Algorithmische Fairness in der polizeilichen Ermittlungsarbeit: Ethische Analyse von Verfahren des maschinellen Lernens zur Gesichtserkennung

Brandner, Lou Therese; Hirsbrunner, Simon David

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Brandner, L. T., & Hirsbrunner, S. D. (2023). Algorithmische Fairness in der polizeilichen Ermittlungsarbeit: Ethische Analyse von Verfahren des maschinellen Lernens zur Gesichtserkennung. *TATuP - Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis / Journal for Technology Assessment in Theory and Practice*, *32*(1), 24-29. https://doi.org/10.14512/tatup.32.1.24

RESEARCH ARTICLE

Algorithmische Fairness in der polizeilichen Ermittlungsarbeit: Ethische Analyse von Verfahren des maschinellen Lernens zur Gesichtserkennung

Lou Therese Brandner*,[1], Simon David Hirsbrunner[1]

Zusammenfassung • Dieser Beitrag diskutiert Fairness in auf künstlicher Intelligenz (KI) basierenden Verfahren der Polizeiarbeit anhand des Beispiels der Gesichtserkennung. Algorithmische Entscheidungen, die auf gesellschaftlichen Diskriminierungsdynamiken beruhen, können Ungerechtigkeiten (re-)produzieren und automatisieren. KI-Fairness betrifft dabei nicht nur die Erstellung und das Teilen von Datensätzen oder das Training von Modellen, sondern auch die Art des Systemeinsatzes in der Realwelt. Die Quantifizierung von Fairness kann davon ablenken, wie Diskriminierung und Unterdrückung sich konkret als soziale Phänomene niederschlagen. Integrative Ansätze können hier dazu beitragen, durch kontinuierliche interdisziplinäre Kollaboration ethische, rechtliche, soziale und wirtschaftliche Faktoren aktiv in die Technikentwicklung einzubeziehen und die Folgen des Einsatzes ganzheitlicher einzuschätzen.

***Algorithmic fairness in investigative policing**: Ethical analysis of machine learning methods for facial recognition*

Abstract • *This article discusses fairness in artificial intelligence (AI) based policing procedures using facial recognition as an example. Algorithmic decisions based on discriminatory dynamics can (re)produce and automate injustice. AI fairness here concerns not only the creation and sharing of datasets or the training of models but also how systems are deployed in the real world. Quantifying fairness can distract from how discrimination and oppression translate concretely into social phenomena. Integrative approaches can help actively incorporate ethical, legal, social, and economic factors into technology development to more holistically assess the consequences of deployment through continuous interdisciplinary collaboration.*

Keywords • *fairness, policing, algorithmic bias, machine learning*

This article is part of the Special topic "Modeling for policy: Challenges for technology assessment from new prognostic methods," *edited by A. Kaminski, G. Gramelsberger and D. Scheer. https://doi.org/10.14512/tatup.32.1.10*

Einleitung

Der Einsatz von so genannter künstlicher Intelligenz (KI) in der polizeilichen Ermittlungsarbeit verspricht effektivere und kostengünstigere Verbrechensprävention und -aufklärung. KI kann dort ansetzen, wo menschliche Fähigkeiten und Kapazitäten potenziell nicht ausreichen; durch automatisierte, papierlose Arbeitsprozesse sollen sowohl die Produktivität als auch die Objektivität polizeilicher Maßnahmen verbessert werden. Diverse technische Lösungen mit KI-Unterstützung für die polizeiliche Ermittlungsarbeit befinden sich auch in Deutschland bereits im Einsatz, bspw. zur Sichtung und Auswertung von kinderpornographischem Bildmaterial (LKA Niedersachsen 2020).

Während die KI-Branche mit großer Geschwindigkeit wächst, wurde erst 2021 mit dem Artificial Intelligence Act der Europäischen Union (EU) der weltweit erste Gesetzentwurf zur KI-Regulierung vorgelegt. Dieser definiert Echtzeit-Gesichtserkennung im öffentlichen Raum als Hochrisikotechnologie, deren Einsatz nur in besonderen Ausnahmefällen wie der Terroris-

* *Corresponding author: lou.brandner@uni-tuebingen.de*
[1] *Internationales Zentrum für Ethik in den Wissenschaften (IZEW), Universität Tübingen, Tübingen, DE*

https://doi.org/10.14512/tatup.32.1.24
Received: 26. 08. 2022; revised version accepted: 17. 01. 2023;
published online: 23. 03. 2023 (peer review)

musbekämpfung oder der Suche nach vermissten Personen erlaubt sein soll (Europäische Kommission 2021). Im Polizeikontext können fehlerhafte Entscheidungen zu gravierenden Grundrechtsverletzungen wie rechtswidrigen Festnahmen führen (Selbst 2017). Um unerwünschte Nebeneffekte und Risiken für Individuen und Personengruppen zu minimieren, sind an diese hochrisikoreichen Anwendungen über technische Ansprüche hinaus besonders strenge Anforderungen auf dem gesellschaftlichen Level zu stellen. Hochrisiko-KI tangiert Fragen sozialer Gerechtigkeit und die Operationalisierung von Prinzipien wie Fairness, Gleichbehandlung und Nichtdiskriminierung.

