sind. Zudem wird von vielen FRTs u.A. die Datenbank „CelebA“ oder „LFW“ als Datengrundlage genutzt (Thom & Hand 2021). Diese Fotos haben meist eine sehr gute Qualität, in denen das Gesicht kaum verdeckt ist. Sie weisen einen Evaluation Bias auf, da in typischen Anwendungsfällen z.B. Bilder von Überwachungskameras genutzt werden, die häufig nur Bilder in schlechter Qualität aufnehmen. Zudem können fehlerhaft gelabelte Daten zu Rauschen führen, was einen Bias zur Folge haben kann (Thom & Hand 2021). Fehlerhafte Labels können auftreten, wenn es sich um subjektive Attribute wie „Attraktivität“ oder „Alter“ handelt. Dabei handelt es sich um einen Measurement Bias.

Allgemein kann der Schluss gezogen werden, dass die Korrektheit von KI Systemen einen großen Einfluss auf die Diskriminierung hat. Sie wird maßgeblich beeinflusst durch die Trainingsdaten, die verschiedene Arten von Bias aufweisen können. Daraus entstehen Fehlentscheidungen, die unterschiedliche Folgen für verschiedene Gruppen haben können. Diese Folgen können bestimmte Gruppen benachteiligen und somit diskriminierend sein.

Dem entgegenwirken kann die Verwendung von Datensätzen, die alle Teile der Bevölkerung abbilden. Buolamwini & Gebru (2018) hat beispielsweise für die FRT einen Datensatz erstellt, der alle verschiedenen Hauttypen abbildet. Zudem sollte die gleiche Anzahl an Daten je Bevölkerungsgruppe vorhanden sein. Wenn eine Gruppe überrepräsentiert ist, kann die Anzahl der Daten auch reduziert werden, wie von Deutscher Bundestag (2020) vorgeschlagen wird. Dies hat dann jedoch einen negativen Einfluss auf die Korrektheit.

3.3.2 Feature Auswahl

In Abschn. 3.2.1.2 wird gezeigt, dass die Auswahl der Features, nach denen die KI klassifiziert, diskriminierend sein kann. In dem exemplarisch vorgestellten Datensatz „CelebA“ hat der Parameter „Geschlecht“ die Ausprägung „Männlich“ oder „Weiblich“, wodurch non-binary oder genderqueere Identitäten nicht beachtet werden. Die Ethnizität im Datensatz „Facetracer“ hat die Ausprägungen „Schwarz“, „Weiß“ und „Asiatisch“. „Schwarz“ und „Weiß“ entspricht jedoch einer gesellschaftlich konstruierten Gruppe, wobei „Asiatisch“ die Herkunft betrifft. Die Daten verfügen durch die unzureichenden Ausprägungen der Features über einen Measurement Bias. Während der Anwendung kann ein Aggregation Bias entstehen, da die Ausprägung der Attribute nicht ausreicht, um die Person zu beschreiben.

Allgemein kann der Schluss gezogen werden, dass die Auswahl der Features einen Einfluss auf Diskriminierung haben kann. Dies ist der Fall, wenn die Features nicht alle möglichen Ausprägungen abbildet und dadurch Personen in Gruppen gedrängt werden, denen sie eigentlich nicht angehören.

3.3.3 Qualitätsmaß

Unter den vorgestellten Anwendungen haben besonders die des „Predictive Policing“ das Problem, keine objektive Messung der Qualität der Systeme zuzulassen. Der Einsatz des Systems selbst beeinflusst schon den Ausgang.

Es liegt eine statistische Diskriminierung vor, da „Verhaftungen“ als Proxy für „Kriminalität“ genutzt werden. Der Anstieg an Verhaftungen lässt einen Anstieg in der Kriminalität vermuten, ist aber durch die Polizeipräsenz selbst beeinflusst, wie in Abschn. 3.1 am Beispiel „Geolitica“ erklärt. Man spricht auch von einem Measurement Bias, wie von Suresh & Guttag (2019) als Beispiel genannt. Ein objektives Qualitätsmaß würde also die Messung der tatsächliche Kriminalität bedeuten, was nicht möglich ist, da hier auch unerkannte Straftaten mit einbezogen werden müssten.

Allgemein kann der Schluss gezogen werden, dass ein objektives Qualitätsmaß für die Bewertung von KI notwendig ist. Wenn dieses nicht vorhanden ist, kann die Qualität der Aussagen des Systems falsch bewertet werden und zu Diskriminierung führen.

3.3.4 Einsatzgebiet

Nicht nur bei der Entwicklung von KI kann Diskriminierung verursacht werden. Am Beispiel „ShotSpotter“ wird deutlich, auch bei der Nutzung der Software kann Diskriminierung auftreten. MacArthur Justice Center (2021) kritisiert „ShotSpotter“, verstärkt in Gegenden mit hohem

Anteil an Schwarzen und Hispanics angewendet zu werden. In anderen Gebieten, in denen der Anteil dieser Gruppen eher gering ist, wird die Software kaum bis gar nicht genutzt. Die Folgen von Ungenauigkeiten (vgl. Abschn. 3.3.1), wie im Fall „ShotSpotter“ die falschen Alarme, sind dadurch ungleich verteilt und können zu Diskriminierung führen.

Allgemein kann der Schluss gezogen werden, dass die Nutzung von KI Diskriminierung hervorrufen kann, wenn sie selektiv angewendet wird. Die negativen Folgen von Ungenauigkeiten treffen dadurch nur diejenigen Menschen, die von der Anwendung betroffen sind.

In diesem Kapitel konnte geklärt werden, dass KI durchaus vorherrschende Diskriminierung reproduzieren. Dies macht sich schon in den Trainingsdatensätzen, also im Entwicklungsprozess der KI, bemerkbar, aber auch in der späteren Anwendung der KI. Nachdem in diesem Kapitel Ursachen von Diskriminierung anhand von Anwendungsbeispielen herausgearbeitet wurden, soll im nächste Kapitel geklärt werden, mit welchen Methoden die Ursachen aufgedeckt werden können. Dafür wird die Analyse von Angwin et al. (2016) vorgestellt. Anschließend sollen die Erkenntnisse aus diesem Kapitel mit denen von Angwin et al. (2016) verglichen werden.

4 Anwendungsbeispiel „COMPAS“

In diesem Kapitel soll die Risk Assessment Technology Software „COMPAS“ untersucht werden. Dafür wird die Analyse von dem non-profit newsroom ProPublica herangezogen, die von Angwin et al. (2016) durchgeführt wurde, um tiefgehende Einblicke in die Software zu bekommen. Es wird wie folgt vorgegangen:

Als Einführung soll die Funktionsweise von „COMPAS“ vorgestellt werden. Anschließend wird die Analyse von Angwin et al. (2016) erläutert. Sie verfolgen einen modell-agnostischen Ansatz, um die Interpretierbarkeit (s. Abschn. 2.4) von „COMPAS“ zu steigern. Die bei der Analyse auftretenden Erkenntnisse sollen dann mit denen aus Abschn. 3.3 verglichen werden. Zum Schluss soll eine Bewertung der Analyseansätze zur Übertragbarkeit auf andere Anwendungen gemacht werden.

