digital | recht

Schriften zum Immaterialgüter-, IT-,
Medien-, Daten- und Wettbewerbsrecht

Susanne Lillian Gössl (Hrsg.)

Diskriminierungsfreie KI

Band 10

Susanne Lilian Gössl (Hrsg.)

Diskriminierungsfreie KI

digital | recht
Schriften zum Immaterialgüter-, IT-, Medien-, Daten- und Wettbewerbsrecht

Herausgegeben von Prof. Dr. Maximilian Becker, Prof. Dr. Katharina de la Durantaye, Prof. Dr. Franz Hofmann, Prof. Dr. Ruth Janal, Prof. Dr. Anne Lauber-Rönsberg, Prof. Dr. Benjamin Raue, Prof. Dr. Herbert Zech

Band 10

Susanne Lilian Gössl ist Professorin an der Rheinischen Friedrich-Wilhelms-Universität Bonn und Direktorin des Instituts für Internationales Privatrecht und Rechtsvergleichung. Der Tagungsband gehört zum Projekt „Geschlechtsneutrale KI“, welches im Rahmen des Digitalisierungsprogramms 2021-2022 des Landes Schleswig-Holsteins durch das Ministerium für Soziales, Jugend, Familie, Senioren, Integration und Gleichstellung gefördert wurde.

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Angaben sind im Internet über http://dnb.d-nb.de abrufbar.
Dieses Buch steht gleichzeitig als elektronische Version über die Webseite der Schriftenreihe: http://digitalrecht-z.uni-trier.de/ zur Verfügung.

Umschlagsgestaltung von Monika Molin

ISBN: 9783757558208
URN: urn:nbn:de:hbz:385-2023060914
DOI: https://doi.org/10.25353/ubtr-xxxx-476a-12bf

Die Schriftenreihe wird gefördert von der Universität Trier und dem Institut für Recht und Digitalisierung Trier (IRDT).
Anschrift der Herausgeber: Universitätsring 15, 54296 Trier.

UNIVERSITÄT TRIER

IRDT Institut für Recht und Digitalisierung Trier

Vorwort

Dieser Tagungsband schließt das Projekt „Geschlechtsneutrale KI“ ab, welches im Rahmen des Digitalisierungsprogramms 2021-2022 des Landes Schleswig-Holsteins durch das Ministerium für Soziales, Jugend, Familie, Senioren, Integration und Gleichstellung gefördert wurde. Im Zentrum des Projekts stand die Erstellung einer Handreichung. Diese erarbeitet die technischen und rechtlichen Hintergründe von Diskriminierungen beim KI-Einsatz und sichtet, analysiert verschiedene nationale und internationale Vorschläge, wie diese Diskriminierungen verhindert oder vermindert werden können. Die Handreichung erschien im Januar 2023. Sie ist auf der Webseite des schleswig-holsteinischen Ministeriums für Soziales, Jugend, Familie, Senioren, Integration und Gleichstellung abrufbar.

Im Anschluss an die Fertigstellung wurde Anfang Mai 2023 eine Tagung in Kiel durchgeführt, welche bestimmte Fragen der Thematik vertiefte und ergänzte. Der Großteil der Vorträge ist in diesem Band verschriftlicht zu finden.

Herzlicher Dank gilt zunächst allen Autor*innen, die ihre Beiträge rechtzeitig eingereicht haben sowie den Herausgeber*innen der Schriftenreihe digital | recht für die Aufnahme in dieselbe. Weiterhin möchte ich mich bei Selen Yakar bedanken, welche bei der Redaktion dieses Tagungsbandes eine unersetzliche Hilfe war, sowie Leon Huchel, ohne dessen Einsatz die Tagung nicht so reibungslos verlaufen wäre.

Herzlicher Dank gilt schließlich – *last but not least* - Janique Brünung und Ino Augsberg, Direktor und Direktorin vom Zentrum für Digitalisierung und Recht in Forschung und Lehre (ZDR) an der CAU zu Kiel, welche nach meinem Weggang aus Kiel einen zentralen Teil der Organisation übernommen haben. Es war und ist wundervoll, Kolleg*innen wie Euch zu haben und mit Euch zusammenzuarbeiten!

Bonn/Kiel, Mai 2023 Susanne Gössl

Inhaltsverzeichnis

Abkürzungsverzeichnis

a.A.	andere Ansicht
aaO.	am angegebenen Ort
Abs.	Absatz
AcP	Archiv für die civilistische Praxis
a.E.	am Ende
a.F.	alte Fassung
AG	Amtsgericht
AGG	Allgemeines Gleichbehandlungsgesetz
Alt.	Alternative
allg.M.	allgemeine Meinung
Anm.	Anmerkung
aPR	allgemeines Persönlichkeitsrecht
arg.	argumentum
Art.	Artikel
Aufl.	Auflage
Az.	Aktenzeichen
Bd.	Band
BDSG	Bundesdatenschutzgesetz
Bearb.	Bearbeiter
BeckOK	Beck'scher Online-Kommentar
betr.	betreffend
BGB	Bürgerliches Gesetzbuch
BGBl.	Bundesgesetzblatt
BGH	Bundesgerichtshof
BGHZ	Amtliche Sammlung der Entscheidungen des Bundesgerichtshofs
BJM	Bundesministerium der Justiz und für Verbraucherschutz
BRegBundesregierung	
BT-Drucks.	Bundestags-Drucksache

BVerfG	Bundesverfassungsgericht
BVerfGE	Amtliche Sammlung der Entscheidungen des Bundesverfassungsgerichts
bzw.	beziehungsweise
COM	European Commission
CR	Computer und Recht
CR	Computer und Recht International
ders.	derselbe
DesignG	Gesetz über den rechtlichen Schutz von Design
d.h.	das heißt
dies.	dieselbe/n
Diss.	Dissertation
DRMS	Digital Rights Management System
DSGVO	Datenschutz-Grundverordnung
DZPhil	Deutsche Zeitschrift für Philosophie
ebd.	ebenda
EGMR	Europäischer Gerichtshof für Menschenrechte
Einl	Einleitung
EMRK	Europäische Menschenrechtskonvention
ErwG.	Erwägungsgrund
etc.	et cetera
EU	Europäische Union
EuG	Gericht der Europäischen Union
EuGH	Europäischer Gerichtshof
EuGHE	Sammlung der Entscheidungen des EUGH
f., ff.	Folgende
Fn.	Fußnote
FS	Festschrift
G-BA	Gemeinsamer Bundesausschuss
GA	Generalanwalt
GebrMG	Gebrauchsmustergesetz
GEMA	Gesellschaft für musikalische Aufführungs- und mechanische Vervielfältigungsrechte
GG	Grundgesetz
ggf.	gegebenenfalls
GRUR	Gewerblicher Rechtsschutz und Urheberrecht

GRUR-Prax	Gewerblicher Rechtsschutz und Urheberrecht, Praxis im Immaterialgüter- und Wettbewerbsrecht
GRUR-RR	GRUR Rechtsprechungs-Report
GVG	Gerichtsverfassungsgesetz
GWB	Gesetz gegen Wettbewerbsbeschränkungen
h.L.	herrschende Lehre
h.M.	herrschende Meinung
Hrsg.	Herausgeber
i.d.F	in der Fassung
i.d.R	in der Regel
i.E.	im Ergebnis
i.H.v.	in Höhe von
IIC	International Review of Intellectual Property and Competition Law
IFG	Informationsfreiheitsgesetz
ImGR	Immaterialgüterrecht
insb	insbesondere
internat.	international
i.S.v./i.S.d.	im Sinne von/des
i.V.m.	in Verbindung mit
JuS	Juristische Schulung
JZ	Juristenzeitung
K&R	Kommunikation & Recht
Kap.	Kapitel
KI	Künstliche Intelligenz
KOM	Europäische Kommission
KUG	Gesetz vom 9.1.1907 betreffend das Urheberrecht an Werken der bildenden Künste und der Photographie
LG	Landgericht
lit.	littera
LLM	large lange models
LS	Leitsatz
LSchR	Leistungsschutzrecht
lt.	laut
LUG	Gesetz vom 19.6.1901 betreffend das Urheberrecht an Werken der Literatur und der Tonkunst

MarkenG	Gesetz über den Schutz von Marken und sonstigen Kennzeichen
MarkenR	Zeitschrift für deutsches, europäisches und internationales Kennzeichenrecht
ML	Maschines Lernen
MMR	MultiMedia und Recht
MStV	Medienstaatsvertrag
MüKoBGB	Münchener Kommentar zum BGB
m.w.N.	mit weiteren Nachweisen
n.F.	neue Fassung
NJW	Neue Juristische Wochenschrift
NJW-RR	NJW-Rechtsprechungs-Report
Nr.	Nummer
OLG	Oberlandesgericht
PatG	Patentgesetz
p.m.a.	post mortem auctoris
RefE	Referentenentwurf
RegE	Regierungsentwurf
RG	Reichsgericht
RGBl.	Reichsgesetzblatt
RGZ	Entscheidungen des Reichsgerichts
RL	Richtlinie
Rn.	Randnummer
Rs.	Rechtssache
Rspr.	Rechtsprechung
S.	Seite
s.a.	siehe auch
Slg.	Sammlung der Entscheidungen des Gerichtshofs der Europäischen Gemeinschaft (Jahr und Seite)
sog.	sogenannte/sogenannter/sogenannten
SortenG	Sortenschutzgesetz
str.	strittig
stRspr.	ständige Rechtsprechung
TMG	Telemediengesetz
u.a.	und andere

UFITA	Archiv für Urheber-, Film-, Funk-, und Theaterrecht ab 2000: Archiv für Urheber- und Medienrecht
UrhG	Gesetz über Urheberrecht und verwandte Schutzrechte
UrhWahrnG	Gesetz über die Wahrnehmung von Urheberrechten und verwandten Schutzrechten
URL	United Resource Locator
Urt.	Urteil
u.v.m.	und vieles mehr
UWG	Gesetz gegen den unlauteren Wettbewerb
v.	versus
VerlG	Gesetz über das Verlagsrecht
VG	Verwertungsgesellschaft
vgl.	vergleiche
VO	Verordnung
Vorb.	Vorbemerkung
WahrnG	Wahrnehmungsgesetz
weit.	weitere
WIPO	World Intellectual Property Organization
WRP	Wettbewerb in Recht und Praxis
WWW	World Wide Web
z.B.	zum Beispiel
ZGE	Zeitschrift für Geistiges Eigentum
Ziff.	Ziffer
zit.	zitiert
ZPO	Zivilprozessordnung
z.T.	zum Teil
ZUM	Zeitschrift für Urheber- und Medienrecht
ZUM-RD	ZUM-Rechtsprechungsdienst
zust.	Zustimmend

Verzeichnis der Autorinnen und Autoren

Jens Ambrock

Dr. jur., Referatsleiter beim Hamburgischen Beauftragten für Datenschutz und Informationsfreiheit und Lehrbeauftragter an der Christian-Albrechts-Universität zu Kiel.

Martin Gerecke

Dr. jur., M.Jur. (Oxford) Partner bei CMS und Lehrbeauftragter an der Christian-Albrechts-Universität zu Kiel

Susanne Lilian Gössl

Prof. Dr., LL.M. (Tulane), Direktorin des Instituts für Internationales Privatrecht und Rechtsvergleichung, Lehrstuhl für deutsches, ausländisches und Internationales Privatrecht und das Recht der Digitalisierung an der Rheinischen Friedrich-Wilhelm-Universität Bonn

Anna Kirchhefer-Lauber

Dr. jur., LL.M. (Bristol), Akademische Rätin an der Universität Münster

Miriam Rateike

B.Sc. and M.Sc., Wissenschaftliche Mitarbeiterin und Doktorandin an der Universität des Saarlandes

Moritz von Rochow
Dr. jur., Wissenschaftlicher Mitarbeiter am Walther-Schücking-Institut für Internationales Recht der Christian-Albrechts-Universität zu Kiel und Fachanwalt für Verwaltungsrecht

Caja Thimm
Prof. Dr., Professorin für Medienwissenschaft und Intermedialität an der Rheinischen Friedrich-Wilhelms-Universität Bonn

Laura Thimm-Braun
LL.M. (Wien), Volljuristin

Selen Yakar
Wissenschaftliche Mitarbeiterin an der Universität zu Köln

Teil 1

Einführung und Grundlagen

Kapitel 1

KI-Systeme und Diskriminierung – Eine Einführung

Susanne Lilian Gössl

Künstliche Intelligenz (KI) kann Empfehlungen aussprechen oder Entscheidungen fällen, die Personen diskriminieren. Dies ist inzwischen im allgemeinen Bewusstsein angekommen.

Die Beiträge in diesem Tagungsband ergänzen und vertiefen die von Selen Yakar und mir im Auftrag des Landes Schleswig-Holstein (Ministerium für Soziales, Jugend, Familie, Senioren, Integration und Gleichstellung) durchgeführte Studie „Geschlechterneutrale KI“[1]. In der Studie wird der aktuelle Rechtsrahmen bezogen auf diskriminierenden KI-Einsatz erläutert, in die technischen Hintergründe eingeführt und insbesondere auf internationaler Ebene nach bereits diskutierten Lösungen gesucht. Hierauf aufbauend werden eigene Lösungsvorschläge entwickelt, um Diskriminierungen beim Einsatz von KI-Systemen zu vermeiden oder zumindest zu verringern.

Es blieb bei diesem bereits recht umfangreichen Programm für spezielle Fragen nicht genügend Raum. Darüber hinaus ist für die Frage einer KI-Regulierung, auch zur Vermeidung von Diskriminierung, zukünftig die gerade in Entwicklung befindliche KI-VO einschlägiges Gesetz. Seit Fertigstellung der Handreichung hat der Vorschlag der KI-VO sich bereits mehrfach geändert, sodass der einschlägige Tagungsbeitrag (Yakar, Diskriminierungsverbote im deutschen und europäischen Recht und die zukünftige KI-VO, S. 53 ff.) eine Aktualisierung zur im Januar 2023 abgeschlossenen Studie darstellt. Die aus Platzgründen und aufgrund der fortschreitenden Entwicklungen entstandenen Lücken versucht dieser Tagungsband zu schließen.

---

[1] Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, 2023, abrufbar unter https://www.schleswig-holstein.de/DE/fachinhalte/G/gleichstellung/geschlechterneutrale_ki.html.

In dieser Einführung wird zunächst der Begriff der „KI-Systeme“ juristisch handhabbar gemacht (A.). Im Anschluss wird aufgezeigt, wie es zu Diskriminierungen beim Einsatz von KI-Systemen kommen kann (B.). Die Einführung schließt mit einem Überblick, was in den folgenden Beiträgen zu erwarten ist (C.)

A. Der Begriff des KI-Systems

I. Was bedeutet „KI-System“?

Der Begriff „KI“ bzw. „KI-Systeme“ ist Schlüsselbegriff für die Handreichung und auch diesen Tagungsband. Eine Definition ist umso schwieriger, abhängig davon, aus welcher Perspektive man sich dem Begriff annähert – philosophisch, anthropologisch oder wie hier: juristisch.
Für die folgenden Beiträge und auch für die genannte Handreichung wird der Begriff soweit definiert, als es erforderlich ist, um ihn für die juristische Arbeitsweise handhabbar zu machen. Es wird die Definition verwendet, welche die Hochrangige Expertengruppe der Europäischen Union (High Level Expert Group on Artificial Intelligence) entwickelt hat und ihren Ausführungen zugrunde legt. KI-Systeme sind demnach

> „[...] vom Menschen entwickelte Softwaresysteme [...], die in Bezug auf ein komplexes Ziel auf physischer oder digitaler Ebene handeln, indem sie ihre Umgebung durch Datenerfassung wahrnehmen, die gesammelten strukturierten oder unstrukturierten Daten interpretieren, Schlussfolgerungen daraus ziehen oder die aus diesen Daten abgeleiteten Informationen verarbeiten, und über das bestmögliche Handeln zur Erreichung des vorgegebenen Ziels entscheiden. KI-Systeme können entweder symbolische Regeln verwenden oder ein numerisches Modell erlernen, und sind auch in der Lage, die Auswirkungen ihrer früheren Handlungen auf die Umgebung zu analysieren und ihr Verhalten entsprechend anzupassen. Als wissenschaftliche Disziplin umfasst die KI mehrere Ansätze und Techniken wie z.B. maschinelles Lernen [...], maschinelles Denken [...] und die Robotik [...].“[2]

[2] *High Level Expert Group on Artificial Intelligence* (HLEG AI) Eine Definition der KI: Wichtigste Fähigkeiten und Wissenschaftsgebiete, Juni 2018, S., S. 6.

Relevant ist bei dieser Definition insbesondere, dass KI-Systeme nicht zu eng zu verstehen ist. Erfasst sind damit nicht nur Systeme, die von menschlichem Handeln unabhängig entscheiden können, sondern auch solche, die noch von Menschen unterstützt werden müssen oder noch menschliche Zwischenschritte bis zu einer endgültigen Entscheidung erfordern.

II. Wo werden KI bereits eingesetzt?

Auch wenn Diskussionen über den Einsatz von KI-Systemen häufig futuristisch scheinen, ist dieser bereits in verschiedensten Lebensbereichen üblich geworden.
In der privaten Wirtschaft werden etwa personalisierte Job-Vorschläge[3], personalisierte Werbung[4], Rankings von Dienstleistungen oder Dienstanbietenden[5] oder Produktvorschläge[6] häufig in der Reihenfolge präsentiert, die von einem KI-System vorher ausgewählt wurde.
Darüber hinaus findet gerade auch in Deutschland eine lebhafte Diskussion zu dem Thema statt, inwieweit ein KI-System in der gerichtlichen oder außergerichtlichen Streitbeilegung eingesetzt werden kann.[7] Während einhellige Meinung ist, dass Art. 92 GG und andere Regelungen der Verfassung davon ausgehen, eine Gerichtsentscheidung müsse stets von einer natürlichen Person, einem Mensch, gefällt werden,[8] sind die Grenzen, inwieweit ein KI-System dem Ge-

---

[3] Lambrecht/Tucker Management Science 65 (2019), 2966 ff.
[4] Dazu z.B. Speicher/Ali/Venkatadri/Ribeiro/Arvanitakis/Benevenuto/Gummadi/Loiseau/Mislove Proceedings of Machine Learning Research 81 (2018) 81.
[5] Mit einem Fokus aufs Scoring Gerberding/Wagner ZRP 2019, 116 (118).
[6] Etwa Wachter Berkeley Technology Law Journal 35 (2020), 367 (369 ff.).
[7] Z.B. Wolff Algorithmen als Richter, 2022; Deichsel Digitalisierung der Streitbeilegung, 2021.
[8] Legal Tech: Herausforderungen für die Justiz, Abschlussbericht der Länderarbeitsgruppe, abrufbar unter https://www.schleswig-holstein.de/DE/landesregierung/ministerien-behoerden/II/Minister/Justizministerkonferenz/Downloads/190605_beschluesse/TOPI_11_Abschlussbericht.pdf?__blob=publicationFile&v=1, 54; Graichen Die Automatisierung der Justiz , 285 ff.; kritisch Wolff Algorithmen als Richter, 2022, S. 125 ff.

richt oder einer Streitbeilegungsstelle zuarbeiten kann, noch nicht vollends austariert.[9] Darüber hinaus werden KI-Systeme etwa auch von staatlichen Einrichtungen eingesetzt, um Sozial[10]- oder Steuerbetrug[11] aufzudecken oder für das „Predictive Policing"[12]. Dies sind nur einige von vielen Beispielen, in denen es in der privaten Wirtschaft oder im öffentlichen Sektor bereits üblich ist, KI-Systeme zur Arbeitserleichterung oder -verbesserung einzusetzen.[13]

B. Die Diskriminierungsproblematik

I. Beispiele für Diskriminierungen durch KI-Einsatz

Dass diese Systemeinsätze zu diskriminierenden Ergebnissen führen können, hat sich in der Vergangenheit vielfach gezeigt, einige Beispiele haben traurige Berühmtheit erlangt. So setzte Amazon.com Inc. eine Software ein, um bei der Personalrekrutierung Lebensläufe durch KI-Systeme vorauswählen zu lassen. Es stellte sich heraus, dass diese Systeme aus verschiedenen Gründen insbesondere Frauen diskriminierten.[14] Das System COMPAS, welches in den USA eingesetzt wird, um die Rückfallwahrscheinlichkeit von Straftäter*innen zu bestimmen, ist immer wieder dem Vorwurf ausgesetzt, Personen afro-amerikanischer Herkunft zu diskriminieren.[15] Auch in weniger grundrechtssensiblen Gebieten zeigt

---

[9] Deichsel Digitalisierung der Streitbeilegung, 2021, S. 115 ff.

[10] Etwa in den Niederlanden, vgl. Amnesty International, Xenophobic Machines, 2021, abrufbar unter https://www.amnesty.org/en/documents/eur35/4686/2021/en/, 22.

[11] Zum Steuerbetrug vgl. von Rochow (in diesem Band), S. 103 ff.

[12] Hierzu ausführlich Thimm/Thimm-Braun (in diesem Band), S. 37 ff.

[13] Weitere Beispiele etwa Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, 2023, insb. S. 53 ff.

[14] Dastin, Amazon scraps secret AI recruiting tool that showed bias against women, 11.10.2018, abrufbar unter https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G; Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, 2023, S. 26, 55 ff.

[15] Dazu etwa Zuiderveen Borgesius Discrimination, artificial intelligence, and algorithmic decision-making, 2018, abrufbar unter https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73, S. 23-25.

sich immer wieder die Diskriminierungsanfälligkeit eines KI-Einsatzes. Von Seifenspendern, Fotografieautomaten oder Gesichtserkennungssoftware bis zu Chatbots,[16] Übersetzungssoftware[17] und Fußball-Videospielen[18] – überall zeigt sich die Gefahr, zu diskriminieren und bestimmte Gruppen auszuschließen.[19]

II. Technische Hintergründe

Ein Algorithmus kommt zu bestimmten Ergebnissen, indem er Korrelationen herstellt. Eine Entscheidungsfindung funktioniert, vereinfacht gesprochen, folgendermaßen: Die Software, die einen Algorithmus einsetzt, wird mit Daten gefüttert. Auf Grundlage dieser Daten werden Korrelationen gefunden. Basierend auf diesen Korrelationen wird eine Vorhersage für einen neuen Fall getroffen, wie dieser zu behandeln sei. Durch diese Vorhersage kommt das System zu einer Empfehlung oder Entscheidung. Bei allen diesen Zwischenschritten können Verzerrungen auftreten, welche infolgedessen auch die Empfehlung oder Entscheidung verzerren – was zu Diskriminierungen führen kann.[20]
Diese Verzerrungen, die in der juristischen Literatur häufig als *bias*[21] bezeichnet werden,[22] entstehen vor allem an drei Stellen: Zum einen kann die Komposition der Trainingsdaten (1.), zum anderen eine unbewusste Vorprägung der Trainingsdaten (2.) zu Verzerrungen führen. Schließlich kann auch das KI-System

---

[16] SocialHax Microsoft Creates AI Bot - Internet Immediately Turns it Racist, 2016, abrufbar unter https://socialhax.com/2016/03/24/microsoft-creates-ai-bot-internet-immediately-turns-racist/.

[17] Dazu Eigenversuch von Yakar in Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, 2023, S. 59.

[18] Maloney/Campbell How the Fifa20 video game reproduces the racial stereotypes embedded within football, 19.01.2023, abrufbar unter https://theconversation.com/how-the-fifa20-video-game-reproduces-the-racial-stereotypes-embedded-within-football-197237.

[19] Beispiele Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, 2023, S. 35 ff.

[20] Ausf. etwa Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, 2023, S. 53 ff.; Alpaydın Machine learning, 2016, S. 16 ff; Gerards/Xenidis Algorithmic discrimination in Europe: Challenges and opportunities for gender equality and non-discrimination law, 2020, S. 32 ff.

[21] Zum teilweise abweichenden Begriff in der Informatik siehe Rateike (in diesem Band), S. 17 ff.

[22] Dazu etwa Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, 2023, S. 54.

oder der Lernprozess desselben so ausgestaltet sein, dass Verzerrungen begünstigt oder sogar erst geschaffen werden (3.).[23]
Insbesondere beim Einsatz von selbstlernenden Systemen kann sich das Diskriminierungspotential noch weiter verstärken. Dies liegt zum einen an der Black-box-Problematik (4.), zum anderen daran, dass *reinforcement learning* die Funktion eines Katalysators zukommen kann (5.). Schließlich stellt sich gerade beim Einsatz von KI-Systemen das Problem der *proxy discrimination* (6.).
Aus technischer Sicht werden diese und weitere Probleme auch im Beitrag von Rateike im nächsten Kapitel beschrieben.[24]

1. Komposition der Trainingsdaten

Die Zusammensetzung und insbesondere die Diversität der Trainingsdaten ist von überragender Bedeutung für das spätere Ergebnis, zu dem das KI-System kommt. Ein berühmtes bereits erwähntes Beispiel dafür, wie die mangelnde Diversität der Trainingsdaten zu Diskriminierung führen kann, ist das Amazon Recruiting-Tool:[25] Die Software sollte objektive Vorhersagen über die Qualität und Passgenauigkeit der Lebensläufe von Jobkandidat*innen treffen. Der Algorithmus wurde aber anhand von Lebensläufen der vorhergegangenen Dekaden trainiert – was dazu führte, dass ein signifikant höherer Anteil der Trainingsdaten von männlichen Personen stammte, da im Informatiksektor Frauen typischerweise unterrepräsentiert waren und sind. Die Trainingsdaten waren also bezogen auf das Geschlecht (und vermutlich noch weitere Daten) nicht sehr divers. Infolgedessen "schlussfolgerte" die Software, dass Frauen weniger qualifiziert für die gesuchten Bereiche waren. Dies führte zu diskriminierenden Vorschlägen, welche Kandidat*innen besser geeignet seien.[26] Ähnliche Unter- und Überrepräsentanzen zeigen sich regelmäßig, wenn die Daten aus der Realität stammen – die Wahrscheinlichkeit, dass Diversität bezogen auf die Personen fehlt, die typischerweise unterrepräsentiert sind, ist hoch, d.h. Frauen insb. in

---

[23] Z.B. Genovesi/Kaesling/Robbins Recommender Systems/Gössl, 2023, (im Erscheinen).
[24] Rateike (in diesem Band), S. 17 ff.
[25] Gershgorn, Amazons "holy grail" recruiting tool was actually just biased against women, Quartz, 10.10.2019, abrufbar unter https://qz.com/1419228/amazons-ai-powered-recruiting-tool-was-biased-against-women.
[26] Ebd.

den MINT-Berufen und in Führungspositionen, Männer in Pflege- oder Erziehungsberufen. Auch bezogen auf Minderheiten ist die Wahrscheinlichkeit hoch, dass Mitglieder bestimmter Gruppen (People of Color, Migrant*innen, LGBTIAQ*, Menschen mit Behinderung...) unterrepräsentiert sind.[27] Ähnliche Beispiele wie jenes des Amazon Recruiting-Tools finden sich im Bereich des Predictive Policing[28] oder im Bereich der Gesichtserkennungssoftware: Software, die vor allem mit Bildern von hellhäutigen und männlichen Personen trainiert wurde, hatte Probleme, dunkelhäutige oder weibliche oder insbesondere dunkelhäutige weibliche Personen zu identifizieren.[29]

2. Vorprägung der Trainingsdaten

Ein weiterer wesentlicher Grund, warum der Einsatz von KI-Systemen zu diskriminierenden Ergebnissen führen kann, ist die Tatsache, dass die Trainingsdaten häufig auf Daten zurückgreifen, die von realen Personen generiert wurden und daher bewusste oder unbewusste Vorurteile dieser Personen enthalten und weitertragen können. Die Ergebnisse der KI-Entscheidung reflektieren diese Vorurteile und können entsprechend verzerrt sein. Zum Beispiel wurde an der Universität Bonn eine empirische Studie zu Geschlechterunterschieden in der Finanzberatung durchgeführt („Gender Differences in Financial Advice“[30]). In dieser Studie wurde untersucht, wie sich Empfehlungen von (menschlichen) Finanzberatern abhängig davon unterscheiden, welches Geschlecht die beratene Person hat. Die Studie zeigte, dass typischerweise Empfehlungen, die Frauen gegenüber abgegeben wurden, kostspieliger waren als Empfehlungen gegenüber Männern. Diese Tatsache lässt sich auf verschiedene Faktoren zurückführen, z.B. dass Frauen risikoaverser sind und daher häufiger die sicheren, aber teuren Anlagemöglichkeiten vorziehen, aber auch, dass Frauen

---

[27] Z.B. Sheltzer/Smith Proceedings of the National Academy of Sciences of the United States of America 111 (2014), 10107 ff.

[28] Dazu Thimm/Thimm-Braun (in diesem Band), S. 37 ff.

[29] Z.B. Buolamwini/Gebru Proceedings of Machine Learning Research 81 (2018) 81 ff.

[30] Bucher-Koenen/ Hackethal/ Koenen/ Laudenbach ECONtribute Discussion Paper No. 095, 2021; siehe auch Cavalluzzo/Cavalluzzo Journal of Money, Credit and Banking 30 (1998), 771 ff.

sich seltener eine zweite Meinung einholen als Männer und daher Empfehlungen weniger gut vergleichen können.[31] Ein KI-System, das von diesen Daten lernt, kann zu der Schlussfolgerung kommen, dass Frauen immer ein teureres Produkt vorgeschlagen werden sollte als Männern, ohne auf die konkrete Person zu schauen.
Ähnlich existieren Webseiten, die professionelle Dienstleistungen in bestimmten Regionen ranken, wobei die Rankings häufig auf Nutzer*innenbewertungen zurückgehen.[32] Dieses Rankingverfahren kann ebenfalls zu Diskriminierungen etwa von Frauen und Minderheiten führen: Studien zeigen, dass Mitglieder dieser Gruppen typischerweise als „schlechter" eingestuft werden als Männer, die nicht zu der Minderheitengruppe gehören, selbst wenn die Qualität der Leistung identisch ist. Zum Beispiel wurden identische Lebensläufe, in denen nur der Vorname geändert werde, unterschiedlich bewertet.[33] Ähnliche Ergebnisse ergeben sich bei der Evaluation von identischen Lehrmaterialien, bei denen nur der Vorname der Lehrperson den Unterschied machte.[34]
Ein Ranking basierend auf vorhergehenden Kundenbewertungen enthält daher mit einer hohen Wahrscheinlichkeit (unbewusste) Vorurteile der Personen, welche die Bewertungen abgegeben haben. Diese Vorurteile führen zu einem weniger vorteilhaften Ranking, was wiederum auf die Kundenakquise und damit die professionellen Erfolge der Person negative Auswirkungen haben kann.

3. Ausgestaltung des KI-Systems oder des Lernprozesses

Schließlich kann das KI-System oder sein Lernprozess so ausgestaltet sein, dass es bzw. er Diskriminierungen verursacht oder bereits vorhandene Verzerrungen in den Daten noch weiter begünstigt. Ein Grund kann die Auswahl der konkreten Variablen sein, nach denen der Algorithmus differenzieren soll. Zum Beispiel in dem Fall, in dem Sozialbetrug durch die niederländischen Behörden

---

[31] Bucher-Koenen/ Hackethal/ Koenen/ Laudenbach ECONtribute Discussion Paper No. 095 2021, 12 f., 14 f.
[32] Z.B. https://www.jameda.de/.
[33] Z.B. Moss-Racusin/Dovidio/Brescoll/Graham/Handelsman Proceedings of the National Academy of Sciences of the United States of America 109 (2012), 16474 ff.; Handley/Brown/Moss-Racusin/Smith Proceedings of the National Academy of Sciences of the United States of America 112 (2015), 13201 ff.
[34] Özgümüs/Rau/Trautmann/König-Kersting Frontiers in psychology 11 (2020), 1074 ff.

ermittelt werden sollte, knüpfte der Algorithmus bei der Auswahl der verdächtigen Personen u.a. an die Staatsangehörigkeit der Person an. Dies führte zu einer Diskriminierung aller Personen mit nicht-niederländischer Staatsangehörigkeit, da die Sozialleistung (Kindergeld) nicht an die Staatsangehörigkeit anknüpft, sondern den Wohnsitz in den Niederlanden[35].[36]

Doch auch die weitere Ausgestaltung des Algorithmus kann einen erheblichen Einfluss auf das Ergebnis haben. Ein weiteres Beispiel, bei dem die Programmierung eine bereits bestehende Diskriminierung verstärkte, findet sich bei der berühmt gewordenen Software COMPAS, die in verschieden US-Staaten eingesetzt wird.[37] Die Software soll die Rückfallwahrscheinlichkeit von Straftätern einschätzen. Problematisch ist, dass COMPAS auf vorhergehenden Verurteilungen und diesen zu entnehmenden Biografien aufbaut und dabei die (unbewussten) Vorurteile der Richter*innen gegenüber afro-amerikanischen Personen enthält. Die Rückfallwahrscheinlichkeit wurde bei afro-amerikanischen Personen als deutlich höher eingeschätzt als bei Personen anderer Hautfarbe. Zugleich wurde dieselbe Wahrscheinlichkeit bei Personen mit weißer Hautfarbe als deutlich geringer eingestuft. Ein Problem in der Programmierung war, dass aber die Fehlerquotenallokation beiden Gruppen gegenüber gleich angesetzt war. Diese Fehlerquotenallokation war aber fehlerhaft, da Personen mit weißer Hautfarbe gleich zwei positive Vorteile für sich hatten, afro-amerikanische einen gegen sich, die Fehlerquoten also ungleich verteilt. Damit verstärkte die

---

[35] *Autoriteit Persoonsgegevens*, Tax Administration fined for discriminatory and unlawful data processing, 19.01.2023, abrufbar unter https://autoriteitpersoonsgegevens.nl/en/news/tax-administration-fined-discriminatory-and-unlawful-data-processing.

[36] Für weitere Beispiele siehe Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, 2023, S. 65 ff. Ali/Sapiezynski/Bogen/Korolova/Mislove/Rieke Proc. ACM Hum.-Comput. Interact. 3 (2019), 1 ff.

[37] Z.B. Larson/Mattu/Kirchner/Angwin How We Analyzed the COMPAS Recidivism Algorithm, 23.05.2016, abrufbar unter https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm; Machine Bias. *ProPublica*, 23.5.2016, abrufbar unter https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; Flores/Bechtel/Lowenkamp Federal Probation 80 (2016), 38 ff.; Martini Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, 2019, S. 2 ff.

Ausgestaltung des KI-Systems die bereits vorhandenen Verzerrungen noch weiter.[38]

4. Blackbox-Problematik

KI-Systeme unterscheiden sich u.a. von linearen Algorithmen dadurch, dass sie „selbstlernend" sind, d.h. dass sie anhand der Trainingsdaten eigene Muster und Regeln aufstellen. Sie können sich dadurch immer weiter fortentwickeln und anpassen. Am deutlichsten ist dies beim *deep learning*. Hier beruhen die Entscheidungen auf sehr kleinschrittigen Verknüpfungen. Die Verknüpfungen ändern sich bei jeder Entscheidung und passen sich so den jeweiligen Daten an. Die Kapazität eines Algorithmus übersteigt dabei um ein Vielfaches die menschliche Wahrnehmungsfähigkeit. [39] Problematisch ist hieran, dass die vom KI-System aufgestellten Muster und Regeln häufig weder für die Personen, die das KI-System programmiert haben, noch für jene, welche das System betreiben, im Vorfeld erkennbar sind und dies auch im Nachhinein häufig bleiben. Diese Eigenständigkeit des KI-Systems stellt einen großen Vorteil im Vergleich zu regelbasierten Systemen dar, da das System sich weiterentwickeln kann. Doch die beschriebene, als „Blackbox-Problematik" bezeichnete Intransparenz der KI-Systeme erschwert die Feststellung und die Vermeidung von Diskriminierungen, da sich das von der KI generierte Ergebnis häufig auch von Fachleuten weder im Vorfeld bestimmen noch im Nachgang nachvollziehen lässt.[40]

5. reinforcement learning *als Katalysator*

Problematisch ist zudem die Wiederholung eigener Fehler des KI-Systems im *reinforcement learning*: Wiederholt das KI-System den gleichen Fehler und schleicht sich dieser dadurch in das erlernte Muster im Rahmen des *reinforce-*

---

[38] Z.B. Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, 2023, S. 68 ff.; Chouldechova Big data 5 (2017), 153 ff.; Rahman. COMPAS Case Study: Fairness of a Machine Learning Model. *Towards Data Science*, 7.9.2020, abrufbar unter https://towardsdatascience.com/compas-case-study-fairness-of-a-machine-learning-model-f0f804108751.

[39] Martini Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, 2019, S., S. 41 ff.

[40] Ausf. Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, 2023, S. 70 ff.

ment learning ein, wiederholt und verstärkt derselbe sich. Es entstehen unendliche Wiederholungen desselben Fehlers. Derselbe wird dadurch noch weiter verstärkt.[41]

6. proxy discrimination

Schließlich ist im Bereich des KI-Einsatzes das Problem der *proxy discrimination* besonders relevant. Dieses Phänomen ergibt sich insbesondere dann, wenn offensichtliche Diskriminierungsmerkmale unterdrückt werden, also gerade nicht verwendet werden können. Das KI-System stellt dann Korrelationen aufgrund von *proxy*-Merkmalen her, die insbesondere in ihrem Zusammenspiel gleiche oder ähnliche Auswirkungen haben wie das eigentliche Merkmal selbst.[42] Problematisch ist hier u.a., dass die Auswahl der *proxy*-Merkmale unvorhersehbar ist, also eine Blackbox darstellt.

C. Offene Fragen und Fahrplan für diesen Band

Die anfangs genannte Handreichung wurde von zwei Juristinnen verfasst. Die juristische Perspektive steht daher im Vordergrund, selbst wenn natürlich die technischen Grundlagen erarbeitet werden mussten. Ein wesentliches Ergebnis der Studie ist u.a., dass im Bereich der Diskriminierungsvermeidung bei KI-Einsatz stärker inter- und intradisziplinär zusammengearbeitet, diskutiert und sich ausgetauscht werden sollte. Kerndisziplin, die sich für eine Zusammenarbeit anbietet, ist selbstverständlich die Informatik. Aus diesem Grund beginnt der Tagungsband mit der Perspektive der Informatik auf die Diskriminierungsproblematik (Rateike, Diskriminierung im Maschinellen Lernen - Ein kurzer Einblick aus der Perspektive der Informatik).[43] Eine andere Disziplin, die zu einem fruchtvollen Austausch beitragen kann und sollte, ist die Soziologie. Aus mediensoziologisch-juristischer Perspektive widmet sich der darauf folgende

---

[41] Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, 2023, S. 69 f.; Ad Hoc Committee on Artificial Intelligence (CAHAI), Feasibility Study, 17.12.2020, abrufbar unter https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da, Rn. 15.

[42] Ali/Sapiezynski/Bogen/Korolova/Mislove/Rieke Proc. ACM Hum.-Comput. Interact. 3 (2019), 1 ff.; Buolamwini/Gebru Proceedings of Machine Learning Research 81 (2018), 12 ff.