Diskriminierung grenzt sich von Ungleichbehandlung dadurch ab, dass sie auf der Zugehörigkeit zu sozial bedeutsamen Gruppen beruht (Lippert-Rasmussen 2013). Das bedeutet, dass Personen aufgrund von Attributen wie Geschlecht, ethnischer Herkunft oder Behinderung ungerechtfertigter, negativer Ungleichbehandlung ausgesetzt sind (Hagendorff 2019). Algorithmische Outputs, die auf ‚Biases' (auf Deutsch Verzerrungen) in KI-Systemen beruhen und sich so auf bestehende gesellschaftliche Diskriminierungsdynamiken stützen, können Ungerechtigkeiten (re-)produzieren und automatisieren (Eubanks 2018). Der Begriff KI-Fairness, auch ‚algorithmische Fairness' genannt, beschreibt Methoden, die durch Verzerrungen in KI-Modellen hervorgerufene soziale Diskriminierung mindern oder ausschließen sollen.

Dieser Beitrag diskutiert Fairness in KI-basierten Verfahren der polizeilichen Ermittlungsarbeit. Im Folgenden wird zunächst kurz der Projektkontext erläutert, aus dem die Autor*innen die Beitragselemente ableiten, bevor verschiedene Quellen von Verzerrungen in KI-Systemen aufgezeigt werden. Daraufhin werden KI-Fairness und ihre Grenzen anhand verschiedener Metriken und deren realweltlichen Auswirkungen problematisiert. Der Beitrag schließt ab mit einer Diskussion zum Bezug zur Technikfolgenabschätzung (TA) und der effektiven Problematisierung von KI-Fairness in interdisziplinären Forschungsprojekten.

Hintergrund

Die Überlegungen in diesem Artikel leiten sich aus ethischen Analysen ab, die im Rahmen zweier interdisziplinärer Projekte im Kontext ziviler Sicherheitsforschung durchgeführt wurden, in denen verschiedene digitale Erkennungs- und Analysetechnologien für die polizeiliche Ermittlungsarbeit entwickelt werden. Die Forschung schließt explizit die Adressierung von Bias und Diskriminierung in Verfahren KI-basierter Gesichtserkennung ein. Die Projekte verfolgen einen integrierten Ansatz der Technikentwicklung (Spindler et al. 2020), bei dem technische Perspektiven mit ethischen, rechtlichen, sozialen und wirtschaftlichen (ELSE) Aspekten in Bezug gesetzt werden. Wie bereits im TA-Kontext diskutiert (Gressel und Orlowski 2019), reichen diese Ansätze über den Anspruch klassischer ethischer Begleitforschung hinaus, da ELSE-Aspekte von Projektbeginn an und über alle Arbeitspakete hinweg einbezogen werden.

Dieser Beitrag hat den Anspruch, KI-Fairness in polizeilichen KI-Systemen über einzelne Projekte hinausgehend zu diskutieren. Die Fokussierung auf die Gesichtserkennung bietet sich jedoch an, da sie als weitverbreitete Anwendung viele ethische Problematiken algorithmischer Polizeiarbeit aufzeigt. Biometrische Erkennungssysteme stellen besondere Herausforderungen an die KI-Modellierung, da individuelle menschliche Charakteristiken erfasst, erkannt und verifiziert werden müssen (Merler et al. 2019). Durch die Entwicklung und Anwendung von KI-basierter Gesichtserkennung ergeben sich auch besondere ethische Problematiken, was sich im bestehenden KI-Regulierungsentwurf (Europäische Kommission 2021) der EU widerspiegelt, der die biometrische Fernidentifizierung als Hochrisikotechnologie einstuft. Verzerrungen und resultierende Diskriminierung werden hier explizit als besondere Risiken genannt (Europäische Kommission 2021, S. 26).

Durch die Entwicklung und Anwendung von KI-basierter Gesichtserkennung ergeben sich auch besondere ethische Problematiken.