4.1 Risk Assessment Technology Software „COMPAS“

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) ist eine Risk Assessment Technology Software der Firma Equivant, die besser bekannt ist unter dem alten Firmennamen Northpointe Inc. Da das Unternehmen zum Zeitpunkt der Studie noch Northpointe Inc. hieß, wird sie im Folgenden mit dem alten Firmennamen benannt.

Die Software wird zur Risikobewertung der Rückfälligkeit von Straftäter*innen genutzt. Sie ist eine der meist verbreitetsten Risk Assessment Softwares in den US, darunter Broward County, Florida, New York State, Wisconsin State, California State (Kirkpatrick 2017).

Welche Daten genau für die Bewertung genutzt werden, wird nicht transparent kommuniziert. Jedoch basiert die Berechnung zu einem großen Teil auf einem Fragebogen mit über 137 Fragen. Es werden Fragen gestellt wie: „War ein Elternteil jeweils im Gefängnis?“, „Wie viele Freunde nehmen illegale Drogen?“, „Waren Angehörige oder Freunde in der Nachbarschaft Opfer von Kriminalität?“ oder „Wie oft sind Sie gelangweilt?“. Weiter mit einbezogen sein sollen u.a. der Bildungsstand, Arbeitslosigkeit und Vorstrafen (Angwin et al. 2016).

Die Straftäter*innen erhalten von „COMPAS“ drei Scores in den Kategorien:

- Risk of Recidivism (Rückfallrisiko)
- Risk of Violence (Gewaltbereitschaft)
- Risk of Failure to Appear (nicht untersucht in der Studie)

Die Scores reichen von 1 (lowest risk)-10 (highest risk) und sind eingeteilt in die Gruppen:

- Low: 1-4
- Medium: 5-7
- High: 8-10

„Rückfall“ bedeutet ein wiederholter Gefängnisaufenthalt, nachdem der letzte Aufenthalt abgeschlossen ist. Nicht als Rückfall bewertet werden Ordnungswidrigkeiten wie Strafzettel, nicht Erscheinen bei der Anhörung oder Straftaten, die nach der Bewertung des Risikoscores aufgedeckt werden, aber vor der Bewertung begangen wurden. Als Gewaltverbrechen zählt Mord, Totschlag, Vergewaltigung, Raub, schwere Körperverletzung. Diese Definitionen wurden von Northpointe Inc. selbst gemacht und von Angwin et al. (2016) für die folgende Analyse übernommen. Sie soll im nächsten Abschnitt erklärt werden.

4.2 Analyse von „COMPAS“ durch Angwin et al. (ProPublica)

Für die Analyse von „COMPAS“ mussten Angwin et al. (2016) zunächst Daten der Straftäter*innen beschaffen, die sie als Grundlage nutzen können. Dafür arbeiteten sie mit dem Broward County Sherrifs Office zusammen und untersuchten 11.757 Bewertungen aus den Jahren 2013 und 2014. Neben den Bewertungen erhalten sie auch Informationen zu den Straftäter*innen wie Ethnizität oder Geschlecht. Broward County nutzt das Tool um zu bewerten, ob ein*e Straftäter*in vor der Verhandlung freigelassen wird oder in U-Haft bleibt.

Angwin et al. (2016) verglichen die durch „COMPAS“ berechnete Rückfälligkeitsrate mit den tatsächlichen Rückfällen in einem Zeitraum von zwei Jahren ab der Bewertung. Zudem untersuchten sie, wie sich der Einfluss von Eigenschaften wie „Geschlecht“ oder „Ethnizität“ auf die Bewertung auswirken. Die Untersuchungsmethoden sollen im Folgenden vorgestellt werden:

Für die Untersuchungen wurden die Risikoklassen „Medium“ und „High“ zusammengefasst, sodass es nur zwei Ausprägungen gibt: „Low“ (Low) oder „High“ (Medium/High). Zunächst wurde die false-positive und false-negative Rate ermittelt:

	FP Rate		**FN Rate**	
	Schwarze	Weiße	Schwarze	Weiße
Risk of Recidivism	45%	23%	28%	48%
Risk of Violence	38%	18%	38%	62%

Tabelle 4.1: FP (false-positive) und FN (false-negative) Raten ermittelt aus der tatsächlichen Rückfälligkeitsrate von Angwin et al. (2016) (eigene Darstellung)

Die Zahlen können rein aus der Auswertung ermittelt werden, wie viele Personen mit hohem Risiko (also Klasse Medium/High) bewertet wurden und wie viele Personen tatsächlich rückfällig geworden sind. Dabei bedeutet FP (false-positive), dass eine Person der jeweiligen Gruppe fälschlich mit hohem Risiko bewertet wurde, obwohl die Person im betrachteten Zeitrum von zwei Jahren nicht rückfällig geworden ist. FN (false-negative) bedeutet also umgekehrt, dass eine Person der jeweiligen Gruppe fälschlich mit niedrigem Risiko bewertet wurde, obwohl sie im betrachteten Zeitraum rückfällig geworden ist. So ergibt sich z.B. für den „Risk of Recidivism“ eine false-positive Rate von 45% für Schwarze, Weiße hingegen eine false-positive Rate von 23%. Zu sehen ist, dass Schwarze in beiden Scores öfter fälschlich mit hohem Risiko eingestuft werden. Zudem werden Weiße in beiden Scores öfter fälschlich mit niedrigem Risiko bewertet. Bei der false-positive bzw. false-negative Rate handelt es sich um ein Qualitätsmaß (vgl. Abschn. 2.3).