[43] S. 17 ff.

Beitrag dem bereits beschriebenen Problem des Predictive Policing bei KI-Einsatz (Thimm/Thimm-Braun, Künstliche Intelligenz und personale Autonomie: Diskriminierende Algorithmen als ethische und rechtliche Herausforderung für die Polizeiarbeit).[44]

Im Anschluss wird der aktuelle und zukünftige Rechtsrahmen dargestellt und darauf analysiert, inwieweit er bereits Vorgaben für den KI-Einsatz enthält und inwieweit diese Vorgaben geeignet sind, Diskriminierungen zu vermeiden (Yakar, Diskriminierungsverbote im deutschen und europäischen Recht und die zukünftige KI-VO).[45] Ein besonderes Augenmerk verdient dabei das Datenschutzrecht (Ambrock, Datenschutzrechtliche Anforderungen an diskriminierungsfreien KI-Einsatz).[46] Auch dem Nichtdiskriminierungsrecht, insbesondere bezogen auf die Proxy-Diskriminierung und die Bildung von unbekannten („nebulösen") Gruppen wird besondere Aufmerksamkeit geschenkt (Kirchhefer-Lauber, Im Nebel: Der Schutz von algorithmischen Gruppen im deutschen Nichtdiskriminierungsrecht).[47] Eine ebensolche Relevanz beim Einsatz von KI-Systemen durch Behörden oder andere öffentliche Einrichtung hat darüber hinaus die Rechtsfolgenseite, weswegen dem Staatshaftungsrecht ein eigenes Kapitel gewidmet ist (von Rochow, Staatshaftung beim KI-Einsatz).[48]

Neben einem interdisziplinären und intradisziplinären Austausch ist darüber hinaus ein Austausch mit der Praxis wichtig. Einen Ausblick auf die – rechtsberatende – Praxis und ihre Perspektive auf KI-Einsätze, bietet das letzte Kapitel (Gerecke, Entscheide Du, KI! – wie uns künstliche Intelligenz in der Anwaltsberatung helfen kann und wo die Grenzen sind).[49]

Der Tagungsband endet mit einer Zusammenfassung der Ergebnisse der Studie, die sich in allgemeine Empfehlungen, Empfehlungen für die KI-nutzende Praxis und Empfehlungen für die Forschung aufteilen.[50] Es bleibt zu hoffen, dass diese Ergebnisse und auch die einzelnen Beiträge dieses Bandes zu weiteren Diskussionen anstoßen und damit dazu beitragen, die Welt des KI-Einsatzes etwas diskriminierungsfreier zu gestalten.

---

[44] S. 37 ff.
[45] S. 53 ff.
[46] S. 69 ff.
[47] S. 89 ff.
[48] S. 103 ff.
[49] S. 129 ff.
[50] S. 139 ff.

Es bleibt zu hoffen, dass diese Ergebnisse und auch die einzelnen Beiträge dieses Bandes zu weiteren Diskussionen anstoßen und damit dazu beitragen, die Welt des KI-Einsatzes etwas diskriminierungsfreier zustalten.

Kapitel 2

Diskriminierung im Maschinellen Lernen - Ein kurzer Einblick aus der Perspektive der Informatik

Miriam Rateike

Maschinelles Lernen (ML) wird zunehmend eingesetzt, um Entscheidungen zu treffen oder zu erleichtern, die Menschen in verschiedenen Bereichen wie Gesundheit und Finanzen betreffen. Dabei ist es in der Vergangenheit immer wieder vorgekommen, dass einige dieser Algorithmen bestimmte Personengruppen aufgrund sensibler Merkmale wie z.B. dem sozialen Geschlecht bevorzugen oder benachteiligen. Dies hat in den letzten Jahren zu einem verstärkten Druck geführt, Diskriminierung durch algorithmische Entscheidungsprozesse zu reduzieren und deren Erklärbarkeit zu gewährleisten. In diesem Artikel werden einige der häufigsten Ursachen für die Diskriminierung durch algorithmische Entscheidungen entlang des Entwicklungsprozesses eines ML-Systems dargestellt. Der Artikel gibt auch einen kurzen Überblick über das Forschungsgebiet der Informatik, das sich mit Fairness (Nicht-Diskriminierung) und Erklärbarkeit von Algorithmen beschäftigt.

A. Einleitung und Begriffsdefinition

Die Entscheidungen von Dritten können den Erfolg und die Chancen eines Menschen wesentlich beeinflussen, z. B. Entscheidungen über die Zulassung zu einem Studium, über die Aufnahme eines Arbeitsverhältnisses oder über die Gewährung eines Darlehens. Werden diese Entscheidungen willkürlich oder fehlerhaft getroffen, kann dies dazu führen, dass Einzelne ihre Ziele nicht erreichen

oder ihre Fähigkeiten nicht voll ausschöpfen können.[1] Damit Entscheidungen auf einer soliden Grundlage erfolgen, ist es wichtig, dass sie sich auf aussagekräftige Faktoren stützen.[2]
Bei der Identifizierung entscheidungsrelevanter Faktoren kann die Statistik helfen.[3] Diese bedient sich der empirischen Tatsache, dass noch unbekannte Ergebnisse, welche in der Zukunft liegen oder unbeobachtet sind, oft auf Regelmäßigkeiten zurückzuführen sind, welche in vergangenen Beobachtungen gefunden werden können.[4] Auf dieser Grundlage können datengestützte Entscheidungen präziser sein als solche, die auf Intuition oder menschlicher Expertise allein beruhen.[5]

I. Maschinelles Lernen zur Entscheidungsfindung

In den letzten Jahren hat sich die Technologie im Bereich der künstlichen Intelligenz (KI) stark entwickelt. Dieser Artikel befasst sich mit dem Maschinellen Lernen (ML), das sich unter die KI unterordnet. ML-Algorithmen sind im Allgemeinen Computeralgorithmen, die ihre Leistung bei einer bestimmten Aufgabe aufgrund ihrer Erfahrung verbessern.[6] Ein Algorithmus ist eine Handlungsanweisung, die Eingaben erhält und Ausgaben erzeugt[7] und zur Lösung einer Klasse von Problemen verwendet wird. Ein typisches Szenario für maschinelles Lernen ist zum Beispiel das Trainieren eines Algorithmus, der handgeschriebene Ziffern erkennen und in Klassen von 0 bis 9 einteilen soll.[8] Dazu werden Trainingsdaten verwendet, welche aus Bildern mit handgeschriebenen Ziffern und der Angabe der jeweils korrekten Ziffer bestehen. Der Algorithmus sammelt Erfahrung, in dem er Bildern aus den Trainingsdaten (Eingabe) einer Klasse (Ausgabe) zuordnet und anschließend seine Leistung daran gemessen

---

[1] Barocas/Hardt/Narayanan Fairness and Machine Learning - Limitations and Opportunities, 2019, abrufbar unter https://fairmlbook.org/.
[2] Barocas/Selbst California Law Review 2016, 671.
[3] Barocas/Hardt/Narayanan Fairness and Machine Learning - Limitations and Opportunities.
[4] Zhang/Bengio/Hardt/Recht/Vinyals Communications of the ACM, 2021, Vol. 64, Issue 3, 107.
[5] Gates/Perry/Zorn Housing Policy Debate 13 (2002), 369.
[6] Mitchell Machine Learning, 2007.
[7] Cormen/Leiserson/Rivest/Stein Introduction to algorithms, 4. Aufl. 2022.
[8] Mitchell Machine Learning.

wird, ob die handgeschriebenen Ziffern richtig klassifiziert wurden. Während des Trainings versucht der Algorithmus dann, die Anzahl der Fehlklassifizierungen zu minimieren.
ML geht damit einen Schritt weiter als die traditionelle Statistik und zielt darauf ab, dass der Algorithmus in der Menge aller beobachteten Merkmale eigenständig diejenigen identifiziert, die statistisch mit dem Ergebnis zusammenhängen.[9] Dabei können ML-Algorithmen mitunter entscheidungsrelevante Faktoren identifizieren, die Menschen aufgrund der Komplexität oder der Kompliziertheit der Beziehungen in historischen Daten übersehen.[10] Verfahren des maschinellen Lernens werden mittlerweile in vielen Bereichen zur Entscheidungsfindung oder -unterstützung eingesetzt, beispielsweise bei der Kreditvergabe,[11] bei der Bewertung des Rückfallrisikos in der Strafjustiz[12] oder bei Personalentscheidungen[13].
Allerdings birgt maschinelles Lernen auch Risiken, insbesondere wenn es in Entscheidungsprozesse mit weitreichenden Konsequenzen für Menschenleben eingebunden wird. Eine wachsende Zahl von Entscheidungsalgorithmen steht in der Kritik, bestimmte Personengruppen zu benachteiligen. Im Jahr 2020 wurde beispielsweise die Lieferung von Gesichtserkennungssoftware großer IT-Unternehmen an die US-Polizei eingestellt, als Bedenken hinsichtlich rassistischer Profilerstellung aufkamen.[14]

II. Diskriminierung im Maschinellen Lernen

Es ist wichtig zu verstehen, unter welchen Umständen beobachtete Ungleichheiten in automatisierten Entscheidungen oder Vorhersagen in Bezug auf

---

[9] Barocas/Hardt/Narayanan Fairness and Machine Learning - Limitations and Opportunities.
[10] Barocas/Hardt/Narayanan Fairness and Machine Learning - Limitations and Opportunities.
[11] Bartlett/Morse/Stanton/Wallace National Bureau of Economic Research: Consumer-lending discrimination in the fintech era, 2019.
[12] Verma/Rubin 2018 IEEE/ACM International Workshop on Software Faierness (FairWare), 1-7.
[13] Dastin/Reuters 2018, abrufbar unter https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.
[14] Greene The Washington Post 2020, abrufbar unter https://www.washingtonpost.com/technology/2020/06/11/microsoft-facial-recognition/.

verschiedene soziale Gruppen als Diskriminierung betrachtet werden können.[15] Während die philosophische, juristische und soziologische Diskriminierungsforschung in diesem Zusammenhang eine wichtige Rolle spielt, würde eine ausführliche Darstellung den Rahmen dieses Artikels überschreiten. In diesem Artikel wird die folgende Definition verwendet: Diskriminierung liegt vor, wenn die soziale Bedeutung von Merkmalen wie Gender oder Race in einer Weise in Handlungen einfließt, die wir als unrechtmäßig empfinden und die sich von unvernünftigem oder böswilligem Verhalten in Bezug auf ein körperliches Merkmal unterscheidet.[16] Bei der sozialen Bedeutung eines Merkmals handelt es sich um kontextabhängige und instabile soziale Konstrukte.[17] Dieser Artikel verwendet sofern notwendig explizit die englischen Worte Gender und Race, um zu verdeutlichen, dass es sich um soziale Bedeutungen oder Kategorien handelt. Diese Unterscheidung ist insbesondere beim Geschlecht von herausragender Bedeutung, da es in biologisches und soziales Geschlecht unterteilt werden kann.[18] Wenn nicht anders angegeben, bezieht sich der Begriff Minderheit im Folgenden auf eine marginalisierte Gruppe, z.B. Frauen, und nicht notwendigerweise auf zahlenmäßig unterlegene statistische Minderheiten.

Der Begriff Bias wird häufig verwendet, um unerwünschte demographische Unterschiede in algorithmischen Systemen im Zusammenhang mit Diskriminierung durch Algorithmen zu beschreiben.[19] Dies ist ein Begriff, der sowohl in der Statistik als auch im maschinellen Lernen verbreitet ist, dort jedoch eine andere Bedeutung trägt. Bias beschreibt eine statistische Verzerrung, welche auftritt auf, wenn ein statistisches Modell einen erwarteten oder durchschnittlichen Wert hat, der sich vom wahren Wert unterscheidet.[20] Zum Beispiel, wenn die

---

[15] Barocas/Hardt/Narayanan Fairness and Machine Learning - Limitations and Opportunities.

[16] Hu/Kohler-Hausmann Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, S. 513.

[17] Hanna/Denton/Smart/Smith-Loud Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, S. 501-512.

[18] Hu/Kohler-Hausmann Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, S.513.

[19] Barocas/Hardt/Narayanan Fairness and Machine Learning - Limitations and Opportunities. Dazu auch Gössl (in diesem Band), S. 7 f.

[20] Hastie/Tibshirani/Friedmann The Elements of Statistical Learning, 2. Aufl. 2009.

statistisch geschätzte Ankunftszeit eines Lieferdienstes immer um einige Stunden zu früh ist.[21] Um Verwirrungen zu vermeiden, wird in diesem Artikel wird daher auf den Begriff Bias weitestgehend verzichtet.

III. Ziel und Strukturierung des Artikels

Ziel dieses Artikels ist es, einen Überblick über einige häufige Ursachen zu geben, die für die Diskriminierung durch Algorithmen verantwortlich sein können. Kapitel B beschreibt den Prozess der Entwicklung maschineller Lernmodelle und stellt in jeder Phase einige der wichtigsten Probleme vor, die zu Diskriminierung durch Algorithmen führen können. Dabei wird kein Anspruch auf Vollständigkeit gestellt. Dieser Artikel legt den Fokus auf die Problemklasse der Klassifizierung, gleichzeitig sind viele der vorgestellten Herausforderungen auch für andere Problemklassen relevant. Kapitel C gibt einen kurzen Einblick in das Forschungsfeld des Vertrauenswürdigen ML und stellt Entwicklungen im Bereich der Fairness und Erklärbarkeit von KI-Systemen vor. Kapitel D schließt mit einer Zusammenfassung und einem Ausblick.

B. Diskriminierung im Maschinellen Lernen

Dieses Kapitel gibt einen Überblick über vier Schritte bei der Entwicklung von Modellen des maschinellen Lernens, welche in der Regel nacheinander durchlaufen werden, wobei eine Revision eines früheren Schrittes jederzeit möglich ist und häufig auch durchgeführt wird. Am Anfang steht die Definition der Fragestellung oder Problemformulierung (I.). Diese führt über die Datensammlung und -aufbereitung (II.) zur Auswahl des maschinellen Lernmodells und der Trainingsmethode (III.) und endet mit der Anwendung des Algorithmus in der Praxis (IV.). In vielen Fällen gibt es eine Feedbackschleife, in der die in der Praxis gesammelten Erfahrungen genutzt werden, um das Modell erneut zu trainieren und damit zu verbessern. Andere Unterteilungen des Prozesses sind ebenfalls möglich.[22]

---

[21] Barocas/Hardt/Narayanan Fairness and Machine Learning - Limitations and Opportunities.
[22] Lehr/Ohm UCDL Rev. 51 (2017), 653.

I. Problemformulierung

Bei der Problemformulierung wird das zu lösende Problem identifiziert. Diskriminierung im Zusammenhang mit maschinellem Lernen kann in Anwendungen, die direkt oder indirekt mit Menschen zu tun haben, gleichermaßen auftreten. In diesem Artikel liegt der Fokus auf Anwendungen, die Daten über Menschen enthalten, welche tendenziell die bestehenden demografischen Unterschiede widerspiegeln.[23] Es ist jedoch wichtig zu beachten, dass auch Daten ohne explizite Personenbezüge demografische Ungleichheiten kodieren können, wie z.B. eine App, die mit Hilfe von Smartphonesensoren Schlaglöcher erkennen soll.[24] Wenn der Besitz von Smartphones in wohlhabenderen Stadtteilen höher ist als in Stadtteilen mit niedrigem durchschnittlichem Einkommen, kann dies dazu führen, dass Schlaglöcher in wohlhabenderen Stadtteilen besser von einem ML-System erkannt und gemeldet werden.
Bei der Problemdefinition geht es um die Entscheidung, welches Problem gelöst werden soll. Sie basiert auf der Erkenntnis, dass es überhaupt ein Problem gibt. Da die Wahrnehmung von Problemen in starkem Maße von den verschiedenen sozialen Gruppen abhängt, denen die Menschen angehören oder denen sie zugeordnet werden, kann die Wahrnehmung der Probleme, die es zu lösen gilt, sehr unterschiedlich sein. Die Forschungslandschaft ist weltweit und auch in Deutschland nach wie vor sehr homogen. Im Jahr 2011 betrug der globale Anteil der Forschungsausgaben auf dem afrikanischen Kontinent, auf dem fast 18% der Weltbevölkerung leben,[25] nur 0,8%.[26] Das Statistische Bundesamt ermittelte, dass im Jahr 2021 nur 27 % der hauptberuflichen Professuren an deutschen Hochschulen von Frauen besetzt waren.[27]
Ein Mangel an Vielfalt in der Gruppe derjenigen, die Probleme definieren, kann dazu führen, dass Probleme von marginalisierten Gruppen übersehen werden.

---

[23] Barocas/Hardt/Narayanan Fairness and Machine Learning - Limitations and Opportunities.
[24] Crawford Harvard Business Review 2013, 1 (4).
[25] UN-DESA, 2022, abrufbar unter https://de.statista.com/statistik/daten/studie/1347674/umfrage/entwicklungder-bevoelkerung-afrikas-an-der-weltbevoelkerung/.
[26] Beaudry/Mounton The next generation of scientists in Africa, 2018.
[27] Statistisches Bundesamt, 2022, abrufbar unter https://www.destatis.de/DE/Presse/Pressemitteilungen/2022/12/PD22_559_213.html .

Caroline Perez zeigt in ihrem Buch „Unsichtbare Frauen“,[28] dass viele Daten, auf deren Grundlage wirtschaftliche, politische oder medizinische Entscheidungen getroffen werden, ausschließlich Männer betreffen, während Daten von und über andere Genderidentitäten fehlen. Sie nennt diese geschlechtsspezifischen Datenlücke den *Gender Data Gap*. Diese fußt jedoch darauf, dass Probleme von marginalisierten Gruppen nicht richtig erkannt und definiert werden, sodass relevante Daten gesammelt werden konnten.
Um die richtigen Fragen stellen und die notwendigen Daten erheben zu können, müssen marginalisierte Gruppen stärker in Führungspositionen und in der Regierung vertreten sein. In den letzten Jahren haben sich mehrere Affinitätsgruppen gebildet, die sich zum Ziel gesetzt haben, die Präsenz und Sichtbarkeit von Minderheiten in der KI-Forschung zu erhöhen und die Vielfalt in der Forschungs- und Entwicklungsgemeinschaft zu vergrößern (siehe z.B. NeurIPS Affinity Workshops (2022)).[29] Gleichzeitig gibt es Vorschläge für partizipatives Design in der (fairen) ML-Forschung und Entwicklung.[30] Dabei werden Nutzer*innen in den Designprozess einbezogen, um sicherzustellen, dass ein ML-System den Bedürfnissen der Personen entspricht, die von seiner Nutzung betroffen sind.[31]

II. Daten

Die meisten ML-Systeme und Algorithmen sind datengetrieben, d.h. sie benötigen Daten zum Lernen. Ziel Maschineller Lernverfahren ist es, Muster und Regelmäßigkeiten aus Daten zu lernen.[32] Wenn die zugrunde liegenden Trainingsdaten demographische Verzerrungen enthalten, lernen die darauf trainierten Algorithmen diese Verzerrungen mit ihren Vorhersagen zu perpetuieren.[33]

---

[28] Perez Invisible Women: Data Bias in a World Designed for Men, 2019.

[29] abrufbar unter https://blog.neurips.cc/2022/08/23/announcing-the-neurips-2022-affinity-workshops/.

[30] Weinberg Journal of Artificial Intelligence Research 74 (2022), 75-109.

[31] Irani/Vertesi/Dourish/Philip/Grinter Proceedings of the SIGCHI conference on human factors in computing systems, 2010, 1311-1320.

[32] Bishop/Nasrabadi Pattern recognition and machine learning, 4. Ausg. 2006.

[33] Makhlouf/Zhioua/Palamidessi Survey on causal-based machin learning fairness nations, 2020, arXiv:2010.09553; Chu/Ilyas/Krishnan/Wang in ‘Proceedings of the 2016 international

Um von Daten zu lernen, ist es zunächst notwendig, Daten zu erheben. Dazu gehört die Identifikation der Zielgruppe (von Menschen oder Dingen) sowie der Merkmale, die für die Problemlösung notwendig und verfügbar sind.[34] Dann wird entschieden wie die Daten zu sammeln sind. Da es in der Regel nicht möglich ist, die Gesamtheit der Zielgruppe in Daten zu erfassen, wird oft eine Stichprobe davon genommen.[35] Auf die Datenerhebung folgt der Datensäuberungsprozess, welcher Fehler in den Daten erkennt und reparieret.[36]

1. Tabellarische Daten

Dieser Artikel konzentriert sich auf tabellarische Daten, in denen eine Person durch einen Vektor von Merkmalen repräsentiert wird.[37] Dabei wird unterschieden zwischen geschützten Merkmalen, aufgrund derer eine Diskriminierung im vorliegenden Problem moralisch verwerflich und/oder rechtlich verboten ist, und ungeschützten Merkmalen, die zur Entscheidungsfindung herangezogen werden können.[38] Welche Merkmale als geschützt gelten, hängt von der praktischen Anwendung ab. In der Literatur wird bisher meist auf rechtlich geschützte Merkmale wie Race oder Gender (vgl. Art. 3 III GG) abgestellt.[39] Das Label ist eine weitere Information über eine Person, die benötigt wird, um einen

---

conference on management of data', 2016, 2201-2206; Tae/Roh/Oh/Kim/Whang in 'Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning', 2019, 1-4.

[34] Association for Computing Machinery/Suresh/Guttag Equity and access in algorithms, mechanisms, and optimization, 2021, 1–9.

[35] Association for Computing Machinery/Suresh/Guttag Equity and access in algorithms, mechanisms, and optimization, 1–9.

[36] Association for Computing Machinery/Suresh/Guttag Equity and access in algorithms, mechanisms, and optimization, 1–9.

[37] Gorishniy/Rubachev/Khrulkov/Babenko Advances in Neural Information Processing Systems 2021, 18932-18943.

[38] Hardt How big data is unfair, abrufbar unter https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de; Zafar/Valera/Gomez Rodriguez/Gummadi Proceedings of the 16th International Conference on World Wide Web, 2017, 1171-1180; Kilbertus/ Gascon/ Kusner/ Veale/ Gummadi/Weller International Conference on Machine Learning, PMLR 80, 2018, 2630-2639.

[39] Hanna/Denton/Smart/Smith-Loud Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 501-512.

Klassifikationsalgorithmus zu trainieren. Das Label ist die vorherzusagende Klasse und könnte z.B. in der automatisierten Kreditvergabe die Rückzahlungsinformation sein.
Wenn Daten verzerrt sind oder Informationen fehlen, können Algorithmen verzerrte Ergebnisse liefern, die im Zweifelsfall bestimmte Gruppen von Menschen diskriminieren. Algorithmen können bestehende Verzerrungen in den Daten nicht nur aufrechterhalten, sondern sogar verstärken. Es gibt verschieden Arten von Problemen in Daten. Im Folgenden werden einige der gängigsten Probleme vorgestellt. Es gibt zahlreiche Literatur mit umfassenden Übersichten.[40]

2. Verzerrte Merkmale

Erhobene Merkmale können für marginalisierte Gruppen weniger aussagekräftig sein als für die Mehrheit. Ist dies der Fall, kann eine Schätzung mit den gesammelten Merkmalen für marginalisierte Gruppen weniger akkurat sein als für die Mehrheit. Die Aussagekraft kann auf Grund von Messverzerrungen abnehmen. Diese liegen vor, wenn Merkmale nicht in der gleichen Art und Weise für alle Untergruppen gemessen werden.[41]
Beispiel.[42] Im COMPAS Datensatz[43] wurden zur Messung der „Kriminalität" eines Individuums unter anderem Verhaftungen von dessen Freunden oder Familienmitgliedern herangezogen. Minderheiten werden jedoch oft häufiger kontrolliert, was zu höheren Verhaftungsraten führt. Ein Messfehler liegt vor, wenn unterschiedlichen Gruppen gleiche Kriminalitätsraten aber unterschiedliche Verhaftungsraten haben können. Dann ist das Merkmal „Verhaftungsraten" für unterschiedliche Gruppen unterschiedlich aussagekräftig sein.

---

[40] Makhlouf/Zhioua/Palamidessi Survey on causal-based machine learning fairness nations; Association for Computing Machinery/Suresh/Guttag, Equity and access in algorithms, mechanisms, and optimization, 1–9; Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, 2023.
[41] Suresh/Guttag Equity and access in algorithms, mechanisms, and optimization, 2021, 1–9.
[42] Suresh/Guttag Equity and access in algorithms, mechanisms, and optimization, 2021, 1–9.
[43] Larson/Atlidakis/Roswell Probublica compas analysis and dataset, abrufbar unter https://github.com/propublica/compas-analysis.

3. Repräsentative Merkmale

Neben der Identifikation welche Merkmale gesammelt werden sollten, ist es auch wichtig, wie Daten gesammelt werden. Da es in der Regel nicht möglich ist, die gesamte Zielgruppe zu erfassen, wird üblicherweise eine Stichprobe genommen.[44] Dabei ist es essenziell, dass die Stichprobe zufällig und aus dem gleichen Umfeld gezogen wird, in dem das ML-Modell in der Praxis eingesetzt wird.[45]
Beispiel.[46] ImageNet[47] ist ein Datensatz mit Fotos, dem es an geografischer Vielfalt mangelt.[48] Dies führt zu einer nachweisbaren Verzerrung des ML-Algorithmus, welcher dann für westliche Kulturen gut funktioniert, aber für andere nicht.

4. Datenblätter für erhöhte Transparenz

Anstatt den Prozess der Datenerhebung zu durchlaufen, verwenden Entwickler von ML-Systemen häufig bereits vorhandene Datensätze.[49] Um die Kommunikation zwischen den Erstellern von Datensätzen und den Nutzer*innen zu verbessern und die Transparenz und Verantwortlichkeit bei der ML-Entwicklung zu erhöhen, wurde von der Literatur vorgeschlagen, mit jedem öffentlichen Datensatz und vortrainierten Modellen Datenblätter zu veröffentlichen.[50] Ein Datenblatt ist ein Dokument, das Informationen darüber liefert, wie ein Datensatz erstellt wurde und welche Merkmale, Motivationen und potenziellen Verzerrungen er repräsentiert.[51]

---

[44] Suresh/Guttag Equity and access in algorithms, mechanisms, and optimization, 1–9.
[45] Varshney. XRDS Crossroads 25/3 (2019), 26-29.
[46] Makhlouf/Zhioua/Palamidessi Survey on causal-based machine learning fairness nations.
[47] Deng/Dong/Socher/Li/Fei-Fei 2009 IEEE conference on computer vision and pattern recognition, 248-255.
[48] Denton/Hanna/Amironesei/Smart/Nicole On the genealogy of machine learning datasets: A critical history of imagenet, Big Data & Society, Vol. 8 Issue 2; Makhlouf/Zhioua/Palamidessi, Survey on causal-based machine learning fairness nations.
[49] Suresh/Guttag Equity and access in algorithms, mechanisms, and optimization, 2021, 1–9.
[50] Gebru/Morgenstern/Vecchione/Vaughan/Wallach/Iii/Crawford Communications of the ACM, Vol. 64 Issue 12, 2021, 86-92.
[51] Gebru/Morgenstern/Vecchione/Vaughan/Wallach/Iii/Crawford Communications of the ACM, Vol. 64 Issue 12, 86-92.

III. Training

Auf die Datenerhebung folgt die Modellauswahl und das Training des ML-Modells mit den Daten. Im Trainingsschritt müssen unter anderem zwei wichtige Entscheidungen getroffen werden: mit welchem Ziel trainiert werden soll und wie die Daten für das Training verwendet werden sollen. In der Regel werden Modelle trainiert, um eine bestimmte Zielfunktion zu optimieren[52]. Bei der Verwendung der Daten stellt sich z.B. die Frage, ob und wie geschützte Attribute für das Training verwendet werden dürfen. Beispielsweise könnte man geschützte Attribute verschlüsseln, d.h. die sensiblen Attribute unlesbar machen.[53] In der Regel werden die Daten zunächst in Trainings-, Test- und Validierungsdaten aufgeteilt. Das Training wird nur auf den Trainingsdaten durchgeführt. Auf den Validierungs-daten werden dann verschiedene Modellparameter, Zielfunktionen und Optimierungsmethoden getestet und verglichen, um die beste Konfiguration auszuwählen.[54] Am Ende wird auf den Testdaten die Güte des Modells bestimmt.

1. Ungleicher Stichprobenumfang

Im Allgemeinen verbessert sich ein Klassifikationsmodell mit der Anzahl der Datenpunkte, die für das Training verwendet werden und weniger Daten führen zu schlechteren Vorhersagen. Statistische Minderheiten verfügen per Definition immer über relativ weniger Daten als statistische Mehrheiten. Das heißt, wenn das Klassifizierungsmodell auf einer repräsentativen Zufallsstichprobe der Population trainiert wird, ist das Modell für statistische Minderheiten in der Regel schlechter als für statistische Mehrheiten.[55]
Es gibt Möglichkeiten, das Problem ungleicher Stichprobengrößen anzugehen und Diskriminierung zu verringern. Beispielsweise könnten relativ mehr Daten von Minderheiten erhoben werden oder die Zielfunktion könnte durch eine hö-

---

[52] Suresh/Guttag Equity and access in algorithms, mechanisms, and optimization, 2021, 1–9.
[53] Kilbertus/Gascon/Kusner/Veale/Gummadi/Weller International Conference on Machine Learning, PMLR 80, 2630-2639.
[54] Suresh/Guttag Equity and access in algorithms, mechanisms, and optimization, 2021, 1–9.
[55] Hardt How big data is unfair, abrufbar unter https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de.

here Gewichtung des Trainingsfehlers auf die Minderheit angepasst werden (*reweighting*) oder Minderheitsdaten relativ häufiger zum Training genutzt werden (*resampling*).[56]

2. Proxymerkmale

Die Vorhersagen oder Entscheidungen eines Algorithmus sind nicht vor Diskriminierung geschützt, wenn die Trainingsdaten keine geschützten Merkmale enthalten. Das Problem sind Proxy-Merkmale, die Informationen über die sensiblen Merkmale enthalten. In einigen Teilen der Welt kann beispielsweise die Postleitzahl einen Hinweis auf Race geben.[57] Ein Problem in der Praxis ist, dass in den verfügbaren Daten häufig fast jedes Merkmal zumindest teilweise mit dem Status einer geschützten Gruppe korreliert.[58] Proxydiskriminierung wird auch in der Rechtswissenschaft definiert und diskutiert.[59] Allerdings ist umstritten, inwieweit solche juristischen Definitionen auf die Bewertung algorithmischer Systeme übertragbar sind.[60] Um Proxy-Diskriminierung zu vermeiden, kann beispielsweise ein fairer ML-Algorithmus verwendet werden. Die Mehrzahl der vorgeschlagenen Algorithmen benötigt dabei Zugriff auf das geschützte

---

[56] Li/Vasconcelos Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, 9572-9581; Kamiran/Calders Knowledge and informations systems, Volume 33, Issue 1, 1-33.

[57] Pope/Sydnor American Economic Journal: Economic Policy, Volume 3, Issue 3, 206-231.

[58] Corbett-Davies/Goel The measure and mismeasure of fairness: A critical review of fair machine learning, 797-806; Chiappa Path-specific counterfactual fairness, Proceedings of the AAAI Conference on Artificial Intelligence, Volume 3.

[59] Johnson/Martinez UC davis L. rev., Volume 33, S. 1227; Alexander/Cole Constitutional Commentary, Vol. 14, No. 3, 453-463; Prince/Schwarcz Iowa L. Rev., Vol. 105, Issue 1257; s. auch Gössl (in diesem Band), 13 und zu weitergehenden Problemen Kirchhefer-Lauber (in diesem Band), 92 ff.

[60] Corbett-Davies/Goel The measure and mismeasure of fairness: A critical review of fair machine learning, 797-806.

Merkmal[61] Einige Ansätze kommen auch mit begrenzter geschützter Information aus.[62]

IV. Einsatz

Nachdem ein Modell fertig trainiert ist, kann es in der Praxis eingesetzt werden. Auch hier gibt es Möglichkeiten, wie es zu Diskriminierung durch automatisierte Entscheidungen kommen kann, da Menschen entscheiden, ob, wo und wie ein ML-Algorithmus eingesetzt wird.

1. Feedbackschleifen

Algorithmen beeinflussen die Entscheidungen der Nutzer*innen und können so zu verzerrten Daten für das Training zukünftiger Algorithmen führen, insbesondere wenn die Schätzungen des Algorithmus diskriminierend sind. Ein Beispiel hierfür ist eine Suchmaschine, die Ergebnisse von nicht-marginalisierten Gruppen an die Spitze ihrer Liste setzt, was die Popularität dieser Ergebnisse erhöht.[63] Nutzer*innen interagieren in der Regel mit den obersten Ergebnissen.[64] Angenommen, die Interaktionen der Nutzer*innen werden von der Suchmaschine gesammelt, um zukünftige Entscheidungen darüber zu treffen, wie

---

[61] Hardt/Price/Srebro Advances in neural information processing systems, Vol. 29; Zafar/Valera/Gomez Rodriguez/Gummadi Proceedings of the 16th International Conference on World Wide Web, 1171-1180; Kilbertus/Gascon/Kusner/Veale/Gummadi/Weller International Conference on Machine Learning, PMLR 80, 2630-2639; Rateike/Majumdar/Mineeva/Gummadi/Valera 2022 ACM Conference on Fairness, Accountability, and Transparency.

[62] Dai/Wang Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 2021, 680-688; Veale/Binns Big Data & Society Issue 4 Vol. 2, 2053951717743530; Yan/Kao/Ferrara Proceedings of the 29thACM International Conference on Information & Knowledge Management, 2020, 1715- 1724; Kilbertus/Gascon/Kusner/Veale/Gummadi/Weller International Conference on Machine Learning, PMLR 80, 2630-2639; Zhao/Dai/Shu/Wang Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, 2022, 1433–1442.

[63] Mehrabi/Morstatter/Saxena/Lerman/Galstyan, ACM Computing Surveys (CSUR), Vol. 54, Issue 6, 1-35.

[64] Lerman/Hogg, PloS one, Issue 9 Vol. 6, e98914.

Informationen präsentiert werden sollen. Dies kann dazu führen, dass die Ergebnisse an der Spitze immer beliebter werden, und zwar nicht aufgrund der Qualität der Ergebnisse, sondern aufgrund der Platzierung durch den Algorithmus.[65]

2. Mensch-Maschine-Interaktion

Wird ein System nicht vollautomatisch eingesetzt, sind in der Regel menschliche Entscheidungsträger*innen involviert (*human in the loop*), die eine algorithmische Vorhersage in ihre eigene Entscheidung einbeziehen können. Doch bei der Interpretation und Einordnung der algorithmischen Vorhersagen können Korrelations- und Bestätigungsfehler auftreten.[66] Unter einem Korrelationsfehler versteht man den Irrtum, aufgrund beobachteter Korrelationen auf eine Kausalität zu schließen. Die meisten ML-Algorithmen lernen Korrelationen, erst in den letzten Jahren hat Kausalität im ML mehr Aufmerksamkeit bekommen,[67] auch in der Forschung zu vertrauenswürdigem ML.[68] Der Bestätigungsfehler beschreibt die menschliche Tendenz, Informationen so zu suchen oder zu interpretieren, dass sie mit den eigenen Vorannahmen übereinstimmen.[69] Zudem neigen Menschen dazu, automatisierten Hilfsmitteln mehr Autorität zuzuschreiben als anderen Informationsquellen (Automatisierungsfehler), was zu Entscheidungen führen kann, die nicht auf einer gründlichen Analyse aller verfügbaren Informationen beruhen.[70]

---

[65] Lerman/Hogg PloS one, Issue 9 Vol. 6, e98914.

[66] Srinivasan/Chander Communications of the ACM, Colume 64, Number 8, 44-49.

[67] Schölkopf Probabilistic and Causal Inference: The Works of Judea Pearl, 765-804.

[68] Madras/Creager/Pitassi/Zemel Proceedings of the conference on fairness, accountability, and transparency, 349-358; Schrouff/Dieng/Rateike/Kwegyir-Aggrey/Farnadi Algorithmic Fairness through the Lens of Causality and Robustness workshop, 1-5; Kusner/Loftus/Russell/Silva Advances in Neural Information Processing Systems (NeurIPS) Vol. 30; Creager/Madras/Pitassi/Zemel International Conference on Machine Learning, 2185-2195; Mhasawqade/Chunara Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics and Society, 784-794; Nilforoshan/Gaebler/Shroff/Goel International Conference on Machine Learning, 16848-16887.

[69] Srinivasan/Chander Commuications of the ACM, Issue 64 Vol. 8, 44-49.

[70] Parasuraman/Manzey Human factors Issue 52, Vol. 3, 381-410.

Eine Gegenmaßnahme können regelmäßige Schulungen über die Funktionsweise und die Grenzen algorithmischer Vorhersagen sowie die damit verbundenen Diskriminierungsrisiken sein. Auch ein zusätzliches Mehraugenprinzip beim Einsatz solcher Systeme kann je nach Anwendungsfall sinnvoll sein.

C. Faires und Erklärbares Maschinelles Lernen

Das vorangegangene Kapitel hat einen Überblick über eine kleine Auswahl von Möglichkeiten gegeben, wie Diskriminierung in jeder Phase des ML-Entwicklungsprozesses auftreten kann, und hat einige Maßnahmen vorgestellt, um dem entgegenzuwirken. Dieses Kapitel stellt kurz zwei Teilgebiete des Forschungsfeldes des vertrauenswürdigen ML vor: Fairness und Erklärbarkeit.