Die Technologie wird trotz der dargestellten Risiken in diversen Kontexten von Überwachung und Kontrolle erprobt (Bundespolizei 2018; Jürs 2022; Monroy 2022); auch in von den Autor*innen begleiteten Projekten werden Demonstratoren von polizeilichen Gesichtserkennungssystemen entwickelt, um bspw. Personen in Überwachungsaufnahmen zu identifizieren. Dies wirft ethisch-normative Fragestellungen hinsichtlich Fairness und Biases auf, die im Rahmen der Technikentwicklung mit technischen, polizeilichen, kommerziellen und rechtlichen Partnern diskutiert werden. Von dieser Zusammenarbeit ausgehend bietet sich eine Evaluierung der Diskriminierungsrisiken unter den Vorzeichen der TA an. Entsprechend sollen in diesem Beitrag mögliche Verzerrungen aufgeführt und im Hinblick auf Fairness-Kriterien reflektiert werden.

Verzerrungen in KI-Verfahren der Gesichtserkennung

Techniken angewandter KI können zur Identifizierung von Gesichtern in fotografischen Bilddaten verwendet werden. Vor der Operationalisierung in konkreten Einsatzgebieten werden solche Systeme anhand riesiger Mengen von Beispieldaten trainiert und lernen so, verschiedene Gesichter anhand verschiedener visuel-

https://doi.org/10.14512/tatup.32.1.24

ler Merkmale zu unterscheiden. Dabei kommen sowohl überwachte als auch unüberwachte Ansätze des KI-Trainings zum Einsatz (Anwarul und Dahiya 2020). Aus unterschiedlichen Gründen variiert die Treffsicherheit des Systems oft zwischen verschiedenen zu identifizierenden Merkmalen, was potenziell zur Diskriminierung sozialer Gruppen führen kann. Im Folgenden werden Auslöser und Mechanismen solcher Verzerrungen in polizeilich genutzter Gesichtserkennung diskutiert.

Qualität der Trainingsdaten

Eine bedeutende Quelle für Verzerrungen in KI-Systemen sind Trainingsdaten.[1] Im Rahmen von Trainingsprozessen etablieren KI-Modelle auf Basis von Datensätzen bestimmte Ähnlichkeiten und Unterschiede zwischen Elementen. In Verfahren des unüberwachten Lernens entstammen die Daten bspw. sozialen Medien oder Bilddatenbanken, die Verzerrungen enthalten können. Bestimmte Gruppen, z. B. Frauen, nichtweiße oder ältere Personen, können in Trainingsdatenbanken unterrepräsentiert sein, was dazu führt, dass ihnen gegenüber die algorithmische Ergebnisgenauigkeit sinkt (Berk et al. 2018). Wenn wenig historische Daten existieren oder aus Gründen des Datenschutzes nicht nutzbar sind, können synthetische Trainingsdaten reale Datensätze ersetzen oder ergänzen. Diese fiktiven, realistische Daten, die gezielt für die KI-Entwicklung hergestellt werden, sind grundsätzlich dazu geeignet, das Problem existierender Verzerrungen in historischen Daten zu umgehen, können Minderheiten aber auch fehlrepräsentieren (Bhanot et al. 2021).

Auch sind Qualität und Auswertbarkeit von Foto- und Videoaufnahmen häufig von schwierigen Konditionen hinsichtlich z. B. Perspektive, Belichtung oder Bildauflösung beeinflusst, die allgemein zur geringeren Genauigkeit von KI-Modellen, z. B. bezüglich der korrekten Identifizierung von Straftäter*innen, beitragen können (Anwarul und Dahiya 2020). Verlässlichkeit und Genauigkeit und daher auch die Fairness der Modelle hängen so zu einem erheblichen Teil von der Verfügbarkeit qualitativ hochwertiger Trainingsdaten ab.

Annotation der Trainingsdaten

Auch die Einschreibung sozialer und kultureller Vorannahmen während der Datenannotation kann zu Verzerrungen in Trainingsdatensätzen führen (Selbst 2017). In überwachten KI-Trainingsprozessen markieren Personen bestimmte Charakteristika in Datensätzen, deren Erkennung anschließend von einem Modell erlernt wird. Das Labeling von Gesichtsdatenbanken wird üblicherweise von KI-Entwickler*innen selbst oder von Dienstleister*innen auf Crowdsourcing-Plattformen wie Amazon Mechanical Turk oder Scale durchgeführt. Dabei identifizieren menschliche Akteure übereinstimmende Aufnahmen von Individuen und kennzeichnen Merkmale wie Gesichtszüge, Mimik und Posen. Diskriminierende Stereotype können hier durch unbeabsichtigte Voreingenommenheit oder absichtliche Beeinflussung Eingang finden (Leslie 2020). Wenn bspw. Individuen bestimmter Ethnien häufiger falsch identifiziert werden, können die Modelle dies in höheren Fehlerraten für diese Gruppen widerspiegeln. Bei aktiven KI-Lernverfahren, in welchen Nutzer*innen Verantwortung bei der Anpassung und Optimierung automatisierter Analysesysteme eingeräumt werden, kann Diskriminierung zudem im laufenden Betrieb eingeschleust werden, wenn Einzelne den Lernprozess durch oft unbewusste Vorurteile beeinflussen (Fischer et al. 2022).