Die eben vorgestellte Auswertung kann eine Aussage darüber treffen, wie korrekt „COMPAS“ vorhersagen trifft und liefert erste Indizien für einen Bias. Um jedoch Zusammenhänge zwischen dem Score und Eigenschaften wie der Ethnizität herstellen zu können, muss ein weiterer Analyseansatz verfolgt werden:

COMPAS fungiert als Blackbox, da der Quellcode der Software nicht ersichtlich ist und lediglich die Eingabe und Ausgabe betrachtet werden kann. Wie in Abschn. 2.5 erklärt, gibt es verschiedene Verfahren, um die Interpretierbarkeit von Blackboxen zu verbessern und dadurch das Verhalten besser zu verstehen. Angwin et al. (2016) verwenden hier eine modell-agnostische Methode, das GSM, wie es von Molnar (2019) genannt wird. Dabei wird ein Ersatzmodell erstellt, welches eine bessere Interpretierbarkeit vorweist, als das Originalmodell (hier die Blackbox „COMPAS“). Angwin et al. (2016) verwenden zwei Ersatzmodelle, die logistische Regression, die im Folgenden genauer betrachtet werden soll, und die *Cox Proportional Hazard Regression.*

Risk of General Recidivism Logistic Model

	Dependent variable:
	Score (Low vs Medium and High)
Female	0.221*** (0.080)
Age: Greater than 45	-1.356*** (0.099)
Age: Less than 25	1.308*** (0.076)
Black	0.477*** (0.069)
Asian	-0.254 (0.478)
Hispanic	-0.428*** (0.128)
Native American	1.394* (0.766)
Other	-0.826*** (0.162)
Number of Priors	0.269*** (0.011)
Misdemeanor	-0.311*** (0.067)
Two year Recidivism	0.686*** (0.064)
Constant	-1.526*** (0.079)
Observations	6,172
Akaike Inf. Crit.	6,192.402
Note:	*p<0.1; **p<0.05; ***p<0.01

(a) Generelles Rückfallrisiko

Risk of Violent Recidivism Logistic Model

	Dependent variable:
	Score (Low vs Medium and High)
Female	-0.729*** (0.127)
Age: Greater than 45	-1.742*** (0.184)
Age: Less than 25	3.146*** (0.115)
Black	0.659*** (0.108)
Asian	-0.985 (0.705)
Hispanic	-0.064 (0.191)
Native American	0.448 (1.035)
Other	-0.205 (0.225)
Number of Priors	0.138*** (0.012)
Misdemeanor	-0.164* (0.098)
Two Year Recidivism	0.934*** (0.115)
Constant	-2.243*** (0.113)
Observations	4,020
Akaike Inf. Crit.	3,022.779
Note:	*p<0.1; **p<0.05; ***p<0.01

(b) Rückfallrisiko für Gewaltstraftaten

Abbildung 4.1: Logistisches Modell zur Einteilung in die Klasse „Low“ oder „Medium and High“ (Angwin et al. 2016)

Letztere soll im Rahmen der Arbeit nicht näher erklärt werden, da es sich um die gleiche modell-agnostische Methodik handelt und die Unterschiede nur im gewählten Ersatzmodell liegen.

Durch die Erstellung einer logistischen Regression (s. Abschn. 2.1.1) kann der Einfluss von den (unabhängigen) Parametern wie „Alter“, „Ethnizität“ oder „Geschlecht“ der Angeklagten auf den (abhängigen) Parameter Risikoscore quantifiziert werden. Der Risikoscore kann, wie weiter oben beschrieben, nur zwei Ausgänge haben: „Low“ und „High“. Dadurch handelt es sich, wie in 2.1.1 erklärt, um eine binäre Klassifizierung. Das Ergebnis der Regression ist in Abb. 4.1 zu sehen.

Wie zu Beginn erklärt, werden zwei Risikoscores für jede*n Straftäter*in berechnet: Einen für das generelle Rückfallrisiko und einen für die Gewaltbereitschaft. Es wurden deshalb zwei Regressionen erstellt, jeweils eine pro Score.

An den Parametern sind die β Koeffizienten zu sehen, die im Zuge des Trainings erlernt wurden. Wie in 2.1.1 erklärt, können diese Koeffizienten nicht direkt interpretiert werden. Sie

müssen erst in die Formel 2.2, der Formel für die logistische Regression, eingesetzt werden und können dadurch in eine Wahrscheinlichkeit umgerechnet werden.

Beispielhaft sei hier der Parameter „Age: Less than 25“ betrachtet. Er soll ins Verhältnis zu den Personen gesetzt werden, die diese Ausprägung nicht haben. Daraus ergibt sich als Wahrscheinlichkeit für $y = 1$, der Einteilung in die Risikoklasse „High“, folgende Rechnung:

$$P(y(x_{AgeLessThan25}) = 1) = \frac{1}{1 + e^{-(\beta_{const} + \beta_{AgeLessThan25} x_{AgeLessThan25})}} * \left(\frac{1}{1 + e^{-(\beta_{const})}} \right)^{-1} \tag{4.1}$$

Alle anderen x außer $x_{AgeLessThan25}$ werden 0 gesetzt. Da $x_{AgeLessThan25}$ ein binärer Parameter ist, gilt $x_{AgeLessThan25} = 1$. Mit $\beta_{const} = -1.526$, $\beta_{AgeLessThan25} = 1.308$ ergibt sich dann für den „Risk of General Recidivism“: $P(y(x_{AgeLessThan25}) = 1) = 2.5$. Das bedeutet, dass Personen, die jünger als 25 sind, eine 2.5 fach höhere Wahrscheinlichkeit haben, mit hohem Risiko bewertet zu werden. Der Effekt der anderen Parameter wurde durch das null setzten eliminiert, sodass es z.B. irrelevant ist, ob die Person Männlich oder Weiblich ist.

Analog zu der Rechnung ergibt sich für Schwarze Angeklagte folgende Rechnung:

$$P(y(x_{Black}) = 1) = \frac{1}{1 + e^{-(\beta_{const} + \beta_{Black} x_{Black})}} * \left(\frac{1}{1 + e^{-(\beta_{const})}} \right)^{-1} \tag{4.2}$$

Mit $x_{Black} = 1$, $\beta_{const} = -1.526$ und $\beta_{Black} = 0.477$ ergibt sich dann für $P(y(x_{Black}) = 1) = 1.45$. Schwarze Angeklagte haben dadurch eine 45% höhere Wahrscheinlichkeit, mit hohem Risiko bewertet zu werden, als Weiße. Für den „Risk of Violence“ ergibt sich mit den Koeffizienten $\beta_{const} = -2.243$ und $\beta_{Black} = 0.659$ für $P(y(x_{Black}) = 1) = 1.77$ eine 77% höhere Wahrscheinlichkeit.

Es wird ersichtlich, dass durch die logistische Regression der Einfluss von Parametern auf eine Zielgröße berechnet werden kann, bei gleichzeitigem eliminieren von anderen Parametern. So könnten auch weitere Fragestellungen beantwortet werden, wie z.B. „Wie viel höher ist die Wahrscheinlichkeit, dass eine Frau mit Vorstrafen, im Vergleich zu einem Mann mit Vorstrafen, mit hohem Risiko bewertet wird?“. Hierfür müssen lediglich die gewünschten Parameter mit ihren β Koeffizienten in die Formel 2.2 eingesetzt werden.