I. Faires Maschinelles Lernen

Stereotype können sich auch ohne explizite Diskriminierung verfestigen und in der Gesellschaft lange Zeit bestehen bleiben.[71] Bei der Entwicklung von ML-Systemen, die Menschen betreffen, stellt sich die zentrale Frage nach dem Ziel der algorithmischen Vorhersage oder Entscheidung. Die verfügbaren Trainingsdaten spiegeln häufig demografische Ungleichheiten in unserer Gesellschaft wider, wie z.B. ein starkes Ungleichgewicht zwischen den Geschlechtern in vielen Berufen.[72] Wenn wir Maschinelles Lernen in die Entscheidungsfindung einbeziehen, ist es wichtig, sich die Frage zu stellen, ob die algorithmische Vorhersage lediglich die vorhandenen Daten widerspiegeln soll oder ob die Daten hinterfragt werden sollen, um Entscheidungen zu treffen, die einer bestimmten Vorstellung von normativ richtigem Verhalten entsprechen, auch und gerade, wenn es um die Perpetuierung von Stereotypen geht.[73]

1. Definitionen von Fairness

Die Fairnessforschung im ML hat eine Vielzahl von philosophischen Fairnessdefinitionen mathematisch beschrieben. Im Folgenden werden drei Arten

---

[71] Barocas/Hardt/Narayanan Fairness and Machine Learning - Limitations and Opportunities.
[72] Barocas/Hardt/Narayanan Fairness and Machine Learning - Limitations and Opportunities.
[73] Barocas/Hardt/Narayanan Fairness and Machine Learning - Limitations and Opportunities.

von Fairness für Klassifikationsprobleme kurz vorgestellt. Für einen detaillierten Überblick sei auf die Literatur verwiesen.[74]

- *Individuelle Fairness.* Formulierungen individueller Fairness drücken das Verständnis aus, dass Personen, die sich ähnlich sind, auch ähnliche Vorhersagen von einem Algorithmus erhalten sollten.[75] Eine Herausforderung besteht darin, ein Abstandsmaß zu definieren, das die Ähnlichkeit zwischen Individuen definiert. Dieses Maß kann von Problem zu Problem variieren.
- *Gruppenspezifische Fairness.* Definitionen der gruppenspezifischen Fairness[76] verlangen, dass die Vorhersagen zwischen geschützten und nicht geschützten Gruppen (oder Untergruppen derer) im Durchschnitt gleich sind. Beispielsweise verlangt die statistische Parität, dass die demographische Verteilung derjenigen, die eine positive (oder negative) Klassifizierung erhalten, mit der Demographie der Gesamtbevölkerung identisch ist.[77]
- *Kausale Fairness.* Die Forschung hat in den letzten Jahren eine Reihe von Kriterien für kausale Fairness vorgeschlagen.[78] Viele davon verbieten, dass geschützte Merkmale algorithmische Vorhersagen kausal beeinflussen.[79]

---

[74] Verma/Rubin 2018 IEEE/ACM International Workshop on Software Faierness (FairWare), 1-7; Mitchell/Potash/Barocas/D'Amour/Lum Annual Review of Statistis and Its Applications, 2021, Vol. 8, 141-163; Mehrabi/Morstatter/Saxena/Lerman/Galstyan ACM Compuling Surveys, Issue 54 Vol. 6, 1-35.

[75] Dwork/Hardt/Pitassi/Reingold/Zemel Proceedings of the 3rd innovations in theoretical computer science conference, 214-226.

[76] Dwork/Hardt/Pitassi/Reingold/Zemel Proceedings of the 3rd innovations in theoretical computer science conference, 214-226; Hardt/Price/Srebro Advances in neural information processing systems, Vol. 29; Zafar/Valera/Gomez Rodriguez/Gummadi Proceedings of the 16th International Conference on World Wide Web, 1171-1180.

[77] Dwork/Hardt/Pitassi/Reingold/Zemel Proceedings of the 3rd innovations in theoretical computer science conference, 214-226.

[78] Kusner/Loftus/Russell/Silva Advances in Neural Information Processing Systems (NeurIPS) Vol. 30; Chiappa Proceedings of the AAAI Conference on Artificial Intelligence Vol. 33.

[79] Nilforoshan/Gaebler/Shroff/Goel International Conference on Machine Learning, 16848-16887.

Eine der am weitesten verbreiteten Definitionen besagt, dass eine Entscheidung in Bezug auf eine Person kontrafaktisch fair (*counterfactually fair*) ist, wenn sie sowohl in der realen Welt als auch in einer kontrafaktischen Welt, in der die Person einer anderen demographischen Gruppe angehört, gleich ist.[80] Während diese Definition auch als Kodifizierung des Rechtsbegriffs der Ungleichbehandlung angesehen wird,[81] stellt das Lernen von Kausalzusammenhängen in der Praxis eine Herausforderung dar.[82]

2. Methoden zur Reduzierung von Diskriminierung

Verschiedene Fairnessziele können oft nicht gleichzeitig erreicht werden, sei es aus mathematischen[83] oder philosophischen[84] Gründen. Darüber hinaus können Entscheidungsträger*innen mehrere widersprüchliche Ziele haben, die Fairness und andere Werte einschließen und möglicherweise eine Abwägung erfordern.[85] Auch in dieser Hinsicht hat die Forschung eine Reihe von Methoden entwickelt, die dazu beitragen können, ein Gleichgewicht zwischen Fairness und anderen Zielen zu finden und Diskriminierung von ML-Systemen zu verringern.
Die Literatur unterscheidet zwischen drei verschiedene Techniken.[86] Bei der Entwicklung eines fairen ML-Systems wird dann die Technik gewählt, die zur Problematik am besten passt. Bei der fairen Vorverarbeitung (*pre-processing*) werden die Daten so bearbeitet, dass sie mit dem geschützten Attribut möglichst

---

[80] Kusner/Loftus/Russell/Silva Advances in Neural Information Processing Systems (NeurIPS) Vol. 30.
[81] Nilforoshan/Gaebler/Shroff/Goel International Conference on Machine Learning, 16848-16887.
[82] Peters/Janzing/Schölkopf Elements of causal inference: Foundations and learning Algorithms.
[83] Kleinberg Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems, S. 40.
[84] Heidari/Loi/Gummadi/Krause Proceedings of the conference on fairness, accountability, and transparency, 181-190.
[85] Barocas/Hardt/Narayanan Fairness and Machine Learning - Limitations and Opportunities.
[86] Barocas/Hardt/Narayanan Fairness and Machine Learning - Limitations and Opportunities.

unkorreliert sind.[87] Bei einem fairen Training (*in-training*), wird der Optimierungsprozess adaptiert.[88] Eine von vielen Möglichkeiten, Diskriminierung in automatisierten Entscheidungen während des Trainings oder der Modellerstellung zu reduzieren, ist ein Entscheidungsmodell für je eine soziale Gruppe zu trainieren.[89] Dazu muss es normativ gerechtfertigt sein die beiden Gruppen unterschiedlichen Entscheidungsregeln zu unterziehen.[90] Bei der Nachbearbeitung (*post-processing*) werden die Vorhersagen eines bereits fertig trainierten Klassifikators im Nachhinein so angepasst, dass sie mit dem geschützten Attribut unkorreliert sind.[91] Die drei Ansätze haben unterschiedliche Stärken und Schwächen.[92] Allen gemeinsam ist in der Regel, dass die Trainingsdaten in der Regel in soziale Gruppen unterteilbar sein müssen und dass ein Zugang zu dem geschützten Merkmal bestehen muss. Diese Annahme kollidiert manchmal mit Datenschutzbestimmungen.[93] Zuletzt ist die Definition sozialer Gruppen oft unscharf und daher nicht eindeutig bestimmbar.[94]

II. Erklärbarkeit von Algorithmen

Vorhersagen mit ML-Algorithmen können zwar präzise sein, sind aber oft undurchsichtig für den Menschen. Dies gilt insbesondere für so genannte Deep-Learning-Modelle, deren Vorhersage von einer Vielzahl von Gewichtungen und Parametern abhängt. Besonders in Bereichen wie Gesundheit und Finanzen ist es wichtig, die Entscheidungsfindung nachvollziehbar und verständlich zu machen,[95] also eine Erklärung zu geben. Eine Erklärung ist die Antwort auf eine

---

[87] Barocas/Hardt/Narayanan Fairness and Machine Learning - Limitations and Opportunities.
[88] Barocas/Hardt/Narayanan Fairness and Machine Learning - Limitations and Opportunities.
[89] Corbett-Davies/Pierson/Feller/Goel/Huq Proceedings of the 23rd ACM SIGKDD international conference on knowledge, discovery and data mining, 797-806.
[90] Hardt How big data is unfair.
[91] Barocas/Hardt/Narayanan Fairness and Machine Learning - Limitations and Opportunities.
[92] Barocas/Hardt/Narayanan Fairness and Machine Learning - Limitations and Opportunities.
[93] Veale/Binns Big Data & Society Issue 4 Vol. 2, 2053951717743530.
[94] Hardt How big data is unfair; Hu/Kohler-Hausmann Proceedings of the 2020 Conference on Fairness, Accountability and Transparency, S. 513.
[95] Burkart/Huber Journal of Artificial Intelligence Research Issue 70, 245-317.

Warum-Frage[96] und ist dann relevant, wenn sie auf die aktuellen Ziele und Bedürfnisse der Nutzer eingeht.[97] Im Folgenden wird ein ausgewählter Erklärungsansatz vorgestellt: Recourse.[98] Mangels einer geeigneten Übersetzung wird der englische Begriff verwendet.

Recourse. Recourse bietet Personen, die eine negative Entscheidung eines ML-Systems erhalten haben, Erklärungen und Empfehlungen, um eine positive Entscheidung zu erwirken.[99] In Bereichen wie Gesundheit und Finanzen wird dies sogar als rechtliche Notwendigkeit angesehen.[100] Recourse gibt nicht nur eine Erklärung für das „Warum?", sondern auch eine Antwort auf die Frage „Welche Handlungen führen zu positiven Entscheidungen?" und schlägt sogenannte „minimale folgenreiche Empfehlungen" vor.[101] Um sicherzustellen, dass Individuen, die eine Recourse-Empfehlung umsetzen, auch eine positive Entscheidung des ML-Algorithmus erhalten, ist es notwendig, die kausalen Zusammenhänge zwischen den Merkmalen zu verstehen.[102]

Handlungsempfehlungen können nur vorgeschlagen werden, wenn Individuen in der Lage sind, durch ihr eigenes Handeln eine positive Entscheidung herbeizuführen und ihre Eigenschaften zu verändern. Wenn jedoch solche Handlungsmöglichkeiten nicht existieren, sind Individuen dem Algorithmus ausgeliefert und haben keinen Einfluss auf Entscheidungen und deren Konsequenzen.[103] Basiert eine Entscheidung beispielsweise allein auf biologischem Alter oder Körpergröße, hat das Individuum keine Möglichkeit, aktiv Einfluss zu nehmen. Die Möglichkeit, eine positive Entscheidung durch eigenes Zutun zu erwirken (Recourse) ist daher auch ein wichtiges Maß für die Fairness eines automatisierten Entscheidungssystems.

---

[96] Miller Artificial Intelligence Issue 267, 1-38.

[97] Burkart/Huber Journal of Artificial Intelligence Research Issue 70, 245-317.

[98] Karimi/Barthe/Schölkopf/Valera ACM Computing Surveys, Issue 55, Vol. 5, 1-29; Karimi/Schölkopf/Valera Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp- 353-362.

[99] Karimi/Barthe/Schölkopf/Valera ACM Computing Surveys, Issue 55, Vol. 5, 1-29.

[100] Karimi/Barthe/Schölkopf/Valera ACM Computing Surveys, Issue 55, Vol. 5, 1-29.

[101] Karimi/Barthe/Schölkopf/Valera ACM Computing Surveys, Issue 55, Vol. 5, 1-29.

[102] Karimi/Von Kügelgen/Schölkopf/Valera Advances in Neural Information Processing Systems, Vol. 33, 265-277.

[103] Venkatasubramanian/Alfano Proceedings of the 2020 conference on fairness, accountability, and transparency, 284-293.

D. Zusammenfassung

Dieser Artikel befasst sich mit Problemen der Diskriminierung und der Nichtberücksichtigung marginalisierter Gruppen bei der Entwicklung von ML-Algorithmen. Diese Probleme können in verschiedenen Phasen der Entwicklung auftreten und verstärken sich gegenseitig. Dabei gibt es Maßnahmen, um diesen Problemen entgegenzuwirken, wie die Zusammenstellung diverser Teams, die sorgfältige Definition und Dokumentation des Datensammlungsprozesses und die Schulung der Anwender*innen algorithmischer Entscheidungshilfen. Die Forschung zu vertrauenswürdigen ML-Systemen hat verschiedene Definitionen von Fairness und Erklärbarkeit hervorgebracht, sowie Techniken entwickelt, um die Diskriminierung und Erklärbarkeit von ML-Systemen zu verbessern. Dabei wurde in den letzten Jahren wurde immer wieder auf die Notwendigkeit interdisziplinärer Forschung hingewiesen.[104] Insbesondere an der Schnittstelle zwischen Rechtswissenschaft und Informatik ist dies essenziell, um praxistaugliche Werkzeuge und Methoden für unterschiedliche Rechtsordnungen zu schaffen.[105] Es gibt immer mehr Arbeiten, die sich mit KI und den damit verbundenen Anforderungen des EU-Rechts befassen.[106] Zur rechtlichen und normativen Einordnung der in diesem Beitrag aufgeworfenen Probleme und Fragen besteht weiterer Forschungsbedarf.

Danksagung
Vielen Dank an Jonas Klesen für das Korrekturlesen und die hilfreichen Anmerkungen.

---

[104] Heidari/Nanda/Gummadi International Conference on Machine Learning PMLR, 2692-2701; Barocas/Hardt/Narayanan Fairness and Machine Learning - Limitations and Opportunities.

[105] Wachter/Mittelstadt/Russell W. Va. L. Rev. Issue 123, S. 735.

[106] Wachter/Mittelstadt/Russell W. Va. L. Rev. Issue 123, S. 735; Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, 2023.

Kapitel 3

Künstliche Intelligenz und personale Autonomie: Diskriminierende Algorithmen als ethische und rechtliche Herausforderung für die Polizeiarbeit

Caja Thimm und Laura Thimm-Braun

A. Einleitung

Wie nur wenige Technologien hat die Datafizierung und Algorithmisierung unseren Alltag verändert. Dies gilt nicht nur für sehr technikaffine Gruppierungen, sondern inzwischen für fast alle Menschen. Angefangen beim Erwerb einer Fahrkarte, über den Onlineeinkauf oder bis hin zu neuen Robotern, die sich als ‚smart robots' nunmehr sogar als Sozialpartner anbieten, scheinen immer nur wenige Momente zwischen den diversen technologischen Sprüngen zu liegen. Das Erscheinen von LLMs, also ‚large lange models', die fast menschenähnlich interagieren und seit ihrem Erscheinen Ende 2022 in Form von ChatGPT ein neues Zeitalter der Datenanalyse eingeläutet haben, markiert den Beginn einer weiteren Durchdringung des Alltages mit Künstlicher Intelligenz (KI).
Dass diese Entwicklung sich über kürzer oder länger in vielen gesellschaftlichen Sektoren und manifestiert und insofern sowohl Chancen als auch Risiken mit sich bringt, ist unbestritten. Je weiter aber die technologischen Entwicklungen in den menschlichen Alltag hineinreichen und auf menschliches Handeln Einfluss nehmen, desto lauter wird der Ruf nach ethischen Rahmenbedingungen und rechtlichen Regulierungsverfahren.[1] Besonders deutlich wurde in den letz-

[1] S. Stellungnahme des Deutschen Ethikrats, Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz, 2023; UNESCO Recommendation on the Ethics of Artificial Intelligence, 2021.

ten Jahren die Kritik an den Datensätzen für die algorithmische Datenverarbeitung.[2] Ausschlaggebend für die Ergebnisse algorithmischer Rechenprozesse sind die Ziele und die einbezogenen Datenbasen und dies ist nicht, wie häufig attribuiert, maschinelles Tun, sondern wird durch die menschlichen Entscheidungen über technische Einsatzzwecke und Datenselektion bestimmt. Wie stark sich Herkunft und Qualität von Daten auf soziale Prozesse, politische Entscheidungen und letztlich auch auf das Autonomie- und Gerechtigkeitsempfinden von Bürger*innen auswirken können, lässt sich anschaulich am Beispiel von „Predictive Policing" bzw. ‚vorausschauender Polizeiarbeit' verdeutlichen.
In diesem Beitrag konzentrieren wir uns dabei auf die Frage, welche ethischen und rechtlichen Rahmenbedingungen für einen verantwortungsvollen Umgang mit diesen Technologien erforderlich erscheinen. Gefragt wird weniger nach den praktischen Umsetzungen der Technologie, als vielmehr nach den möglichen Problemfeldern, die der Nutzung solcher prädiktiver Analyseverfahren inhärent sind. Vor dem Hintergrund diverser Forschungsbefunde darüber, dass rassistische und sexistische Diskriminierung sowie Intransparenz der Prozesse sich auf die Polizeiarbeit auswirken können, hat sich um Predictive Policing eine polarisierende Debatte entwickelt.[3] Insofern kommt dieser Debatte eine über den engeren Nutzungsbereich hinausgehende gesellschaftliche Rolle und Akzeptanz von KI zu. Die Geschichte technischer Entwicklungen hat vielfältig gezeigt, dass der Mensch und seine Institutionen mit Technologien nicht Schritt halten können. Aktuell sind Mensch und Gesellschaft noch nicht darauf vorbereitet, mit einer Technologie des 21. Jahrhunderts umzugehen, die eine bisher nicht abschätzbare Wirkungsmacht hat. Schon bald werden neue Generationen von KI-Anwendungen zum alltäglichen Standard gehören. Unter dieser Prämisse erscheint es notwendig, auch technische Entwicklungen in der Po-

---

[2] Floridi/Taddeo Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2016, 374 (2083); Mittelstadt/Allo/Taddeo/Wachter/Floridi Big Data & Society 2020, 1.

[3] Degeling/Berendt AI & Society, 2018, 347; Lum/William Significance, 13/5 (2016), 14; Richardson/Schultz/Crawford New York University Law Review Online 2019, 192.

lizeiarbeit immer neu kritisch zu reflektieren, um eine Technologie der „automatisierten Ungleichheit" zu verhindern [4]. Dazu gehört zunächst eine ethische Perspektivierung der Rolle von Daten, deren Generierung und Verarbeitung.[5]

B. Datenethik: Algorithmen als Akteure?

Zentral für die Perspektivierung von Daten ist die Frage, wie sich bestehende soziale und kulturelle Machtordnungen, normative und ethische Perspektiven in ihnen kondensieren. Dies schließt diverse Prozesse mit ein, so Erzeugung, Aufzeichnung, Pflege, Verbreitung, Weitergabe und Nutzung von Daten, aber auch ihre algorithmische Verarbeitung. KI-basierte Prognosesysteme und prädiktive Analytik beruhen auf einer Datenselektion, die notwendigerweise gesellschaftliche Machtverhältnisse widerspiegelt, auch in Bezug auf „fehlende Daten". Zentraler Ausgangspunkt der Überlegungen ist die Tatsache, dass Daten als Grundlage von Informationsprozessen nicht neutral sind. Mit dem Ansatz des „Values in Design" [6] werden die in Praktiken ihrer Herstellung bereits eingeschriebenen Werteordnungen reproduziert, indem bestimmte Wissensformen durch menschliche Programmierer*innen selektiert und dadurch bestehende Realitätsentwürfe und Werturteile fortgeschrieben werden. Diese Problematik stellt sich für alle Datenanalysen gleichermaßen und wurde vielfach im Hinblick auf Fragen der Ergebnissicherheit von Datenauswertungen hin verdeutlicht.[7]

Dass sich diese Erkenntnis über die Macht von Algorithmen als globale gesellschaftspolitische Herausforderung ansehen lässt, zeigen u.a. die Reaktionen diverser internationaler Organisationen. So verabschiedeten die 193 Mitgliedstaaten auf der UNESCO-Generalkonferenz 2021 die Empfehlungen zur Ethik der künstlichen Intelligenz, das erste globale, normative Instrument zu diesem

---

[4] Mühlhoff DZPhil 2020, 867.

[5] Zu Fragen der Staatshaftung u.a. beim Einsatz von KI durch die Polizei siehe von Rochow (in diesem Band), S. 103 ff.

[6] Simon/Wong/Rieder Internet Policy Review 2020, 9 (4).

[7] Dazu auch Gössl (in diesem Band), S. 7 ff. und Rateike (in diesem Band), S. 25 ff.; ausf. auch Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, 2023, abrufbar unter https://www.schleswig-holstein.de/DE/fachinhalte/G/gleichstellung/geschlechterneutrale_ki.html, S. 53 ff.

Thema überhaupt. Es soll nicht nur die Menschenrechte und die Menschenwürde schützen, sondern auch fördern und als „ethischer Leitkompass" ein richtungsgebendes Fundament sein, das die Achtung der Rechtsstaatlichkeit in der digitalen Welt stärkt und informiert.[8] Im Jahr 2022 folgte die EU Fundamental Rights Agency (FRA), die dezidiert auf die diskriminierende Funktion vieler Algorithmen hinweist.[9] Diese Regulierungsbestrebungen verdeutlichen ein anderes, paradox anmutendes Phänomen. So fürchtet man einerseits, dass "data relations enact a new form of data colonialism, normalizing the exploitation of human beings through data, just as historic colonialism appropriated territory and resources and ruled subjects for profit" [10], andererseits bedeutet die mangelhafte Präsenz von Bevölkerungsgruppen im Datenset, wie z.B. von wenig datafizierten Ländern, dass Trainingsdaten für die diversen KI-Anwendungen ohne ihre digitalen Praxen erstellt werden. Damit werden die alltäglichen Digitalkulturen von Millionen von Menschen ausgeblendet.

Algorithmenbasierte *agency* ist in jedem Falle die grundlegende Voraussetzung für die Mustererkennung großer Datenmengen, wie dies auch für Prognosen in der Polizeiarbeit genutzt wird. Wenn es aber nicht mehr allein die Programmierer*innen sind, die wünschenswerte Handlungsskripte in Algorithmen einschreiben, sondern diese vielmehr in der Lage sind, eigenständige Aktions- und Wissensstrukturen zu entwickeln, werden die „autonomen moralischen Agenten" – so die Befürchtung – zu einer neuen, nicht-menschlichen Stimme in der Auseinandersetzung über das ‚richtige und gute Handeln'. Dieser nächste Schritt in der Entwicklung scheint seit dem Ende 2022 veröffentlichen LLM System ‚ChatGPT' bereits erfolgt. Diese KI stellt insofern eine neue Herausforderung dar, als die Technologie hier als (scheinbar) eigenständiger sozialer Akteur Normen perpetuiert. Die einstmals populäre Aussage, dass die Daten für sich selbst sprechen, hat sich längst als falsch erwiesen und muss daher auch für die Polizeiarbeit neu gedacht werden.

---

[8] UNESCO Recommendation on the Ethics of Artificial Intelligence.

[9] FRA (European Union Agency for Fundamental Rights), Bias in Algorithms. Artificial Intelligence and Discrimination, 2022; zu weiteren Leitlinien auf nationaler, trans- und internationaler Ebene Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, 2023, abrufbar unter https://www.schleswig-holstein.de/DE/fachinhalte/G/gleichstellung/geschlechterneutrale_ki.html, S. 76 ff.

[10] Couldry/Mejas Television & New Media 2019, 336.

C. Datengestützte Polizeiarbeit

Der Einsatz von Algorithmen und KI in der Polizeiarbeit hat sich in den letzten Jahren durch die immer einfachere und günstigere Verfügbarkeit digitaler Technologien und Daten gesteigert.[11] Die Anwendungsfälle reichen von Live-Gesichtserkennungstechnologien zur Identifizierung gesuchter Personen, bis zur ‚vorausschauenden Polizeiarbeit' und Risikoeinschätzungsverfahren.
„Predictive Policing", genauer auch als prognosebasierte Polizeiarbeit definiert, beschreibt die polizeiliche Anwendung von analytisch-technischen Verfahren, um operative Prognosen bezüglich wahrscheinlicher Ursprünge bzw. Zeiten und Orte zukünftiger Kriminalität zu generieren und umzusetzen,[12] d.h. um „vor die Lage" zu kommen.[13] Dadurch soll die präventive Polizeiarbeit unterstützt und mittels Prognosen künftiger Straftaten, straffälliger Personen und Tatorte der Verhinderung von Verbrechen dienen.[14] Diese Berechnungen basieren in der Regel auf historischen Daten über Zeit, Ort und Art der begangenen Straftaten, die überwiegend aus offiziellen Quellen stammen. Solche Daten können durch Umgebungsvariablen wie Bevölkerungsdichte, Vorhandensein bestimmter öffentlicher Orte oder Dienstleistungen sowie wichtige Ereignisse oder Feiertage ergänzt werden.[15]
Bei der Erhebung der relevanten Daten werden verschiedene Formen unterschieden. „Predictive mapping" oder „georäumliche Kriminalitätsvorhersage" zielt darauf ab, durch Datenaggregation zu ermitteln, wann und wo Straftaten

---

[11] Knobloch Vor die Lage kommen: Predictive Policing in Deutschland. Chancen und Gefahren datenanalytischer Prognosetechnik und Empfehlungen für den Einsatz in der Polizeiarbeit 2018, S.13; Handbuch Cyberkriminologie/Ruppert, 2023, Big Data und Algorithmen im Rahmen der Kriminalitätsbegegnung.

[12] Hofmann Predictive Policing. Methodologie, Systematisierung und rechtliche Würdigung der algorithmusbasierten Kriminalitätsprognose durch die Polizeibehörden, 2020.

[13] Knobloch Vor die Lage kommen: Predictive Policing in Deutschland. Chancen und Gefahren datenanalytischer Prognosetechnik und Empfehlungen für den Einsatz in der Polizeiarbeit 2018, S. 13.

[14] Egbert Räume der Unfreiheit 2018, S. 241.

[15] Handbuch Cyberkriminologie/Ruppert, 2023, Big Data und Algorithmen im Rahmen der Kriminalitätsbegegnung; Knobloch Vor die Lage kommen: Predictive Policing in Deutschland. Chancen und Gefahren datenanalytischer Prognosetechnik und Empfehlungen für den Einsatz in der Polizeiarbeit 2018, S. 13.

stattfinden könnten.[16] Die Vorhersagen können sich somit sowohl auf Orte wie auch auf Personen beziehen. Im Falle personenbezogener Prognosearbeit wird auch von „predictive profiling"[17] oder „person-based predictive targeting"[18] bzw. allgemeiner vom personenbezogenen Predictive Policing gesprochen. Pearsall[19] definiert dies umfassend als „[...] taking data from disparate sources, analysing them and then using results to anticipate, prevent and respond more effectively to future crime."
Überblickshaft lassen sich die Methoden dabei wie folgt systematisieren:

- Methoden zur Vorhersage von Ort und Zeit von Straftaten
- Methoden zur Vorhersage von Straftätern und zur Identifizierung von Personen, die wahrscheinlich Verbrechen begehen werden.
- Methoden zur Vorhersage der Identität von Tätern/Täterinnen
- Methoden zur Vorhersage von Opfern von Straftaten

Mit Predictive Policing verbinden sich ausgesprochen heterogene, ja konfliktärgegensätzliche Haltungen und Einstellungen.[20] Einerseits wird mit diesen Verfahren die Erwartung verbunden, verfügbare Ressourcen optimiert, zielbewusst und ökonomisch einzusetzen und potentielle Opfer zu schützen; Egbert sieht Predictive Policing sogar als Treiber rechtlicher Innovation".[21] Auf der anderen Seite lässt sich ein grundsätzliches Spannungsfeld zwischen Datenmacht, Datenbesitz, staatlichem Handeln und der Sorge um die Bedrohung von Persönlichkeitsrechten konstatieren. So wird vor allem der Zusammenhang zwischen der Datenqualität und den sozialen sowie politischen Folgen von algorithmischen Zuschreibungen problematisiert sowie Datenerhebung und -prozessierung kritisch reflektiert.[22]Gefragt wird dabei u.a. inwiefern Bürgerrechte, Privatheit und

---

[16] Lynskey International Journal of Law in Context 2019, 162 (167).
[17] Sommerer Neue Kriminalpolitik 2017, 147(149).
[18] Ferguson University Pennsylvania Law Rev 2015, 327.
[19] Pearsall NIJ Journal 2010, 16.
[20] Deutscher Ethikrat 2023, S. 241 ff.
[21] Egbert Zeitschrift für Rechtssoziologie 2020, 26.
[22] Degeling/Berendt AI & Society 2018, 347; Lum/William Significance 13/5 (2016), 14-19, DOI: https://doi.org/10.1111/j.1740-9713.2016.00960.x; Richardson/Schultz/Crawford

personale Autonomie durch solche Verfahren eingeschränkt werden können. So verweisen z.B. Thun und Egbert auf die Risken psychologisch-psychiatrischer Diagnosen für personalisierte Verfahren, anhand derer das Gewaltrisiko von relevanten Personen und Gefährdern anhand eines wissenschaftlich geprüften Verrechnungsmodells individuell konkretisiert und polizeiliche Kontrollmaßnahmen entsprechend priorisiert werden.[23] Potentiell ermöglichen Techniken wie Gesichtserkennung und Kennzeichenleser, wie sie heute bereits in den angelsächsischen und in einigen asiatischen Ländern üblich sind, ein System, das Zuboff als ‚surveillance state' bezeichnet hat.[24]

Ethisch besonders problematisch sind personenbezogene Ansätze des Predictive Policing, die aktuell in den USA an Bedeutung gewinnen. Diese erstellen Risikoprognosen oder -profile, welche den Blick auf zukünftige Täter*innen oder Opfer von Kriminalität richten.[25] Dazu werden neben verschiedenen polizeilichen Daten auch Vorstrafen, der Wohnort, aber auch Social Media Daten über das soziale Umfeld der Betroffenen einbezogen. Verschiedene Studien aus den USA verdeutlichen, wie sich diese datenethischen Probleme zu Diskriminierungen verdichten können.[26] Eine bedeutende Gefahr des Dateneinsatzes besteht in der möglichen Potenzierung von solchen Stigmatisierungseffekten. Nachdem Algorithmen keine Kausalitäten evaluieren, sondern allenfalls Korrelationen aufzeigen können, dürfte bislang gelebte, in den Datensätzen beinhaltete Diskriminierung festgeschrieben und durch den Algorithmus potenziert werden, so dass sogar „dirty data".[27] entstehen können. Dies droht nicht nur bei der Einspeisung personenbezogener Daten, sondern bereits, wenn in einer raumbezogenen Prognosesoftware ein bestimmtes Terrain als Risikogebiet markiert wird und daraufhin Personen dort kontrolliert werden sollen.[28] Predictive Policing kann also durch verzerrte Datengrundlagen, durch das Design oder durch die

---

New York University Law Review Online 2019, 192; Predictive Policing - Eine Bestandsaufnahme für den deutschsprachigen Raum/Rolfes 2020, S. 215.

[23] Thurn/EgbertVorgänge 227: Polizei und Technik 2019, 71.

[24] Zuboff The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power, 2019.

[25] Ferguson University Pennsylvania Law Review 2015, 327 (373).

[26] Lum/William Significance 13/5 (2016), 14-19.

[27] Richardson/Schultz/Crawford New York University Law Review Online 2019, 192.

[28] Beck/Kusche/Valerius/Eisele/Böhm Digitalisierung, Automatisierung, KI und Recht, 2020, S. 529 f.

‚queries' selbst diskriminierende Effekte nach sich ziehen. Für zukünftige Entwicklungen scheint zudem das Risko zu steigen, dass ein „automation bias" eintritt.[29] Damit wird das Phänomen beschrieben, dass Menschen ihre eigenen Kompetenzen denen von technologischen Systemen unbewusst unterordnen, sodass am Ende die Prognose nicht mehr kritisch hinterfragt, sondern als zutreffend angesehen wird.

D. Rechtliche Anforderungen

Auch in Deutschland gehört Predictive Policing mehr und mehr zur Realität polizeilicher Arbeit. Mittlerweile ist der Gebrauch der Predictive Policing Technik in ganz Deutschland weit verbreitet. Dabei soll „der Computer", also die Predictive Policing Technik, automatisiert arbeiten und menschliches Handeln – hier vor allem durch die Polizeibeamt*innen – Schritt für Schritt ersetzen, so die Idee.[30] Wie diese neue Form von Polizeiarbeit rechtlich einzuordnen ist, wird inzwischen breit diskutiert. Einige der entsprechenden Überlegungen sollen kurz skizziert werden.

Predictive Policing bewegt sich zwischen Polizeirecht und Strafrecht. Auch wenn sich die Technik gleichzeitig diverser Daten bedient, die abgeschlossene Lebenssachverhalte beinhalten und damit rückblickend beurteilen, ist dieses Verfahren aus kriminologischer Sicht eine neue Form der Prognose.[31] Es geht um die Antizipation und Verhinderung noch bevorstehender Delikte, um *vorgelagerte* Prävention. So erfolgt ein Eingreifen durch die Polizei in das Geschehen noch vor Gefahr und Verdacht, da eine berechnete Wahrscheinlichkeit künftiger Deliktsbegehung, nicht eine unmittelbar bevorstehende Rechtsgutsgefährdung, den Ausschlag gibt. Ob und wie das von der Software berechnete Risiko im konkreten Einzelfall kurz vor dessen Eintritt steht, bleibt bei der Anwendung von Predictive Policing offen. So kann es auch zu sogenannten „false

---

[29] Hofmann Predictive Policing. Methodologie, Systematisierung und rechtliche Würdigung der algorithmusbasierten Kriminalitätsprognose durch die Polizeibehörden, 2020, S. 270 ff.; dazu auch Rateike (in diesem Band), S. 30 f.

[30] Rademacher/Perkowski JuS 2020, 713 (714).

[31] Singelnstein NStZ 2018, 1 (3).

positives" kommen, d.h. Personen wird fälschlicherweise eine Gefährlichkeit zugeschrieben, auf die mit einschneidenden, aber letztlich unberechtigten Maßnahmen reagiert wird.[32]

Zunächst muss zwischen *personenbezogenen* und *nicht personenbezogenen* Daten unterschieden werden. Personenbezogenen Daten, also Daten, die eine Person identifiziert oder identifizierbare Personen betreffen, Art. 4 Nr. 1 DSGVO, werden auf verschiedenen Ebenen geschützt, so durch die DSGVO und das Bundesdatenschutzgesetz (BDSG),[33] die Strafverfolgungsrichtlinie[34], die Charta der Grundrechte der EU (EUGrCh) und die europäische Menschenrechtskonvention (EMRK). Aus verfassungsrechtlicher Perspektive geht es dabei vor allem um das Recht auf informationelle Selbstbestimmung aus Art. 2 Abs. 1 i.V.m. Art. 1 Abs. 1 GG sowie die Gleichheitsrechte aus Art. 3 Abs. 3 GG. Predictive Policing Technologien dürfen grundsätzlich nicht eingesetzt werden, wenn sie bekanntermaßen nach den in Art. 3 Abs. 3 GG genannten Merkmalen oder vergleichbaren Merkmalen differenzieren.[35] Daneben können der Zugriff auf laufende Kommunikation gem. Art. 10 GG, der Zugriff auf Daten innerhalb der Wohnung gem. Art. 13 GG und das IT-Grundrecht bei der Erhebung von Daten von Festplatten, aus Cloud-Speichern abgeleitet aus Art. 2 Abs.1 i.V.m. Art. 1 Abs. GG relevant sein.[36] Ebenso wichtig sind die Grundrechte aus der EU-Grundrechtecharter, Art. 7 und 8 EUGrCh.

Zusätzlich muss zwischen der Datenerhebung und der Datenverarbeitung unterschieden werden.[37] Letzteres bedeutet im Rahmen von Predictive Policing, dass immer mehr Informationen in Form gespeicherter Daten vorliegen und für die Polizeiarbeit herangezogen werden. Es hängt entscheidend davon ab, in welchem Umfang welche Arten von Daten verarbeitet werden können.

---

[32] Singelnstein NStZ 2018, 1 (5).

[33] Zu den datenschutzrechtlichen Anforderungen beim Einsatz von KI-Systemen ausführlich Ambrock (in diesem Band), S. 69 ff.

[34] RL (EU) 2016/680 vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten durch die zuständigen Behörden zum Zwecke der Verhütung, Ermittlung, Aufdeckung oder Verfolgung von Straftaten oder der Strafvollstreckung sowie zum freien Datenverkehr und zur Aufhebung des Rahmenbeschlusses 2008/977/JI des Rates, ABl. EU 2016 L 119/89.

[35] Rademacher/Perkowski JuS 2020, 713 (717).

[36] Vgl. BVerfG NJW 2008, 822; Rademacher/Perkowski JuS 2020, 713 (717).

[37] Vgl.u.a. BVerfG NJW 2016, 1781.

Für die Datenerhebung hat das BVerfG in der ständigen Rechtsprechung folgende Maßstäbe entwickelt: Je weiter der Informationseingriff in das Vorfeld möglicher Rechtsgutsbeeinträchtigungen vorverlagert wird, umso höher sind die Anforderungen an die Informationseingriffe.[38] Im Bereich der Verhütung von Straftaten gelten diese Anforderungen folglich auch für den Einsatz von Predictive Policing Techniken.
Hinsichtlich der Verarbeitung von Daten ist festzuhalten, dass die Anforderungen im Vergleich gering sind, sofern diese Daten bereits zu Zwecken der Strafverfolgung oder Gefahrenabwehr erhoben wurden und zulässig von der Polizei gespeichert wurden (z.B. §§ 42 Abs. 3, 43 ASOG Bln).[39] Daraus folgt, dass sich regelmäßig aus diesen polizeigesetzlichen Regelungen auch Befugnisse zur Weiterverarbeitung ergeben können, da sie die Nutzung von der im Rahmen der Strafverfolgung gewonnen Daten auch zur Gefahrenabwehr und umgekehrt gestatten.[40] Im Gegenzug ergibt sich daraus, dass sofern kein solcher Anlass wie aus beispielsweise §§ 42 Abs. 3, 43 ASOG Bln besteht, auch strengere Anforderungen gelten. Die erleichterten Maßstäbe aus dem ASOG sind im hiesigen Fall jedoch nicht gegeben, denn die Predictive Policing Technik speist sich gerade aus öffentlichen und privaten Daten, z.B. von Nachrichtendiensten. Für die automatisierte Verarbeitung umfangreicher Datensätze hat das BVerfG bereits für die präventive Rasterfahndung hohe Hürden aufgestellt[41]. Im Vorfeld der Gefahrenabwehr scheidet eine solche Rasterfahndung aus.[42]
Daneben stellt sich zwingend die Frage nach der Zuverlässigkeit und damit nach der Qualität der verarbeiteten Datensätze sowie nach der Transparenz bei der Anwendung entsprechender Algorithmen. Es braucht dafür nach wie vor den Vergleich zum Entscheidungsfindungsprozess einer*s Polizeibeamten*in, diese:r wägt nach bekannten, einschlägigen Kriterien ab, ob beispielsweise eine Gefahr gegeben ist (z.B. Häufigkeit und Intensität vergangener Vorfälle, Drohungen der Betroffenen). Diese Entscheidung begründet und dokumentiert er:sie entsprechend. Auf diese Weise kann die Entscheidung auch gerichtlich überprüft werden. Dieses Grundprinzip der Überprüfbarkeit muss auch von automatisierten Systemen, wie Predictive Policing, gewährleistet werden, um

---

[38] Rademacher/Perkowski JuS 2020, 713 (717).
[39] Singelnstein, NStZ 2018, 1 (6).
[40] Kuhlmann/Trute GSZ 2021, 103.
[41] BVerfG NJW 2006, 1939 (1943f.).
[42] BVerfG NJW 2006, 1939 (1943f.).

eine gerichtliche Kontrolle zu garantieren.[43] Nur so lassen sich die rechtlichen Maßstäbe einhalten. Folge der Auswertung von Daten aufgrund des Algorithmus sind Handlungsmaßnahmen, also eine auf der Algorithmenprognose basierende Entscheidung tätig zu werden. Dann stellt sich die Frage, welche Maßnahmen durch die Polizei zulässig und verhältnismäßig sind. Diese Maßnahmen können von erhöhter Polizeipräsenz bis zu konkreten Maßnahmen wie Durchsuchungen oder sogar Ingewahrsamnahmen gehen. Zwar erstellt die Predictive Policing Technik auch eine Gefahrenprognose, aber eine abstrakte, sodass auch hier der menschliche Faktor einbezogen werden muss, denn bei der polizeilichen Prognose geht es um eine individuelle und normativ geprägte Beurteilung einer konkreten Sachlage.[44] Folglich bedarf es zusätzlicher Einschätzung durch die Polizeibeamt*innen selbst bezogen auf die konkrete Sachlage. Allein auf eine abstrakte Prognose der Predictive Policing Software kann kein grundrechtsintensiver Eingriff gestützt werden. Insoweit ist festzuhalten, dass es die Prüfung und Bewertung der Ergebnisse durch eine:n zuständige:n Amtsträger:in bedarf.[45]
Eine erhöhte Polizeipräsenz kann auch andere Auswirkungen haben. So hat die FRA in einem Bericht untersucht, wie die in Algorithmen erzeugten Verzerrungen auf die vorausschauende Polizeiarbeit angewendet werden und wie die neue KI-Verordnung der EU diese Probleme ansprechen soll.[46] Wie bereits für den US-amerikanischen Raum angesprochen, entstehen rechtliche Probleme durch übermäßige Polizeipräsenz in bestimmten Vierteln oder Regionen aufgrund von abstrakten Gefahrenprognosen durch die Predictive Policing Technologien. Unter der sog. Überpolizeilichkeit wird die im Verhältnis zur tatsächlichen Kriminalitätsrate unverhältnismäßige „Über"-Präsenz der Polizei in einem bestimmten Gebiet verstanden.[47] Überpolizeiliches Vorgehen kann beispielsweise darin bestehen, dass konkrete Maßnahmen gegen bestimmte Personen ergriffen werden, einschließlich polizeilicher Kontrollen und Durchsuchungen oder Identitätskontrollen. Das Ergebnis von Überpolizeilichkeit stellt oft eine benachteiligende Behandlung von Personen aufgrund bestimmter Merkmale dar. Hier wird der Ort selbst zu einem Stellvertreter für die ethnische Herkunft. Dies führt zu einer übermäßigen Polizeipräsenz, einer weniger strengen Schwelle bei

---

[43] Martini JZ 2017, 1017 (1019 ff.).
[44] Rademacher AöR 2017, 366 (383 f.).
[45] Singelnstein NStZ 2018, 1 (8); Rademacher AöR 2017, 366 (383 f.).
[46] FRA Bias in Algorithms. Artificial Intelligence and Discrimination.
[47] FRA Bias in Algorithms. Artificial Intelligence and Discrimination, S. 31.