Ergebnisgenauigkeit der KI-Verfahren

Ein bekanntes Phänomen in der KI-Gesichtserkennung sind signifikant höhere Fehlerquoten für weibliche[2] und nicht-weiße Gesichter und somit eine schlechtere Ergebnisgenauigkeit. Abseits der Repräsentation in Trainingsdaten können hier auch Faktoren wie Kameraeinstellungen (Roth 2009), Gesichtsmorphologie oder Makeup (Albiero et al. 2022) eine Rolle spielen. Diese Umstände bilden gesellschaftliche Dynamiken ab, die weiße und männliche Personen als den menschlichen Standard positionieren und andere Gruppen als Abweichungen marginalisieren. Bestehende Diskriminierung wird so im KI-Kontext fortgeführt, insbesondere gegenüber von intersektionalen Unterdrückungen betroffenen Gruppen wie Schwarzen Frauen (Buolamwini und Gebru 2018).

Nicht nur Unter-, sondern auch Überrepräsentation kann zu Biases führen, wenn z. B. Schwarze Individuen häufiger polizeilich erfasst und somit häufiger durch Datenbankabgleiche – korrekt wie inkorrekt – identifiziert werden (Bacchini und Lorusso 2019). Da eine häufigere Erfassung nicht zwingend eine tatsächliche höhere Straffälligkeit bedeutet, sondern auch durch diskriminierende Polizeipraktiken wie unverhältnismäßige Überwachung oder Verfolgung bedingt ist (Garvie und Frankle 2016; Selbst 2017), werden so existierende rassistische Dynamiken verstärkt.

Überprüfung algorithmischer Fairness

Erschwerend kommt hinzu, dass der Analyseprozess und die identifizierten Datenmuster im maschinellen Lernen, z. B. bei der Identifizierung eines Individuums, wie beschrieben häufig hochkomplex und für Personen kognitiv kaum interpretierbar sind. Gerade der Beitrag, den die gewählten Trainingsdatensätze auf die Outputs eines solchen Blackbox-Modells haben, ist selten konkret nachvollziehbar. Die Opazität und wahrgenommene Autorität von Computerprozessen können dazu führen, dass Biases unerkannt bleiben, die zusätzlich stabilisiert werden, wenn polizeiliche Ermittler*innen ohne tiefergehendes Vorwissen über KI-Systeme scheinbar neutrale algorithmische Entscheidungen unter dem Einfluss subjektiver Vorannahmen interpretieren (Helm und Hagendorff 2021).

1 Der Begriff der Trainingsdaten schließt hier Trainingsdaten (Vorlagen zum Erlernen von Mustern), Validierungsdaten (Optimierung des Modells) und Testdaten (Überprüfung des Modellverhaltens) mit ein.

2 Gängige KI-Modelle kategorisieren biologisches Geschlecht binär als männlich–weiblich und ziehen komplexere soziale Genderkonstruktionen nicht in Betracht. Da dieser Beitrag sich mit der Praxis der Programme befasst, wird die dichotome Einteilung übernommen.

https://doi.org/10.14512/tatup.32.1.24

Lösungsansätze der KI-Fairness

Die Europäische Kommission definiert Fairness im KI-Kontext als den Schutz vor (algorithmischer) Verzerrung, Diskriminierung und Stigmatisierung von Personen und Gruppen (Europäische Kommission 2018). Diese Problematiken werden auf verschiedenen Wegen von Fairness-Ansätzen adressiert. Dort, wo hohe Fehlerquoten durch Unterrepräsentation in Trainingsdaten bedingt sind, können Datensätze mit höherer Diversität hinsichtlich Hautfarbe und Geschlecht wie z. B. Diversity in Faces (Merler et al. 2019) oder Pilot Parliaments Benchmark (Buolamwini und Gebru 2018) Abhilfe schaffen. Bereits formulierte Anforderungen an fairness-sensible Daten (Le Quy et al. 2022) bieten sich zum Vergleich verwendeter Datensätze an. Doch Diversität in Datensätzen ist zwar eine Grundlage, aber kein alleiniger Garant für Fairness.

Statistische Methoden der KI-Fairness sollen durch algorithmische Verzerrungen in KI-Modellen hervorgerufene soziale Diskriminierung mindern oder ausschließen. Gängige informatische Definitionen unterscheiden dabei insbesondere zwischen individueller Fairness und Gruppenfairness. Erstere soll sicherstellen, dass statistische Messungen der Ergebnisse für Individuen mit denselben Merkmalen gleich sind. Bei Verfahren der Gruppenfairness werden Ergebnisse eines Modells so angeglichen, dass sie für verschiedene vordefinierte Gruppen von Datensubjekten mit geschützten Merkmalen ähnlich oder gleich sind (Mahoney et al. 2020, S. vii).