Wie geschildert, haben Angwin et al. (2016) als GSM auch eine Cox Proportional Hazard Regression verwendet. Diese soll im Rahmen der Arbeit nicht näher erklärt werden, da es sich um die gleiche modell-agnostische Methodik handelt und die Unterschiede im Ersatzmodell

liegen. Mit der Analyse wurde u.A. die Genauigkeit des Modells bestimmt. Dafür wurde die so genannte Konkordanz (s. Abschn. 2.3) bestimmt und ein Wert von 63.6% für den „Risk of Recidivism“ und 65.1% für den „Risk of Violence“ berechnet. Diese sind beide laut Angwin et al. (2016) schlechter als die von Northpointe Inc. angegebenen 68%.

4.3 Bewertung von „COMPAS“

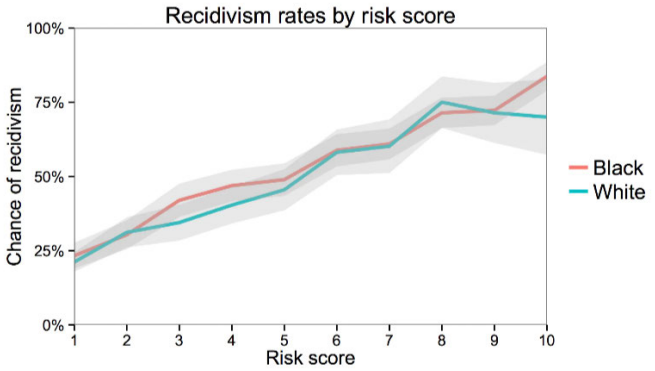
Nachdem in den vorigen Abschnitten die Software „COMPAS“ und die Analyse durch Angwin et al. (2016) vorgestellt wurde, soll in diesem Abschnitt überprüft werden, welche der Erkenntnisse aus Abschn. 3.3 in der Analyse wiedergefunden werden können.

Die Analyse von Angwin et al. (2016) hat u.A. untersucht, wie korrekt das System arbeitet, indem die vorhergesagten Risikoscores mit den tatsächlichen Rückfällen übereinstimmt. Dabei wurde festgestellt, dass die false-positive Rate für Schwarze höher ist, also Schwarze häufiger falsch mit einem hohem Risiko bewertet werden, als Weiße. Zudem ist die false-negative Rate für Weiße geringer, sodass Weiße häufiger fälschlich mit niedrigem Risiko bewertet werden, als Schwarze. Beide Fehlerraten fallen zugunsten von Weißen aus. Somit konnte auch bei „COMPAS“ eine Diskriminierung auf die Ungenauigkeit der Software zurück geführt werden, wie in Abschn. 3.3.1 beschrieben.

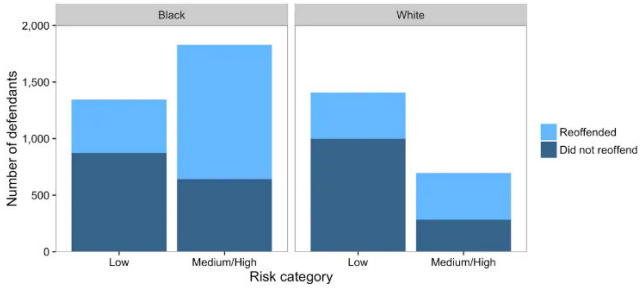
Es liegt also ein Bias vor, jedoch ist schwer zu bewerten, wo dieser genau auftritt. So könnte ein „Representation Bias“ in den Trainingsdaten von „COMPAS“ vorliegen. Wenn z.B. vorrangig Schwarze in den Trainingsdaten vorhanden sind, könnte die Zugehörigkeit zu der Gruppe als Ursache für Kriminalität erlernt werden. Mit der logistischen Regression wurde zwar sichtbar gemacht, dass die Ethnizität einen Einfluss auf die Risikobewertung hat, jedoch kann nicht geklärt werden, warum das so ist. Das GSM verbessert die Interpretierbarkeit durch Sichtbarmachung des Verhaltens von „COMPAS“, die Erklärbarkeit kann aber nur durch die Einsicht der Daten und des Modells, also durch Transparenz, gelöst werden. Hier wird deutlich, warum sich die Erklärbarkeit aus der Interpretierbarkeit und Transparenz zusammensetzt (vgl. Abschn. 2.4) und jeweils in den Ethik Richtlinien aus Abschn. 2.2 gefordert wird.

Angwin et al. (2016) hat durch den Abgleich der prognostizierten Rückfallquote mit der tatsächlichen Rückfallquote die Korrektheit der Software überprüft und damit zwei Qualitätsmaße definiert: Zum einen die false-positive und false-negative Rate und zum Anderen die Konkordanz. Auch hier besteht jedoch die Problematik, dass keine vollständige Objektivität gegeben sein kann (vgl. Abschn. 3.3.3). Die Bewertung durch „COMPAS“ beeinflusst vermutlich den Ausgang des Urteils über die Straftäter*innen, sonst wäre die Nutzung obsolet. Das Urteil kann dann wiederum die Handlungen von Straftäter*innen beeinflussen, wie auch Zweig

(2019) in ihrem Buch kritisiert. Die Bewertung mit hohem Risiko könnte somit demotivierend auf die Angeklagten wirken und eine Rückfälligkeit begünstigen. In Kombination mit der höheren false-positive Rate gegen Schwarze würde der eben beschriebene Effekt Schwarze mehr betreffen als Weiße und damit die Diskriminierung weiter verstärken.



(a) Rückfallquote in Prozent je Klasse



(b) Anzahl der Rückfälle nach Ethnizität

Abbildung 4.2: Vergleich der Rückfallquoten nach Risikoklasse und Ethnie (Corbett-Davies et al. 2016)

Abschließend soll noch auf den Konflikt zwischen Angwin et al. (2016) und Northpointe Inc. über das Fairnessmaß eingegangen werden. Northpointe Inc. verteidigt sich gegen die Vorwürfe von Angwin et al. (2016), diskriminierend gegen Schwarze zu sein, mit einer eigenen Studie, erstellt von Dieterich et al. (2016). Dieterich et al. (2016) argumentiert, dass „COMPAS“ unabhängig von der Ethnizität bewertet und eine nahezu gleiche prozentuale Rückfallquote je Klasse herrscht. Dies ist in Abb. 4.2 zu sehen, welche auf der selben Datenbasis, wie die Analyse von Angwin et al. (2016) basiert:

Im oberen Bild ist die Rückfallquote je Risikoscore (hier 1-10 statt „Low“ und „High“), im Vergleich zwischen Schwarzen und Weißen, aufgetragen. Ein höherer Risikoscore bedeutet demnach für beide Gruppen ein ähnlich hohes Rückfallrisiko. Im unteren Bild ist die Anzahl

der Angeklagten je Risikoklasse („Low“ oder „Medium/High“) und der Anteil der rückfällig gewordenen Angeklagten (hellblau) innerhalb dieser Risikoklasse, für Schwarze (links) und Weiße (rechts) Angeklagte zu sehen.