Durchsuchungen und Verhaftungen und sogar zu einem „Schuldig-nach-Algorithmus". Beobachtet wurden eine Reihe sozialer Folgen, so. z.B. eine Wertminderung von Immobilien und eine geringere Investitionsneigung in die Infrastruktur von Seiten der Behörden.[48]

Neben der EU-Ebene gibt es auch zentrale Entscheidungen auf völkerrechtlicher Ebene. So hat der Europäische Gerichtshof für Menschenrechte (EGMR) in ständiger Rechtsprechung festgestellt, dass der Staat verpflichtet ist, einen möglichen Kausalzusammenhang zwischen der mutmaßlichen rassistischen Einstellung von Polizeibeamt*innen und der Misshandlung von Personen durch diese zu untersuchen.[49] Der Gerichtshof bekräftigt, dass die staatlichen Behörden bei der Untersuchung gewalttätiger Vorfälle zusätzlich verpflichtet sind, alle angemessenen Schritte zu unternehmen, um rassistische Motive zu entlarven und festzustellen, ob ethnischer Hass oder Vorurteile bei den Ereignissen eine Rolle gespielt haben könnten.[50] Die gleiche Überprüfung muss dann auch für *biased data* und deren Algorithmen gelten.

Es lässt sich feststellen, dass neue Technologien in besonders grundrechtssensiblen Bereichen eigenständige Rechtsgrundlagen, die der Rechtsprechung des BVerfG entsprechen, bedürfen. Auch wenn bereits einige Länder die rechtlichen Grundlagen für einen Gebrauch geschaffen haben, die womöglich den gesetzlich vorsetzten Rahmen der BVerfG Rechtsprechung einhalten, bleiben offene Fragen. Dies betrifft in herausragender Weise die Frage, wie mit voreingenommenen Daten umzugehen ist, aber auch Probleme mit fehlerhaften polizeilichen Systemen. Um diese Fragen zu adressieren, und betroffene Personen zu schützen, bedarf es neben dem rechtlichen Rahmen auch einer überprüfenden Infrastruktur, wie Ombudsstellen und anfängliche „fundamental rights assessments".[51] Natürlich muss auch der Nutzen der Technologie mit betrachtet werden. Konkret geht es darum, dass die Polizei auf der Basis dieser automatisch generierten Informationen Straftaten verhindern könnte, bevor sie begangen werden. Nur zu welchem Preis für Grundrechte ist die große Frage.

---

[48] Zur Staatshaftung in diesen Fällen von Rochow (in diesem Band), S. 106 ff.

[49] EGMR, Boacă u. a. gegen Rumänien, Nr. 40355/11, 12. Januar 2016, Rn. 108.

[50] EGMR, Boacă u. a. gegen Rumänien, Nr. 40355/11, 12. Januar 2016, Rn. 105.

[51] FRA Bias in Algorithms. Artificial Intelligence and Discrimination, S. 15.

E. Prädikative Technologien: Bedrohte Bürgerrechte, bedrohte Autonomie? Fazit und Versuch eines Ausblicks

Angesichts der historischen und gesellschaftlichen Verbreitung von Stereotypen, sozialer Ungleichheit und struktureller Diskriminierung wäre die Annahme, dass digitale Systeme frei von solchen Bias oder politischen Vorgaben, kulturellen Schemata oder gesellschaftspolitischen Werten sind, sicher unrealistisch. Eher lässt sich feststellen, dass immer neue Entwicklungen in Bezug auf Entscheidungsmöglichkeiten und -kontexte automatisierter, bzw. (teil-)autonomer Systeme im Konflikt mit menschlichen Praxen stehen und in einigen Fällen sogar die körperliche Unversehrtheit betreffen, wie z.B. dies für Gesichtserkennungstechnologien der Fall sein kann.[52] Wenn der Einsatz von Technologien zu einer „Automatisierung von Ungleichheit" und neuen Formen der sozialen Selektion führen kann,[53] so werden solche Praxen zu einem Werkzeug für ein weitereichendes algorithmisches Bevölkerungsmanagement. Dies kommt nicht nur, wie voranstehend skizziert, im Feld der prädiktiven Polizeiarbeit zum Tragen, sondern auch im Gesundheits- und Versicherungssystem, im Bildungsbereich, oder auch auf dem Arbeitsmarkt.[54] Eubanks beispielweise spricht in diesem Zusammenhang von der „Automatisierung von Ungleichheit".[55]
Aber prädiktive Technologien im Kontext von Polizeiarbeit sind, wie viele KI-basierte Anwendungen auch, nur ein Puzzleteil in einer größeren Transformation. Durch die komplexen Verschränkungen von menschlichen und technischen Systemen entstehen vielfältige Problemstellungen, die sich aktuell im zeitgenössischen Diskurs um das Verhältnis zwischen menschlicher Autonomie und autonomen technischen Systemen abbilden lassen,[56] die Maschine wird zu

---

[52] FRA, Facial recognition technology: Fundamental Rights Considerations in the Context of Law Enforcement 2019.
[53] Mühlhoff DZPhil 2020, 867.
[54] Deutscher Ethikrat 2023, S. 140ff.
[55] Eubanks Automating inequality: how high-tech tools profile, police, and punish the poor, 2017.
[56] Rötzer/Misselhorn Programmierte Ethik. Brauchen Roboter Regeln oder Moral?, 2016, S. 9; Rath/Krotz/Karmasin/Thimm /Bächle Maschinenethik – Normative Grenzen autonomer Systeme, 2019, S. 73; Rössler Autonomy: An essay on the life well-lived, 2021.

Freund und Feind gleichermaßen.[57] Damit einher geht die Bobachtung, dass in immer mehr Bereichen des alltäglichen Lebens handlungsförmige Aktionen in zunehmender Zahl von technischen Systemen vollzogen werden. Der alltägliche Handlungsraum, der vormals nur mit anderen Menschen sozial geteilt werden musste, trägt zunehmend auch technische Anteile und hat mit Technologien wie KI oder Robotern neue ‚Mitspieler' erhalten. Setzt man den Ausdruck Autonomie in den Kontext der aktuellen Debatte um die Technologisierung des Alltags, so verbindet sich damit einerseits die positive Wertung als unabhängig und selbstgesteuert, andererseits kommt damit auch die Furcht vor (zu) mächtiger Technologie sowie damit verbundener Macht- und Kontrollreduzierung des Menschen zum Ausdruck. Dies ist nicht zuletzt der Tatsache geschuldet, dass die Informatik und die Kognitionswissenschaft in einem technizistischen Verständnis so genannte intelligente, selbstlernende Algorithmen oder vermeintlich eigenständig agierende Roboter „autonom" nennen, wodurch der Begriff häufig zusammen mit den geteilten Gegenstandsbereichen auch in sozial- und geisteswissenschaftlichen Fragestellungen importiert wird.
Viele KI-Technologien werden mit Misstrauen beobachtet, die Debatte um die Ethik der Daten bzw. die Ethik der KI hat gerade erst begonnen.[58] Umso höher ist der Umgang mit solchen Systemen zu bewerten, die für die Bürger*innen in ihrem konkreten Alltag und für ihr Rechtsempfinden relevant werden. Aus dieser Sicht kommt der vorausschauenden Polizeiarbeit eine hohe politische und ethische Verantwortung zu, denn ob und welche „Kollateralschäden" Bürger*innen in Bezug auf ihre Grundrechte in Kauf nehmen, wird von hoher Relevanz für die weitere Entwicklung sein. Hier sind sowohl die Urteile der höchsten Gerichte maßgebend, als auch die Gesetzgebung(-svorhaben) auf EU-Ebene, die für alle Institutionen, auch die Polizeien, verpflichtend bzw. handlungsleitend sein sollten.

---

[57] Thimm/Bächle/Thimm Die Maschine - Freund oder Feind? Mensch und Technologie im digitalen Zeitalter, 2019, S. 17.
[58] Deutscher Ethikrat 2023.

Teil 2

Der rechtliche Rahmen

Kapitel 4

Diskriminierungsverbote im deutschen und europäischen Recht und die zukünftige KI-VO

Selen Yakar

A. Einleitung

Durch die immer größere Verbreitung der Einsetzung von KI-Systemen stellt sich zunehmend die Frage nach einem ausreichenden Schutz der Rechte der von diesen Systemen betroffenen Endverbraucher*innen. Auch wenn Algorithmen nicht in der Lage sind, bewusst zu diskriminieren, so zeigt die Bildung sogenannter *bias* zum Beispiel durch die Übernahme gesellschaftlicher Strukturen wie Vorurteile und Stereotypen durch präexistente *bias* oder die Über- oder Unterrepräsentation bestimmter gesellschaftlicher Gruppen in den Datensätzen, dass durch die Verwendung algorithmischer Entscheidungssysteme eine erhebliche Gefahr der Diskriminierung besteht.[1] Algorithmische Entscheidungen haben im Gegensatz zu menschlichen Entscheidungen eine größere Folgenwirkung dadurch, dass zum Beispiel menschliche Vorurteile aus der Vergangenheit unendlich fortgeschrieben und verstärkt werden können.[2]

---

[1] Im Detail Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, 2023, abrufbar unter: https://www.schleswig-holstein.de/DE/fachinhalte/G/gleichstellung/Downloads/handreichung_geschlechterneutrale_ki_lang.html?nn=9e7751e3-75ce-45f8-858b-3c234800fb27, S. 54-67; auch Gössl (in diesem Band), S. 7 ff.

[2] Dritter Gleichstellungsbericht der Bundesregierung „Digitalisierung geschlechtergerecht gestalten", BT-Drs. 19/30750, S. 22.

Ein Beispiel für das tiefgreifende Diskriminierungspotential dieser Systeme stellt das von Amazon eingesetzte Recruiting-Tool dar:[3] Die Idee Amazons war, ein KI-System zu entwickeln, dass beim Einstellungsprozess neuer Mitarbeiter*innen unterstützen sollte. Dafür wurde das Tool mit Bewerbungsunterlagen angenommener Mitarbeiter*innen der letzten 10 Jahre trainiert, um orientiert an diesen Beispielen ein Muster der zu bevorzugenden Eigenschaften neuer Mitarbeiter*innen zu bilden. Problematisch war, dass der Algorithmus schnell ein Muster formte, das männliche Bewerber bevorzugte. Hintergrund dessen war, dass die alten Bewerbungsunterlagen überwiegend aus männlichen Bewerbern bestanden, und der Algorithmus daraus den Schluss zog, dass die Einstellung dieser wünschenswerter ist. Auch als das Merkmal des Geschlechts aus dem verwendeten Datensatz gelöscht wurde, konnte der Algorithmus durch die sog. *proxy discrimination*[4] weiterhin durch Anknüpfung an scheinbar neutrale Merkmale auf Grundlage des Geschlechts diskriminieren.
Ein weiteres Beispiel bietet das Programm „Google Translate". Dieses führt bei der Übersetzung von Sätzen aus Sprachen ohne geschlechtsspezifischen Pronomen in eine Sprache mit geschlechtsspezifischen Pronomen zu einer Fortführung der in der Gesellschaft verankerten Stigmata: Während zum Beispiel der türkische Satz „o güzel" (korrekt übersetzt: „er/sie/es ist schön") in „sie ist schön" übersetzt wird, wird der Satz „o çalışkan" (korrekt übersetzt: er/sie/es ist fleißig") in „er ist fleißig" übersetzt.[5] Damit übernimmt der Algorithmus die in der Gesellschaft verankerten Stigmata und baut sie in sein Muster ein.
Das Verbot geschlechterbezogener Diskriminierung ist bereits vielfach rechtlich ausgestaltet. Nachfolgend wird untersucht, ob das Problem der algorithmenbasierten Diskriminierung wegen des Geschlechts durch das bestehende Recht bereits abgedeckt wird oder ob in diesem Bereich Regulierungsbedarf besteht. In diesem Zusammenhang wird der Begriff der Diskriminierung wegen des Geschlechts anhand der Regulierungen des Grundgesetzes (GG), der Landesverfassung Schleswig-Holsteins (VerfSH), der Europäischen Menschenrechtskon-

---

[3] Dastin Amazon scraps secret AI recruiting tool that showed bias against women, Reuters, 2018, abrufbar unter: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[4] Im Detail Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, S. 72 f.

[5] Überprüft von der Autorin.

vention (EMRK), des Übereinkommens der Vereinten Nationen (VN) zur Beseitigung jeder Form von Diskriminierung der Frau (CEDAW), der Grundrechte-Charta der EU (EUGrCh), des Digital Markets Acts (DMA)[6], des Digital Services Acts (DSA)[7] und der Entwurf einer Verordnung zur Festlegung Harmonisierter Vorschriften für Künstliche Intelligenz (KI-VO-E)[8] analysiert und herausgearbeitet, inwiefern durch diese Vorschriften dem Problem der algorithmenbezogenen Diskriminierung abgeholfen werden kann.[9]

B. Überblick

Im nationalen und internationalen Rechtsraum bestehen bereits unterschiedliche Regulierungen zur Ungleichbehandlung aufgrund des Geschlechts und der Formulierung eines Verbots diesbezüglich. Es zeigt sich jedoch, dass diese Vorschriften die algorithmenbezogene Diskriminierung nicht umfassend zu regulieren vermögen. Zum einen wird kein ausreichender Schutz des privatrechtlichen Verhältnisses formuliert, zum anderen fehlt es auch an einem das KI-System mit einbeziehenden Anwendungsbereiches der existierenden Normen.

I. Verfassungsnormen

Das Grundgesetz enthält in Art. 3 Abs. 2 GG das Gleichbehandlungsgebot der Geschlechter („Männer und Frauen sind gleichberechtigt.“, S. 1) und außerdem eine Verpflichtung des Staates, diese auf allen Ebenen durchzusetzen („Der Staat fördert die tatsächliche Durchsetzung der Gleichberechtigung von Frauen

---

[6] Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates über bestreitbare und faire Märkte im digitalen Sektor (Gesetz über digitale Märkte), COM/2020/842 final.

[7] Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates über einen Binnenmarkt für digitale Dienste (Gesetz über digitale Dienste) und zur Änderung der Richtlinie 2000/31/EG, COM/2020/825 final.

[8] 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 – Allgemeine Ausrichtung (6. Dezember 2022), ST 15698 2022 INIT.

[9] Vgl. dazu auch Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung.

und Männern und wirkt auf die Beseitigung bestehender Nachteile hin.“, S. 2).[10] Einen solchen Staatsauftrag formuliert auch Art. 9 VerfSH („Die Förderung der rechtlichen und tatsächlichen Gleichstellung von Frauen und Männern ist Aufgabe des Landes [...]“).[11]
Beide Artikel beziehen sich jedoch nur auf den Schutz der Bürger*innen vor Diskriminierungen vonseiten des Staates.[12] Nicht geschützt werden die von den Entscheidungen der Algorithmen betroffenen Personen vor KI-Systemen, die vonseiten privater Akteur*innen eingesetzt werden. Lediglich in Situationen eines „soziale[n] Machtverhältnis[ses]“[13] zum Beispiel aufgrund einer Monopolstellung auf dem Markt wird auch ein Schutz des Art. 3 Abs. 2 GG in privatrechtlichen Rechtsverhältnissen angenommen.[14] Hier wird argumentiert, dass aufgrund eines Missbrauchs privater Macht besondere Notwendigkeit eines Schutzes der diesem Monopol ausgelieferten Person besteht.[15] Als Beispiel für ein solches Machtverhältnis können große Online-Plattformen angeführt werden, die aufgrund ihrer Monopolstellung auf dem internationalen Markt ein Abhängigkeitsverhältnis bilden.[16] In einem solchen Fall sollen sich die diese Plattformen nutzenden Personen auf ihr Grundrecht auf Gleichbehandlung berufen können.[17] Diese eng ausgelegte Ausnahme, die eine Drittwirkung der Grundrechte zwischen Privaten ermöglicht, bietet jedoch keinen umfassenden Schutz vor jeglicher algorithmenbasierter Diskriminierung aufgrund des Geschlechts. Ein solcher Schutz gilt beispielsweise nur für Grundrechtsverletzungen innerhalb des Geltungsbereichs des Art. 3 GG. Vor einer Grundrechtsverletzung durch eine Online-Plattform, die international betrieben wird, stellt diese Ausweitung des Anwendungsbereichs des Art. 3 Abs. 2 GG daher keinen ausreichenden Schutz dar.

---

[10] Jarass/Pieroth/Kment/Jarass GG, 17. Aufl. 2022, Art. 3 Rn. 107.
[11] Becker/Brüning/Ewer/Schliesky/Welti ShVerf, 1. Aufl. 2021, Art. 9 Rn. 16.
[12] Jarass/Pieroth/Kment/Jarass GG Art. 3 Rn. 104; Becker/Brüning/Ewer/Schliesky/Welti ShVerf Art. 9 Rn. 3.
[13] BGH NJW 2013, 1519 Rn. 27.
[14] Dreier/Heun GG, 3. Aufl. 2013, Art. 3 Rn. 70; Jarass/Pieroth/Kment/Jarass GG Art. 3 Rn. 17.
[15] Dreier/Heun, Art. 3 Rn. 70.
[16] BVerfG NJW 2019, 1935.
[17] Vgl. ausführlich *Gössl/Yakar,* Geschlechterneutrale KI. Eine Handreichung, S. 24.

II. Völkerrechtliche Übereinkommen

Die Europäische Menschenrechtskonvention (EMRK) und das Übereinkommen der Vereinten Nationen zur Beseitigung jeder Form von Diskriminierung der Frau (CEDAW) sind völkerrechtliche Übereinkommen und regulieren die Ungleichbehandlung aufgrund des Geschlechts auf internationaler Ebene. So verpflichtet Art. 14 EMRK die Vertragsstaaten dazu, die in der EMRK anerkannten Rechte und Freiheiten ohne Diskriminierung „wegen des Geschlechts, der Rasse, der Hautfarbe, der Sprache, der Religion, der politischen oder sonstigen Anschauung, der nationalen oder sozialen Herkunft, der Zugehörigkeit zu einer nationalen Minderheit, des Vermögens, der Geburt oder eines sonstigen Status zu gewährleisten". Gewährleistet wird dieser Schutz allen Menschen, unabhängig von ihrer Staatsangehörigkeit. Voraussetzung ist allein, dass die Ungleichbehandlung „von der Hoheitsgewalt eines Mitgliedstaates" ausgeht.[18] Die CEDAW befasst sich dagegen „exklusiv mit der Aufhebung der Diskriminierung von Frauen und Mädchen aufgrund ihres Geschlechts"[19] und stellt ebenfalls eine Verpflichtung der Vertragsstaaten zur Verhinderung der Diskriminierung, hier explizit der Frau, auf allen Ebenen auf. Diese Vorschriften wiederum verpflichten ausschließlich die Vertragsstaaten und gehen damit ebenfalls nicht über den Schutz vor Handlungen des jeweiligen Staates hinaus.[20]

III. Die Grundrechtecharta

Die Grundrechte-Charta der Europäischen Union stellt in Art. 23 Abs. 1 EUGrCh ähnlich wie Art. 3 Abs. 2 GG und Art. 9 VerfSH den Auftrag, „die Gleichheit von Frauen und Männern [...] in allen Bereichen [...] sicherzustellen". Dieser Auftrag bezieht sich ausschließlich auf die Durchführung von Unionsrecht und ist sowohl von den Organen der Union selbst als auch von den Mitgliedstaaten zu gewährleisten (Art. 51 Abs. 1 EUGrCh) und stellt daher ebenfalls nur einen Schutz vor staatlichem Handeln dar.[21]

---

[18] Grabenwarter/Pabel EMRK/Grabenwarter/Pabel, 7. Aufl. 2021, § 17 Rn. 2.

[19] Lembke/Rodi Menschenrechte und Geschlecht, 2014, S. 53.

[20] HK-EMRK/Meyer-Ladewig/Lehner, 4. Aufl. 2017, Art. 14 Rn. 15; Grabenwarter/Pabel EMRK/Grabenwarter/Pabel, § 26 Rn. 34 ff.

[21] Geiger/Khan/Kotzur/Kirchmair/Kotzur EUV/AEUV 7. Aufl. 2023, Einführung zur Charta der Grundrechte Rn. 11.

IV. Nationale Vorschriften

Im Gegensatz zu den oben genannten Vorschriften formulieren das Allgemeine Gleichbehandlungsgesetz (AGG) und die Datenschutzgrundverordnung (DSGVO) Verpflichtungen in privaten Rechtsverhältnissen.[22]
Das AGG formuliert Diskriminierungsverbote zwischen Privaten in beschäftigungsrechtlichen Verhältnissen (§§ 6-18 AGG) und im allgemeinen Zivilrechtsverkehr (§§ 19-21 AGG) und stellt damit eine Einschränkung der grundsätzlichen Vertragsfreiheit dar.[23] § 7 Abs. 1 AGG formuliert damit ein Benachteiligungsverbot wegen eines in § 1 genannten Grundes, womit beispielsweise auch die Benachteiligung eines/einer Beschäftigten wegen des Geschlechts untersagt ist. Das AGG schließt dabei den Schutz vor algorithmenbasierter Diskriminierung grundsätzlich nicht aus seinem Anwendungsbereich aus, nennt ihn aber auch nicht explizit. In diesem Zusammenhang wurde vor kurzem ein offener Brief an die Bundesregierung mit der Forderung geschrieben, das AGG in Bezug auf einen Schutz vor algorithmenbasierter Diskriminierung zu reformieren. Die unterzeichnenden Organisationen weisen darauf hin, dass zum einen der Anwendungsbereich des AGG um teil- und vollautomatisierte Entscheidungsverfahren ausgedehnt werden muss und zum anderen eine nicht abschließende Auflistung an Diskriminierungskategorien enthalten sollte, darüber hinaus Tatbestände sowohl der intersektionalen Diskriminierung als auch solcher, in denen das Opfer nicht eindeutig identifizierbar ist, aufgenommen werden sollten. Schließlich fordert der Brief die Möglichkeiten der Verbandsklage und der Prozessstandschaft.[24] Auch im aktuellen Koalitionsvertrag wird festgeschrieben,

---

[22] Zum AGG auch Kichhefer-Lauber (in diesem Band), S. 89 ff.

[23] Martini Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, 2019, S. 77.

[24] AlgorithmWatch/AlgorithmWatch CH/Antidiskriminierungsverband Deutschland u.a. Offener Brief: Jetzt algorithmenbasierte Diskriminierung erkennen und Schutzlücken schließen! 2023, abrufbar unter: https://algorithmwatch.org/de/offener-brief-diskriminierung-allgemeines-gleichbehandlungsgesetz/?utm_campaign=2023-01-17%20Open%20Letter%20Fundraising%20AGG,%20DE.

dass bestehende Schutzlücken im AGG geschlossen werden sollen und der Anwendungsbereich ausgeweitet werden soll.[25]
Auch die DSGVO[26] greift zum Schutz natürlicher Personen in die Vertragsfreiheit dergestalt ein, dass sie insbesondere die Verarbeitung personenbezogener Daten reguliert. Sie formuliert dabei nur die Regulierung von personenbezogenen Daten und klammert anonymisierte Daten aus.[27] Das Problem der *proxy discrimination*[28] zeigt aber, dass auch durch diese Daten verzerrte Ergebnisse entstehen können. Damit wird auch hier kein ausreichender Schutz geschaffen.

V. EU-Verordnungen

Im Lichte der neuen Herausforderungen durch die Digitalisierung des europäischen Marktes wurden außerdem neue EU-Verordnungen erlassen: der Digital Markets Act (DMA), der Digital Services Act (DSA) und die KI-Verordnung im Entwurfsstadium (KI-VO-E). Der DMA hat zum Ziel, faire Wettbewerbsbedingungen auf dem Markt zu schaffen und unterwirft dafür die sogenannten „*gatekeeper*", also Unternehmen mit einer Monopolstellung auf dem Markt, im Rahmen der Nutzung von KI-Systemen bestimmten Regeln und Verboten.[29] Der DSA reguliert die Haftung von Vermittlungsdiensten bei der Zwischenspeicherung und Verwendung von personenbezogenen Informationen. Sowohl der DMA als auch der DSA stellen zumindest im Rahmen ihres Anwendungsbereichs geeignete Schutzmechanismen für die Gewährleistung eines Schutzstandards der Endnutzer*innen dieser Art von Systemen.
Die KI-VO liegt derzeit noch als Entwurf vor. Ziel dieser Verordnung ist es, einen europäischen Rechtsrahmen für eine vertrauenswürdige KI zu schaffen. Außerdem sollen auch die Grundrechte der betroffenen Personen gestärkt werden. Die Effektivität des KI-VO-E wird weiter unten detailliert thematisiert.

---

[25] Koalitionsvertrag 2021-2025 Mehr Fortschritt wagen, abrufbar unter: https://www.spd.de/fileadmin/Dokumente/Koalitionsvertrag/Koalitionsvertrag_2021-2025.pdf, S. 96.
[26] dazu ausführlich Ambrock (in diesem Band), S. 69 ff.
[27] Lauscher/Legner, ZfDR 2022, 367 (379).
[28] dazu auch Gössl (in diesem Band), S. 13 f.
[29] Genovesi/Kaesling/Robbins Recommender Systems/Gössl, 2023, (im Erscheinen); Zimmermann/Heinzel Der Digital Markets Act, Januar 2022, abrufbar unter: https://www.germanwatch.org/sites/default/files/digital_markets_act_hintergrundpapier_2.pdf.

C. Diskriminierungsbegriff

Die bestehenden Diskriminierungsverbote basieren alle auf einem ähnlichen Verständnis des Diskriminierungsbegriffs. Für eine genaue Regulierung der Diskriminierungsrisiken durch KI-Systeme ist eine Differenzierung nicht nur zwischen einer Ungleichbehandlung und einer Diskriminierung, sondern auch zwischen unmittelbarer und mittelbarer Ungleichbehandlung bzw. Diskriminierung notwendig.

I. Ungleichbehandlung und Diskriminierung

Auch wenn die Begriffe Ungleichbehandlung und Diskriminierung häufig als Synonyme verwendet werden, sind sie nicht deckungsgleich. Eine Ungleichbehandlung besteht zum einen dann, wenn zwei vergleichbare Sachverhalte aufgrund eines geschützten Unterscheidungsmerkmals unterschiedlich behandelt werden. Eine unterschiedliche Behandlung aufgrund des Geschlechts liegt damit zunächst dann vor, wenn eine Person aufgrund ihrer Geschlechtszugehörigkeit anders behandelt wird als andere Personen derselben Vergleichsgruppe.[30] Von einer Ungleichbehandlung wird aber zum anderen auch dann gesprochen, wenn wesentlich ungleiche Sachverhalte gleich behandelt werden. Dies wird damit begründet, dass durch eine solche Gleichbehandlung unterschiedlicher Sachverhalte im Ergebnis die größere Vergleichsgruppe ungleich behandelt wird.[31] Im Rahmen der geschlechtsspezifischen Ungleichbehandlung bedeutet dies, dass eine Frau oder ein Mann aufgrund ihrer Zugehörigkeit zu einem Geschlecht anders behandelt wird als andere, nicht diesem Geschlecht zugehörige Personen.[32] Eine Diskriminierung hingegen liegt erst dann vor, wenn diese Ungleichbehandlung nicht gerechtfertigt werden kann. Die verfassungs- und europarechtlichen Vorschriften folgen alle dem gleichen Ansatz der Verhältnismäßigkeit. Danach ist eine Ungleichbehandlung dann gerechtfertigt, wenn sie einem legitimen Zweck folgt, die zur Verfolgung dieses Zwecks angewandten

---

[30] Jarass/Pieroth/Kment/Jarass GG Art. 3 Rn. 10 f; Grabenwarter/Pabel EMRK/Grabenwarter/Pabel, § 26 Rn. 8.

[31] Jarass/Pieroth/Kment/Jarass GG Art. 3 Rn. 12.

[32] Jarass GRCh 4. Aufl. 2021, Art. 23 Rn. 10.

Mittel erforderlich und geeignet sind, und insgesamt der Zweck nicht außer Verhältnis zum geschützten Rechtsgut, hier der Gleichbehandlung der Geschlechter, steht.[33]

II. Unmittelbare und mittelbare Ungleichbehandlung bzw. Diskriminierung

Unterschieden wird außerdem zwischen unmittelbarer und mittelbarer Ungleichbehandlungen bzw. Diskriminierungen. Bei einer unmittelbaren Ungleichbehandlung aufgrund des Geschlechts erfolgt die unterschiedliche Behandlung durch direkte Anknüpfung an das Merkmal des Geschlechts. Personen werden also gerade wegen ihrer Geschlechtszugehörigkeit anders behandelt als der Rest der Vergleichsgruppe. Mittelbar hingegen sind Ungleichbehandlungen, die an scheinbar neutrale Merkmale anknüpfen, im Ergebnis die Person mit einer bestimmten Geschlechtszugehörigkeit jedoch anders behandeln als andere derselben Vergleichsgruppe.[34] Relevant wird dies insbesondere im Rahmen der bereits erwähnten *proxy discrimination*: Zu überlegen ist, ob der bestehende Schutz vor mittelbarer Diskriminierung ausreicht, um die *proxy discrimination* zu verhindern, oder ob es diesbezüglich neuer Ansätze bedarf.[35]

III. Das Diskriminierungsverständnis der EU, insbesondere der KI-VO-Entwurf

Im Gegensatz zu den verfassungs- und europarechtlichen Normen formulieren die EU-Verordnungen keinen neuen Diskriminierungsbegriff. Der Anknüpfungspunkt des KI-VO-E ist das Risiko, dem die betroffenen Personen durch Nutzung des KI-Systems ausgesetzt sind. Der KI-VO-E unterwirft die KI-Systeme je nach Risiko für Grundrechte und Sicherheit unterschiedlichen Regulierungen (s. unten D. I.). Im Übrigen bezieht sich der Entwurf auf das Grundrechtsverständnis der EUGrCh, indem er auf eine „nach dem Unionsrecht verbotenen Diskriminierung“ verweist (Art. 10 Abs. 2 lit. f) KI-VO-E).

---

[33] HK-EMRK/Meyer-Ladewig/Lehner EMRK Art. 14 Rn. 9; Dreier/Heun GG Art. 3 Rn. 24; Jarass GRCh Art. 23 Rn. 18 f; Grabenwarter/Pabel EMRK/Grabenwarter/Pabel, § 26 Rn. 16.

[34] Sachs/Nußberger Grundgesetz, 9. Aufl. 2021, Art. 3 Rn. 248; Becker/Brüning/Ewer/Schliesky/Welti ShVerf Art. 9 Rn. 25 ff; Jarass GRCh Art. 23 Rn. 10; ErfK/Schlachter-Voll, 23. Aufl. 2023, AGG § 3 Rn. 9.

[35] So Ad Hoc Committee on Artificial Intelligence Feasibility Study, 2020, abrufbar unter https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da.

Der DSA und der DMA hingegen basieren ihren Ansatz nicht auf der Wahrung der Grundrechte, sondern auf einer Harmonisierung der Rechtslage für Unternehmen auf dem europäischen Markt.

IV. Zwischenfazit

Zusammenfassen lässt sich, dass die verfassungs- bzw. europarechtlichen Vorschriften auf die Verhinderung einer Ungleichbehandlung der Bürger*innen durch den Staat oder andere Träger*innen öffentlicher Gewalt ausgerichtet sind, wohingegen sich die neuen EU-Verordnungen auf eine einheitliche Regulierung des europäischen Binnenmarkts und dort insbesondere auf die Herausforderungen der Digitalisierung fokussieren.
Zu fragen ist, ob die neuen Verordnungen damit einen ausreichenden Schutz der Grundrechte dort liefern können, wo die verfassungs- und europarechtlichen Normen zu kurz kommen, da sie sich nicht auf die speziellen Gefahren durch den Einsatz digitaler Systeme beziehen. Hier kann auch auf die durch den offenen Brief zur Reformierung des AGG aufgeworfenen Lücken in dessen Anwendungsbereich verwiesen werden.[36] Die dort genannten Forderungen verdeutlichen die Lücke, die sowohl im nationalen als auch im internationalen Recht in Bezug auf die Diskriminierung durch Algorithmen besteht. Auch der dritte Gleichstellungsbericht der Bundesregierung betont die Notwendigkeit der Aufnahme „konkrete[r] Regelungen für den Einsatz [algorithmischer] Systeme in das AGG".[37] Die verfassungs- und europarechtlichen Vorschriften formulieren zwar einen Gleichbehandlungsgrundsatz, schreiben diesen aber nur für staatliche Organe vor. Außerdem nehmen sie die algorithmenbasierte Diskriminierung bisher nicht mit auf. Die neuen Verordnungen hingegen beschränken sich auf die Diskriminierung durch Algorithmen, formulieren aber keinen neuen Diskriminierungsbegriff.
Nachfolgend soll daher der KI-VO-E dahingehend untersucht werden, ob er die bestehenden Lücken schließen und einen umfassenden Schutz vor algorithmenbezogener Diskriminierung wegen des Geschlechtes formulieren kann.

---

[36] AlgorithmWatch/AlgorithmWatch CH/Antidiskriminierungsverband Deutschland u.a. Offener Brief: Jetzt algorithmenbasierte Diskriminierung erkennen und Schutzlücken schließen!.

[37] BT-Drs. 19/30750, S. 168.

D. Der KI-Verordnungsentwurf

Der KI-VO-E stellt den ersten internationalen Versuch einer einheitlichen europäischen Regulierung der künstlichen Intelligenz dar.[38] Er bringt in seinem Begründungstext vier Ziele zum Ausdruck:[39] Zunächst soll die Verordnung gewährleisten, dass sichere KI-Systeme auf den europäischen Markt gebracht werden und dadurch die bestehenden Grundrechte und Werte der EU gewahrt werden. Des Weiteren soll im Zuge der Förderung von Investitionen und Innovationen rund um die KI Rechtssicherheit gewährleistet werden. Außerdem ist Ziel der Verordnung „die wirksame Durchsetzung des geltenden Rechts zur Wahrung der Grundrechte" und die Stärkung der Sicherheitsanforderungen an KI-Systeme. Schließlich soll der Marktfragmentierung durch die „Entwicklung eines Binnenmarkts für rechtskonforme, sichere und vertrauenswürdige KI-Anwendungen" entgegengewirkt werden.[40]

I. Risikobasierter Ansatz

Der Verordnungsentwurf verfolgt einen risikobasierten Ansatz. Danach werden die KI-Systeme auf Grundlage des von ihnen ausgehenden Risikos für die Grundrechte der Betroffenen unterschiedlichen Verfahren unterworfen: verbotene Praktiken (Art. 5 KI-VO-E), hochriskante KI-Systeme, die einem Konformitätsverfahren unterworfen werden (Art. 6 ff. KI-VO-E), und Systeme mit geringen (Art. 52 KI-VO-E) bis gar keinen Risiken (Art. 69 KI-VO-E).[41]
Die Entwicklung und Einsetzung von KI-Systemen mit verbotenen Praktiken ist untersagt. Dazu gehören beispielsweise das Inverkehrbringen, die Inbetriebnahme oder die Verwendung eines KI-Systems zur Beeinflussung des Verhaltens einer Person oder zur Ausnutzung der Schwäche, Schutzbedürftigkeit oder Behinderung einer Person, um dieser oder einer anderen Person Schaden zuzufügen (Art. 5 Abs. 1 lit. a) und b) KI-VO-E). Außerdem ist die Verwendung von

---

[38] Vgl. dazu ausführlich Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, S. 48 ff.
[39] Begründung des ersten Entwurfs, Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Feststellung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union vom 06.12.2022, S. 3.
[40] Begründung des ersten Entwurfs, ebd., S. 3.
[41] Müller EuZ 2022 (2022), A1–A25 (A7 ff.).