Es existiert eine Vielzahl statistischer Methoden, mit denen sich KI-Fairness evaluieren lässt (Mehrabi et al. 2021). Die Identifizierung des geeigneten Verfahrens in einem konkreten Fall operationalisiert über technische Fragestellungen hinaus verschiedene Vorstellungen von Fairness und Gerechtigkeit. Angenommen, eine KI selektiert erfolgsversprechende Kandidat*innen für ein Informatikstudium: Bei der Optimierung gegenüber dem Attribut binäres Geschlecht auf ‚demographic parity' soll die Verteilung sozialer Gruppen in einer Gesamtbevölkerung abgebildet werden, sodass jeweils die Hälfte der Studienplätze an Frauen und an Männer vergeben wird, unabhängig von der durchschnittlichen Qualität der jeweiligen Bewerbungen (Quotengerechtigkeit). Optimiert auf die Metrik der ‚equalized odds' hingegen, zielt das System darauf, dass Bewerber*innen mit ähnlichen Qualifikationen gleiche Chancen haben, während die letztendliche Geschlechterquote im Studiengang gleichgültig ist (Chancengerechtigkeit).

KI-Verfahren zur Fairness in der Gesichtserkennung werden gerne auf Metriken der Chancengerechtigkeit mit gleichen ‚true-positive'-Raten für geschützte und ungeschützte Gruppen trainiert. Es wird dabei angenommen, dass eine Erkennung durch das System die favorable Option darstellt, was jedoch nicht immer der sozialen Realität entspricht. In der polizeilichen Ermittlungsarbeit bedeuten ‚false positives' in einer Gesichtserkennungs-Software, dass Unbeteiligte als Verdächtige fehlidentifiziert werden; die Folgen von falsch-positiven Ergebnissen sind hier besonders schwerwiegend bis hin zu unbegründeten Festnahmen und Gefährdungen der körperlichen Unversehrtheit. ‚False negatives' bedeuten, dass Täter*innen nicht identifiziert werden und dadurch der Strafverfolgung entgehen. Während der Suche nach einem vermissten Kind wiederum spielen ‚false positives' eine weniger große Rolle, da sie direkt nach der Fehlentscheidung des Systems durch menschliche Akteure überprüfbar sind und daher keine Folgen für das fälschlich identifizierte Kind haben. Ein ‚false negative' resultiert dagegen in einer Nicht-Identifizierung des vermissten Kindes und hat damit potenziell schwerwiegende Nachteile für Betroffene und Ermittlungsbehörden, deren Suche durch den Fehler des Systems verlängert wird oder sogar erfolglos bleibt.

Niedrige Fehlerraten sind in der algorithmischen Polizeiarbeit daher von hoher Priorität, was insbesondere für sozial benachteiligte Gruppen relevant ist, da für diese wie dargelegt Fehlerraten häufig signifikant höher sind. Eine künstliche Angleichung der Raten verringert dabei zwangsweise die Treffgenauigkeit des technischen Systems (Kleinberg et al. 2016) und Fairnessformeln lassen sich oft nicht parallel optimieren, da sie über gemeinsame Variablen verbunden sind (Ruf und Detyniecki 2021). Das Ausklammern geschützter Merkmale (‚fairness through unawareness') (Gajane und Pechenizkiy 2018) kann keine allgemeine Lösung darstellen, da es die Nichterkennung von Biases zur Folge haben kann. So ließe sich beim Ausschluss des Attributs Geschlecht in der Gesichtserkennung eine erhöhte Fehlerrate gegenüber Frauen schlechter feststellen und daher nicht beheben. Zudem können nichtgeschützte Attribute als Stellvertreter für geschützte Merkmale fungieren. Z. B. hat sich gezeigt, dass die an sich unproblematische Variable ‚Postleitzahl' unter Umständen auf das geschützte Merkmal ‚Ethnie' (‚race') schließen lässt (Datta et al. 2017).

Jede Quantifizierung und Abstrahierung von Fairness kann davon ablenken, wie KI-Systeme in strukturellen Diskriminierungsdynamiken verankert sind.