Angenommen, Schwarze haben generell eine höhere Rückfallquote, als Weiße (dies mag historisch bedingt und ebenfalls unfair sein). Wenn nun die Rückfallquote je Gruppe und Risikoscore gleich sein soll, wie es in Abb. 4.2a zu sehen ist, dann muss der (absolute) Anteil der fälschlich mit hohem Risiko bewerteten Schwarzen Angeklagten zwangsläufig größer sein. Dies wird in Abb. 4.2b deutlich: Der (absolute) Anteil der nicht-rückfällig gewordenen Angeklagten ist unter Schwarzen größer als unter Weißen, so wie es Angwin et al. (2016) kritisieren (vgl. Corbett-Davies et al. (2016)).

Bei der Frage, welches der beiden Fairnessmaße fairer ist, handelt es sich um eine interdisziplinäre Fragestellung. Sie kann deshalb nicht im Rahmen dieser Arbeit bewertet werden. Dennoch zeigt der Konflikt eine generelle Schwierigkeit von Fairness auf: Die Quantifizierung von Fairness ist nicht trivial und kann je nach Perspektive anders bewertet werden. Wie Kleinberg et al. (2016) detaillierter beschrieben, führt die Erfüllung eines Fairnessmaßes (mit sehr wenigen Ausnahmen) zwangsläufig zu der Verletzung eines anderen.

In diesem Abschnitt konnten einige Punkte aus Abschn. 3.3 wiedergefunden und eine Diskriminierung erkannt werden. Die konkrete Ursache der Diskriminierung bleibt allerdings mangels Transparenz und eines objektiven Qualitätsmaßes unentdeckt. Im Folgenden soll die Anwendbarkeit des GSMs auf weitere Anwendungen, wie die in Kapitel 3 genannten, bewertet werden.

4.4 Bewertung der Übertragbarkeit der Analyse von Angwin et al. (ProPublica) auf weitere Anwendungen

Abschließend soll in diesem Abschnitt bewertet werden, wie das GSM für weitere Anwendungen genutzt werden kann. Dafür sollen auch die Anwendungsbeispiele aus Kapitel 3 herangezogen werden:

Mit Hilfe der logistischen Regression als GSM konnte das Verhalten von „COMPAS“ für einige Fragestellungen sichtbar gemacht werden. Diese lautet z.B. „Wie stark wird der Risikoscore dadurch beeinflusst, dass die Person Schwarz ist?“. Oder mittels der Cox Regression: „Wie häufig liegt das System richtig?“. Sie können aber nicht folgende Frage erklären: „Warum ist der Risikoscore so stark abhängig von der Ethnizität?“. Mit Hilfe des GSM kann nur das generelle Verhalten erklärt werden, weniger aber individuelle Entscheidungen (warum hat Datensatz x das Ergebnis y), wie auch Molnar (2019) beschreibt.

Das GSM eignet sich grundsätzlich für alle Anwendungen, in denen tabellarische Daten (auch strukturierte Daten genannt) verarbeitet werden. Das sind Daten, die in Tabellenform dargestellt werden können, also z.B. Features in Spalten und Instanzen in Zeilen gespeichert sind (Molnar 2019).

Eine Software wie „Geolitica“, die für Predictive Policing genutzt wird, könnte daraufhin untersucht werden, wie sich verschiedene Faktoren auf die Vorhersage auswirken. Hierfür würde man Daten benötigen, die jeweils die Gegebenheiten zu einem Zeitpunkt abbilden, z.B. durch Uhrzeit, Wetter, ermittelte Straftaten nach Wohneinheiten, Migrantenanteil. Zu jedem Datenpunkt muss außerdem die Vorhersage bekannt sein, also die zugehörige Risikostufen je Wohneinheit. Eine Wohneinheit könnte z.B. ein Bezirk sein. Mit den Daten kann anschließend ein Modell mit guter Interpretierbarkeit trainiert werden, wie z.B. die logistische Regression. Anhand der Parameter kann dann, analog zur Analyse von Angwin et al. (2016), der Einfluss der Faktoren auf die Vorhersage ermittelt werden. So könnte z.B. herausgefunden werden, wie stark eine hohe Risikostufe vom Migrantenanteil abhängig ist.

Die Software „ShotSpotter“ ließe sich mit der GSM z.B. daraufhin überprüfen, wie stark die Erkennung eines Schusses von der Umgebungslautstärke abhängt. Hierfür würde man einen Datensatz benötigen, der Geräusche von echten Schüssen und unterschiedlichen Waffentypen beinhaltet, sowie Geräusche, die Waffen ähneln. Diese müssten bei unterschiedlicher Umgebungslautstärke vorhanden sein. Weiterhin benötigt man für jeden Datensatz die Vorhersage von SShotspotter". Die Umgebungslautstärke kann dann durch einen Parameter (z.B. skalar in Dezibel) abgebildet und der Einfluss auf die Erkennung quantifiziert werden.

Für Anwendungen, in denen Bilder oder Texte verarbeitet werden, eignet sich das GSM weniger (Molnar 2019), weshalb für die FRT andere Methoden im Fokus stehen. Hierzu gehört z.B. die *Sensitivity Map*. Damit ist es möglich, die Teile eines Bildes zu markieren, welche für die Klassifizierung relevant sind (Molnar 2019). So ließe sich überprüfen, welche Teile des Bildes relevant für die Identifizierung einer Person oder für die Zuordnung eines Attributes, wie das Geschlecht, ausschlaggebend sind.

5 Fazit und Ausblick

Im Rahmen dieser Literaturarbeit wurde nach Belegen für Diskriminierung durch KI am Beispiel der Strafverfolgung gesucht. Dabei wurde erkannt, dass KI keineswegs frei von Vorurteilen entscheidet, sondern vorherrschende Diskriminierung reproduziert oder sogar verstärken kann. Nahezu alle gefundenen Anwendungen standen in der Kritik, diskriminierend zu sein.

Diskriminierung kann während des gesamten KI Prozesses, von der Entwicklung bis zur Anwendung, entstehen: So kann sich schon bei der Erstellung des Trainingsdatensatzes ein Bias manifestieren, indem Daten falsch gelabelt oder die Bevölkerung nur unzureichend abgebildet wird. Es wurden Anwendungen wie „Geolitica“ vorgestellt, die keine objektive Bewertung der Korrektheit zulassen, da die Anwendung selbst die Vorhersage beeinflusst. Daraus resultiert eine falsche Bewertung der Qualität, was die Entwicklung von qualitätsverbessernden Maßnahmen, wie z.B. gegen Diskriminierung, verhindert, da kein Problembewusstsein herrscht. Weiterhin verfügt jede KI über Ungenauigkeiten, die häufig nicht alle Menschen gleichermaßen betrifft, sondern vorrangig diejenigen, die schon ohne die Anwendung von KI benachteiligt sind.