KI-Systemen zur Klassifizierung von Personen (Art. 5 Abs. 1 lit. c) KI-VO-E) und die Verwendung von Echtzeit-Fernidentifizierungssystemen, außer in eng umrissenen Grenzen (Art. 5 Abs. 1 lit. d) KI-VO-E), untersagt.
Dahingegen wird für solche KI-Systeme, die nur ein geringes bis gar kein Risiko für die Grundrechte und die Sicherheit der Personen darstellen, lediglich Transparenzpflichten formuliert (Art. 52 KI-VO-E) bzw. angeraten, Verhaltenskodizes aufzustellen (Art. 69 Abs. 1 KI-VO-E).
Die Kategorie der Hochrisiko-KI-Systeme stellt den Kern der Verordnung dar. KI-Systeme sind zum einen dann hochriskant, wenn sie ein Produkt (Art. 6 Abs. 1 KI-VO-E) oder eine Sicherheitskomponente eines Produkts (Art. 6 Abs. 2 KI-VO-E) im Sinne einer Harmonisierungsvorschrift (aufgezählt in Anhang II) darstellen oder unter einen der in Anhang III aufgelisteten Bereiche fallen (sog. „*stand-alone* KI-Systeme", Art. 6 Abs. 3 KI-VO-E). Die Liste des Anhang III enthält Bereiche, in denen die Nutzung von KI-Systemen ein potentiell hohes Risiko für die Grundrechte und die Sicherheit der von dem KI-System betroffenen Personen darstellen: Biometrik; kritische Infrastruktur; allgemeine und berufliche Bildung; Beschäftigung; Personalmanagement und Zugang zur Selbstständigkeit; Zugänglichkeit und Inanspruchnahme grundlegender privater und öffentlicher Dienste und Leistungen; Strafverfolgung; Migration, Asyl und Grenzkontrolle und Rechtspflege und demokratische Prozesse. Die Liste kann insoweit erweitert werden, als dass das hinzugefügte KI-System unter eine der acht bereits formulierten Bereiche fällt und ein vergleichbares Risiko für die Grundrechte oder die Sicherheit der betroffenen Personen darstellt (Art. 7 Abs. 1 KI-VO-E). Eine über diese acht Bereiche hinausgehende Erweiterung der Liste ist nicht möglich. Zu regulierende KI-Systeme werden damit dann als hochriskant bewertet und einem Konformitätsverfahren des Art. 43 KI-VO-E unterworfen, wenn es unter eines der genannten Bereiche subsumiert werden kann.

II. Das Konformitätsverfahren gem. Art. 43 Abs. 2 i.V.m. Anhang VI KI-VO-E

Der KI-VO-E formuliert unterschiedliche Konformitätsverfahren, die von den Anbieter*innen der Hochrisiko-KI-Systeme aufgestellt und durchgeführt werden sollen.
KI-Systeme, die als Hochrisiko-KI-Systeme bewertet werden, müssen den Anforderungen des Titel III Kapitel 2 des KI-VO-E genügen. Die Anbieter*innen dieser Systeme sind zum Nachweis der Einhaltung dieser Anforderungen verpflichtet und müssen ein Konformitätsverfahren durchlaufen (Art. 40 ff. KI-

VO-E). Die Ausführungen dieses Beitrags beschränken sich auf das Konformitätsverfahren für die sogenannten *stand-alone* KI-Systeme, die keinen harmonisierten Normen gem. Art. 40 KI-VO-E oder besonderen Spezifikationen gem. Art. 41 KI-VO-E unterliegen.
Die Vorschriften der Art. 8 ff. KI-VO-E formulieren Anforderungen zur Gewährleistung von KI-Systemen, die die Grundrechte und die Sicherheit der betroffenen Personen wahren sollen. So müssen sie beispielsweise ein Risikomanagementsystem aufbauen, mithilfe dessen sowohl bekannte und vorhersehbare Risiken ermittelt und analysiert werden als auch eine Bewertung möglicher weiterer, bisher unbekannter Risiken durchgeführt werden soll (Art. 9 Abs. 2 KI-VO-E). Darüber hinaus sollen die Datensätze, die den Ursprung der *bias*-Bildung[42] darstellen, mithilfe eines Daten-Governance-Systems reguliert werden (Art. 10 KI-VO-E). In Art. 10 Abs. 2-5 KI-VO-E werden besondere Kriterien für Trainings-, Validierungs- und Testdatensätze formuliert. So müssen diese beispielsweise gem. Abs. 3 „relevant, repräsentativ und soweit wie möglich fehlerfrei und vollständig" ausgestaltet sein, um einer Diskriminierung entgegenzuwirken. Außerdem muss eine technische Dokumentation der KI-Systeme gem. Art. 11 KI-VO-E vorgenommen werden.
Die Anbieter*innen der Hochrisiko-KI-Systeme müssen im Rahmen eines Konformitätsverfahrens die Einhaltung dieser Anforderungen der Art. 8 ff. KI-VO-E sicherstellen. Gem. Art. 48 KI-VO-E verpflichten sich die Anbieter*innen im Zuge einer EU-Konformitätserklärung dazu, die in der KI-VO festgelegten Vorschriften hinsichtlich der Hochrisiko-KI-Systeme einzuhalten. Mit dieser Konformitätserklärung erhält das KI-System eine CE-Kennzeichnung gem. Art. 49 KI-VO-E.

III. Effektivität des KI-VO-E hinsichtlich der Verhinderung von algorithmenbasierter Diskriminierung aufgrund des Geschlechts

Die Regulierung von KI-Systemen auf Grundlage der von ihnen ausgehenden Risiken stellt grundsätzlich einen geeigneten Ansatz für eine effektive Bekämpfung der algorithmenbasierten Diskriminierung wegen des Geschlechts dar. Der derzeitige Entwurf der Verordnung weist aber noch einige Lücken auf, die für einen umfassenden Schutz vor algorithmenbasierter Diskriminierung geschlossen werden müssen.

---

[42] dazu ausführlich s. Gössl (in diesem Band), S. 7 ff.

Zunächst ist anzumerken, dass der KI-VO-E die zu regulierenden KI-Systeme in starre Risikogruppen einteilt. Problematisch ist dabei zum einen, dass aufgrund der Natur selbstlernender Systeme mögliche Risiken erst im Laufe der Einsetzung des Systems entstehen und demnach zu Beginn nicht sichtbar werden. Zum anderen führt die nicht erweiterbare Liste des Anhang III dazu, dass neue KI-Systeme möglicherweise nicht den vorgegebenen Bereichen zugeordnet werden können.[43]
Auch das nur intern stattfindende ex-ante-Kontrollverfahren führt dazu, dass die KI-Systeme nur zu Beginn von den Anbieter*innen selbst überprüft werden, und keine regelmäßige Kontrolle durch externe Aufsichtsbehörden vorausgesetzt wird, was ebenfalls nicht mit der Natur selbstlernender Systeme vereinbar ist.[44]
Diese beiden Aspekte können zur Folge haben, dass KI-Systeme, die zunächst nur geringe Risiken für die Grundrechte der Betroffenen aufweisen, aber später im Laufe der Weiterentwicklung der Algorithmen hochriskante Praktiken entwickeln, keinem Konformitätsverfahren gem. Art. 43 KI-VO-E unterworfen werden. Zum anderen müssen KI-Systeme, die den in Anhang III aufgelisteten Bereichen vergleichbare, aber dort nicht aufgelistete Praktiken aufweisen, kein solches Verfahren durchlaufen.
Schließlich bietet der Verordnungsentwurf keinen ausreichenden Schutz vor Grundrechtsverletzungen. Der Entwurf basiert auf dem Prinzip der Ursachenverhinderung, d.h. es wird versucht, die Risikopotentiale, die zu einem verzerrten Ergebnis führen können, bereits zu Beginn zu minimieren. Nicht reguliert ist jedoch der Umstand, was geschieht, wenn KI-Systeme dennoch zu verzerrten Ergebnissen führen. Um den Betroffenen die Durchsetzung subjektiver Rechte wie Schadensersatz und Entschädigung zu ermöglichen, müssen Rechte wie ein Recht auf nichtautomatisierte Entscheidung und das Recht auf ausreichende Erläuterung gewährleistet werden.[45] Auch könnte überlegt werden, ob darüber hinausgehende Grundrechtsprüfungen vor dem Hintergrund des ungleichen

---

[43] Vgl. auch AlgorithmWatch An EU Artificial Intelligence Act for Fundamental Rights, 2021, Nr. 1.
[44] Vgl. auch AlgorithmWatch An EU Artificial Intelligence Act for Fundamental Rights, Nr. 3.
[45] Vgl. auch AlgorithmWatch An EU Artificial Intelligence Act for Fundamental Rights, Nr. 5.

Machtverhältnisses zwischen Anbieter*innen der KI-Systeme und den diesen Entscheidungen ausgesetzten Endnutzer*innen eingeführt werden könnten.

E. Fazit und Ausblick

Der KI-VO-E stellt einen ersten Schritt zu einer internationalen Regulierung algorithmenbasierter Entscheidungsmechanismen dar, ist aber noch an einigen Stellen auszubauen:
Statt starren Zuordnungen sollten flexible Risikogruppen formuliert werden, um durch einen dynamischen risikobasierten Ansatz sicherzustellen, dass KI-Systeme, sollten sie im Laufe ihrer Verwendung neue Risiken für die Grundrechte und die Sicherheit der Betroffenen bilden, einer anderen Risikogruppe zugeordnet und deren Voraussetzungen unterworfen werden können.
Um das von den KI-Systemen ausgehende Risiko entsprechend beurteilen zu können, sind außerdem regelmäßige und externe Kontrollmechanismen zu etablieren, mithilfe derer objektive und regelmäßige Konformitätsverfahren durchgeführt werden können.
Schließlich ist ein umfassender Grundrechtsschutz aufzubauen, um mithilfe einer Prüfung möglicher Verletzungen der Rechte der Betroffenen subjektive Rechte auf Schadensersatz und Entschädigung formulieren zu können.
Im Ergebnis bildet der KI-VO-E damit eine Grundlage für eine Regulierung der neuen Gefahren, die sich durch die algorithmenbasierten Entscheidungen der neuen Technologien für die Gleichberechtigung und die Grundrechte der Menschen gebildet haben. Werden die erwähnten Lücken geschlossen, kann damit ein umfassenderer Schutz als der der bereits bestehenden nationalen und internationalen Diskriminierungsverbote formuliert werden

.

Kapitel 5

Datenschutzrechtliche Anforderungen an diskriminierungsfreien KI-Einsatz

Jens Ambrock

Die Funktionsweise der KI steht im kaum auflösbaren Spannungsverhältnis mit den wesentlichen Grundprinzipien des Datenschutzrechts.[1] Damit steht dieses Rechtsgebiet zwar nicht alleine – auch Urheberrecht, Straßenverkehrsrecht, allgemeines Gleichbehandlungsrecht usw. sind nicht wirklich kompatibel mit den bahnbrechenden neuen technischen Möglichkeiten. Der Datenschutz ist jedoch das aktuelle Thema unserer Zeit, weil (nur) in diesem Rechtsgebiet momentan eine echte Regulierung künstlicher Intelligenz stattfindet. Das momentan bedeutsamste Tool ChatGPT wurde in der Folge in Italien wegen Datenschutzverstößen für rund einen Monat gesperrt. Wesentlicher Grund für dieses robuste Enforcement waren bereits Unvereinbarkeiten mit grundlegenden Anforderungen der Datenschutz-Grundverordnung (DSGVO). Ein weiterer Antrieb für die staatliche Untersagung bildeten diskriminierende Ergebnisse sowie diskriminierungsfördernde Intransparenz.

A. Grundlegende Anforderungen des Datenschutzrechts am Beispiel ChatGPT

Ende März 2023 hat die Datenschutzbehörde Garante per la Protezione dei Dati Personali (GPDP) angeordnet, dass der Dienst in Italien vorerst nicht mehr nutzbar gemacht werden durfte.[2] Die verantwortliche Stelle OpenAI ist der Verpflichtung nachgekommen, indem sie den Zugang gesperrt hat für Rechner, die

---

[1] Raji KI im öffentlichen Sektor, Verfassungs- und Datenschutzrechtlicher Rahmen für den Einsatz intelligenter Technologien durch den Staat, 2023 (im Erscheinen), S. 242.

[2] Pressemitteilung der GPDP v. 31.3.2023, abrufbar unter https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9870847.

anhand ihrer IP-Adresse als italienisch erkannt wurden. Die Anordnungsgründe lauteten:

1. Keine klar erkennbare Rechtsgrundlage für massenhafte Datenerhebung und Verarbeitung zum Training der KI
2. Verarbeitung unrichtiger Daten, da mitunter unrichtige Ergebnisse bei Anfragen zu natürlichen Personen ausgegeben werden
3. Keine Datenschutzinformationen, weder für Nutzer*innen noch für Personen, deren Daten im Datenbestand sind
4. Unzureichende Beachtung von Betroffenenrechten z.B. auf Auskunft, Löschung und Berichtigung
5. Verstöße gegen Jugendschutz wegen teilweise unangemessener Antworten des Chatdienstes (die italienische Datenschutzaufsicht ist zugleich für Jugendschutz im Internet zuständig)[3]

Nach rund einem Monat nahm die GPDP die Anordnung vorläufig zurück, nachdem OpenAI zahlreiche Sofortmaßnahmen ergriffen hatte.[4] Mit der Rücknahme ist weder eine Bestätigung der Rechtmäßigkeit verbunden noch ist gewährleistet, dass ChatGPT damit dauerhaft angeboten werden darf. Vielmehr hat die Aufsichtsbehörde damit bestätigt, dass ein für die Dauer der Detailermittlungen zunächst tolerabler Zustand hergestellt wurde. Die Zugeständnisse von OpenAI umfassen öffentliche Datenschutzinformationen inklusive einer detaillierten Beschreibung der Funktionsweise, Widerspruchsformulare gegen die Verwendung der Nutzeranfragen zu Trainingszwecken sowie gegen die Verwendung der Daten von EU-Bürger*innen für die Antworten des Chatdienstes sowie die Möglichkeit der Löschung unrichtiger Rohdaten.
Jede europäische Datenschutzbehörde ist in ihrem Territorium zuständig, weil der Anbieter keine Niederlassung im Europäischen Wirtschaftsraum hat.[5]

---

[3] In Deutschland sind die Landesmedienanstalten Aufsichtsbehörden für Medienjugendschutz; Ziff. 5 ist keine datenschutzrechtliche Anforderung, sodass der Beitrag auf Jugendschutz nicht weiter eingeht.

[4] Pressemitteilung der GPDP v. 28.04.2023, abrufbar unter https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9881490.

[5] Freund/Schmidt/Heep/Roschek/Ambrock DSGVO, 2023, Art. 55 Rn. 24.

Ebenfalls Prüfungen eingeleitet haben mehrere europäische Datenschutzaufsichten sowie Behörden in den USA und Kanada. Die deutschen Landesdatenschutzbehörden haben eine gemeinsame Prüfaktion mit einheitlichen Fragen an OpenAI gestartet und der Europäische Datenschutzausschuss hat eine Taskforce zur koordinierten Rechtsdurchsetzung eingesetzt.[6]

I. Rechtsgrundlage für die Verarbeitung

Im Rahmen ihrer zunächst summarischen Prüfung geht die italienische GPDP davon aus, dass keine Rechtsgrundlage im Sinne des Art. 6 DSGVO existiert, die den Dienst ChatGPT legitimiert. Künstliche Intelligenz nennt die DSGVO zwar an keiner Stelle, sie ist jedoch zumeist auf solche Dienste anwendbar.[7] Soweit eine KI personenbezogene Daten automatisiert verarbeitet, ist der Anwendungsbereich eröffnet und eine Rechtsgrundlage notwendig.[8] Dies betrifft sowohl das Training der KI als auch die spätere Nutzung.[9] Dient also eine KI lediglich der Wettervorhersage oder beispielsweise der Auswertung von Wellenbewegungen zur Tsunami-Warnung,[10] sind keine personenbezogenen Daten involviert. Dann bedarf es keiner weiterer datenschutzrechtlicher Überlegungen. Auch anonyme Trainingsdaten unterfallen nicht der DSGVO,[11] weshalb teilweise synthetische Daten für die Entwicklung von KI-Systemen verwendet werden, um den Anwendungsbereich zu umgehen.[12] Zumeist erfolgt das Training jedoch anhand von Echtdaten bzw. solchen, die zumindest einen mittelbaren Personenbezug aufweisen.[13]

---

[6] Pressemitteilung des EDSA vom 13.04.2023, S. 3, abrufbar unter https://edpb.europa.eu/news/news/2023/edpb-resolves-dispute-transfers-meta-and-creates-task-force-chat-gpt_en.

[7] Joos NZA 2020, 1216 (1216 f.); Sydow/Sydow/Marsch DSGVO BDSG, 3. Auf. 2022, Einl. Rn. 173.

[8] Tillmann/Vogt VuR 2018, 447 (449).

[9] Raji KI im öffentlichen Sektor S. 238 ff.

[10] Beispiel nach Tischbriek ZfDR 2021, 307 (315).

[11] Freund/Schmidt/Heep/Roschek/Strassemeyer/Quiel Art. 22 Rn. 20; Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, 2023, S. 39, 41; Lauscher/Legner ZfDR 2022, 367 (379).

[12] Kaulartz/Braegelmann Rechtshandbuch Artificial Intelligence und Machine Learning/Meents, 2020, Kap. 8.8. Rn. 45 ff; Raji DuD 2021, 303.

[13] Rostalski Künstliche Intelligenz/Hornung, 2022, S. 91 (92).

Ist der Anwendungsbereich eröffnet, greift das datenschutzrechtliche Verbotsprinzip.[14] Wenn eine KI personenbezogene Daten verarbeitet, ist das vom Grundsatz her verboten, solange keine Ausnahme greift.[15] Eine solche Ausnahme kann in einer Einwilligung oder einer gesetzlichen Grundlage bestehen. Es existiert keine spezifische Rechtsgrundlage für künstliche Intelligenz, sodass letztlich auf die allgemeinen Erlaubnistatbestände des Art. 6 DSGVO zurückgreifen muss,[16] die eine umfassende Einzelfallbetrachtung voraussetzen.[17]
Die Hambacher Erklärung der deutschen Datenschutzbehörden[18] unterstreicht die Bedeutung des Zweckbindungsgebots der jeweiligen Rechtsgrundlage. Danach haben Verwender*innen von Daten bereits bei Erhebung festzulegen und zu dokumentieren, wofür konkret die Daten benötigt werden. Von dieser Zweckbestimmung darf nur im Ausnahmefall anhand der Kriterien des Art. 6 Abs. 4 DSGVO abgewichen werden. Die Zweckbindung gehört zu den „Tragenden Säulen des Datenschutzrechts“,[19] stellt dabei aber eine besondere Herausforderung für KI dar.[20] Sie ist schon konzeptionell unvereinbar mit dem Grundkonzept von Big Data, bei dem zunächst eine große Datenbasis aufgebaut wird, ohne dabei festzulegen, welche Schlüsse daraus gezogen werden sollen.[21] Schließlich ist es in der Entwicklungsphase unvorhersehbar, welche Fragestellungen ein Dienst wie ChatGPT in der späteren Nutzungsphase zu bearbeiten hat. Art. 6 Abs. 4 DSGVO ermöglicht durchaus eine Zweckänderung, wenn der neue Zweck mit dem bisherigen vereinbar ist. Das kann jedoch nicht pauschal bejaht werden, sondern erfordert Einzelfallbetrachtung der konkreten Daten und der konkreten Zwecke.[22] Dieses differenzierte Vorgehen ist regelmäßig

---

[14] Karg DuD 2013, 75; Raji KI im öffentlichen Sektor S. 223.
[15] Malorny RdA 2022, 170 (173).
[16] Niemann/Kevekordes CR 2020, 17 (22 ff.); Rostalski Künstliche Intelligenz/Hornung, S. 91 (99).
[17] Zu Schadensersatz bei Verstößen gegen die DSGVO von Rochow (in diesem Band), S. 120 f.
[18] DSK Hambacher Erklärung zur Künstlichen Intelligenz v. 3.4.2019, abrufbar unter https://www.datenschutzkonferenz-online.de/media/en/20190405_hambacher_erklaerung.pdf.
[19] Bizer DuD 2007, 350 (352).
[20] Raji KI im öffentlichen Sektor S. 235.
[21] Jandt/Steidle Datenschutz im Internet/Ambrock, 2018, Teil A Rn. 13; Sydow/Sydow/Marsch DSGVO BDSG, 3. Auf. 2022, Einl. Rn. 174.
[22] Rostalski Künstliche Intelligenz/Hornung S. 91 (102).

problematisch, wenn eine Anwendung jedermann zur freien Verfügung steht und zudem nicht transparent ist, woher und aus welchem Zusammenhang die Daten stammen.

II. Richtigkeit

Art. 5 Abs. 1 lit. d DSGVO untersagt Diskriminierung durch verzerrte Datensätze.[23] Auch unwahre Daten zu einer Person sind personenbezogene Daten.[24] Dabei ist die informationelle Selbstbestimmung in besonderem Maße verletzt, wenn nicht nur Daten, sondern zugleich auch Unwahrheiten über ein Individuum im Umlauf sind. Falsch sind Ergebnisse, wenn sie objektiv nicht der Wahrheit entsprechen; wenn Ergebnisse unangemessen und dadurch diskriminierend sind, hilft das Gebot der Datenrichtigkeit nicht. Zudem verlangt die Vorschrift auch Aktualität des Datenbestands. Verantwortliche Stellen sind verpflichtet, angemessene Maßnahmen zur Gewährleistung der Datenqualität zu ergreifen, sowohl aktiv durch Überprüfung und Aktualisierung als auch passiv durch Berücksichtigung von Hinweisen.[25] Bekanntermaßen ist der Dienst ChatGPT-4 nicht durchgehend in der Lage, korrekte Aussagen zu Personen zu machen und baut für seine Antworten auf einem Datenbestand von Ende 2021 auf.[26]

Eine Umsetzung des Gebots der Datenrichtigkeit ist im Nachhinein kaum möglich, weil sich selbstlernende Systeme eigenständig weiterentwickeln. Wenn fehlerhafte Ergebnisse ausgeworfen werden, lässt sich bei einer KI nicht immer nachvollziehen, welcher Entwicklungsschritt zu korrigieren wäre.[27] Es besteht allenfalls die Option, die Rohdaten zu überprüfen und gegebenenfalls zu korri-

---

[23] Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, S. 39.

[24] Malorny RdA 2022, 170 (17).

[25] Freund/Schmidt/Heep/Roschek/Schmidt Art. 5 Rn. 37 f.; Raji KI im öffentlichen Sektor S. 241.

[26] Jahn Was die KI von ChatGPT alles kann, Handelsblatt v. 21.04.2023, abrufbar unter https://www.handelsblatt.com/technik/it-internet/chatgpt-was-die-ki-von-openai-alles-kann-/28941524.html.

[27] Spiecker gen. Döhmann/Papakonstantinou/Hornung/de Hert/Roßnagel/Richter GDPR, 2023 (im Erscheinen), Art. 5 Rn. 118.

gieren. Fehler passieren aber nicht notwendigerweise durch unrichtige Rohdaten, sondern auch im Rahmen des Trainings etwa durch ungeeignete Parameter oder nicht repräsentative Gewichtungen.[28]

III. Transparenz

ChatGPT hielt bis zu den Verhandlungen mit der italienischen Aufsicht keine brauchbaren Datenschutzinformationen vor. Wenn jedoch Daten zu Trainingszwecken übermittelt und genutzt werden – und das ist bei KI-Anwendungen so gut wie immer der Fall –[29] dann ist das klar zu benennen. Die Hambacher Erklärung[30] der deutschen Datenschutzbehörden zeigt daher die auf Art. 5 Abs. 1 lit. a, Art. 12 ff. DSGVO folgende Anforderung auf, dass KI transparent, nachvollziehbar und erklärbar sein muss.
Diese sehr eindeutig im Datenschutzrecht verankerte Anforderung ist für Betreiber*innen von KI kaum umsetzbar. Dies ist in ihrer Funktionsweise begründet. Wesensmerkmal einer KI ist die sehr selbständige Analyse, ohne dass Datenanalysten und Betreiber exakt nachvollziehen können, wie ein Ergebnis zustande gekommen ist.[31] Letztlich ist eine komplexe KI eine „black box", bei der das Zustandekommen einer Entscheidung wenig oder gar nicht erklärbar ist.[32] Der große Mehrwert solcher Maschinen ist gerade das Erkennen von Zusammenhängen im Datenbestand, die Menschen nicht ohne weiteres gesehen hätten.[33] Sie gibt jedoch für gewöhnlich nur das Ergebnis aus, nicht notwendigerweise die Herleitung dorthin. Zudem ist die rechtlich gebotene Information „zum Zeitpunkt der Erhebung" (Art. 13 Abs. 1 DSGVO) kaum umsetzbar. In

---

[28] Rostalski Künstliche Intelligenz/Hornung S. 91 (92).

[29] Joos NZA 2020, 1216 (1218).

[30] DSK Hambacher Erklärung zur Künstlichen Intelligenz, siehe oben Fn. 18.

[31] Vgl. Sydow/Sydow/Marsch Einl. Rn. 174; zu diesem Problem auch Gössl (in diesem Band), S. 12.

[32] Malorny RdA 2022, 170 (172); Schweighofer/Sorge/Borges/Schäfer/Weltl/Grabmair/Krupka Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, 2018, S. 30 f.; Spiecker gen. Döhmann/Papakonstantinou/Hornung/de Hert/Roßnagel/Richter GDPR, 2023 (im Erscheinen), Art. 5 Rn. 118; von Walter MMR-Beil. 2021 Heft 08, 22 (23).

[33] Ambrock Die Übermittlung von S.W.I.F.T.-Daten an die Terrorismusaufklärung der USA, 2013, S. 155; Chamoni/Budde Methoden und Verfahren des Data Mining, 1997, S. 9; Tischbriek ZfDR 2021, 307 (313).

der Regel ist dieser Zeitpunkt während der Trainingsphase des KI-Modells, in dem nicht feststeht, welche Schlüsse die KI künftig durch Verkettung von Rohdaten ziehen wird.[34]

Auch die inhaltlichen Anforderungen an die Datenschutzinformationen einer KI sind hoch. So ist in einfachen Worten adressatengerecht zu erklären, wie die KI arbeitet. Dabei ist nicht erforderlich, dass die Funktionsweise so verstanden wird, dass der Leser*innen die Software nachbauen könnten. Es geht vielmehr darum, die involvierte Logik abstrakt zu erklären, sodass die Gefahren deutlich werden. Raji vergleicht dies anschaulich mit der Funktionsweise des elektrischen Stroms:[35] Dessen Grundidee, Gefahren und Vorsichtsvorkehrungen lassen sich auch für Laien erklären, ohne dass auf sich abstoßender Elektronen eingegangen werden muss. Ähnlich sind die Anforderungen an die Erklärung einer KI. Es muss nachvollziehbar sein, warum ein bestimmter Input zu einem bestimmten Output führt.[36] Nutzer*innen und Betrachtungsobjekte der KI müssen daraufhin abschätzen können, welche Verhaltensänderungen beispielweise einer KI zur Erkennung der Kreditwürdigkeit zu einer positiveren Einschätzung verhelfen würden.[37]

Neben der öffentlichen Datenschutzinformation ist eine interne Datenschutz-Folgenabschätzung erforderlich. Dieses Risiko-Assessment ist unter anderem verpflichtend bei der Einführung neuer Technologien (Art. 35 Abs. 1 Satz 1 DSGVO). Dies beinhaltet insbesondere Konstellationenen, bei denen die gesellschaftlichen Auswirkungen noch nicht voll absehbar sind.[38] Auch an sich etablierte Techniken sind neu, wenn sie verbessert oder auf neue Art eingesetzt werden.[39] Zudem ist eine Datenschutz-Folgenabschätzung notwendig bei systematischer und umfassender Bewertung von Individuen auf Grundlage einer automatisierten Vereinbarung (Art. 35 Abs. 3 lit. a DSGVO). Dabei ist nicht

---

[34] Vgl. Lauscher/Legner ZfDR 2022, 367 (382).

[35] Raji KI im öffentlichen Sektor S. 226.

[36] Burrell How the machine „thinks", Understanding opacity in machine learning algorithms, 2016, Big Data & Society, 1 (1), 1; Raji KI im öffentlichen Sektor S. 228.

[37] Zur Perspektive der Informatik hierzu Rateike (in diesem Band), S. 34.

[38] Artikel-29-Datenschutzgruppe, WP 248 rev.01, S. 10; Spiecker gen. Döhmann/Papakonstantinou/Hornung/de Hert/Ambrock GDPR, 2023 (im Erscheinen), Art. 35 Rn. 17.

[39] Gellert The Article 29 Working Party's Provisional Guidelines on Data Protection Impact Assessment, 2 EDPL (2017), 212 (213).

entscheidend, ob Profile in Form von Datensätzen zu einzelnen Personen vorgehalten werden. Es genügt, dass während der einzelnen Verarbeitung verschiedene Angaben zu einer Person zusammengeführt werden.[40]

IV. Betroffenenrechte

Herr seiner Daten ist nur, wer nachvollziehen kann, was an welchem Ort über ihn gespeichert ist, und wer dabei Unwahrheiten korrigieren kann.[41] Hinsichtlich der Datenrichtigkeit sind Löschung und Berichtigung besonders wichtig. Erneut stößt das Grundkonzept der KI unauflösbar auf klare Anforderungen des Datenschutzrechts. Sind die Rohdaten in ein KI-Modell eingeflossen, ist kaum mehr nachvollziehbar, an welcher Stelle ein Fehler passiert ist, der durch nachträgliche Löschung oder Berichtigung getilgt werden könnte.[42] Bereits die Auskunft auf alles, was über die Trainings-Rohdaten hinaus geht, ist in der Umsetzung problematisch. Auskünfte zu personenbezogenen Daten nach erfolgten Verarbeitungen sind nicht umsetzbar, wenn nicht nachvollziehbar ist, wie die KI zu ihren Ergebnissen gelangt.[43] Dadurch sind die Betroffenenrechte nicht ausreichend auf das Szenario künstlicher Intelligenz abgestimmt.[44]
Nach Art. 15 Abs. 1 lit. h DSGVO umfasst der Auskunftsanspruch auch Aussagen über involvierte Logik. Dies betrifft jedoch nur die abstrakte Funktionsweise, sodass die Risiken und grundlegenden Entscheidungsfaktoren nachvollziehbar sind (siehe oben 3.). Nicht offengelegt werden muss der Algorithmus in der Weise, das wirklich nachvollzogen und vorhergesagt werden kann, welche Entscheidung die KI bei welcher Frage treffen wird.[45] Diese Detailtiefe wäre ein geschütztes Geschäftsgeheimnis[46] entsprechend der langjährigen Rechtsprechung zur Schufa.[47] Diese Einschränkung ist auch mit der DSGVO vereinbar,

---

[40] Spiecker gen. Döhmann/Papakonstantinou/Hornung/de Hert/Ambrock GDPR, 2023 (im Erscheinen), Art. 35 Rn. 23.
[41] Jandt/Steidle Datenschutz im Internet/Ambrock, Teil A Rn. 11.
[42] Sydow/Sydow/Marsch Einl. Rn. 175.
[43] Sydow/Sydow/Marsch Einl. Rn. 175.
[44] Kumkar/Roth-Isigkeit JZ 2020, 277.
[45] Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, S. 44; Lauscher/Legner ZfDR 2022, 367 (382); Lorenz VuR 2019, 213 (219).
[46] von Walter MMR-Beil. 2021 Heft 08, 22 (23).
[47] BGH 28.1.2014 – VI ZR 156/13, BGHZ 200, 38.

weil deren ErwGr 63 Satz 5 Geschäftsgeheimnisse und Urheberrechte an Software als Grenze anerkennt.[48] Einblick kann zwar in die Rohdaten verlangt werden, aber nicht in den Entscheidungsprozess ansich.[49] Zudem besteht nur eine Auskunftsanspruch über die eigenen Daten. Um die Arbeitsweise einer KI wirklich nachzuvollziehen, müssten alle Rohdaten vorliegen, aber die Gesamtheit der Rohdaten unterliegt dem Vertraulichkeitsinteresse der übrigen Betroffenen und der Geschäftsgeheimnisinhaber*innen.[50]

B. Diskriminierungsbeschränkende Regelungen

Die Ausführungen unter A. zeigen, dass KI-Dienste wie ChatGPT bereits mit den grundlegenden Kernanforderungen des Datenschutzrechts im kaum auflösbaren Konflikt stehen. Soweit Datenrichtigkeit und Transparenz betroffen sind, hat dies schon mittelbare Bezüge zu Diskriminierung. Die DSGVO weist jedoch auch konkrete Regelungen auf, die auf die Verhinderung von Diskriminierung gerichtet sind, jedoch nicht Gegenstand der italienischen Anordnung zu ChatGPT gewesen sind. Diskriminierungsbekämpfung ist zwar nicht der zentrale Gesetzeszweck des Datenschutzrechts, aber dennoch an diversen Stellen punktuell mitgeregelt.[51]

I. Ausstrahlung anderweitiger Diskriminierungsverbote auf das Datenschutzrecht

Zunächst haben Diskriminierungsverbote aus anderen Rechtsbereichen direkte Auswirkungen auf die Rechtmäßigkeit der zugrunde liegenden Datenverarbeitungen. Die meisten datenschutzrechtlichen Erlaubnisnormen enthalten Tatbestandsmerkmale wie „erforderlich“ oder „notwendig“ oder setzen wie Art. 6 Abs. 1 lit f DSGVO eine direkt benannte Güterabwägung voraus. Wenn eine KI diskriminiert, dann steht das der Abwägung im Rahmen der Rechtsgrundlage entgegen.[52] Ist eine Verarbeitung beispielsweise diskriminierend im Sinne

---

[48] Joos NZA 2020, 1216 (1218).

[49] Joos NZA 2020, 1216 (1219); Hoeren/Niehoff RW 2018, 48 (58).

[50] Rostalski Künstliche Intelligenz/Hornung S. 91 (109).

[51] Lauscher/Legner ZfDR 2022, 367 (379).

[52] DSK Hambacher Erklärung zur Künstlichen Intelligenz, siehe oben Fn. 18.

des § 1 AGG, darf sie nicht erfolgen. Sie ist damit auch nicht erforderlich, notwendig oder verhältnismäßig im Sinne der jeweiligen datenschutzrechtlichen Rechtsgrundlage.[53] Dasselbe gilt für Diskriminierungsverbote aus Grundrechten. Da die DSGVO unionsrechtliches Sekundärrecht ist, sind für ihre Auslegung die Grundrechte der GRCh heranzuziehen,[54] konkret Art. 20-23. GRCh.[55]
Die im vorangegangenen Beitrag von Yakar dargestellten Diskriminierungsverbote[56] haben daher unmittelbar auch datenschutzrechtliche Verarbeitungsverbote zur Folge. Wenn eine ohnehin verbotene Diskriminierungshandlung zugleich auch nach der DSGVO verboten ist, folgt daraus ein entscheidender Mehrwert: Im Datenschutzrecht existieren funktionierende, staatliche Regulierungsstrukturen, sodass eine flächendeckende Rechtsdurchsetzung möglich ist. Anderen Bereichen wie dem AGG oder den Gleichheitsgrundrechten stünde ohne die Annexfolge der DSGVO nur die individuelle Rechtsdurchsetzung im Einzelfall offen.

II. Schutz besonderer Datenkategorien

Besonders diskriminierungsgeeignete Daten sind besonders geschützt.[57] Art. 9 Abs. 1 DSGVO statuiert ein nur unter engen Voraussetzungen zu durchbrechendes Verbot für die Verarbeitungen von Informationen zur rassischen und ethnischen Herkunft, zu politischen Meinungen, religiösen oder weltanschaulichen Überzeugungen, Gewerkschaftszugehörigkeiten, genetischen Daten, biometrischen Daten, Gesundheitsdaten, zum Sexualleben und zur sexuellen Orientierung. Nur die direkte Anknüpfung an die Artikel-9-Daten ist verboten, nicht aber die mittelbare Anknüpfung.[58] Scheinbar neutrale Kriterien, die in engem Zusammenhang mit den eigentlich verbotenen Kategorien stehen, aber

---

[53] Forgó/Helfrich/Schneider Betrieblicher Datenschutz/Hanloser, 3. Aufl. 2019, Teil V Kap. 1 Rn. 60.
[54] Jandt/Steidle Datenschutz im Internet/Ambrock Teil A Rn. 4 f.
[55] Dazu Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, S. 33 f.
[56] Yakar (in diesem Band), S. 55 ff.
[57] Simitis/Hornung/Spiecker gen. Döhmann Datenschutzrecht/Petri, 2019, Art. 9 Rn. 10.
[58] Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, S. 42; Lauscher/Legner ZfDR 2022, 367 (380).

nicht unmittelbar die Kategorie selbst betreffen, dürfen unter den normalen Voraussetzungen der DSGVO verwendet werden. So ist es unzulässig, Angaben zur Zugehörigkeit einer ethnischen Volksgruppe zu verarbeiten, während Informationen zum Geburtsort oder der Staatsangehörigkeit möglich sind.[59] Das Verarbeitungsverbot klingt nachvollziehbar, um Diskriminierungen zu verhindern, bringt aber auch sinnvolle Dienste regelmäßig an ihre Grenzen. Soll beispielsweise eine KI wirksam Hate Speech erkennen – zum wünschenswerten Zweck der Eindämmung von Diskriminierung – dann muss sie Gesprächsinhalte inhaltlich analysieren, anstatt nur Schlüsselworte zu suchen. Hate Speech enthält jedoch regelmäßig Angaben über politische und weltanschauliche Anschauungen[60] oder zur (vermeintlichen) sexuellen Orientierung oder Ethnie.

1. Erlaubnistatbestände des Art. 9 Abs. 2 DSGVO

Art. 9 Abs. 2 DSGVO enthält einen Katalog von Erlaubnistatbeständen für die Verarbeitung von Daten der besonders geschützten Kategorien, doch die Ausnahmen sind eng und der Katalog abschließend. Danach ist stets die Erforderlichkeit für im Gesetz spezifisch benannte Zwecke Voraussetzung. Das kann bei einem frei verwendbaren KI-System, dem jede:r Fragen stellen kann, nicht gewährleistet werden.

2. Einwilligung

Der einzige für alle Zwecke und Bereiche breit einsetzbare Ausnahmetatbestand des Art. 9 Abs. 2 DSGVO ist die spezifische Einwilligung. Willigen die Betroffenen einer KI-Verarbeitung sensibler Daten ein, erlaubt das die Entscheidungsfindung durch die KI auch anhand diskriminierungsgeeigneter Kategorien. Auch die ggfs. ungewollte Diskriminierung ist dann von der legalisierenden Wirkung der Einwilligung umfasst.[61] Wenn also beispielsweise in ein Analysetool zum Recruiting eingewilligt wurde, dann verhindert Art. 9 nicht mehr die mögliche Ausländerdiskriminierung der KI.
Es genügt jedoch nicht die Einwilligung der Person, die der KI Fragen stellt. Auch alle Betroffenen müssten einwilligen. Das ist der entscheidende Punkt, warum die Einwilligung für KI-Systeme meist als Rechtfertigungstool nicht in

---

[59] Taeger/Gabel/Buchner DSGVO BDSG TTDSG, 4. Aufl. 2022, § 3 BDSG Rn. 59.
[60] von Walter MMR-Beil. 2021 Heft 08, 22 (26).
[61] Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, S. 42.