Fairness tangiert auch die Art des Systemeinsatzes in der Realwelt. Angenommen, ein KI-gestütztes System für Gesichtserkennung wird in einem Stadtteil mit mehrheitlich Schwarzer Bevölkerung installiert: Auch wenn das eingesetzte KI-System gleiche Fehlerraten für alle Ethnien ausgibt, stellen sich hier Fragen der Diskriminierung durch KI – eben deshalb, weil eine Schwarze Nachbarschaft von einem besonders präzisen (sogar algorithmisch ‚fairen') System überwacht wird, während andere Stadtteile unbeobachtet bleiben. Jede versuchte Quanti-

https://doi.org/10.14512/tatup.32.1.24

fizierung und Abstrahierung von Fairness kann davon ablenken, wie KI-Systeme in strukturellen Diskriminierungsdynamiken verankert sind (John-Mathews et al. 2022) und wie diese sich konkret im Leben von unterprivilegierten Personen niederschlagen (Birhane et al. 2022). Das tangiert auch andere Voraussetzungen für Fairness wie Transparenz und Erklärbarkeit, da Betroffene der Entscheidungen von Hochrisiko-Systemen z. B. wissen können sollten, auf welche Metrik hin optimiert wurde und welche Ergebnisse bei den verwendeten Verfahren erreicht wurden.

Diskussion und Ausblick

Wie sich im Kontext der hier diskutierten polizeilichen Gesichtserkennung zeigt, ist statistisch verstandene KI-Fairness alleine nicht geeignet, um Nicht-Diskriminierung im realen Betrieb sicherzustellen. Aus ethischer Perspektive darf der Blick nicht von historischen und strukturellen Machtdynamiken abrücken, da eine eng gefasste Auffassung von KI-Fairness zu sehr auf bloße Daten anstatt auf realweltliche Konsequenzen gerichtet sein kann (John-Mathews et al. 2022). Normative Überlegungen sind hier von großer Bedeutung, um die Folgen der Anwendung von Fairnesskriterien und -metriken kontextspezifisch einzuschätzen. Ob und wie ein ‚fairer' Einsatz von Gesichtserkennungssystemen und anderer polizeilicher KI möglich ist und wie die Risiken dieser Anwendungen minimiert werden können, kann nicht allgemeingültig beantwortet werden; Fragen wie diese müssen im Kontext der jeweiligen Szenarien ganzheitlich und aus transdisziplinärer Sicht von den beteiligten Stakeholdern verhandelt werden. Das bekannte Collingridge-Dilemma, demzufolge in frühen Entwicklungsstadien noch Einfluss auf Technologien genommen werden kann, aber noch Unsicherheit über die tatsächliche Nutzung herrscht, während in späten Stadien das Gegenteil gilt, ist im Zusammenhang mit sich häufig noch in der Forschungsphase befindenden KI eine besondere Herausforderung der TA (Humm et al. 2021).

Integrierte Ansätze der Technikentwicklung können dieses Dilemma und verwandte Problematiken effektiv adressieren, da technische Perspektiven von Anfang an und durchgehend mit ELSE-Aspekten in Bezug gesetzt werden. Die für die TA notwendige kontinuierliche und ganzheitliche Beobachtung der Technikentwicklung und Beurteilung einzelner Einsatzformen (Albrecht und Kellermann 2020) werden so ermöglicht. Im Rahmen der Projektarbeit der Autor*innen zeigt sich, dass dieses Vorgehen dazu geeignet ist, normative Probleme wie algorithmische Diskriminierung frühestmöglich zu thematisieren; in den Polizeiprojekten wurde das ethische Teilprojekt bereits zu Anfang in die Formulierung technischer Spezifikationen der Vorhaben zur Gesichtserkennung involviert. So konnten die beschriebenen Quellen von Diskriminierung sowie vorhandene Lösungsansätze und ihre Grenzen kontinuierlich mit technischen, polizeilichen und rechtlichen Partnern diskutiert und neu ausgehandelt werden.

Funding • This article is based on the work in the research projects PEGASUS and VIKING, funded by the German Federal Ministry of Education and Research (BMBF) as part of the federal government's program "Research for Civil Security".

Competing interests • The authors declare no competing interests.

Literatur

Albiero, Vítor; Zhang, Kai; King, Michael; Bowyer, Kevin (2022): Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. In: Transactions on Information Forensics and Security 17, S. 127–137. https://doi.org/10.1109/TIFS.2021.3135750

Albrecht, Thorben; Kellermann, Christian (2020): Künstliche Intelligenz und die Zukunft der digitalen Arbeitsgesellschaft. Konturen einer ganzheitlichen Technikfolgenabschätzung. Working Paper Forschungsförderung, Nr. 200. Düsseldorf: Hans-Böckler-Stiftung.