Die Liste der Ursachen ist lange nicht vollständig. Es wurde keine eigene Untersuchung von KI-Modellen gemacht, sondern lediglich auf Analysen anderer Personen zurückgegriffen. Da Anwendungen in der Strafverfolgung i.d.R. kommerziell und nicht open source sind, können sie von außen nur als Blackbox betrachtet werden. Dies erschwert eine genaue Ursachenanalyse erheblich. Hier setzen Transparenzforderungen an, die in allen ethischen Richtlinien vorkommen. Transparenz gilt als Voraussetzung für die Erklärbarkeit von KI und damit die Realisierung ethischer KI. Es muss häufig viel Aufwand betrieben werden, um durch z.B. externe Daten eine Überprüfung zu ermöglichen, da die Trainingsdaten selbst nicht eingesehen werden können. Buolamwini & Gebru (2018) erstellten z.B. selbst ein Datenset, welches die verschiedenen Hauttöne und Geschlechter gleichermaßen abbildet und verwendeten dieses für die Überprüfung von FRT. Ein weiterer vergleichbarer Datensatz wurde von Hazirbas et al. (2021) erarbeitet. In der Arbeit wurde gezeigt, dass ein Bias durch externe Prüfung zwar sichtbar gemacht werden kann, jedoch ohne Transparenz nur schwer Rückschlüsse auf die tatsächliche Ursache gemacht werden können.

Transparenz alleine reicht jedoch nicht aus, um ein System erklärbar zu machen. Häufig sind KI-Modelle so komplex, dass sie vom Menschen nur noch schwer verstanden werden können. Oder sie stellen eine Blackbox dar, weil es sich um eine nicht quelloffene Software handelt. Hier setzt XAI an. Durch die Disziplin wird es möglich, die Interpretierbarkeit von KI zu steigern. Im Rahmen der Arbeit wurde der Ansatz des GSM vorgestellt und an einem praktischen Anwendungsbeispiel gezeigt, wie dieser Ansatz konkret bei der Aufdeckung von Diskriminierung helfen kann. Dies soll als Motivation für weitere Arbeiten verstanden werden, da XAI der Schlüssel für Vertrauen in KI-Anwendungen ist. Nur so ist es möglich, kausale Zusammenhänge zwischen Verhalten und Eingabedaten zu schaffen, Diskriminierung sichtbar zu machen und eine faire und ethische KI realisieren zu können. Einen umfassenden Überblick über die Methoden geben z.B. Gunning & Aha (2019), Burkart & Huber (2021) und Molnar (2019).

In dieser Arbeit wurden *Ursachen* und *Aufdeckung* von Diskriminierung behandelt. Am Ende nützen diese Erkenntnisse jedoch nichts, wenn sie ohne Konsequenzen bleiben. Es muss also auch überlegt werden, wie eine *Vermeidung* der Ursachen erreicht wird. Es handelt sich um eine stark interdisziplinäre Aufgabe, da Diskriminierung auch schon lange vor der Entwicklung von KI existierte. Hier müssen Politiker*innen, Informatiker*innen, sowie Personen aus den Fachbereichen Soziologie und Philosophie zusammen arbeiten. Ein Versuch war u.A der Einsatz einer Enquete-Kommission durch Deutscher Bundestag (2020), die über Auswirkungen von KI auf verschiedene Gesellschafsbereiche diskutiert hat. Auch die EU Komission (2020) hat hierzu ein Weißbuch veröffentlicht.

Diskriminierung wurde definiert als eine nachteilige Behandlung von Gruppen aufgrund von sensitiven Eigenschaften wie dem Geschlecht, der Ethnizität oder sexuellen Orientierung. Ein Ansatz ist deshalb, sensitive Eigenschaften für die Verwendung von Entscheidungsprozessen zu verbieten. Der Ansatz nennt sich *fairness through unawareness*, bei dem sensitive Eigenschaften nicht explizit für algorithmische Entscheidungen verwendet werden dürfen (Hagendorff 2019). Dieser Ansatz scheint jedoch wenig zielführend zu sein, wie Haeri & Zweig (2020) heraus finden. So könne das Weglassen von sensitiven Eigenschaften die Fairness sogar verringern.

Eine weit verbreitete Maßnahme sind Regularien für KI. So wird z.B. von Tutt (2016) eine FDA (U.S. Food and Drug Administration) für Algorithmen vorgeschlagen, also eine Behörde, die nur für Algorithmen zuständig ist. Wichtig ist dabei die Erfüllung der Regularien *vor* der Markteinführung, nicht anders herum. Das Frauenhofer Institut arbeitet deshalb an einem Zertifikat „made in Germany“, welches ein Maß an Qualität u.A. in den Bereichen „Ethik und Recht“, „Fairness“, „Transparenz“ und „Datenschutz“ sicherstellen soll (Frauenhofer IAIS 2022). Auch der TÜV hat ein „AI Lab“ (TÜV AI Lab 2022) in dem z.B. Prüfverfahren erarbeitet werden.

Zudem wollen sie die Einteilung von KI in Risikoklassen vorantreiben, denn nicht jede KI erfordert die selben Regulierungen.

Die Einteilung von KI in Risikoklassen wird auch von Zweig (2019) beschrieben. In Abb. 5.1 ist eine Risikomatrix für die Beurteilung der notwendigen Regularien zu sehen. Das Risiko ist einerseits abhängig von der Monopolstellung. Je größer die Monopolstellung und dadurch geringe Ausweichmöglichkeit auf andere Systeme, desto mehr müsse ein System reguliert werden. Anderseits ist das Risiko abhängig vom Schadenspotential: Je größer dieses ist, desto mehr müsse das System reguliert werden. Jede Risikoklasse habe dann eigene Anforderungen an Transparenz und Interpretierbarkeit.