Frage kommt. Nur in abgegrenzten Projekten mit klarem Zweck und erreichbarem Betroffenenkreis ist die Einwilligung praktikabel. Wenn beispielsweise Patient*innen in die KI-gestützte Tumorerkennung einwilligt, dann lässt sich dies gut auf eine Einwilligung stützen. In einer solchen Konstellation ist es auch unproblematisch möglich, nach der Erlaubnis zu fragen, ob Röntgenbilder zum Training der KI per Einwilligung gespendet werden sollen. Wenn aber ein Programm wie ChatGPT sich autonom in frei verfügbaren Quellen bedient, dann müssten die dort Betroffenen einwilligen. Das sind im Zweifel alle noch lebenden Menschen, über die Ende 2021 Informationen im Internet abrufbar gewesen sind. Das wird nicht umsetzbar sein. Ein weiteres Praxisproblem der Einwilligung ist ihre jederzeitige und voraussetzungsfreie Widerrufbarkeit. Entscheiden sich einzelne Betroffene, ihre Einwilligung zurückzuziehen, sind deren Daten *ex nunc* nicht mehr zu verwenden. Das ist in der Umsetzung problematisch, weil in einem KI-Modell kaum nachvollziehbar ist, inwiefern die Trainingsdaten tatsächlich eingeflossen sind.[62]

3. Verzicht auf besondere Datenkategorien

Wenn kein Ausnahmetatbestand des Art. 9 Abs. 2 DSGVO greift, weil insbesondere die Einwilligung keine praktische Option darstellt, dann dürfen die sogenannten besonderen personenbezogenen Daten nicht verarbeitet werden. Diese simpel anmutende Lösung ist in der Praxis jedoch ebenfalls kaum oder nur mit hohem Aufwand realisierbar, weil viele besonders geschützte Daten in harmlos erscheinenden Datensätzen mitschwingen. So enthalten beispielsweise Fotografien von Personen stets auch implizit die Angabe, ob es sich um Brillenträger*innen handelt oder welche Hautfarbe sie aufweisen. Daraus lassen sich Informationen zum Sehvermögen, also zum Gesundheitszustand, sowie zur ethnischen Herkunft ableiten.[63] Als Beiwerk ist dies nicht zu beanstanden. Die Verwendung von Portraitbildern durch eine KI ist deshalb nicht pauschal verboten. Wichtig ist aber, dass die KI dieses Informations-Beiwerk nicht zur Grundlage seiner Mustererkennung und Entscheidung macht. Wenn die besonderen Kategorien in einem Kontext verwendet werden, in dem es speziell auf sie ankommt (z.B. Werbung an bestimmte Volksgruppe) oder speziell auf das Detail abgestellt

---

[62] Raji KI im öffentlichen Sektor, S. 223.

[63] Sydow/Marsch/Kempfert Art. 9 Rn. 6.

wird (z.B. Liste mit Einwohner*innen dunkler Hautfarbe), ist dies eine Verletzung des Art. 9 Abs. 1 DSGVO. Die KI muss davon abgehalten werden, die geschützten Merkmale für eine Mustererkennung zu verwenden, um daraus Schlüsse zu ziehen, dass z.B. Brillenträger*innen bevorzugt einzustellen sind oder die Hautfarbe ein Kriterium für die Bonität sein könnte. Es kommt also beim „Beifang" auf den Kontext der Verarbeitung an, während direkte Angaben der besonderen Kategorien wie z.B. Diagnosen (Sehstärke, Erkältung, HIV) steht Gesundheitsdaten sind.
Die KI hat folglich sehr komplexe Differenzierungen vorzunehmen, wenn sie die unzulässige Verarbeitung von Daten besonders geschützter Kategorien zu unterlassen hat. Ihr dies anzutrainieren, ist durchaus möglich. Es erfordert jedoch engmaschige Steuerung beim Training.

III. Automatisierte Entscheidungsfindung

Art. 22 DSGVO richtet sich gegen automatisierte Entscheidungen mit Rechtswirkungen für Individuen. Zweck ist eher die Verhinderung von Diskriminierung als der Schutz personenbezogener Daten.[64] Die Vorschrift richtet sich bei aller Technikoffenheit[65] insbesondere an KI-Systeme.[66] Es ist nicht der Inhalt einer Entscheidung, den Art. 22 DSGVO reguliert, sondern nur das Verfahren, wie die Entscheidung zustande kommt.[67] Zugleich wird auch nicht das komplette Verfahren Schranken unterworfen, sondern nur die abschließende Entscheidung selbst, also der letzte Schritt der Entscheidungsfindung. Insofern beschränkt Art. 22 DSGVO nicht die KI selbst, sondern nur das, was mit dem Output der KI geschieht.[68] Eine KI wird durch die Vorschrift also nicht daran gehindert, Personen zu diskriminieren. Wenn sie aber schon diskriminiert werden, dann soll das durch den Willen des Gesetzgebers nicht ausschließlich durch eine „kalte" Maschine erfolgen, sondern durch einen Menschen „mit Herz".

---

[64] Freund/Schmidt/Heep/Roschek/Strassemeyer/Quiel Art. 22 Rn. 18.
[65] Taeger/Gabel/Taeger Art. 22 DSGVO Rn. 28.
[66] Ienca/Pollicino/Liguori/Stefanini/Andorno/Bygrave, The Cambridge Handbook of Information Technology, Life Sciences and Human Rights, 2022, S. 166, 175; Raji KI im öffentlichen Sektor, S. 205.
[67] Lauscher/Legner ZfDR 2022, 367 (380).
[68] Vgl. von Walter MMR-Beil. 2021 Heft 08, 22 (24).

Der Unionsgesetzgeber bringt mit Art. 22 DSGVO ein reines Unbehagen gegenüber nichtmenschlichen Entscheidungen zum Ausdruck, regelt dabei aber nur seltene Extremfälle der völlig autonomen Entscheidung mit Rechtswirkung für einzelne Menschen.[69] Er greift damit eine empirisch nicht belegte, aber doch angenommene weit verbreitete Algorithmenphobie auf.[70] Dabei lässt sich durchaus infrage stellen, ob menschliche Entscheidungen tatsächlich vorzugswürdig sind gegenüber von Menschen trainierte künstliche Intelligenzen. Ob eine KI vorurteilsfreier agieren kann als Einzelpersonen, hängt entscheidend von der Qualität der Trainingsdaten und dem Engagement beim KI-Training ab. Art. 22 DSGVO differenziert insofern jedoch nicht, sondern erklärt alle automatisierten Entscheidungen für problematisch, soweit sie den Rechtsstatus von Menschen betreffen.

1. Verbot automatisierter Entscheidung mit Rechtswirkung

Aus Art. 22 Abs. 1 DSGVO folgt das subjektive Recht, nicht einer ausschließlich automatisierten Entscheidung unterworfen werden zu müssen.[71] Jedenfalls darf eine solche Entscheidung nicht ungeprüft und unkorrigierbar erfolgen.[72] Entscheidendes Kriterium ist, ob sie gegenüber einzelnen Menschen rechtliche Wirkung entfaltet oder sie in ähnlicher Weise erheblich beeinträchtigt. Beispiele für solche Rechtswirkungen sind der Abschluss, die Ablehnung oder die Kündigung eines Vertrags, nicht aber die automatisierte Preisdifferenzierung.[73] Auch staatliche Entscheidungen haben rechtliche Wirkung, wenn beispielsweise eine KI eine Baugenehmigung erteilt bzw. ablehnt, über eine Einbürgerung beschließt, Sozialleistungen gewährt oder ablehnt[74] oder ein gerichtliches Urteil fällt.

---

[69] Gola/Heckmann/Schulz DSGVO BDSG, 2. Aufl. 2022, Art. 22 Rn. 2; Schindler ZD-Aktuell 2019, 06647; Taeger/Gabel/Taeger Art. 22 Rn 9.

[70] Raji KI im öffentlichen Sektor, S. 24 ff.

[71] Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, S. 43.

[72] Joos NZA 2020, 1216 (1217).

[73] Lauscher/Legner ZfDR 2022, 367 (380).

[74] Spiecker gen. Döhmann/Papakonstantinou/Hornung/de Hert/Tambou GDPR, 2023 (im Erscheinen), Art. 22 Rn. 19.

Praktisch relevante Anwendungsfälle ergeben sich oft bei der Ablehnung eines Vertrags im Internet, weil eine KI einen Betrugsversuch vermutet.[75] So nehmen KI-Systeme mitunter diskriminierende Wertungen vor, indem sie beispielsweise Auslandsabrufe oder auch ungewöhnliches Surfverhalten bedingt durch gesundheitliche Einschränkungen oder hohes Alter als kriminelle Attacke einstufen. Wird zunächst nur der Zugang gesperrt und per Mail aufgefordert, das Passwort neu zu vergeben, ist der rechtliche Status unangetastet. Wird hingegen ein Nutzungsvertrag dauerhaft gekündigt oder eine Leistung verweigert, ist Art. 22 DSGVO einschlägig. Dasselbe gilt bei Alarmierungen wegen auffälliger Kamerabilder,[76] bei denen eine Analyse-KI kriminelles Verhalten vermutet. Soweit lediglich menschliche Security alarmiert wird, ist der rechtliche Status unverändert; wenn eine Tür automatisiert verschlossen wird, mag das im Einzelfall anders sein.

2. Ausnahmen für zulässige automatisierte Entscheidung

Automatisierte Entscheidungen sind nicht in jedem Fall verboten. So sind beispielsweise Smart Contracts zumeist erlaubt, weil sie im B2B-Bereich keine Entscheidungen über Individuen beinhalten. Darüber hinaus sieht Art. 22 Abs. 2 DSGVO Ausnahmen vom Verbot vor, wenn die Automatisierung durch Einwilligung, durch Vertrag oder durch mitgliedstaatliche Rechtsgrundlage vorgesehen ist. Diese Ausnahmen sind in ihrer Wirkung weitreichend, indem sie algorithmische Diskriminierung alleine dadurch erlauben, dass sie von vorneherein vorgesehen ist.[77]

Praktisch wichtigster Ausnahmefall ist die Erforderlichkeit der automatisierten Entscheidungsfindung zur Vertragsbegründung. Dies ist hoch relevant für durch KI abgelehnte Vertragsabschlüsse beim im Massengeschäft im Internet. Die automatisierte Entscheidung muss auch nicht den Hauptleistungsgegenstand der Vereinbarung bilden, sondern kann die Vertragsdurchführung erleichtern.[78] Das Kriterium ist dahingehend eng auszulegen, ob weniger invasive Verfahren möglich sind. Ein typisches Beispiel ist die Zahlartensteuerung, ob im Einzelfall Rechnungskauf angeboten wird, oder auch die Betrugsprävention im

---

[75] von Walter MMR-Beil. 2021 Heft 08, 22 (25).

[76] Tischbiek ZfDR 2021, 307 (313).

[77] Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, S. 44.

[78] von Walter MMR-Beil. 2021 Heft 08, 22 (25).

Internet. Es wäre nicht im Kundeninteresse, wenn der Kauf erst getätigt werden kann, nachdem die nächste freie Servicekraft verfügbar ist und Gelegenheit hatte, sich ein ernsthaftes Bild von der Ausfallwahrscheinlichkeit eines Kredits oder der Redlichkeit einer Käuferin zu machen. Wenn hier eine KI binnen Sekunden eine Entscheidung trifft, ist das zunächst im beiderseitigen Interesse. Wer durchs Raster fällt, dem bleibt das Recht auf Nachprüfung durch ein Individuum gemäß Art. 22 Abs. 3 DSGVO. In dem zwingend anzubietenden Verfahren besteht die Möglichkeit, dem fallbearbeitenden Menschen einen ergänzenden Standpunkt vorzutragen.

3. Menschliche Letztentscheidung

Die Praxisrelevanz des Art. 22 DSGVO ist bislang überschaubar. Die strengen Anforderungen lassen sich nämlich relativ leicht aushebeln, indem zur Letztentscheidung stets ein Mensch eingesetzt wird. Vorbereitende Systeme fallen nicht in den Anwendungsbereich der KI, solange sie nur eine Entscheidung empfehlen, ohne sie abschließend automatisiert zu fällen.[79] In der Praxis vertrauen weder Staat noch Unternehmen gewichtige Entscheidungen z.B. über das Personal ausschließlich einer Maschine an, ohne eine Fachkraft zumindest zu involvieren.[80]

Wichtig ist jedoch, dass die Mitwirkung des Menschen keine bloße Förmelei sein darf.[81] Es muss genug Spielraum für den Menschen geben, von der Maschine abzuweichen. Dafür muss die Person, die eine inhaltliche Bewertung des maschinellen Vorschlags vornimmt, entsprechend befugt und fachkompetent sein.[82] Sie benötigt auch die notwendigen zeitlichen Ressourcen, um den Entscheidungsfall tatsächlich überprüfen zu können. Wer nur wenige Minuten pro Sachverhalt zur Verfügung hat, wird keine andere Wahl haben, als stets auf „ok" zu klicken.

Selbst dann, wenn die Entscheidungskraft qualifiziert und entscheidungsbefugt ist, unterliegt sie oftmals dem sogenannten Phänomen „automation bias", bei dem der Maschine ein so hohes Vertrauen entgegengebracht wird, dass die

---

[79] Raji KI im öffentlichen Sektor, S. 207.

[80] Vgl. Joos NZA 2020, 1216; Lauscher/Legner ZfDR 2022, 367 (381).

[81] von Walter MMR-Beil. 2021 Heft 08, 22 (24).

[82] Taeger/Gabel/Taeger Art. 22 Rn. 14.

Individuen ihre Ergebnisse kaum anzweifeln.[83] Das Phänomen wird verstärkt, wenn der Lösungsweg einer KI nicht offen gelegt wird, sondern aus Big Data eine nicht näher begründete Schlussfolgerung abgeleitet und ausgegeben wird. Dann fehlt es der letztentscheidenden Person faktisch an den notwendigen Informationen, um das Ergebnis qualifiziert anzuzweifeln.[84] Kommt beispielsweise eine KI nach Analyse des kompletten Internets zu der Annahme, ein Kreditantragsteller werde trotz gutem Schufa-Score voraussichtlich seine Schulden nicht vollständig tilgen, dann hat die KI dafür eine größere Datenlage zur Verfügung als die sachbearbeitende Person jemals wird kognitiv erfassen können. Wenn die Person weiß, auf welche Information die Annahme fußt, seien es gesundheitliche Mutmaßungen, seien es ein bevorstehender Militäreinsatz, Drogenkonsum oder eine bevorstehende Scheidung – dann kann sie sich eine eigene Meinung bilden. Auch diskriminierende Kriterien kann die Person dann wieder „herausrechnen". Wenn ihr hingegen die Entscheidungsgrundlage fehlt, wird sie gegebenenfalls das KI-gestützte Ergebnis mittragen müssen.

Der Anwendungsbereich des Art. 22 DSGVO könnte sich durch eine bevorstehende Entscheidung des EuGH gegenüber der bisherigen Praxis deutlich ausweiten. Hintergrund sind die Schlussanträge des Generalanwalts Pikamäe hinsichtlich der Einstufung der Schufa.[85] Bisweilen war allgemein davon ausgegangen worden, dass die Errechnung eines prozentualen Wahrscheinlichkeitswertes für einen Zahlungsausfall für sich genommen nicht unter Art. 22 Abs. 1 DSGVO fallen kann, weil die Auskunftei dabei nichts entscheidet.[86] Sie gibt nicht einmal eine Empfehlung ab, ob ein Kredit vergeben werden soll. Jedes Kreditinstitut, das die Schufa nutzt, entscheidet eigenständig, bei wieviel Prozent seine „Schmerzgrenze" liegt. Der Generalanwalt hat nun seinen Standpunkt mitgeteilt, der Schufa-Score beeinträchtige betroffene Personen in einer in Weise, die einer rechtlichen Wirkung ähnlich sei.[87] Der Wert, den die Schufa errechnet, sei demnach auch eine Entscheidung, die nicht zwingend rechtlich sein

---

[83] Bahner, Übersteigertes Vertrauen in Automation, 2008, S. 40; Skitka/Moiser et al., Does automation bias decision-making? International Journal of Human-Computer Studies, 1999, 991; Raji KI im öffentlichen Sektor, S. 207.

[84] Vgl. Tischbriek ZfDR 2021, 307 (316).

[85] Schlussanträge v. 16.3.2023 – C-634/21, ECLI:EU:C:2023:220.

[86] Taeger/Gabel/Taeger Art. 22 Rn. 12.

[87] Schlussanträge (o. Fn. 84), Rn. 36.

müsse, sondern auch wirtschaftlicher Natur sein könne.[88] Das Scoring der Auskunftei bestimme die spätere Kreditablehnung vor und sei damit bereits als eigenständige Entscheidung einzustufen.[89] Folgt der EuGH der Wertung seines Generalanwalts, ist die Score-Berechnung der Schufa künftig als automatisierte Entscheidung anzusehen, weil dritte Wirtschaftsteilnehmer*innen nach ständiger gelebter Praxis den Wert für die Begründung, Durchführung oder Beendigung eines Vertragsverhältnisses zugrunde legen, ohne ihn nennenswert zu hinterfragen.
Daraus folgt für KI-Dienste, dass sie durchaus erlaubt sind, man ihnen aber nicht ohne weiteres die Letztentscheidung über Individuen überlassen darf. Bereits dann, wenn eine intransparente KI faktisch großen Einfluss hat, nimmt sie Entscheidungen über Menschen vor, die nur andere Menschen treffen dürfen.

IV. Scoring

Nach § 31 Abs. 1 BDSG darf ein Wahrscheinlichkeitswert über künftige Verhaltensweisen zur Vertragsbegründung oder -beendigung nur unter besonderen Bedingungen genutzt werden. Faktisch geht es um Bonitätsauskünfte von Auskunfteien mit Punktwerten für die Zahlungsausfallwahrscheinlichkeit eines Kredits oder Rechnungskaufs.[90] Aber auch andere Prognosen sind tatbestandlich umfasst. Damit sind genau die Mustererkennungen vom Anwendungsbereich des § 31 BDSG umfasst, für die KI prädestiniert ist. Es handelt sich um ein zusätzliches Verbot, eine Rechtsgrundlage (vgl. oben A.I.) wird darüber hinaus benötigt. Normadressat ist nicht die Stelle, die den Score berechnet, sondern wie schon bei Art. 22 DSGVO die verwendende Stelle, die darauf rechtliche Schritte aufbaut.[91] Auch hier handelt es sich also um keine Beschränkung dessen, was eine KI machen darf sondern lediglich eine Beschränkung dessen, wie Menschen das Ergebnis nutzen.
Der Wahrscheinlichkeitswert muss unter Zugrundelegung eines wissenschaftlich anerkannten mathematisch-statistischen Verfahrens ermittelt worden sein.

---

[88] Schlussanträge (o. Fn. 84), Rn. 38.
[89] Schlussanträge (o. Fn. 84), Rn. 42.
[90] Kühling/Buchner/Buchner DSGVO BDSG, 3. Aufl. 2020, § 31 Rn. 7.
[91] Abel ZD 2018, 103 (104); Plath/Kamlah DSGVO BDSG, 3. Aufl. 2018, § 31 Rn. 8, 13; Sydow/Marsch/Guggenberger § 31 Rn. 7.

Die Datenlage hat also inhaltlich richtig und aktuell zu sein[92] und das darauf aufbauende Verfahren hat nachweisbar korrekte Prognosen zu erstellen. Das wird eigentlich als sehr niedrige Hürde angesehen, die nur eine Selbstverständlichkeit abbildet.[93] Es ist aber hoch problematisch für „black boxes" und fehleranfällige, nicht ausgereifte KI.
Konkret verboten ist nach § 31 Abs. 1 Nr. 3 BDSG darüber hinaus reines Geoscoring ausschließlich auf Basis der Wohnanschrift einer Person.[94] Verstöße sind selten, weil das Verbot leicht umgangen werden kann. Schon das Hinzufügen eines weiteren Attributs genügt. Wenn Auskunfteien zu einer Person kaum weitere Merkmale als Namen und Anschrift kennen, dann wird im Zweifel der Wohnort mit dem Alter und Geschlecht kombiniert. Und wenn Alter und/oder Geschlecht unbekannt sind, werden beide Merkmale aus dem Vornamen abgeleitet. Der Vorname ist dafür zwar nur bedingt geeignet, aber es ist handelt sich auch nur um ein Wahrscheinlichkeitswert in Form einer geschätzten Prognose, für die eine KI, die ausschließlich aus Erfahrungswerten Schlüsse zieht, technisch gut geeignet ist. Diskriminierungen aufgrund des Vornamens werden durch § 31 BDSG nicht verhindert; hier besteht Reformbedarf.

C. Fazit

In besonderen Konstellationen kann das Datenschutzrecht Diskriminierung verhindern. Dies betrifft spezifische Analysen mittels Geoscoring, Gesundheitsdaten oder rein maschinellen Entscheidungen. Dabei handelt es sich jedoch nur um enge Schlaglichter in außergewöhnlichen Fällen; flächendeckende Diskriminierungsverhinderung kann der Datenschutz nicht gewährleisten.[95] Ob das

---

[92] BeckOK BDSG/Krämer, 43. Ed. 2023, § 31 Rn. 8i.
[93] Geberding/Wagner ZRP 2019, 116 (118).
[94] Zu sanktionierten Praxisfällen AG Hamburg v. 16.3.2017 – 233 OWi 12/17;
Paal/Pauly/Frenzel DSGVO BDSG, 3. Aufl. 2021, § 31 Rn. 7.
[95] Vgl. Lauscher/Legner ZfDR 2022, 367 (383).

Recht auf menschliche Entscheidung tatsächlich vorurteilsfreiere Ergebnisse erzielt,[96] kann auch bezweifelt werden.[97]
Datenschutzrecht ist schlichtweg nicht primär als Waffe gegen Diskriminierung gedacht.[98] Zum aktuellen Stand ist es das schlagkräftigste Rechtsgt, das gegen diskriminierende KI existiert. Ausreichend ist es jedoch nicht. Der Gesetzgeber ist hier, insbesondere bei der Ausgestaltung der KI-Verordnung, gefragt, adäquate Regelungen für faire intelligente Systeme zu schaffen.

---

[96] So aufgrund einer angenommenen Fehleranfälligkeit von KI Schwartmann/Jaspers/Thüsing/Kugelmann/Atzert DSGVO BDSG, 2. Aufl. 2020, Art. 22 Rn. 6; Taeger/Gabel/Taeger Art. 22 Rn. 9; s.a. Paal FS Teager, 2020, S. 331 (332 f.).
[97] So Freund/Schmidt/Heep/Roschek/Strassemeyer/Quiel Art. 22 Rn. 18.
[98] Rostalski Künstliche Intelligenz/Hornung, S. 91 (113).

Kapitel 6

Im Nebel: Der Schutz von algorithmischen Gruppen im deutschen Nichtdiskriminierungsrecht

Anna Kirchhefer-Lauber

A. Einleitung

Künstliche Intelligenz, genauer algorithmische Entscheidungssysteme, auf Englisch Algorithmic Decision-Making Systems (ADM-Systeme), die auf Verfahren maschinellen Lernens beruhen, werden sowohl im Vorfeld als auch bei Vertragsschluss immer mehr eingesetzt, um Entscheidungen über Menschen zu treffen. Diese Entscheidungen können tiefgreifende Auswirkungen auf deren persönliches Leben haben.[1] Im Rahmen der Kreditvergabe schätzen sie die Kreditwürdigkeit einer Person ein, bei Einstellungen beurteilen sie die Qualifikation und Geeignetheit der Bewerber*innen für die Stelle und im Bereich der personalisierten Werbung und Vertragsangebote erstellen sie Profile, um dann möglichst passgenaue, attraktive Angebote in bestimmten Situationen zu präsentieren, in denen die Annahmewahrscheinlichkeit optimal vorliegt.[2]
Es ist legitim, wenn Banken – ihrerseits den Aktionären und dem System verpflichtet – keine Kredite vergeben, die mit hoher Wahrscheinlichkeit nicht zurückgezahlt werden. Gleichermaßen verständlich ist die Absicht der Arbeitgeber*innen, die best-qualifizierten und persönlich passenden Kandidat*innen zu finden. Auch, dass Anbieter mit ihrer Werbung direkt die richtige Zielgruppe ansprechen möchten, ist nachvollziehbar. Die ADM-Systeme, die sie dabei einsetzen, treffen ihre Entscheidungen auf der Basis einer enormen Quantität an

---

[1] Orwat Diskriminierungsrisiken durch Verwendung von Algorithmen, 2019, S. 34-75.
[2] Xenidis/Gerards Algorithmic discrimination in Europe 2020, S. 83 ff.; diese und weitere Beispiele auch bei Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, 2023, S. 35 ff.

Daten, die sie durchforsten, um dann eine Vielzahl einzelner (unverdächtiger) Variablen zu korrelieren und auf Grundlage dieser „potenzierten Korrelation" Aussagen darüber zu treffen, ob oder zu welchen Konditionen einer Person ein Vertrag angeboten bzw. mit ihr geschlossen werden sollte oder nicht.[3] Das Potenzial der ADM-Systeme ist dabei enorm: Die Algorithmen nehmen Korrelationen vor und entdecken Muster, die für Menschen nicht oder nur in engen Grenzen nachvollziehbar sind.[4]

Dadurch entstehen als Produkt der angeforderten Prognosen zu den Zielfragestellungen neue algorithmische Personengruppen,[5] auch „ad hoc groups"[6] genannt. Innerhalb der algorithmisch entstandenen Personengruppen kann man wie folgt unterscheiden: zum einen Gruppen, deren gemeinsames Merkmal am Ende des Korrelationsprozesses von Menschen nachvollzogen werden kann, z.B. „Hundebesitzer", „traurige Teenager", „Alleinerziehende", „Videospieler" oder „Arme".[7] Diesen Gruppen ist gemein, dass sie nicht durch eine verpönte Kategorie im Sinne des europäischen oder deutschen Nichtdiskriminierungsrechts geschützt sind; zum anderen Gruppen, deren gemeinsames Merkmal im Nebel der potenzierten Korrelationen verbleibt und für das keine menschlich-sprachliche Ordnungsstruktur oder Begrifflichkeit existiert.[8] Solche Gruppen können sich beispielsweise aus der Gesamtschau und Korrelation von technischen Daten wie der Art der Mausbewegung über die Bildschirmfläche, Pixelanordnungen im Porträtfoto oder Ähnlichem ergeben, deren Aussagegehalt zur eingegebenen Zielfrage sich dem menschlichen Geist entzieht.[9] Entscheidend ist im Rahmen der Gruppenentstehung, dass sie für die Gesellschaft sozial unsichtbar

---

[3] Hoffman/Shahrian/Freitas Proceedings of Machine Learning Research 2014, 365 (371); Pohlmann et al. ZVersWiss 2022, 135 (144).

[4] Vgl. Kolleck/Orwat Mögliche Diskriminierung durch algorithmische Entscheidungssysteme und maschinelles Lernen – ein Überblick, 2020, S. 17 ff.

[5] Wachter 97 Tulane Law Review 2022, 149 (157).

[6] Mittelstadt 30 Philosophy and Technology 2017, 475.

[7] Federal Trade Commission Data brokers: A call for transparency and accountability, 2014.

[8] Wachter 35 Berkely Tech LJ 2020, 367 (414–418); De Vries 12 Ethics and Information Technology 2010, 71 (81).

[9] Burrell How the Machine "Thinks:" Understanding Opacity in Machine Learning Algorithms, Big Data & Society, 2016; Molnar Interpretable Machine Learning, 2020.

bleibt und daher auch keine demokratische Kontrolle der Gründe für die Differenzierung erfolgen kann. Das gilt umso mehr, weil sich die technischen Schritte der Gruppenbildung auch nicht *ex post* in menschlich fassbare Konzepte und Sprache überführen lassen.[10] Daraus ergibt sich ein grundlegendes Erklärbarkeits-, Rechtfertigungs- und Kontrollproblem.

Dass auch das technisch-mathematische Vorgehen von Algorithmen erhebliche Diskriminierungsrisiken birgt, ist mittlerweile bekannt[11] und hat einen umfassenden Diskurs nach sich gezogen.[12] Die europäische Kommission erachtet in ihrem Verordnungsvorschlag für einen „AI-Act" (= KI-VO-E) die Diskriminierungsfreiheit von ADM-Systemen als eine der zentralen Herausforderungen[13] ohne dabei auf die Rolle, Kompatibilität oder Eigenrationalität des europäischen Nichtdiskriminierungsrechts einzugehen.[14] Eine berechtigte Frage lautet nämlich: Welche Rolle soll und *kann* das deutsche (und auch das europäische) Nichtdiskriminierungsrecht überhaupt in seiner jetzigen Form bei der Regulierung diskriminierender ADM-Systeme übernehmen?[15]

Für das deutsche Nichtdiskriminierungsrecht enthält § 7 des Allgemeinen Gleichbehandlungsgesetzes (AGG) ein arbeitsrechtliches Benachteiligungsverbot und § 19 AGG ein allgemein-zivilrechtliches Benachteiligungsverbot. Beide verbieten Benachteiligungen aus rassistischen Gründen oder wegen der ethnischen Herkunft, des Geschlechts, der Religion oder Weltanschauung, einer Behinderung, des Alters oder der sexuellen Identität. Sowohl unmittelbare Benachteiligungen, wenn die Ungleichbehandlung direkt an einem der in § 1 AGG

---

[10] Mittelstadt 30 Philosophy and Technology 2017, 475; Grindrod Mathematical Underpinnings of Analytics: Theory and Applications, 2014.

[11] Association for Computing Machinery FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency/Suriyakumar et al., 2021, 723; Orwat Diskriminierungsrisiken durch Verwendung von Algorithmen.

[12] Hierzu vgl. Kolleck/Orwat Mögliche Diskriminierung durch algorithmische Entscheidungssysteme und maschinelles Lernen – ein Überblick; Koch Zeitschrift für Praktische Philosophie 2020, 265; Lauscher/Legner ZfDR 2022, 367; Oswald/Borucki Demokratietheorie im Zeitalter der Frühdigitalisierung/Langer/Weyerer, 2020, S. 219; Schnegg/Tschuggnall/Voithofer/Auer Inter- und multidisziplinäre Perspektiven der Geschlechterforschung/Horwath, 2022, S. 71.

[13] COM (2021) 206 final, 2.

[14] Vgl. auch Sesing/Tschech MMR 2022, 24.

[15] Eine ähnliche Frage stellt BMUV/Rostalski Künstliche Intelligenz/Müller, 2022, 205 (207).

genannten Merkmale ansetzt, als auch mittelbare Benachteiligungen, die aus scheinbar neutralen Kriterien resultieren, sind verboten. Direkte Diskriminierungen und indirekte Diskriminierungen durch ADM-Systeme sind in der Literatur bereits Gegenstand ausführlicher Erörterung gewesen und sollen daher hier nicht wiederholt beleuchtet werden.[16]
In den folgenden Zeilen möchte ich mich vielmehr auf die ethischen und rechtlichen Implikationen der Erzeugung und Verwendung *algorithmischer Gruppen* zur Entscheidungsfindung im allgemeinen Zivilrechtsverkehr konzentrieren, nachdem ich aufgezeigt habe, dass das deutsche Nichtdiskriminierungsrecht zum jetzigen Zeitpunkt nicht in der Lage ist, die damit einhergehenden Gefahren für die (Privat)Autonomie einzuhegen.

B. Die Differenzierung nach algorithmischen Gruppen de lege lata

Nach Art. 22 DSGVO hat eine Person das Recht, keiner ausschließlich auf einer automatisierten Verarbeitung beruhenden Entscheidung unterworfen zu werden, die ihr gegenüber rechtliche Wirkung entfaltet oder sie in ähnlicher Weise erheblich beeinträchtigt. Davon umfasst ist z.B. das sog. Profiling bei Bonitätsprüfungen für Kreditvergaben.[17] Die Vorschrift stellt jedoch nur sicher, dass nachteilige Entscheidungen nicht ausschließlich auf einer automatisierten Verarbeitung personenbezogener Daten beruhen, sondern es bei einer mensch-

---

[16] Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, 2023, S. 25 ff und 72 ff.; Koch Zeitschrift für Praktische Philosophie 2020, 265; Wagener/Stark Die Digitalisierung des Politischen/Rentsch, 2023, 23; Hilgendorf/Tiz Vom richtigen Umgang mit den „Anderen"/Lobe, 2022, 147; Brandstetter/Dobler/Ittstein Mensch und Künstliche Intelligenz/Schiedermeier, 2021, 11; Damar Chancen und Risiken von Künstlicher Intelligenz und Algorithmen aus antidiskriminierungsrechtlicher Perspektive, 2021; Hacker 55 Common Market Law Review 2018, 1143; zur Entstehung des Forschungsfeldes „Algorithmic Fairness": Hoeren/Sieber/Holznagel MMR-HdB/Kevekordes, 58. EL März 2022, Rn. 41.
[17] Scheer Algorithmen und ihr Diskriminierungsrisiko, 2019, 15; ausf. bereits Abrock (in diesem Band), S. 81 ff.

lichen Letztentscheidung bleibt: Der Einzelne soll nicht zum bloßen Objekt algorithmenbasierter Automatisierung werden.[18] Wird eine Algorithmus nur zur Vorbereitung einer menschlichen Entscheidung eingesetzt, greift die Verbotsnorm nicht.[19]

Durch die datenschutzrechtliche Regelung wird jedoch keine Aussage zum erlaubten Inhalt einer Entscheidung getroffen, die durch Algorithmen vorbereitet wurde.[20] Die Rechtmäßigkeit der Letztentscheidung kann de lege lata nur an den Anforderungen des AGG gemessen werden. Dabei wird schnell deutlich, dass für die meisten der in der Einleitung erwähnten Konstellationen bereits die Anwendbarkeit des AGG problematisch ist (I.) und auch die bereitgestellten Instrumente die Haftung für Entscheidungen nach unsichtbaren Merkmalen nicht erfassen (II.).

I. Anwendungsbereich des AGG

Die Haftung für Entscheidungen auf der Grundlage von *algorithmischen Gruppen* fällt schon nicht ohne Probleme in den Anwendungsbereich des AGG, da es gem. § 19 I Nr. 1 AGG nur auf Verträge Anwendung findet, die „typischerweise ohne Ansehen der Person zu vergleichbaren Bedingungen in einer Vielzahl von Fällen zustande kommen [. . .] oder bei denen das Ansehen der Person [. . .] eine nachrangige Bedeutung hat und die zu vergleichbaren Bedingungen in einer Vielzahl von Fällen zustande kommen". Das ist der Fall, wenn der Anbieter im Rahmen seiner Kapazitäten grundsätzlich mit jedermann abzuschließen bereit ist.[21] Bei der potenzierten Korrelation und algorithmischer Gruppenbildung handelt der Anbieter aber nicht ohne Ansehung der Person, weil ADM-Systeme auf der Basis immenser Datenmengen gerade dazu eingesetzt werden, bestimmte Personen aus der Masse an Menschen herauszufiltern oder durch Korrelationen ein „Mehr" über den Menschen zu erfahren als die reine „Ansehung der Person" ergeben würde. Durch die algorithmische Sortierung und Beurteilung wird versucht, die Person genau „kennenzulernen" und

---

[18] Martini/Nink NVwZ 2017, 681.

[19] Hoeren/Niehoff RW 2018, 48 (54).

[20] Ambrock (in diesem Band), S. 69 ff.; vgl. auch zum Ganzen Lauscher/Legner ZfDR 2022, 367 (380 ff).

[21] HK-AGG/Franke/Schlichtmann, 5. Aufl. 2022, AGG § 19 Rn. 30.

einzuschätzen.[22] Wenn Kreditgeschäfte, die meist auf einer individuellen Risikoprüfung beruhen, keine Massengeschäfte im Sinne des Gesetzes sein sollen,[23] dann erst recht nicht Geschäfte, deren Entscheidungen auf einer algorithmischen Einschätzung und Zuordnung der Person fußen. Dies gilt umso mehr, wenn man die jüngste Rechtsprechung des BGH beachtet. Danach gilt: „Enthält die Prüfung des Vertragsschlusses ein stark individualisiertes, personales Element, verzichtet das Gesetz im Rahmen des § 19 I Nr. 1 AGG zugunsten der persönlichen Willensbildung des Anbieters auf eine Benachteiligungskontrolle."[24] Der Anwendungsbereich ist daher in vielen Fällen, in denen ADM-Systeme eingesetzt werden, nicht eröffnet. Auf diese Problematik des fehlenden Schutzes vor Diskriminierung bei zunehmender Personalisierung ist schon früh aufmerksam gemacht worden[25] und sie ließe sich durch das Streichen des Merkmals „typischerweise ohne Ansehen der Person" beheben. Sodann ließe sich weiterfragen, ob bei einer Erweiterung des Anwendungsbereichs Schutz für unsichtbare Gruppen bestünde.

II. Entscheidungen auf Grundlage unsichtbarer Gruppen

§ 19 I AGG enthält ein allgemeines zivilrechtliches Diskriminierungsverbot in Bezug auf die in § 1 AGG genannten Merkmale bei der Begründung, Durchführung und Beendigung privatrechtlicher Schuldverhältnisse. Im Umkehrschluss ist es zulässig, für eine differenzierende Entscheidung auf Kriterien abzustellen, die weder unmittelbar noch mittelbar an die in § 19 I genannten Merkmale anknüpfen, solange ihnen nicht §§ 138, 242 BGB entgegenstehen. Ob ein Vertragsschluss mit einer bestimmten Person attraktiv und erstrebenswert ist, kann neben der Zahlungswilligkeit und Zahlungsfähigkeit von einer

---

[22] Kolleck/Orwat Mögliche Diskriminierung durch algorithmische Entscheidungssysteme und maschinelles Lernen – ein Überblick, S. 13.

[23] Amtliche Begründung BT-Drs. 16/1780, 42.

[24] BGH NJW 2020, 852; BGH NJW 2021, 2514: „Enthält die Prüfung des Vertragsschlusses ein stark individualisiertes, personales Element, verzichtet das Gesetz im Rahmen des § 19 I Nr. 1 AGG zugunsten der persönlichen Willensbildung des Anbieters auf eine Benachteiligungskontrolle."

[25] Grünberger Personale Gleichheit. Der Grundsatz der Gleichbehandlung im Zivilrecht, 2013, S. 609 – 614 mit umfassenden Nachweisen.

Vielzahl an weiteren spezifischen Umständen abhängen, die sich nicht nur aus Treu und Glauben, aus der Verkehrssitte oder der Natur des Schuldverhältnisses, sondern gerade auch aus der Person der potentiellen Vertragspartner*innen ergeben können.

Zulässige Differenzierungskriterien sind zunächst grundsätzlich alle Merkmale, die von § 19 I AGG nicht erfasst sind und die auf sozial üblichen sowie gesellschaftlich akzeptierten Vorstellungen beruhen.[26] Für den Bereich statistischer Vorprüfungen ist daher bislang gefordert worden, dass bei Prognoseverfahren im Rahmen der Entscheidungsfindung, ob eine Person für den Vertragsschluss in Frage kommt, ein rechtmäßiges Ziel vorliegen und das Verfahren zur Zielerreichung angemessen und erforderlich sein muss.[27] Für den Bereich der Kreditvergabe ist beispielsweise der Ausschluss vermuteter Bonitätsrisiken als rechtmäßiges Ziel anzusehen. Gleichwohl muss der Anbieter darlegen, inwieweit die Berücksichtigung und Gewichtung bestimmter Kriterien im Rahmen solcher Prognoseverfahren zur Bonitätsfeststellung geeignet und erforderlich sind.[28] Der Grund liegt auf der Hand: In der Auswahl dieser Variablen können Diskriminierungsursachen verborgen liegen.[29] Selbst wenn keine verbotenen Variablen im Sinne des § 1 AGG gewählt werden, können Diskriminierungen entstehen, wenn die ausgewählten Variablen Korrelationen zu verbotenen Variablen aufweisen.[30] Dann ist eine mittelbare Benachteiligung nach § 3 II AGG zu prüfen.