Anwarul, Shahina; Dahiya, Susheela (2020): A comprehensive review on face recognition methods and factors affecting facial recognition accuracy. In: Pradeep Singh, Arpan Kar, Yashwant Singh, Maheshkumar Kolekar und Sudeep Tanwar (Hg.): Proceedings of ICRIC 2019. Cham: Springer, S. 495–514. https://doi.org/10.1007/978-3-030-29407-6_36

Bacchini, Fabio; Lorusso, Ludovica (2019): Race, again. How face recognition technology reinforces racial discrimination. In: Journal of Information, Communication and Ethics in Society 17 (3), S. 321–335. https://doi.org/10.1108/jices-05-2018-0050

Berk, Richard; Heidari, Hoda; Jabbari, Shahin; Kearns, Michael; Roth, Aaron (2018): Fairness in criminal justice risk assessments. The state of the art. In: Sociological Methods & Research 50 (1), S. 3–44. https://doi.org/10.1177/0049124118782533

Bhanot, Karan; Qi, Miao; Erickson, John; Guyon, Isabelle; Bennett, Kristin (2021): The problem of fairness in synthetic healthcare data. In: Entropy 23 (9), S. 1–21. https://doi.org/10.3390/e23091165

Birhane, Abeba et al. (2022): The forgotten margins of AI ethics. In: FACCT'22. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, S. 948–958. https://doi.org/10.1145/3531146.3533157

Bundespolizei (2018): Projekt zur Gesichtserkennung erfolgreich. Testergebnisse veröffentlicht. Online verfügbar unter www.bundespolizei.de/Web/DE/04Aktuelles/01Meldungen/2018/10/181011_abschlussbericht_gesichtserkennung.html, zuletzt geprüft am 17. 01. 2023.

Buolamwini, Joy; Gebru, Timnit (2018): Gender shades. Intersectional accuracy disparities in commercial gender classification. In: Proceedings of Machine Learning Research 81, S. 1–15. Online verfügbar unter https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf, zuletzt geprüft am 17. 01. 2023.

Datta, Anupam; Fredrikson, Matthew; Ko, Gihyuk; Mardziel, Piotr; Sen, Shayak (2017): Use privacy in data-driven systems. Theory and experiments with machine learnt programs. In: Proceedings of the 2017 ACM SIGSAC Conference, S. 1193–1210. https://doi.org/10.1145/3133956.3134097

Eubanks, Virginia (2018): Automating inequality. How high-tech tools profile, police, and punish the poor. New York: St. Martin's Press. https://doi.org/10.5204/lthj.v1i0.1386

Europäische Kommission (2021): Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union. 2021/0106 (COD). Brüssel: Europäische Kommission.

Europäische Kommission (2018): Ethik-Leitlinien für eine vertrauenswürdige KI. Brüssel: Europäische Kommission. https://doi.org/10.1007/978-3-663-09857-7_27

Fischer, Maximilian; Hirsbrunner, Simon; Jentner, Wolfgang; Miller, Matthias; Keim, Daniel; Helm, Paula (2022): Promoting ethical awareness in communication analysis. Investigating potentials and limits of visual analytics for intelligence applications. In: FACCT'22. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, S. 877–889. https://doi.org/10.1145/3531146.3533151

Gajane, Pratik; Pechenizkiy, Mykola (2018): On formalizing fairness in prediction with machine learning. In: arxiv.org. https://doi.org/10.48550/ARXIV.1710.03184

Garvie, Clare; Frankle, Jonathan (2016): Facial-recognition software might have a racial bias problem. In: The Atlantic, 07. 04. 2017. Online verfügbar unter www.theatlantic.com/technology/archive/2016/04/the-underlying-bias-of-facial-recognition-systems/476991/, zuletzt geprüft am 17. 01. 2023.

Gressel, Céline; Orlowski, Alexander (2019): Integrierte Technikentwicklung. Herausforderungen, Umsetzungsweisen und Zukunftsimpulse. In: TATuP – Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis 28 (2), S. 71–72. https://doi.org/10.14512/tatup.28.2.s71

Hagendorff, Thilo (2019): Maschinelles Lernen und Diskriminierung. Probleme und Lösungsansätze. In: Österreichische Zeitschrift für Soziologie 44 (S1), S. 53–66. https://doi.org/10.1007/s11614-019-00347-2

Helm, Paula; Hagendorff, Thilo (2021): Beyond the prediction paradigm. Challenges for AI in the struggle against organized crime. In: Law and Contemporary Problems 84 (3), S. 1–17. https://scholarship.law.duke.edu/lcp/vol84/iss3/2

Humm, Bernhard; Lingner, Stephan; Schmidt, Jan; Wendland, Karsten (2021): KI-Systeme. Aktuelle Trends und Entwicklungen aus Perspektive der Technikfolgenabschätzung. In: TATuP – Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis 30 (3), S. 11–16. https://doi.org/10.14512/tatup.30.3.11

John-Mathews, Jean-Marie; Cardon, Dominique; Balagué, Christine (2022): From reality to world. A critical perspective on AI fairness. In: Journal of Business Ethics 178, S. 945–959. https://doi.org/10.1007/s10551-022-05055-8

Jürs, Martin (2022): Hamburg Airport. Lufthansa setzt auf Gesichtserkennung. In: fvw Travel Talk, 29. 04. 2022. Online verfügbar unter www.fvw.de/touristik/verkehr/hamburg-airport-lufthansa-setzt-auf-gesichtserkennung-225741, zuletzt geprüft am 17. 01. 2023.