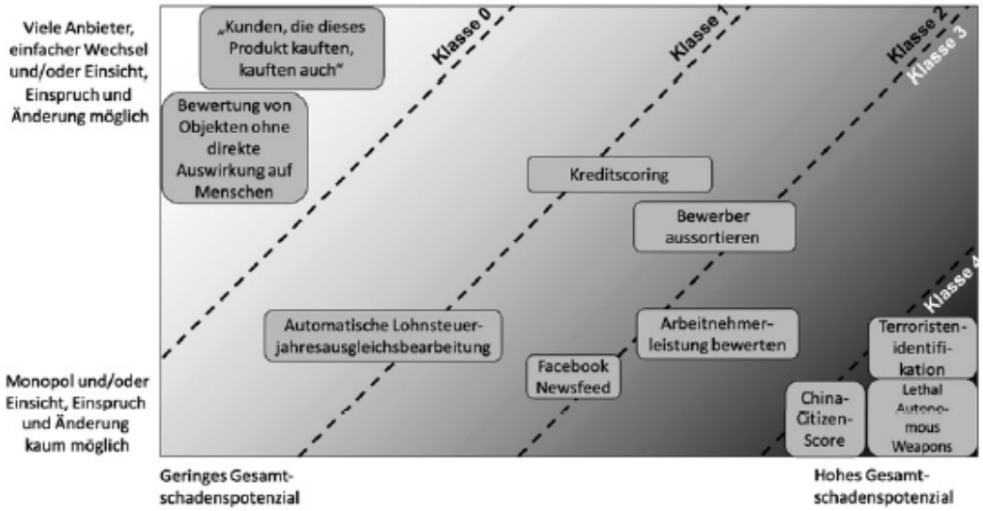


Abbildung 5.1: Risikomatrix zur Beurteilung der notwendigen Regulierung (Zweig 2019)

Da in dieser Arbeit die Diskriminierung von KI Systemen im Fokus stand, wurden vor allem negative Folgen der KI sichtbar. Das Sichtbarmachen von Diskriminierung durch KI kann aber auch umgekehrt betrachtet eine fairere Gesellschaft voran treiben. Voraussetzung ist allerdings, dass von allen Akteuren der KI dieses Interesse besteht. Das bedeutet auch schon im Vorfeld den Nutzen einer KI in Frage zu stellen und das Schadenspotential zu bewerten. Dressel & Farid (2018) finden z.B. heraus, dass die Risikobewertungssoftware „COMPAS" kaum genauer oder fairer entscheidet als Menschen, die wenig bis gar kein Wissen über Strafjustiz haben. Die Arbeit sollte deshalb aufzeigen, dass KI, so wie jedes Softwareprodukt, fehlerbehaftet ist. Durch die Nutzung kann Diskriminierung verursacht oder verstärkt werden und das muss mit in die Bewertung einer KI einbezogen werden. Als generellen Leitsatz hat die EU Komission (2018) dazu formuliert, dass „[...] die Vorteile von KI-Systemen insgesamt die vorhersehbaren individuellen Risiken erheblich überwiegen" sollen.

Literaturverzeichnis

A. Bansal et al. (2021), *Deep Learning-Based Face Analytics*, Springer International Publishing.
URL: *https://www.ebook.de/de/product/41380837/deep_learning_based_face_analytics.html*

ACM Code 2018 Task Force (2018), 'ACM Code of Ethics and Professional Conduct'.

Amnesty International (2021), 'Ban the scan'. Accessed 01.12.21.
URL: *https://banthescan.amnesty.org/*

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016), 'How We Analyzed the COMPAS Recidivism Algorithm'. Accessed: 02.03.22.
URL: *https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm*

Barocas, S., Hardt, M. & Narayanan, A. (2019), *Fairness and Machine Learning*, fairmlbook.org.
URL: *http://www.fairmlbook.org*

Brühl, J. (2018), 'Wo die Polizei alles sieht'. Accessed: 26.03.22.
URL: *https://www.sueddeutsche.de/digital/palantir-in-deutschland-wo-die-polizei-alles-sieht-1.4173809*

Buolamwini, J. & Gebru, T. (2018), Gender shades: Intersectional accuracy disparities in commercial gender classification, *in* S. A. Friedler & C. Wilson, eds, 'Proceedings of the 1st Conference on Fairness, Accountability and Transparency', Vol. 81 of *Proceedings of Machine Learning Research*, PMLR, pp. 77–91.
URL: *https://proceedings.mlr.press/v81/buolamwini18a.html*

Burkart, N. & Huber, M. F. (2021), 'A survey on the explainability of supervised machine learning', *Journal of Artificial Intelligence Research* **70**, 245–317.

Buzzfeed News (2021), 'How a facial recognition tool found its way into hundreds of us police departments, schools, and taxpayer-funded organizations'. Accessed: 26.01.22.
URL: *https://www.buzzfeednews.com/article/ryanmac/clearview-ai-local-police-facial-recognition*

Chen, S. (2021), 'Chinese scientists develop AI 'prosecutor' that can press its own charges'. Accessed: 02.03.22.
URL: *https://www.scmp.com/news/china/science/article/3160997/chinese-scientists-develop-ai-prosecutor-can-press-its-own*

CNN Business (2021), 'A false facial recognition match sent this innocent black man to jail'. Accessed: 26.01.22.
URL: *https://edition.cnn.com/2021/04/29/tech/nijeer-parks-facial-recognition-police-arrest/index.html*

Corbett-Davies, S., Pierson, E. & Goel, A. F. S. (2016), 'A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.'. Accessed: 31.03.22.
URL: *https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/*

Deutscher Bundestag (2020), 'Bericht der enquete-kommission künstliche intelligenz – gesellschaftliche verantwortung und wirtschaftliche, soziale und ökologische potenziale', Drucksache 19/23700.

Dieterich, W., Mendoza, C. & Brennan, T. (2016), 'COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity'.

Dressel, J. & Farid, H. (2018), 'The accuracy, fairness, and limits of predicting recidivism', *Science Advances* **4**(1).

Döbel et al. (2018), 'MASCHINELLES LERNEN - EINE ANALYSE ZU KOMPETENZEN; FORSCHUNG UND ANWENDUNG', Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V.

EU (2000), 'Charta der Grundrechte der Europäischen Union'.

EU Komission (2018), 'Ethik-Leitlinien für eine Vertrauenswürdige KI'.

EU Komission (2020), 'Weissbuch zur künstlichen intelligenz – ein europäisches konzept für exzellenz und vertrauen (COM(2020) 65 final)'.

Faceception (2022), 'Corporate homepage'. Accessed: 16.01.22.
URL: *www.faception.com*

Frauenhofer IAIS (2022), 'Prüfverfahren für eine KI-Zertifizierung »made in Germany«'. Accessed: 02.04.22.
URL: *https://www.iais.fraunhofer.de/de/forschung/kuenstliche-intelligenz/ki-zertifizierung.html*

Gunning, D. & Aha, D. (2019), 'DARPA's explainable artificial intelligence (XAI) program', *AI Magazine* **40**(2), 44–58.

Haeri, M. A. & Zweig, K. A. (2020), The crucial role of sensitive attributes in fair classification, *in* '2020 IEEE Symposium Series on Computational Intelligence (SSCI)', IEEE.

Hagendorff, T. (2019), 'Maschinelles lernen und diskriminierung: Probleme und lösungsansätze', *Österreichische Zeitschrift für Soziologie* **44**(S1), 53–66.

Han, H. & Jain, A. K. (2014), Age , gender and race estimation from unconstrained face images.