Wenn jedoch das ADM-System selbstlernend die Variablen auswählt und korreliert, die es für das rechtmäßige Ziel als relevant errechnet, bleibt im Nebel, welche Kriterien zum Ergebnis geführt haben und ob sie eine Korrelation zu den verbotenen Variablen aufweisen. Das durch den Einsatz von ADM-Systemen bekannte Transparenzrisiko verwirklicht sich ausnahmslos.[31] Ihm wäre nur durch eine umfassende Offenlegungspflicht des Verwenders zu begegnen, die

---

[26] HK-AGG/Franke/Schlichtmann AGG § 19 Rn. 33.

[27] Freyler NZA 2020, 284 (287).

[28] HK-AGG/Franke/Schlichtmann AGG § 19 Rn. 33 a.E.

[29] Barocas/Selbst California Law Review 2016, 671 (681); Dzida/Groh NJW 2019, 1917.

[30] Scheer Algorithmen und ihr Diskriminierungsrisiko, S. 12.

[31] BMUV/Rostalski Künstliche Intelligenz/Müller 205 (219).

jedoch bei den selbst-lernenden ADM-Systemen in der Regel an der fehlenden Erklärbarkeit der Entscheidungen scheitert.[32]

III. Normativ-ethische Bedenklichkeit *algorithmischer Gruppen*

Um zu ermitteln, wie diese Schieflage zu lösen ist, möchte ich mich auf die Vorfrage konzentrieren: Warum ist das beschriebene Szenario eigentlich vor dem weiteren Begründungshorizont des Nichtdiskriminierungsrechts bedenklich? Und aus welchen Gründen sollte ein solches Vorgehen nicht möglich sein? Dafür lohnt sich ein Blick in die Begründung und Rechtfertigung von Nichtdiskriminierungsrecht. Übergeordnete Zwecke können im Gleichheitsideal, dem Persönlichkeitsschutz, der Kompensation für historische Nachteile und der Verhinderung von Stereotypisierung gesehen werden.[33] Bei allen Legitimationsaspekten kommen unterschiedliche Gesichtspunkte zum Ausdruck, die in den Rechtswissenschaften und der Philosophie in der Vergangenheit unterschiedlichen Konzeptionen zugeordnet worden sind.[34] Man kann sie auch danach unterscheiden, ob sie primär dem Schutz vor symbolischer Ungleichheit dienen oder ob es ihnen um den Schutz vor Ungerechtigkeit im Einzelfall mit Folgeproblemen geht.

1. Sozial unsichtbare Differenzierung

Durch die potenzierte, selbst erlernte Korrelation unsichtbarer Merkmale können ADM-Systeme „verborgene Gesellschaftsstrukturen" nutzen und operationalisierbar machen, ohne sie dabei für den Menschen nachvollziehbar aufzudecken. Der Soziologe Armin Nassehi hat daher von einer „dritten Entdeckung der Gesellschaft" gesprochen.[35] Das ist problematisch, wurde doch bisher

---

[32] Mainzer Philosophisches Handbuch Künstliche Intelligenz/Waltl, 2020, S. 15 ff.; zur technischen Seite siehe Rateike (in diesem Band), S. 34 f

[33] Britz Einzelfallgerechtigkeit versus Generalisierung, 2008, S. 120 ff., 138 ff.

[34] Vgl. die Überblicke bei Grünberger Personale Gleichheit. Der Grundsatz der Gleichbehandlung im Zivilrecht, S. 536-559; Fredman Discrimination Law, 2. Aufl. 2011, S. 1 ff.; Lauber Paritätische Vertragsfreiheit durch reflexivem Diskriminierungsschutz, 2010, S. 89-94; Hepple/Szyszczak Discrimination: the Limits of the Law/Gardner, 1992, 148 (155); Mahlmann Elemente einer ethischen Grundrechtstheorie, 2008, S. 412 ff.

[35] Nassehi Muster: Theorie der digitalen Gesellschaft, 2019, 69.

die Aufgabe und das Ziel des Nichtdiskriminierungsrecht vor allem darin gesehen, bestehende hierarchische soziale Strukturen – sozial bedeutete in diesem Kontext auch überwiegend gesellschaftlich sichtbar – in Frage zu stellen und zu bekämpfen.[36] Der ethische Unwert der Diskriminierung ergab sich gerade aus der Perpetuierung *bestehender sichtbarer gesellschaftlicher* Strukturen der Exklusion und Hierarchisierung. [37] Mit Perpetuierungsrisiko beschreibt Jan-Laurin Müller den Effekt, dass ADM-Systeme bestimmte bestehende soziale Gesellschaftsstrukturen *aufgreifen*, ihre eigene technisch-normative Handlungslogik nach diesen Strukturen ausrichten und diese damit verfestigen.[38]
Die von den ADM-Systemen durch potenzierte Korrelationen erschaffenen Muster und Gruppen sind jedoch gesellschaftlich zu diesem Zeitpunkt nicht sichtbar. Eine auch von Lippert-Rasmussen als zentrales Merkmal der ethischen Bedenklichkeit von Benachteiligungen angesehene Voraussetzung ist damit nicht erfüllt: Die Ungleichbehandlung bezieht sich gerade nicht auf eine *socially salient group*.[39] Eine „sozial sichtbare Gruppe" liegt nach Lippert-Rasmussen vor „if perceived membership (...) is important to the structure of social interactions across a wide range of social contexts."[40] Die soziale Sichtbarkeit wird von ADM-Systemen bei der Vornahme der Korrelationen jedoch nicht berücksichtigt. Sie orientieren sich nicht an gesellschaftlichen Gruppen, sondern erschaffen *algorithmische Gruppen* rein aus Korrelationsergebnissen unüberschaubar vieler Variablen. Tal Zarsky hat formuliert:

> „The individuals are indeed part of a group, but one that is synthetic by nature, of unclear boundaries and structured by an algorithm. It is not a group that the individual feels a string affinity to, or that the public easily identifies as such."[41]

Das Risiko einer gesellschaftlichen Stereotypisierung oder Stigmatisierung einer Gruppe oder der Verfestigung einer solchen besteht somit zu dem Zeitpunkt der

---

[36] Grünberger Personale Gleichheit. Der Grundsatz der Gleichbehandlung im Zivilrecht, S. 530 ff.

[37] BMUV/Rostalski Künstliche Intelligenz/Müller 205 (213).

[38] BMUV/Rostalski Künstliche Intelligenz/Müller 205 (212).

[39] Lippert-Rasmussen Born Free and Equal? A philosophical inquiry into the nature of discrimination, 2014, S. 13-53.

[40] Ebd., S. 30.

[41] Zarsky Washington Law Review 2014, 1375 (1407).

Differenzierung nicht. Welche Risiken und normativen Bedenken birgt aber dann eine solche „Entdeckung" einer Gruppe ohne Gruppen*identität*[42] jenseits der menschlichen Erkenntnismöglichkeiten?

2. Sichtbare Folgeerscheinungen

Auch wenn die Bildung *algorithmischer Gruppen* im Zeitpunkt der Differenzierung nicht mit einer Stigmatisierung einhergeht, so könnte es doch sein, dass die Gruppe zu einem späteren Zeitpunkt Sichtbarkeit erlangt. Das hat folgenden Grund: Im Bereich der Entwicklung von ADM-Systemen arbeiten die Programmierer in den meisten Fällen mit Basismodellen, die dann nur leicht für den Kontext variiert werden.[43] Das birgt die Gefahr, dass Menschen von verschiedenen Anbietern und vor allem Kontext-übergreifend in bestimmte *algorithmische Gruppen* „sortiert" werden. Daraus resultieren verminderte Teilhabemöglichkeiten, die dann zu einem späteren Zeitpunkt auch gesellschaftlich sichtbar werden können: Kein Kredit, kein Eigenheim; kein Leasingvertrag, kein Auto; keine Zusatzversicherung, keine Zähne.
Beeinträchtigungen können daher insbesondere als Folgeerscheinungen in der Verminderung (ökonomischer) Entfaltungsmöglichkeiten bestehen. Insbesondere besteht diese Gefahr unabhängig davon, ob auch ein Risiko der symbolischen Ungleichheit im Zeitpunkt der Diskriminierung vorliegt. Wenn Lippert-Rasmussen fordert, dass sich eine Person ihrer Gruppenzugehörigkeit bewusst sein muss,[44] dann gewichtet er den Schutz vor symbolischer Ungleichheit höher als den Schutz vor folgender materieller Ungleichheit. Der europäische Gerichtshof hingegen erachtet weder die tatsächliche Zugehörigkeit zu einer Gruppe noch eine Gruppen*identität* als notwendige Voraussetzung, um eine Benachteiligung anzunehmen, weil er Teilhabemöglichkeiten des Individuums in den Vordergrund stellt.[45]

---

[42] Vgl. auch Hellman/Moreau Philosophical Foundations of Discrimination Law/Khaitan, 2013, 145.

[43] Liang et al. On the Opportunities and Risks of Foundation Models, 3. EL Juli 2022.

[44] Lippert-Rasmussen Born Free and Equal? A philosophical inquiry into the nature of discrimination, S. 34; a.A. Eidelson Discrimination and Disrespect, 2015.

[45] Z.B. EuGH ECLI:EU:C:2015:480 = BeckEuRS 2015, 452029 Rz. 60.

3. Autonomiedefizit

Bei der Benachteiligung durch die Mitgliedschaft in einer *algorithmischen Gruppe* geht es also nicht im engeren Sinne um die Gruppenzugehörigkeit an sich und die Gerechtigkeitsform der *group justice* als Ausprägung distributiver Gerechtigkeit, sondern um *individual justice* für eine Person, die – den Korrelationen durch ein ADM-System ausgeliefert – in ihrer Vertragsabschlussfreiheit oder bei den Vertragskonditionen benachteiligt wird.[46] Die ethische Bedenklichkeit ergibt sich somit aus einem durch den Output des ADM-Systems verursachten Autonomiedefizit der Person. Nur wer eine Auswahl an Handlungsmöglichkeiten hat, kann als autonome Person durchs Leben gehen.[47] Zu Recht wird daher auch in der deutschen Literatur darauf hingewiesen, dass Diskriminierungsverbote auch die Funktion haben, die tatsächlichen Voraussetzungen für die Inanspruchnahme der Vertragsfreiheit zu sichern.[48] Wenn Auswahlmöglichkeiten wegfallen oder beschränkt werden, weil ADM-Systeme eine Person aufgrund von Korrelationen in eine Gruppe einordnen, mit der ein Vertragsschluss vermieden wird oder zumindest zu schlechteren Konditionen stattfinden soll, resultiert dies in einem Autonomiedefizit der betroffenen Person. Gardner und Raz haben die Pflicht eines Menschen, nicht zu diskriminieren, aus der Autonomie und der Freiheit des anderen hergeleitet.[49] Das Nichtdiskriminierungsrecht hat zum Ziel, Menschen zu ermöglichen, ihre Autonomie und damit einhergehend ihre Vertragsfreiheit zu leben. Genau hier werden nun Individuen und aufgrund der systematischen Wiederholung – algorithmisch gesehen – ganze Gruppen vulnerabel, ohne dass dies menschlich bemerkt und kontrolliert wird. Man kann sich zum Beispiel eine Konstellation vorstellen, in der ADM-Systeme die Gruppenzugehörigkeit als relevant für die Zielfragestellung erachten und darüber hinaus weitere unsichtbare Korrelationen vornehmen: Ein Algorithmus wird mit dem Ziel eingesetzt, Werbung für ein Diätmittel in

---

[46] Zu group justice und individual justice aus deutscher Perspektive vgl. Lauber 2010; zu den Zielen des AGG, die Freiheit aller zu sichern, Verträge abschließen zu können: Amtliche Begründung BT-Drs. 16/1780, 39; Eichenhofer Teil II – Anlagen zum Wortprotokoll 171, S. 173, Protokoll Nr. 15/51.

[47] Raz Morality of Freedom, 1986, S. 203.

[48] HK-AGG/Franke/Schlichtmann, AGG § 19 Rn. 12; Looschelders JZ 2012, 105 (106).

[49] Hepple/Szyszczak Discrimination: the Limits of the Law/Gardner 148 (155); Gardner 9 Oxford J. Legal Stud. 1989, 1 (22); Raz, Morality of Freedom.

dem Moment der höchsten Kaufbereitschaft zu schalten. Der Algorithmus hat die Zielgruppe „traurige Teenager" erlernt (diese wäre für den Menschen noch nachvollziehbar) und korreliert nun in weiteren Schritten die entscheidenden Variablen der Gruppe mit weiteren Faktoren, z.B. dem Gebrauch der Wörter „hässlich" und „dick" in den Posts der Gruppenangehörigen, der Quantität negativer Emojis, der Art der Musik, die gestreamt wird, usw. so wie schließlich der Tageszeit, in der die Kaufbereitschaft der „traurigen Teenager + Variablen unbekannt" am höchsten ist. Durch diese Vorgehensweise entsteht eine neue Form der Macht- und Informationsasymmetrie. Die Quantität der verfügbaren Daten ermöglicht eine Personalisierung durch Korrelation, die einem Vertragspartner ein Informationsübergewicht verschafft, von dem die andere Partei keine Kenntnis hat und das zur Folge haben kann, dass ihre Wahlmöglichkeiten und damit auch ihre Autonomie eingeschränkt werden.

4. Einzelfallgerechtigkeit

Die unsichtbare Zuordnung zu *algorithmischen Gruppen* ist darüber hinaus mit Blick auf die Einzelfallgerechtigkeit problematisch. Das sind zunächst alle statistischen Differenzierungs- und Prognoseverfahren. Unter dem Stichwort des sog. Generalisierungsrisikos sind daher auch schon für sichtbare Gruppen die Kontextualisierungs- und Individualisierungsdefizite algorithmischer Entscheidungsprozesse diskutiert worden.[50] Neu ist, dass für Individuen nun nicht aufgrund bestimmter Merkmale die Zugehörigkeit zu einer *sozialen* Gruppe statistisch konstruiert wird, sondern dass nicht nachvollziehbare Konstruktionen von Gruppen und die Einordnung von Individuen anhand einer Vielzahl von Variablen erfolgt, die die Person dennoch nicht zutreffend abbilden oder charakterisieren müssen. Dieses Risiko wird dadurch verstärkt, dass die Methoden der meisten ADM-Systeme auf der Herstellung von Korrelationen statt Kausalität basieren. Korrelationen sind ein Hinweis aber kein Beweis für Ursachen- und Wirkungszusammenhänge. Es ist daher immer möglich, dass ADM-Systeme ihre Entscheidungsfindung auf eine sog. Scheinkorrelation stützen, auf einen statistischen Sachverhalt, bei dem es einen scheinbaren kausalen Zusammenhang zwischen korrelierenden Variablen gibt, der jedoch gerade nicht auf

[50] Britz 2008, S. 79 ff., 133 ff.; Zarsky Washington Law Review 2014, 1375 (1409); BMUV/Rostalski Künstliche Intelligenz/Müller 205 (214).

ein Ursache-Wirkungsprinzip zurückgeführt werden kann.[51] Der Zusammenhang kann demnach auf einer *conditio sine qua non* beruhen oder aber auch rein zufälliger Natur sein.[52] Diese Art der Entscheidungsfindung ist gegenläufig zu menschlichen Entscheidungs- und Rationalitätsstandards, denen in der Regel irgendeine Form der kausalen Begründung zugrunde liegt.[53] Für den einzelnen Menschen mutet die Entscheidung willkürlich an, weil er sie nicht nachvollziehen kann. Dem Prozess der Beurteilung durch das ADM-System kann nicht durch eine Eigendarstellung entgegengetreten werden. In dieser Hinsicht stellen *algorithmische Gruppen* auch eine Herausforderung für die Ziele des Persönlichkeitsschutzes und der Selbstbestimmung dar:

> "Algorithmic classification must therefore be considered a threat to data subjects' capacity to shape and control identity."[54]

IV. Jenseits von Demokratie

Schließlich wirft die unsichtbare Gruppenbildung durch Algorithmen immer das Problem auf, dass es keinen demokratischen Diskurs über Ungleichbehandlungen geben kann, deren Gründe im Nebel bleiben. Ist beispielsweise offengelegt, dass eine Benachteiligung bei der Kreditvergabe aufgrund eines niedrigen Bildungsabschlusses stattfindet, so ist es der Gesellschaft möglich, Erwägungen darüber anzustellen, ob diese Benachteiligung sachgerecht und normativ angemessen ist oder nicht.[55] Bei intransparenten Benachteiligungen durch ADM-Systeme können diese Reflexionen gar nicht vorgenommen werden. Da die Merkmale unklar sind, kann auch der Zusammenhang zwischen ihnen und dem Zweck der Ungleichbehandlung nicht beurteilt werden.
ADM-Systeme arbeiten – vereinfacht gesprochen – mit einem potenzierten statistischen Zusammenhang. In der Regel ist es durchaus möglich, dass bei einem statistischen Zusammenhang zwischen dem Zweck der Ungleichbehandlung und dem Merkmal der Ungleichbehandlung, die Ungleichbehandlung sachlich

---

[51] Zum Begriff der Korrelation und Scheinkorrelation: Online-Statistik-Lexikon: www.statista.com.
[52] Kim 58 William & Mary Law Review 2017, 857 (875).
[53] Lauscher/Legner ZfDR 2022, 367.
[54] Mittelstadt 30 Philosophy and Technology 2017, 475 (480).
[55] Beispiel nach Koch Zeitschrift für Praktische Philosophie 2020, 265 (269).

angemessen ist.[56] Diese beurteilt sich aber nicht nur nach der Verwendung signifikanter Korrelationen, sondern erfordert eine menschliche Bewertung der Angemessenheit, die zumindest eine Kenntnis der statistischen Zusammenhänge und der verwendeten Merkmale voraussetzt. Fehlt eine solche, ist jeder politische, gesellschaftliche und rechtliche Diskurs über die Angemessenheit von Differenzierungen von vornherein unmöglich.

V. Fazit und Ausblick

Das allgemeine zivilrechtliche Benachteiligungsverbot in § 19 AGG ist nicht geeignet Benachteiligungen aufgrund der Zuordnung zu *algorithmischen Gruppen* zu erfassen und eine Haftung für Benachteiligungen zu begründen. Die Benachteiligung wegen der Zuordnung zu einer bestimmten *algorithmischen Gruppe* wirft dennoch erhebliche ethisch-normative Bedenken auf, die sich insbesondere auf das Autonomiedefizit der durch das ADM-System beurteilten Person, ihren verminderten Personalitätsschutz und potenziell verminderte (ökonomische) Entfaltungsmöglichkeiten beziehen. Dem Grundproblem, dass die korrelierten Variablen unsichtbar sind, kann zunächst technisch begegnet werden, indem – soweit möglich – eine interpretierbare Künstliche Intelligenz bzw. „explainable artificial intelligence" (XAI)[57] entwickelt wird.

---

[56] Koch Zeitschrift für Praktische Philosophie 2020, 265 (270).

[57] Kamath/Liu Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning, 2021, S. 1-26; Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, S. 92.

Kapitel 7

Staatshaftung beim KI-Einsatz

Moritz von Rochow

A. Diskriminierende KI im Staatsdienst

Die Anwendungsbereiche für Künstliche Intelligenz[1] im Staatsdienst sind mannigfach – von Bauämtern[2] über Schulen[3] und Finanzämtern[4] bis hin zu Polizeibehörden.[5]
2022 hat eine Künstliche Intelligenz in neun französischen Regionen Satellitenbilder ausgewertet und so 20.356 illegale grundsteuerpflichtige Swimming-

---

[1] Zu Definition und Aufbau von KI vgl. Gössl/Yakar Geschlechterneutrale KI: Eine Handreichung, 2023, abrufbar unter: https://www.schleswig-holstein.de/DE/fachinhalte/G/gleichstellung/Downloads/handreichung_geschlechterneutrale_ki_lang.pdf?__blob=publicationFile&v=1; kurz auch Gössl (in diesem Band), S. 4 f.
[2] Zur Automatisierung im Baugenehmigungsverfahren, vgl. Roth-Isigkeit NVwZ 2022, 1253.
[3] Adams/Weale/Barr A-level results: almost 40% of teacher assessments in England downgraded, The Guardian 2020, abrufbar unter: https://www.theguardian.com/education/2020/aug/13/almost-40-of-english-students-have-a-level-results-downgraded.
[4] Pannett France uses AI to spot (and tax) undeclared swimming pools, Washington Post 2022, abrufbar unter: https://www.washingtonpost.com/world/2022/08/30/france-undeclared-swimming-pools-artificial-intelligence/; Klenk/Nullmeier/Wewer Handbuch Digitalisierung in Staat und Verwaltung/Djeffal, 2020, 51 (57).
[5] Sankin/Mehrotra/Mattu/Gilbertson Crime Prediction Software Promised to Be Free of Biases. New Data Shows It Perpetuates Them, The Markup 2021, abrufbar unter: https://themarkup.org/prediction-bias/2021/12/02/crime-prediction-software-promised-to-be-free-of-biases-new-data-shows-it-perpetuates-them.

Pools entdeckt, was sich auf entgangene und nun durchzusetzende Steuereinnahmen i.H.v. 10 Millionen US-$ summiert.[6] Auch im deutschen Steuerverfahren ist der Einsatz von KIs inzwischen Standard.[7]
Im Bereich behördlicher Auskünfte kann KI in Gestalt von ChatBots wertvolle Dienste leisten.[8] Der Bund hat für die Bundesverwaltung mehrere solcher Chatbots (Bundesbots) entwickelt,[9] die als „Basisdienst Chatbot" anderen Behörden bereitgestellt werden sollen. Die Auskünfte solcher Chatbots können allerdings nicht nur falsch und unvollständig sein, sie können auch diskriminieren.[10]

Gibt man etwa bei Kira, dem Karrierebot des ITZ Bund die Frage ein:

„Gibt es allgemeine Anforderungen für eine Tätigkeit beim ITZ?",

generiert der Bot die folgende Antwort:

„Neben dem erfolgreichen Bestehen der Probezeit als Tarifbeschäftigte oder Tarifbeschäftigter und dem Besitz der deutschen bzw. einer EU-Staatsbürgerschaft gehören hierzu [...]."

Außer Acht lässt der Chatbot Kira, dass gemäß § 7 Abs. 1 Nr. 1 b BBG der angestrebte Beamtenstatus nicht nur Deutschen und EU-Bürger*innen offensteht, sondern auch Personen aus dem EWR, also aus Norwegen, Island und Liechtenstein. Sieht eine Isländerin aufgrund der falschen und diskriminieren-

---

[6] Pannett France uses AI to spot (and tax) undeclared swimming pools, Washington Post 2022, abrufbar unter: https://www.washingtonpost.com/world/2022/08/30/france-undeclared-swimming-pools-artificial-intelligence/.
[7] Klenk/Nullmeier/Wewer/Djeffal Handbuch Digitalisierung in Staat und Verwaltung, 2020, 51 (57).
[8] Ebd.
[9] Mehr dazu auf der Seite des Informationstechnikzentrums Bund, Informationstechnikzentrum Bund, Bundesbots erleichtern die Kommunikation, 2022, abrufbar unter: https://www.itzbund.de/DE/itloesungen/standardloesungen/chatbots/chatbots.html.
[10] Ebd, 19 ff.

den Äußerung des ChatBots von einer Bewerbung ab, können sich bei Vorliegen weiterer Voraussetzungen Schadensersatzforderungen, z.B. aus § 15 AGG ergeben.[11]

Aufgrund der Corona-Pandemie wurden im Jahr 2020 die A-Level (Abitur) und GCSE (Mittlere Reife) Prüfungen in Großbritannien abgesagt. Infolgedessen entschied man sich dafür, die von den Lehrer*innen vergebenen Vornoten zugrunde zu legen und diese sodann algorithmisch an den Mehrjahresnotenschnitt anzupassen. Hatte eine Schule in der Vergangenheit zehn „A"s vergeben, nun aber zwölf Schüler*innen mit „A", vorbenotet, wurden die schlechtesten beiden automatisch auf „B" oder sogar auf „C" herabgestuft.[12] Im Ergebnis wurden so 39,1% der Prüfungsresultate[13] herabgestuft. Zahlreichen Schüler*innen wurde durch eine KI der Zugang zu ihrer Wunschhochschule verbaut.[14]

Im Ergebnis fielen die KI-berechneten Abschlussnoten an traditionell leistungsstarken Privatschulen doppelt so hoch aus wie an öffentlichen Gesamtschulen.[15] Der Unterschied zwischen den von den Lehrer*innen vergebenen Vornoten und dem algorithmenbasierten Examensergebnis traf damit v.a. Schüler*innen mit niedrigem sozioökonomischen Status und Migrationshintergrund.[16] Ist in Folge einer solchen KI-Benotung der Wunschstudienplatz nach erfolgreicher Notenanfechtung anderweitig vergeben, stellen sich Fragen der Staatshaftung.

Ein weiteres Beispiel, welches Fragen hinsichtlich der Staatshaftung für diskriminierende KI aufwirft ist die Verwendung von Pre-Policing-KI.[17] Mit Urteil

---

[11] Zu Äußerungen im Chat als Indiz für eine Diskriminierung, vgl. LAG Schleswig-Holstein NJW 2022, 2951.

[12] Lamont The student and the algorithm: how the exam results fiasco threatened one pupil's future, The Guardian 2021, abrufbar unter: https://www.theguardian.com/education/2021/feb/18/the-student-and-the-algorithm-how-the-exam-results-fiasco-threatened-one-pupils-future.

[13] Adams/Weale/Barr The Guardian 2020.

[14] Lamont The Guardian 2021.

[15] Naughton From viral conspiracies to exam fiascos, algorithms come with serious side effects, The Guardian 2020, abrufbar unter: https://www.theguardian.com/technology/2020/sep/06/from-viral-conspiracies-to-exam-fiascos-algorithms-come-with-serious-side-effects.

[16] Adams/Weale/Barr The Guardian 2020.

[17] Zum Predictive Policing ausführlich Thimm/Thimm-Braun (in diesem Band), S. 37 ff.

vom 16. Februar 2023 hat das Bundesverfassungsgericht der Verwendung von Pre-Policing-KI in Deutschland enge Grenzen gesetzt:

> „Besonderes Eingriffsgewicht kann je nach Einsatzart die Verwendung lernfähiger Systeme, also Künstlicher Intelligenz (KI), haben. Deren Mehrwert, zugleich aber auch ihre spezifischen Gefahren liegen darin, dass nicht nur von den einzelnen Polizistinnen und Polizisten aufgegriffene kriminologisch fundierte Muster Anwendung finden, sondern solche Muster automatisiert weiterentwickelt oder überhaupt erst generiert und dann in weiteren Analysestufen weiter verknüpft werden. Mittels einer automatisierten Anwendung könnten so über den Einsatz komplexer Algorithmen zum Ausweis von Beziehungen oder Zusammenhängen hinaus auch selbstständig weitere Aussagen im Sinne eines „predictive policing" getroffen werden. So könnten besonders weitgehende Informationen und Annahmen über eine Person erzeugt werden, deren Überprüfung spezifisch erschwert sein kann. Denn komplexe algorithmische Systeme könnten sich im Verlauf des maschinellen Lernprozesses immer mehr von der ursprünglichen menschlichen Programmierung lösen, und die maschinellen Lernprozesse und die Ergebnisse der Anwendung könnten immer schwerer nachzuvollziehen sein (vgl. EuGH, Urteil vom 21. Juni 2021, Ligue des droits humains, C-817/19, ECLI:EU:C:2022:491, Rn. 195). Dann droht zugleich die staatliche Kontrolle über diese Anwendung verloren zu gehen."[18]

B. Schäden durch Staats-KI

Die vorstehenden Beispiele zeigen, dass die staatliche Verwendung von KI zu Schäden führen kann[19]: Ein fehlerhaft beschiedener Bauantrag, lässt einen Grundstückseigentümer vom Bau eines Gewerbebetriebs Abstand nehmen, wodurch ihm Einnahmen entgehen; die Anfechtung eines französischen Grundsteuerbescheids erzeugt Anwaltskosten; der BundesBot hält eine Bewerberin davon ab, sich auf eine Stelle zu bewerben; eine fehlerhaft erteilte Auskunft führt für ein Unternehmen zu einem kostspieligen Bürokratiemehraufwand; ein Studienplatz wird wegen einer KI-Schulnote anderweitig vergeben; der Wert einer Immobilie sinkt, weil Pre-Policing-motivierte Polizei-Patrouillen einen „Chilling-Effect" erzeugen.[20] Daneben sind auch immaterielle Schäden denkbar.[21]

---

[18] BVerfG NJW 2023, 1196.

[19] Staudenmayer NJW 2023, 894, 895.

[20] Hierzu auch Thimm/Thimm-Braun (in diesem Band), S. 44 ff.

[21] Paal NJW 2022, 3673.

Unter den zahlreichen rechtlichen und ethischen Fragen, welche der staatliche Einsatz von KI aufwirft, ist daher die Frage nach der Haftung für technisches Versagen und KI-induzierte Diskriminierung eine ganz zentrale.[22] Im Völkerrecht vollzieht sich eine ähnliche Diskussion um den Einsatz autonomer bewaffneter Drohnen.[23]
Um den Staat für die von seiner KI begangenen Fehler und Diskriminierungen in Haftung zu nehmen, sind im deutschen Staatshaftungsrecht im Wesentlichen zwei Pfade vorgesehen: Zum einen lässt sich ein Haftungsanspruch gegen den Staat auf den Amtshaftungsanspruch nach § 839 BGB i.V.m. Art. 34 GG stützen. Zum anderen können allgemeine und spezielle Enteignungs- und Aufopferungsansprüche herangezogen werden.
Beide Herangehensweisen haben Vor- und Nachteile und beide Herangehensweisen erfordern ein gewisses Maß an KI-spezifischer Rechtsfortbildung.

C. Amtshaftungsanspruch § 839 BGB i.V.m. Art. 34 GG

Nach dem Preußischem Recht waren Beamt*innen für Rechtsbrüche grundsätzlich persönlich haftbar, wie dies der Formulierung des § 839 BGB noch heute zu entnehmen ist.[24] Durch die Haftungsüberleitung des Art. 34 GG wurden jene Ansprüche auf die handelnde Körperschaft übergeleitet. Der Regelungszweck einer solchen Haftungsüberleitung ist es, die Entscheidungsfreude

---

[22] Gössl/Yakar Geschlechterneutrale KI: Eine Handreichung, abrufbar unter: https://www.schleswig-holstein.de/DE/fachinhalte/G/gleichstellung/Downloads/handreichung_geschlechterneutrale_ki_lang.pdf?__blob=publicationFile&v=1, 74.

[23] Frau, Vereinte Nationen, 61/3, 2013, 99; Frau, HuV 1 (2018) 5; Frau/Frau Drohnen und das Recht, 2014, 235; Wagner, Miami Legal Studies Research Paper, No 15-12, 2015, 1(3-7); Boothny/Boothy New Technologies and the Law in War and Peace, 2019, 137; Heyns/Akande/Hill-Cawthrone/Chengeta International and Comparative Law Quarterly, 65/4, 2016, 791; Cebul New Perspectives in Foreign Policy 14 (2017) 43; Schellenberger StudZR-WissOn 20 (2019), 291; Geiss Die völkerrechtliche Dimension autonomer Waffensysteme, 2015; Rauch Autonome Waffensysteme und Völkerrecht, 2020; Frau/Moewes Drohnen und das Recht, 193.

[24] Conrad/Grünewald/Kalscheuer/Milker Öffentlich-rechtliches Äußerungsrecht-HdB/von Rochow, 2022, 334.

der Beamt*innen zu fördern, damit die Angst vor persönlicher Haftung nicht das staatliche Handeln dirigiert. Dieser Zweck kann mit Blick auf KI nur dann verfolgt werden, wenn zwischen KI und Bürger:in ein Mensch tritt, welcher die KI mediatisiert. Dies ist v.a. dann der Fall, wenn die KI lediglich intern unterstützend verwendet wird. Pre-Policing-KI etwa erfordert immer noch eine*n physische*n Beamt*in, welche:r die KI-Erkenntnisse z.B. durch Platzverweise und Meldeauflagen umsetzt. Anders ist dies im Falle behördlich verwendeter ChatBots: Die Interaktion findet hier unmittelbar zwischen KI und Bürger:in statt, ohne dass eine natürliche Person zwischengeschaltet wäre.

I. Beamte

Erste dogmatische Schwierigkeiten, die unmittelbar mit Bürger*innen interagierende KI dem Amtshaftungsanspruch zu unterwerfen, begegnen hinsichtlich des Begriffs des „Beamten". So verlangt § 839 BGB ausdrücklich:

> „Verletzt ein Beamter vorsätzlich oder fahrlässig die ihm einem Dritten gegenüber obliegende Amtspflicht, so hat er dem Dritten den daraus entstehenden Schaden zu ersetzen."

Unter den haftungsrechtlichen Beamtenbegriff fallen nicht nur Statusbeamt*innen, sondern auch Personen, die in einem sonstigen öffentlich-rechtlichen Dienstverhältnis oder in einem privatrechtlichen Dienstverhältnis zu einer öffentlich-rechtlichen Körperschaft stehen.[25] Können aber auch Computer hierunter subsummiert werden?
Gleich ob nach statusrechtlichem oder haftungsrechtlichem Beamtenbegriff, scheint auf den ersten Blick eine natürliche Person zwingend erforderlich zu sein. Daher halten Teile der Kommentarliteratur das Amtshaftungsrecht für unanwendbar, wenn sich die öffentliche Gewalt technischer Einrichtungen bedient.[26]
Klar ist, dass der Beamtenbegriff jedenfalls dann nicht zum Problem wird, wenn es sich bei der KI lediglich um behördeninterne entscheidungsunterstützende Systeme handelt.[27]

---

[25] MüKoBGB/Papier/Shirvani, 8. Aufl. 2020, BGB § 839 Rn. 183.
[26] Ebd. Rn. 191.
[27] Martini/Ruschemeier/Hain VerwArchiv 2021, 1 (18).

Aber auch in Fällen der direkten Interaktion zwischen KI und Bürger:in greift ein allzu wörtliches Verständnis des Begriffs „Beamter" angesichts solcher zum Zeitpunkt des Inkrafttretens des BGB nicht vorstellbaren technischen Entwicklungen zu kurz. § 839 BGB muss heute stets im Zusammenhang mit Art. 34 GG gelesen und verstanden werden:[28]

> „Verletzt jemand in Ausübung eines ihm anvertrauten öffentlichen Amtes die ihm einem Dritten gegenüber obliegende Amtspflicht, so trifft die Verantwortlichkeit grundsätzlich den Staat oder die Körperschaft, in deren Dienst er steht."

Art. 34 GG macht mehr, als nur die Passivlegitimation vom Amtswalter auf den Amtsträger zu verlagern.[29] § 839 BGB und Art. 34 GG sind derart eng verknüpft, dass sie sich wechselseitig bedingen.[30] Die Norminterpretation muss daher stets in der Zusammenschau beider Regelungen erfolgen, welche auf diese Weise eine einheitliche Anspruchsgrundlage bilden.[31]

Nach Art. 34 GG ist es gerade nicht erforderlich, dass ein „Beamter" handelt, sondern dass „jemand in Ausübung eines ihm anvertrauten öffentlichen Amtes die ihm einem Dritten gegenüber obliegende Amtspflicht verletzt." Vor diesem Hintergrund macht es für den letztlich haftenden Staat keinen Unterschied, ob diesem das Handeln eines*r „echten" Beamt*in zugerechnet wird, oder einer künstlichen Intelligenz. Der Begriff „jemand" in Art. 34 GG ist deutlich interpretationsoffener als der Begriff „Beamter".
Die Argumentation, der Begriff „jemand" verlange zwingend eine natürliche Person, verkennt die Entwicklung dieses Begriffs in der Praxis. Denn schon lange versteht die Rechtspraxis unter „jemand" jegliche öffentlich-rechtliche und privatrechtlichen Dienstverhältnisse jenseits des Beamtenstatus'. Dieser haftungsrechtliche Beamt*innenbegriff ist mit einem strengen Wortlautverständnis des § 839 BGB, der ursprünglich den Statusbeamten im Blick hatte, kaum zu vereinbaren und lässt sich dogmatisch nur durch eine verfassungskonforme Gesamtschau der haftungsrechtlichen Vorschriften begründen.

---

[28] Von Mangoldt/Klein/Starck/Danwitz, 7. Aufl. 2018, GG Art. 34 Rn. 55.
[29] Ebd.
[30] Ebd.
[31] Ebd.

Sinn und Zweck der Norm des § 839 BGB i.V.m. Art. 34 GG ist es, den Bürger*innen gegen den rechtswidrig handelnden und diesen schädigenden Staat einen Anspruch zu verleihen. Mit diesem Schutzzweck verträgt es sich nicht, den Staat hinter einem anachronistischen Wortlautschleier zu verstecken. Zumindest eine analoge Übertragung auf Künstliche Intelligenzen ist angezeigt, angesichts einer offensichtlichen KI-betreffenden Regelungslücke bei vergleichbarer Interessenlage. „Jemand" im Sinne des § 839 BGB i.V.m. Art. 34 GG ist damit die handelnde Körperschaft, ungeachtet dessen, ob sie sich einer physischen Person oder einer künstlichen Intelligenz bedient.