Kleinberg, Jon; Mullainathan, Sendhil; Raghavan, Manish (2016): Inherent trade-offs in the fair determination of risk scores. In: arxiv.org. https://doi.org/10.48550/arXiv.2203.05051

Le Quy, Tai; Roy, Arjun; Iosifidis, Vasileios; Zhang, Wenbin; Ntoutsi, Eirini (2022): A survey on datasets for fairness-aware machine learning. In: WIREs Data Mining and Knowledge Discovery 12 (3), S. 1–59. https://doi.org/10.1002/widm.1452

Leslie, David (2020): Understanding bias in facial recognition technologies. An explainer. In: SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3705658

Lippert-Rasmussen, Kasper (2013): Born free and equal? A philosophical inquiry into the nature of discrimination. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199796113.001.0001

LKA Niedersachsen (2020): Künstliche Intelligenz. LKA Niedersachsen stellt Software zur Bekämpfung von Kinderpornografie bundesweit zur Verfügung. Online verfügbar unter www.lka.polizei-nds.de/a/presse/pressemeldungen/kuenstliche-intelligenz-lka-niedersachsen-stellt-software-zur-bekaempfung-von-kinderpornografie-bundesweit-zur-verfuegung-114750.html, zuletzt geprüft am 17. 01. 2023.

Mahoney, Trisha; Varshney, Kush; Hind, Michael (2020): AI fairness. Sebastopol: O'Reilly Media. Online verfügbar unter https://krvarshney.github.io/pubs/MahoneyVH2020.pdf, zuletzt geprüft am 17. 01. 2023.

Mehrabi, Ninareh; Morstatter, Fred; Saxena, Nripsuta; Lerman, Kristina; Galstyan, Aram (2021): A survey on bias and fairness in machine learning. In: ACM Computing Surveys 54 (6), S. 1–35. https://doi.org/10.1145/3457607

Merler, Michele; Ratha, Nalini; Feris, Rogerio; Smith, John (2019): Diversity in faces. In: arxiv.org. https://doi.org/10.48550/arXiv.1901.10436

Monroy, Matthias (2022): DNA, Gesichtsbilder und Fingerabdrücke. Biometrische BKA-Systeme enthalten Datenblätter zu zehn Millionen Personen. In: Netzpolitik.org, 09. 03. 2022. Online verfügbar unter https://netzpolitik.org/2022/dna-gesichtsbilder-und-fingerabdruecke-biometrische-bka-systeme-enthalten-datenblaetter-zu-zehn-millionen-personen, zuletzt geprüft am 17. 01. 2023.

Roth, Lorna (2009): Looking at Shirley, the ultimate norm. Colour balance, image technologies, and cognitive equity. In: Canadian Journal of Communication 34 (1), S. 111–136. https://doi.org/10.22230/cjc.2009v34n1a2196

Ruf, Boris; Detyniecki, Marcin (2021): Towards the right kind of fairness in AI. In: arxiv.org. https://doi.org/10.48550/arXiv.2102.08453

Selbst, Andrew (2017): Disparate impact in big data policing. In: Georgia Law Review 52, S. 109–195. http://dx.doi.org/10.2139/ssrn.2819182

Spindler, Mone; Booz, Sophia; Gieseler, Helya; Runschke, Sebastian; Wydra, Sven; Zinsmaier, Judith (2020): How to achieve integration? Methodological concepts and challenges for the integration of ethical, legal, social and economic aspects into technological development. In: Bruno Gransche und Arne Manzeschke (Hg.): Das geteilte Ganze. Horizonte Integrierter Forschung für künftige Mensch-Technik-Verhältnisse. Wiesbaden: Springer, S. 213–240. https://doi.org/10.1007/978-3-658-26342-3

DR. LOU THERESE BRANDNER
ist wissenschaftliche Mitarbeiterin am Internationalen Zentrum für Ethik in den Wissenschaften der Universität Tübingen. Sie hat Soziologie an der Universität von Amsterdam studiert und an der Universität La Sapienza promoviert. Sie forscht zu algorithmischer Überwachung, digitalem Kapitalismus und räumlichen Fragen.

DR. SIMON DAVID HIRSBRUNNER
leitet am IZEW Projekte zu den Themen KI-Ethik, Datenethik, sowie algorithmischer Polizeiarbeit. Er ist Sozial- und Medienwissenschaftler und forschte bisher im Bereich Human-Centered Computing an der Freien Universität Berlin, sowie bei der Wikimedia, am Potsdam Institut für Klimafolgenforschung, der Universität Potsdam und der Universität Siegen.

https://doi.org/10.14512/tatup.32.1.24