Haskins, C. (2020), 'Scars, Tattoos, And License Plates: This Is What Palantir And The LAPD Know About You'. Accessed: 26.03.22.
URL: *https://www.buzzfeednews.com/article/carolinehaskins1/training-documents-palantir-lapd*

Hazirbas, C., Bitton, J., Dolhansky, B., Pan, J., Gordo, A. & Ferrer, C. C. (2021), 'Towards measuring fairness in AI: the casual conversations dataset', *IEEE Transactions on Biometrics, Behavior, and Identity Science* pp. 1–1.

Johannson, M. (2020), 'Parlamentarische Anfragen, Bezugsdokument: E-003872/2020'. Accessed: 26.03.22.
URL: *https://www.europarl.europa.eu/doceo/document/E-9-2020-003872-ASW_DE.html*

Kirkpatrick, K. (2017), 'It's not the algorithm, it's the data', *Communications of the ACM* **60**(2), 21–23.

Kleinberg, J., Mullainathan, S. & Raghavan, M. (2016), 'Inherent trade-offs in the fair determination of risk scores'.

Krafft, T. D. & Zweig, K. A. (2019), 'Transparenz und Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse', Verbraucherzentrale Bundesverband e.V.

Kumar, N., Belhumeur, P. N. & Nayar, S. K. (2008), FaceTracer: A Search Engine for Large Collections of Images with Faces, *in* 'European Conference on Computer Vision (ECCV)', pp. 340–353.

Kumar, N., Berg, A., Belhumeur, P. N. & Nayar, S. (2011), 'Describable visual attributes for face verification and image search', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**(10), 1962–1977.

Li, F. & Du, Y. (2018), 'From AlphaGo to power system AI: What engineers can learn from solving the most complex board game', *IEEE Power and Energy Magazine* **16**(2), 76–84.

Lipton, Z. C. (2018), 'The mythos of model interpretability', *Communications of the ACM* **61**(10), 36–43.

Liu, Z., Luo, P., Wang, X. & Tang, X. (2015), Deep learning face attributes in the wild, *in* 'Proceedings of International Conference on Computer Vision (ICCV)'.

MacArthur Justice Center (2021), 'ShotSpotter is deployed overwhelmingly in Black and Latinx neighborhoods in Chicago.'. Accessed: 25.03.22.
URL: *https://endpolicesurveillance.com/burden-on-communities-of-color/*

Marciano, M. A. & Sweder, K. S. (2016), 'Hybrid Machine Learning Approach for DNA Mixture Interpretation'.

Marcus Smith, S. M. (2021), *Biometric Identification, Law and Ethics*, Springer International Publishing.
URL: *https://www.ebook.de/de/product/41631324/marcus_smith_seumas_miller_biometric_identification_law_and_ethics.html*

Marks, P. (2021), 'Can the biases in facial recognition be fixed; also, should they?', **64**(3), 20–22.

Microsoft (2021), 'Face detection and attributes'. Accessed: 02.02.22.
URL: *https://docs.microsoft.com/en-us/azure/cognitive-services/face/concepts/face-detection*

Molnar, C. (2019), *Interpretable Machine Learning*.

NBC News (2019), 'How facial recognition became a routine policing tool in america'. Accessed: 21.01.22.
URL: *https://www.nbcnews.com/news/us-news/how-facial-recognition-became-routine-policing-tool-america-n1004251*

OECD (2019), 'Empfehlung des Rats zur Künstlichen Intelligenz'.

OpenAI (2019), 'OpenAI Five Defeats Dota 2 World Champions'. Accessed: 02.03.22.
URL: *https://openai.com/blog/openai-five-defeats-dota-2-world-champions/*

Orwat, C. (2019), 'Diskriminierungsrisiken durch Verwendung von Algorithmen', Institut für Technikfolgenabschätzung und Systemanalyse (ITAS), Karlsruher Institut für Technnologie (KIT).

Peteranderl, S. (2016), 'Predictive Policing: Dem Verbrechen der Zukunft auf der Spur'. Accessed: 26.03.22.
URL: *https://www.bpb.de/themen/medien-journalismus/netzdebatte/238995/predictive-policing-dem-verbrechen-der-zukunft-auf-der-spur/*

Rosenbach, M. & Sarovic, A. (2021), 'NRW-Datenschutzbeauftragte hält Einsatz von Palantir-Software für unzulässig', Spiegel Netzwelt. Accessed: 26.03.22.

Saka, E. (2020), 'Big data and gender-biased algorithms'.

Scherr, A. (2016), 'Diskriminierung/Antidiskriminierung - Begriffe und Grundlagen', Politik und Zeitgeschichte (APUZ 9/2016).

Stanley, J. (2021), 'Four Problems with the ShotSpotter Gunshot Detection System'. Accessed: 25.03.22.
URL: *https://www.aclu.org/news/privacy-technology/four-problems-with-the-shotspotter-gunshot-detection-system/*

Suresh, H. & Guttag, J. V. (2019), 'A framework for understanding sources of harm throughout the machine learning life cycle'.

Thom, N. & Hand, E. M. (2021), Facial attribute recognition: A survey, *in* 'Computer Vision', Springer International Publishing, pp. 447–459.

Tutt, A. (2016), 'An FDA for algorithms', *SSRN Electronic Journal* .

TÜV AI Lab (2022), 'Risikoklassen für Künstliche Intelligenz'. Accessed: 02.04.22.
URL: *https://www.tuev-verband.de/digitalisierung/kuenstliche-intelligenz/tuev-ai-lab*

United States. Governement Accountability Office (2020), 'Facial recognition technology: Privacy and accuracy issues related to commercial uses, report to congressional requesters gao-20-522'.

Vaquero, D. A., Feris, R. S., Tran, D., Brown, L., Hampapur, A. & Turk, M. (2009), Attribute-based people search in surveillance environments, *in* '2009 Workshop on Applications of Computer Vision (WACV)', IEEE.

Verma, S. (2019), 'Weapons of math destruction: How big data increases inequality and threatens democracy', *Vikalpa: The Journal for Decision Makers* **44**(2), 97–98.

Waltl, D. B. (2019), Erklärbarkeit und Transparenz im Machine Learning, *in* 'Springer Reference Geisteswissenschaften', Springer Fachmedien Wiesbaden, pp. 1–23.

Wang, Y. & Kosinski, M. (2018), 'Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.', *Journal of Personality and Social Psychology* **114**(2), 246–257.

Wu, X. & Zhang, X. (2016), 'Automated inference on criminality using face images'.

Yang, M.-H. (2009), Face detection, *in* 'Encyclopedia of Biometrics', Springer US, pp. 303–308.

Zweig, K. (2019), *Ein Algorithmus hat kein Taktgefühl*, Penguin Random House.
URL: *https://www.ebook.de/de/product/36506085/katharina_zweig_ein_algorithmus_hat_kein_taktgefuehl.html*

Erklärung zur selbstständigen Bearbeitung einer Abschlussarbeit

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich gemacht.

Ort Datum Unterschrift im Original