II. Ausübung eines öffentlichen Amtes

Eine weitere Voraussetzung für die Haftung des Staates aus § 839 BGB i.V.m. Art. 34 GG ist die Ausübung eines öffentlichen Amtes. Ob eine KI-generierte Äußerung in Ausübung eines öffentlichen Amtes erfolgt, richtet sich nach dem damit verfolgten Zweck und Funktionszusammenhang.[32] Ist eine eindeutige Zuordnung nicht möglich, kann die Ausübung eines öffentlichen Amtes widerleglich vermutet werden.[33]

III. Verletzung einer drittbezogenen Amtspflicht

Der Begriff der drittbezogenen Amtspflichten ist weit zu verstehen und umfasst auch das Verwaltungsinnenrecht[34] und damit Konstellationen, in denen die KI ohne direkte Interaktion mit Bürger*innen verwaltungsunterstützend tätig wird. Teilweise wird vertreten, dass es sich bei reinem Innenrecht formal nicht um eine Amtspflicht i.S.d. Staatshaftungsrechts handeln könne.[35] Diese Ansicht verkennt jedoch, dass es nach dem Schutzzweck nicht auf den formalen Status als Innen- oder Außenrecht ankommen kann, sondern allein auf den materiellen Inhalt.[36] So kann z.B. auch eine interne Weisung, die formal den Status

---

[32] Conrad/Grünewald/Kalscheuer/Milker/ Öffentlich-rechtliches ÄußerungsR-HdB/von Rochow, 336.
[33] Ebd. 337.
[34] Ebd.
[35] MüKoBGB/Papier/Shirvani, BGB § 839 Rn. 245; Erman/Mayen, 16. Aufl. 2020, BGB § 839 Rn. 48, 50.
[36] Conrad/Grünewald/Kalscheuer/Milker Öffentlich-rechtliches ÄußerungsR-HdB/von Rochow, 338.

von Innenrecht hat, eine drittgerichtete Amtspflichtverletzung sein, wenn sie z.B. die Weitergabe von Daten Dritter zum Gegenstand hat. Auch der verwaltungsinterne Output informationstechnischer Systeme kann die Qualität einer dienstlichen Weisung haben[37] und im Falle einer materiellen Drittwirkung Amtshaftungsansprüche auslösen.
Es kann also nicht darauf ankommen, ob die KI formal mit Außenwirkung eingesetzt wird oder nicht. Maßstab muss vielmehr sein, ob ihr Einsatz materiell Außenwirkungen entfaltet. Solche materiellen Außenwirkungen generiert z.B. das von einer Pre-Policing-KI geübte Racial Profiling, ungeachtet dessen, dass die KI-generierten Informationen zunächst lediglich behördenintern ausgegeben werden, woraufhin ein Mensch dann konkrete Maßnahmen, wie z.B. Personenkontrollen, einleitet.
Insbesondere bei der Verwendung von ChatBots kommen Amtspflichten zum Tragen: Eine ungeschriebene Amtspflicht ist etwa die Pflicht, Auskünfte, Mitteilungen und Bescheinigungen richtig, klar, unmissverständlich und vollständig zu erteilen.[38] Selbst durch eine Hinterlegung inhaltlich korrekter Textbausteine könnten die behördlichen Auskunfts- und Mitteilungspflichten nur unzureichend erfüllt werden, da die Behörden auch eine positive Pflicht haben, Bürger*innen davor zu bewahren, sehenden Auges „ins offene Messer“ zu laufen. Die Pflicht, nicht wissentlich Falsches zu sagen, wandelt sich dann in eine positive Hinweis- und Belehrungspflicht,[39] der auch ein ChatBot gerecht werden muss.

IV. Verschulden

§ 839 BGB verlangt für die Staatshaftung grundsätzlich Vorsatz oder Fahrlässigkeit. Dieses Erfordernis gilt als eines der schwierigsten Probleme bei der Staatshaftung für Machine-Learning und neuronale Netze. Ohne weiteres ist ein Verschulden gegeben, wenn Amtswalter*innen ein Überwachungsverschul-

---

[37] Martini/Ruschemeier/Hain VerwArchiv 2021, 1 (20); der dort vertretenen Ansicht, dass der formal interne Charakter einer dienstlichen Weisung einer Amtspflichtverletzung widerspreche, ist aus genannten Gründen nicht zuzustimmen. Näher dazu: Conrad/Grünewald/Kalscheuer/Milker Öffentlich-rechtliches ÄußerungsR-HdB/von Rochow, 338.
[38] Ebd, 341.
[39] Ebd.

den trifft oder wenn bei der Entscheidung, ob ein KI-System überhaupt verwendet werden darf, Abwägungsfehler passiert sind.[40] In letzterem Fall besteht die Amtspflicht im Sinne einer Erforschungspflicht[41] in der Garantenstellung, den Sachverhalt und die technische Wirkungsweise der KI zu erkunden und zu verstehen, dergestalt, dass die Beurteilungs- und Entscheidungsgrundlage nicht in wesentlichen Punkten zum Nachteil der Betroffenen unvollständig bleibt. Im Rahmen der erstmaligen Zulassungs- oder Einsatzentscheidung können Zuverlässigkeitsüberprüfungen und Testläufe geboten sein, deren Unterlassen ein Verschulden begründet. In besonders grundrechtssensiblen Bereichen kann der Einsatz von KI gänzlich ausgeschlossen sein.[42]

Nach der Erstentscheidung über den grundsätzlichen Einsatz bestehen Verkehrssicherungs- und Überwachungspflichten.[43] Es handelt sich hierbei um eine Daueraufgabe zu überprüfen, wie ein technisches System die Aufgabe verrichtet.[44] Diese Verkehrssicherungspflicht soll dem Bedürfnis Schadensfolgen nach Risikosphären zuzurechnen, Rechnung tragen.[45]

Eine haftungsrechtliche Anknüpfung nicht an den generellen KI-Einsatz, sondern an die einzelne KI-generierte Entscheidung gestaltet sich nach klassischem Amtshaftungsrecht schwieriger, ist aber für den Rechtsschutz der Bürger*innen unerlässlich. Diese wissen nämlich nur, wer für Outputfehler passivlegitimiert ist, nicht aber, wer den Input generiert hat.

Bei komplizierten Formen von KI, vor allem denen, die auf maschinellem Lernen beruhen, ist es außerdem sehr schwierig, nachzuvollziehen, wie es zu einem schadensträchtigen Output gekommen ist.[46] Im Gegensatz zu einfacheren Formen von KI, die nur nach bestimmten Vorgaben lernen, sind höher entwickelte KIs in der Lage, sich diese Vorgaben selbst zu kreieren.[47]

---

[40] Martini/Ruschemeier/Hain VerwArchiv 2021, 1 (13).

[41] Ebd.; BGH NJW 1989, 99.

[42] Martini Grundlinien eines Kontrollsystems für algorithmenbasierte Entscheidungsprozesse, 2019, 20.

[43] Martini/Ruschemeier/Hain VerwArchiv 2021, 1 (14).

[44] Ebd.

[45] Ebd.

[46] Staudenmayer NJW 2023, 894 (895).

[47] Ebd.

Bei Anwendung des klassischen Amtshaftungsrechts, würde diese „Blackbox“[48] zum Freibrief für Behörden werden, die sich immer darauf berufen könnten, für die computerinternen selbstlernenden Prozesse kein Verschulden zu tragen. Bei selbstlernenden Systemen stößt der Verschuldensnachweis damit an Grenzen. Solche Systeme zeichnen sich gerade dadurch aus, Operationen vorzunehmen, die die Systementwickelnden nicht vorhergesehen haben und konnten.[49] Gerade die die diskriminierungsanfälligen[50] Feedback-Schleifen, etwa bei rekurrenten neuronalen Netzen,[51] sind derart dynamisch, dass nachträgliche Kontrollen sich zwangsläufig auf veraltete Momentaufnahmen beziehen müssen.[52]
Diesem Umstand kann nur durch eine verschuldensunabhängige Haftung oder jedenfalls eine widerlegliche Verschuldensvermutung begegnet werden. Dogmatisch konstruiert wird eine solche teilweise durch das im Enteignungs- und Aufopferungsrecht anerkannte Institut der Gefährdungshaftung (Dazu in Abschnitt D). Da das Aufopferungsrecht aber einen äußerst engen Anwendungsbereich hat und überdies begriffliche Unschärfen aufweist, ist eine KI-bezogene Rechtsfortbildung des Amtshaftungsrechts vorzugswürdig.
Zu klären ist schließlich die Frage der Beweislast. Entsprechend dem für die Produzent*innen- und Ärzte*innenhaftung entwickelten Rechtsgedanken der Verantwortungssphären kann es den betroffenen Bürger*innen nicht aufgebürdet werden, dem Staat Fehler in der KI nachzuweisen, insbesondere, wenn sie es mit lernenden Systemen zu tun haben, in deren „Blackbox“ selbst die verwendende Behörde keinen Einblick mehr hat. Andernfalls würde die Beweislastverteilung den Einsatz von IT-Systemen und undurchschaubaren Systemarchitekturen zu einem faktisch nicht überprüfbaren Spielraum der Behörde transformieren.[53]

---

[48] Zur Problematik auch Gössl (in diesem Band), S. 12.
[49] Martini/Ruschemeier/Hain VerwArchiv 2021, 1 (31); Staudenmayer, Ebd.
[50] Steege MMR 2019, 715 (716).
[51] Als künstliche neuronale Netze bezeichnet man ein mathematisches Modell, das sich durch Eingaben selbst verbessern kann, vgl. Klenk/Nullmeier/Wewer/Djeffal Handbuch Digitalisierung in Staat und Verwaltung, 2020, 51 (52).
[52] Martini/Ruschemeier/Hain VerwArchiv 2021, 1 (31).
[53] Ebd (29).

V. Subsidiarität der Amtshaftung

Eine weitere Hürde für Bürger*innen, gegenüber dem Staat Amtshaftungsansprüche geltend zu machen, ist die in § 839 Abs. 1 S. 2 BGB normierte Subsidiarität der Amtshaftung. Diese besagt, dass ein Anspruch gegen den Staat ausgeschlossen ist, wenn der Verletzte auf andere Weise Ersatz zu erlangen vermag und dem Beamten nur Fahrlässigkeit zur Last fällt. Verlockend ist es für den Staat, KI-geschädigte Bürger*innen auf die meist zivilen Softwarehersteller zu verweisen, in deren Sphäre der Fehler fallen mag. Als Anspruchsgrundlagen stehen den Geschädigten gegen diesen z.B. § 1 Abs. 1 S. 1 ProdHaftG, § 823 Abs. 1 S. 1 BGB in der Ausprägung der Produzent*innenhaftung[54] und spezialgesetzliche Normen wie z.B. §§ 33 ff. LuftVG zur Verfügung.[55] Ein solcher Verweis auf das hoheitliche Haftungsprivileg greift aber zu kurz, da das Softwareunternehmen als Verwaltungshelfer – ähnlich einem Bauunternehmen im Straßenbau – dieses Haftungsprivileg gleichfalls für sich in Anspruch nehmen kann. Wenn aber auch der Konkurrenzanspruch dem Haftungsprivileg unterliegt, kann sich der Staat hierauf nicht haftungsbefreiend berufen.

D. Enteignungs- und Aufoperungsansprüche

Zu den Enteignungs- und Aufopferungsansprüchen gehören v.a. die spezialgesetzlich normierten Ansprüche im Polizei- und Ordnungsrecht. Durch diese Haftungsnormen sind die richterrechtlich entwickelten, allgemeinen Entschädigungsinstitute wegen enteignungsgleichen und aufopferungsgleichen Eingriffs für den speziellen Bereich des Polizei- und Ordnungsrechts normiert worden.[56] Dort, wo spezielle landesrechtlichen Haftungsnormen fehlen, greift der enteignungsgleiche Eingriff als Anspruchsgrundlage ein.[57]

---

[54] Die europäische Produkthaftungsrichtlinie soll nach einem Kommissionsvorschlag an die Digitalisierung allgemein, und KI im Besonderen angepasst werden.

[55] Giemulla/van Schyndel/Hajda Haftung und künstliche Intelligenz (KI) in der Luftfahrt, ZLW 3/2022, 349 (354).

[56] MüKoBGB/Papier/Shirvani BGB § 839 Rn. 117.

[57] Ossenbühl/Cornils, Staatshaftungsrecht, 5. Teil. Der Anspruch wegen rechtswidriger Eigentumsverletzung (enteignungsgleicher Eingriff), beck-online.

Eine solche Anspruchsgrundlage - § 39 OBG NRW – war Grundlage der vom BGH entwickelten Rechtsprechungslinie zur Haftung für fehlerhafte technische Einrichtungen. Bekannt ist sie als „feindliches-Grün-Rechtsprechung“[58]: Eine automatisierte Verkehrsampel an einer Kreuzung zeigt in alle Richtungen „grün“, woraufhin es zum Unfall kommt. Hatte der BGH in früheren Jahren eine Staatshaftung zu Gunsten der Verkehrsteilnehmenden mangels Unmittelbarkeit noch abgelehnt, so akzeptiert er seit dem Jahr 1986 im Falle fehlerhafter technischer Einrichtungen eine verschuldensunabhängige Gefährdungshaftung des Verwenders technischer Systeme.[59]

Gegenüber dem Amtshaftungsanspruch haben Enteignungs- und Aufopferungsansprüche damit für Geschädigte den Vorteil, verschuldensunabhängig zu sein. Es handelt sich um eine „öffentlich-rechtliche Gefährdungshaftung“.[60] Das früher streng gehandhabte Unmittelbarkeitserfordernis der Schädigung wurde mit der Rechtsprechung zum „feindlichen Grün“[61] gelockert, dahingehend, dass es bei technischen Einrichtungen genügt, wenn sich im Schadenseintritt eine für die konkrete hoheitliche Betätigung typische Gefahrenlage konkretisiert hat.[62] Das Merkmal des finalen Eingriffs wurde durch das Kriterium der unmittelbaren Auswirkung ersetzt.[63]

Der Schutzumfang der verschuldensunabhängigen Enteignungs- und Aufopferungsansprüche ist allerdings äußerst begrenzt. Der im Falle des „feindlichen Grün“[64] relevante § 39 OBG NRW, bzw. dessen Entsprechungen in anderen Bundesländern gelten nur im Ordnungsrecht. Neben spezialgesetzlichen Normen kommen als Gefährdungshaftungsnormen der enteignende bzw. enteignungsgleiche, sowie der allgemeine Aufopferungs- bzw. aufopferungsgleiche Anspruch in Betracht.[65] Anerkannt wurden entsprechende Ansprüche z.B. im

---

[58] BGH NJW 1987, 1945

[59] Ebd.

[60] MüKoBGB/Papier/Shirvani BGB § 839 Rn. 191; Kment NVwZ 2015, 927 (927).

[61] BGH NJW 1987, 1945.

[62] MüKoBGB/Papier/Shirvani, ebd. Rn. 192.

[63] Kment NVwZ 2015, 927 (929).

[64] BGH NJW 1987, 1945.

[65] Ebd.

Falle der rechtswidrigen Bescheidung einer Bauvoranfrage[66] oder Baugenehmigung[67] – einer Tätigkeit, die ohne weiteres durch KIs erledigt werden können[68] und die regelmäßig hohe Schadenssummen generiert.
Während ein enteignungsgleicher bzw. aufopferungsgleicher Anspruch verlangt, dass der hoheitliche Eingriff rechtswidrig war, erfordern Ansprüche aus Aufopferung oder enteignendem Eingriff ein spezifisches Sonderopfer auf Seiten der Betroffenen.
Grundsätzlich ist für die Rechtswidrigkeit kein menschliches Handeln erforderlich, sodass auch technisches Versagen ausreicht, um einen enteignungs- oder aufopferungsgleichen Anspruch zu begründen.[69] Ungeklärt ist indes, ob Bezugspunkt der Rechtmäßigkeitsprüfung das schädigende Ereignis oder der schädigende Erfolg ist.[70] Ein Anknüpfen an die Rechtmäßigkeit der Handlung selbst kommt jedenfalls bei KI nicht in Betracht,[71] da dies die Gefährdungshaftung in eine Verschuldenshaftung umwandeln würde.
Jedenfalls rechtswidrig ist es, wenn bereits der generelle Einsatz der KI unzulässig ist. Eine Rechtmäßigkeitshürde ist hierbei zum einen Art. 22 DSGVO[72], wonach eine betroffene Person das Recht hat, nicht einer ausschließlich auf einer automatisierten Verarbeitung beruhenden Entscheidung unterworfen zu werden.[73] Im Ergebnis verlangt die Norm somit, dass am Entscheidungsprozess stets ein Mensch beteiligt ist.
Eingeschränkt wird dieses grundsätzliche Verbot der vollautomatischen Entscheidung durch Art. 22 Abs. 2 DSGVO, wenn die Behördenentscheidung aufgrund von Rechtsvorschriften der Union oder der Mitgliedstaaten, denen der:die Verantwortliche unterliegt, zulässig ist

---

[66] BGH NVwZ 2011, 1150; BGH NJW 1983, 215; BGH NJW 1979, 36.
[67] BGH NJW 1985, 1338.
[68] Roth-Isigkeit NVwZ 2022, 1253.
[69] Kment NVwZ 2015, 927 (928).
[70] Ebd. 927.
[71] Ebd. Fn. 24.
[72] Zu Art. 22 DSGVO als sozio-technischer Gestaltungsnorm, vgl. Djeffal DuD 2021, 529; vgl. auch Ambrock (in diesem Band), S. 81 ff.
[73] Die Ausnahme des Art. 22 Abs. 3 DSGVO ist in Deutschland mangels geregelter Staatshaftung nicht einschlägig. Dazu siehe unten.

„...und diese Rechtsvorschriften angemessene Maßnahmen zur Wahrung der Rechte und Freiheiten sowie der berechtigten Interessen der betroffenen Person enthalten [...].

Angesichts der haftungsrechtlichen Unschärfen im richterrechtlich entwickelten Staatshaftungsrecht mit Blick auf KI ist die Annahme von „Rechtsvorschriften, die angemessene Maßnahmen zur Wahrung der Rechte und Freiheiten sowie der berechtigten Interessen der betroffenen Person enthalten" fernliegend. Erforderlich ist mindestens ein materielles Gesetz.[74] Auch § 35a VwVfG genügt diesen Anforderungen für sich genommen nicht, da die Norm keinerlei Regelungen zur Wahrung der Rechte und Freiheiten der Betroffenen enthält. Ohne spezifisches Haftungsregime ist eine vollautomatisierte Datenverarbeitung daher nicht mit Art. 22 Abs. 1, 2 DSGVO vereinbar.

Im Geiste seiner Wesentlichkeitsrechtsprechung hat das Bundesverfassungsgericht darüber hinaus strenge Maßstäbe für die gesetzlichen Grundlagen des KI-Einsatzes entwickelt. Ein KI-Einsatz ohne eine gesetzliche Grundlage, welche diesen Anforderungen genügt, ist stets rechtswidrig:[75]

„Eine spezifische Herausforderung besteht darüber hinaus darin, die Herausbildung und Verwendung diskriminierender Algorithmen zu verhindern. Daher dürften selbstlernende Systeme in der Polizeiarbeit nur unter besonderen verfahrensrechtlichen Vorkehrungen zur Anwendung kommen, die trotz der eingeschränkten Nachvollziehbarkeit ein hinreichendes Schutzniveau sichern."[76]

Generell bereitet es große Schwierigkeiten technische Fehlfunktionen an den Maßstäben von „rechtmäßig" und „rechtswidrig" zu messen, weshalb das Merkmal des Sonderopfers im Anspruch aus enteignendem oder Aufopferungsanspruch automatisch eine besondere Bedeutung erlangt.[77]

Ein Sonderopfer soll jedenfalls dann vorliegen, wenn Nachteile aufgetreten sind, die nach der gesetzlichen Wertung über das hinausgehen, was Betroffene hinzunehmen haben, die also die allgemeine Opfergrenze überschreiten. Damit soll der Konstellation Rechnung getragen werden, dass ein Eingriff in betroffene Grundrechte zwar grundsätzlich aufgrund von Gemeinwohlinteressen gerechtfertigt werden kann, jedoch nur bei Gewährung eines kompensatorischen Ausgleichs verhältnismäßig ist.

---

[74] Gola/Heckmann DS-GVO/BDSG/Schulz, 3. Aufl. 2022, DS-GVO Art. 22 Rn. 30.

[75] BVerfG NJW 2023, 1196.

[76] Ebd.

[77] Kment NVwZ 2015, 927 (928 f.).

Die Konturenlosigkeit des Sonderopfergedankens führt für Betroffene zu Rechtsschutzdefiziten.[78] Darüber hinaus wurden Enteignungs- und Aufopferungsansprüche bislang nur für Eingriffe in Art. 14 GG (enteignender Eingriff), den eingerichteten und ausgeübten Gewerbebetrieb[79] und Art. 2 Abs. 2 GG (Aufopferungsanspruch)[80] anerkannt. Ob auch ein Eingriff in das allgemeine Persönlichkeitsrecht einen Aufopferungsanspruch begründet, hat der BGH bislang offengelassen.[81] Dies abzulehnen, wäre allerdings mit der Gleichwertigkeit des Art. 2 Abs. 2 GG mit Art. 2 Abs. 1 i.V.m. Art. 1 Abs. 1 GG kaum zu vereinbaren.[82]

Bei behördlichen Auskunftserteilungen dient der öffentlich-rechtliche Aufopferungsanspruch als dogmatische Grundlage für eine Gefährdungshaftung, wenn die Auskunfterteilung automatisiert oder durch künstliche Intelligenz erfolgt.[83] Im Sozialrecht greift der auch ChatBot-Auskünfte umfassende[84] sozialrechtliche Herstellungsanspruch, nach dem ein Sozialversicherter, der aufgrund pflichtwidrig unterlassener oder fehlerhafter Beratung zur Dispositionen veranlasst worden ist, einen Anspruch gegen den Leistungsträger auf Herstellung des Zustands hat, der ohne den Beratungsfehler bestünde.[85]

E. Kritik

Die Unklarheit der Haftungskriterien ist kritikwürdig. So legen die ungeschriebenen und von der Rechtsprechung nur grob umrissenen Tatbestandsmerkmale der Aufopferungsansprüche ein rechtsstaatliches Defizit mit Blick auf Art. 20 Abs. 3 und Art. 19 Abs. 4 GG offen.[86] Die Rechtsfolgen sind ange-

---

[78] Ebd.

[79] VGH Kassel NVwZ 1995, 611.

[80] BGH NJW 1976, 186; BGH NJW 1953, 857.

[81] BGH NJW 1968, 989.

[82] Conrad/Grünewald/Kalscheuer/Milker Öffentlich-rechtliches ÄußerungsR-HdB/von Rochow, 346.

[83] Ebd. 2022, 347.

[84] Martini/Ruschemeier/Hain VerwArchiv 2021, 1 (8).

[85] MüKoBGB/Papier/Shirvani BGB § 839 Rn. 140.

[86] Kment NVwZ 2015, 927 (929).

sichts der Konturenlosigkeit der Tatbestandsvoraussetzungen für die Bürger*innen nicht mehr absehbar.[87] Aus diesem Grund besteht das Erfordernis einer gesetzgeberischen Regelung, wie sie einst § 1 Abs. 2 StHG angestrebt hat.[88] Bis dahin muss allein aus Gründen der Rechtsstaatsgarantie das Sonderopferkriterium weit ausgelegt werden, da sich im Schadenseintritt stets eine für die konkrete hoheitliche Betätigung typische Gefahrenlage konkretisiert, wenn sich der Staat zur Erfüllung seiner Aufgaben unkontrollierbarer technischer Einrichtungen, wie z.B. selbstlernender Systeme und neuronaler Netze bedient. Ein derart weites Verständnis deckt sich mit der BGH-Rechtsprechung zum „feindlichen Grün",[89] die den Staat verschuldensunabhängig für technische Fehler automatisierter Systeme zur Verantwortung zieht.

Wenn dies aber bereits für die rudimentäre Form der Verwaltungsautomatisierung durch eine Ampelschaltung gilt, muss es erst recht gelten, wenn der Staat lernfähige Systeme einsetzt.[90] Diesen werden derart erhöhte Entscheidungs- und Verhaltensspielräume eingeräumt, dass sich das Risiko unvorhersehbarer Grundrechtseingriffe deutlich erhöht hat. In diesen Eingriffen realisiert sich gerade das dem Sonderopfergedanken zugrunde liegende Risiko.[91]

Die Aufopferungsansprüche bleiben gleichwohl lückenhaft, da sie sich im Wesentlichen auf den Ausgleich von Eingriffen in Körper und Eigentum beschränken. Für andere Schutzpositionen und insbesondere für Diskriminierungsverbote i.S.d. Art. 3 GG gewähren diese Rechtsinstitute keinen Schutz.[92]

Schließlich besteht rechtspolitischer Handlungsbedarf dahingehend die prozessuale Passivlegitimation gegenüber den Bürger*innen klar zu regeln. Entwickelt eine übergeordnete Behörde eine KI und verpflichtet andere Behörden zu ihrer Verwendung, kann es nicht den Betroffenen aufgebürdet werden, die Schadensverursachung im jeweiligen Aufgaben- und Verantwortungsbereich zu

---

[87] Kment, ebd.

[88] MüKoBGB/Papier/Shirvani BGB § 839 Rn. 193. Das StHG war 1982 kurzzeitig in Kraft, bevor das BVerfG es aufgrund fehlender Bundeskompetenz für nichtig erklärt hat. Die Bundeskompetenz ist im Rahmen der Föderalismusreform inzwischen hergestellt worden. Zu einer erneuten Gesetzesinitiative hinsichtlich eines Staatshaftungsgesetzes kam es seitdem aber nicht.

[89] BGH NJW 1987, 1945.

[90] Martini/Ruschemeier/Hain VerwArchiv 2021, 1 (6).

[91] Ebd.

[92] Ebd.

lokalisieren. Für die Praxis empfiehlt sich bisweilen vorsorglich eine Streitverkündung nach §§ 72, 74 ZPO an alle Beteiligten, sofern sie bekannt sind.
Den genannten Problemen und Unschärfen versucht in letzter Zeit v.a. die EU legislativ zu begegnen. Haftungsrechtlich relevant sind hier die Bestimmungen der Datenschutzgrundverordnung (DSGVO), sowie de lege ferenda der Artificial Intelligence Liability Directive – AILD.

F. Europarechtliche Haftungsharmonisierung

I. DSGVO

Erste europarechtliche Regelungen zur Haftung für KIs enthält die DSGVO.[93] Nach Art. 82 DSGVO ist ein Ersatzanspruch möglich, wenn spezifische Risiken automatisierter Entscheidungsfindung in einen Schaden der Bürger*innen umschlagen, v.a. wenn die Behörde die Automatisierungsschranken des Art. 22 Abs. 2 b DSGVO i.V.m. § 35 a VwVfG überschreitet.[94]
Die Reichweite des Art. 82 DSGVO ist mit Blick auf KI begrenzt, denn Art. 22 DSGVO[95] verbietet es nur, die Bürger*innen einer ausschließlich auf einer automatisierten Verarbeitung beruhenden Entscheidung zu unterwerfen. ChatBots, die lediglich Auskünfte erteilen stellen zwar ein rechtsrelevantes Handeln i.S.d. öffentlich-rechtlichen Äußerungsrechts dar und können so Haftungsansprüche auslösen. Sie treffen aber keine Entscheidungen. Auch wenn eine behördeninterne KI-generierte Entscheidungsempfehlung durch eine:n Behördenmitarbeiter:in geprüft und unterschrieben wird, dann dürfte es zumindest schwer zu beweisen sein, dass die Entscheidung ausschließlich auf einer vollautomatisierten Datenverarbeitung beruht. Darüber hinaus sind Maßnahmen zur Verhütung,

---

[93] Verordnung (EU) 2016/679 vom 27.04.2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr (Datenschutz-Grundverordnung); Nach Art. 2 Abs. 2 d) ist der Anwendungsbereich jedoch ausgeschlossen für durch die Verarbeitung personenbezogener Daten durch zuständige Behörden zum Zwecke der Verhütung, Ermittlung, Aufdeckung oder Verfolgung von Straftaten oder der Strafvollstreckung, einschließlich des Schutzes vor und der Abwehr von Gefahren für die öffentliche Sicherheit.
[94] Martini/Ruschemeier/Hain VerwArchiv 2021, 1 (7).
[95] Dazu Ambrock (in diesem Band), S. 81 ff.

Ermittlung, Aufdeckung und Verfolgung von Straftaten, sowie die Strafvollstreckung nach Art. 2 Abs. 2 DSGVO vom Anwendungsbereich der Verordnung ausgenommen.
Die Haftungsregelung des § 82 Abs. 1 DSGVO gewährt einen Schadensersatzanspruch gegen den „Verantwortlichen" oder „Auftragsverarbeiter" (Art. 4 Nr. 7, Nr. 8 DSGVO). Art. 82 Abs. 3 DSGVO eröffnet diesen eine Exculpationsmöglichkeit, welche die Verknüpfung der Haftung aus Art. 82 DSGVO mit dem Rechtsverstoß aus Art. 22 DSGVO nahezu wirkungslos werden lässt. Der Verantwortliche oder der Auftragsverarbeitende wird nämlich von der Haftung gemäß Absatz 2 befreit, wenn er:sie nachweist, dass er:sie in keinerlei Hinsicht für den Umstand, durch den der Schaden eingetreten ist, verantwortlich ist. Für den Schaden, den die Opfer von KI-Outputs erleiden, sind die Verarbeitenden der Input-Daten von selbstlernenden Systemen regelmäßig nicht verantwortlich.
Die DSGVO regelt v.a. den unsachgemäßen Umgang mit personenbezogenen Daten und gewährt denjenigen Ansprüche, deren Daten betroffen sind. Angreifbare KI-Entscheidungen hingegen schaffen neue Opfer, basierend auf den Daten unbekannter Dritter. KI-Geschädigte deren eigene Daten nicht bereits vom Input betroffen sind, sind damit regelmäßig auf das richterrechtlich ausgestaltete deutsche Staatshaftungsrecht verwiesen.

II. AILD

Im Oktober 2020 hat das EU-Parlament nach Art. 225 AEUV die Kommission zur Gesetzesinitiative im Bereich der KI-Haftung aufgefordert und einen eigenen Verordnungsentwurf entwickelt.[96] Die Kommission hat daraufhin im September 2022 den Vorschlag einer KI-Haftungsrichtlinie (Artificial Intelligence Liability Directive – AILD) unterbreitet.[97] In den Erwägungsgründen

[96] Europäisches Parlament, Entschließung des Europäischen Parlaments vom 20. Oktober 2020 mit Empfehlungen an die Kommission für eine Regelung der zivilrechtlichen Haftung beim Einsatz künstlicher Intelligenz (2020/2014(INL)), Abl. EU 2021 Nr. C 404/107.
[97] Europäische Kommission, Vorschlag für eine Richtlinie des Europäischen Parlaments und des Rates zur Anpassung der Vorschriften über außervertragliche zivilrechtliche Haftung an künstliche Intelligenz (Richtlinie über KI-Haftung), 496 final vom 28.09.2022; erläutert bei Staudenmeyer NJW 2023, 894.

kommt die Kommission zu dem Schluss, dass die nationalen Haftungsvorschriften für die Bearbeitung von Haftungsansprüchen, die durch KI-gestützte Produkte und Dienstleistungen verursacht werden, ungeeignet sind. Dies betrifft insbesondere die dem Geschädigten auferlegten Beweislasten vor dem Hintergrund von Komplexität, Autonomie und Undurchsichtigkeit („Blackbox-Effect“)[98].

Zwar erkennt die Kommission an, dass die Gerichte – wie in diesem Beitrag hinsichtlich § 839 BGB vorgeschlagen – ihre Auslegung der bestehenden Vorschriften KI-gemäß *ad hoc* anpassen können, um zu einem gerechten Opferschutz zu kommen. Dies wird aber – so die Kommission – zu erheblicher Rechtsunsicherheit führen.[99]

Mit Art. 3 geht die Richtlinie zunächst das Beweisproblem an: Die Mitgliedstaaten sollen sicherstellen, dass nationale Gerichte befugt sind, sich Beweismittel zu einem Hochrisiko-KI-Systeme offenlegen zu lassen, welches im Verdacht steht, einen Schaden verursacht zu haben, wobei der:die Geschädigte weiterhin die Plausibilität seines Schadensersatzanspruchs durch die Vorlage von Tatsachen und Beweismitteln ausreichend belegen muss. Eine solche Offenlegungsanordnung soll auch in Bezug auf Geschäftsgeheimnisse, vertrauliche Informationen und in Bezug auf die öffentliche oder nationale Sicherheit verhältnismäßig sein. Ausnahmetatbestände zu Gunsten der öffentlichen und nationalen Sicherheit haben freilich in der Vergangenheit dazu geführt, dass sich Staaten, wie die Bundesrepublik Deutschland, weitgehende faktische Bereichsausnahmen des europäischen Rechts für ihre Hoheitlichen Tätigkeiten gewährt haben. Die Einstufung von Verschlusssachen etwa ist nach wie vor wenig justiziabel.[100]

Nach Artikel 4 der Richtlinie soll keine pauschale Verschuldensvermutung geschaffen werden. Vielmehr wird das Verschulden nur bei Vorliegen von drei kumulativen Kriterien vermutet.

---

[98] Gössl/Yakar Geschlechterneutrale KI: Eine Handreichung, abrufbar unter: https://www.schleswig-holstein.de/DE/fachinhalte/G/gleichstellung/Downloads/handreichung_geschlechterneutrale_ki_lang.pdf?__blob=publicationFile&v=1, 70 ff.

[99] Europäische Kommission, Vorschlag für eine Richtlinie des Europäischen Parlaments und des Rates zur Anpassung der Vorschriften über außervertragliche zivilrechtliche Haftung an künstliche Intelligenz (Richtlinie über KI-Haftung), 496 final vom 28.09.2022.

[100] von Rochow VerfBlog, 2022/1/11, https://verfassungsblog.de/wer-whistleblower-nicht-schutzt-muss-haften/.

1. Der Kläger weist nach oder das Gericht vermutet wegen rechtswidriger Vorenthaltung von Beweismitteln, dass ein Verschulden vorliegt, da gegen eine als Sorgfaltspflicht ausgestaltete Schutznorm verstoßen worden ist;
2. es kann nach vernünftigem Ermessen davon ausgegangen werden, dass das Verschulden das Ergebnis beeinflusst hat;
3. der Betroffene kann die Kausalität zwischen KI-Ergebnis und Schaden nachweisen.

Bei Hochrisiko-KI-Systemen gelten Sonderregelungen hinsichtlich der Verschuldensvermutung. Hier muss der*die Kläger*in gegenüber dem Anbieter einen Verstoß gegen die Qualitätskriterien von Trainings-, Validierungs- und Testdatensätze nachweisen oder, dass das KI-System nicht den Transparenzanforderungen entspricht oder nicht wirksam von natürlichen Personen beaufsichtigt werden kann. Auch Verstöße gegen Genauigkeit, Robustheit und Cybersicherheit kann der:die Betroffene bei Hochrisikosystemen aktiv nachzuweisen haben. Gegenüber den KI verwendenden Behörden genügt der Nachweis einer Verletzung der Pflicht zur Verwendung oder Überwachung des KI-Systems entsprechend der beigefügten Gebrauchsanweisung oder gegebenenfalls zur Aussetzung oder Unterbrechung seiner Verwendung oder der Nachweis, dass Eingabedaten verwendet worden sind, die nicht der Zweckbestimmung des KI-Systems entsprechen.
Handelt es sich nicht um ein Hochrisiko-System gilt die Verschuldensvermutung nur, wenn es für den:die Kläger:in übermäßig schwierig ist, den ursächlichen Zusammenhang nachzuweisen – eine Bestimmung deren unbestimmte Rechtsbegriffe viel Auslegungsspielraum lassen und der angestrebten Rechtsklarheit widersprechen.

G. Fazit

In Deutschland ist die Staatshaftung generell und in Bezug auf KI besonders durch rechtliche Unsicherheit und eine dem Rechtsstaatsprinzip aus Art 20 Abs. 3 GG kaum gerecht werdende Unschärfe geprägt. Dieses Grundproblem

des deutschen Staatshaftungsrechts beseitigt auch die AILD nicht, die nur spezifische Haftungs- und Beweisaspekte, die typisch für den Umgang mit KI sind, harmonisiert.[101]

Die existierenden deutschen Haftungsregime sind im Einzelfall anwendbar, insgesamt jedoch lückenhaft. Viele dieser Lücken lassen sich durch Rechtsfortbildung schließen. Dies betrifft v.a. den Amtshaftungsanspruch aus § 839 BGB i.V.m. Art. 34 GG: Da beide Normen als einheitliche Anspruchsgrundlage in der Gesamtschau ausgelegt werden müssen, ist ein Anknüpfen am Wortlaut des „Beamten" i.S.e. zwingend natürlichen Person heute nicht mehr geboten. Das Verschuldenserfordernis lässt sich i.S.d. „feindliches-Grün-Rechtsprechung" des BGH beim Einsatz technischer Systeme zur Gefährdungshaftung weiterentwickeln.

Dogmatisch wird eine solche Gefährdungshaftung bisher allerdings meist statt auf die Amtshaftungsansprüche über die Enteignungs- und Aufopferungsrechtsprechung hergeleitet – entweder unter Rückgriff auf spezialgesetzliche Normen des Polizei- und Ordnungsrechts oder aber über die allgemeinen richterrechtlichen Ansprüche zur Enteignung und Aufopferung. Diese Herleitung ist insbesondere deshalb ungeeignet, weil sie bislang nur Eingriffe in Art. 14 GG, Art. 2 Abs. 2 GG und ggf. das allgemeine Persönlichkeitsrecht erfasst. Das der künstlichen Intelligenz immanente Diskriminierungspotential i.S.d. Art. 3 GG bleibt so außen vor.

Die Defizite im Staatshaftungsrecht führen zu einem rechtspolitisch und wirtschaftlich kaum tragbaren *status quo*, der auch europarechtlichen Ansprüchen nicht mehr gerecht wird. Nach Art. 22 DSGVO hat der*die Bürger*in ein Recht, nicht einer ausschließlich auf einer automatisierten beruhenden Entscheidung unterworfen zu werden, wenn nicht der Gesetzgeber Rechtsvorschriften schafft, die angemessene Maßnahmen zur Wahrung der Rechte und Freiheiten sowie der berechtigten Interessen der betroffenen Personen enthalten. Bis in diesem Sinne ein rechtsklares und bürger*innenfreundliches Staatshaftungsrecht geschaffen worden ist, ist es dem Staat verwehrt, seine Entscheidungen ausschließlich auf vollautomatische Prozesse zu stützen.

Das im Jahr 1982 kurzzeitig gültige dann aber vom Bundesverfassungsgericht mangels Bundeskompetenz für nichtig erklärte Staatshaftungsgesetz enthielt bereits eine Regelung, wonach das Versagen einer technischen Einrichtung als

---

[101] Staudenmeyer NJW 2023, 894, 895.

Pflichtverletzung gelten sollte, wenn der Träger öffentlicher Gewalt anstatt durch Personen durch diese Einrichtung öffentliche Gewalt selbstständig ausüben lässt und das Versagen einer Pflichtverletzung dieser Personen entsprechen würde.[102] Im Jahr 2004 hatte die Bundesregierung auf eine kleine Anfrage der FDP-Fraktion mitgeteilt, dass eine Neuordnung des Staatshaftungsrechts in diesem Sinne nicht geboten sei. Mit dem Einzug von KI-Systemen in die hoheitliche Verwaltung und dem stetig wachsenden Druck der EU zur Harmonisierung der KI-bezogenen Haftungsregelungen ist diese Einschätzung heute nicht mehr aufrechtzuerhalten. Art. 22 DSGVO ebenso wie die zukünftige AILD enthält einen Gestaltungsauftrag, dessen Zweck sich auf die Sicherung der Rechte der Adressierten automatischer Entscheidungen bezieht.[103] Insbesondere dort, wo der Einsatz von KI normative Wirkungen entfaltet, fordern die Grundrechte Maßnahmen zu ihrer effektiven Verwirklichung.[104] Hierzu gehört mit Blick auf Art. 19 Abs. 4 und Art. 20 Abs. 3 GG auch die Schaffung eines KI-gerechten europäisch harmonisierten Staatshaftungsrechts.

---

[102] MüKoBGB/Papier/Shirvani BGB § 839 Rn. 193.

[103] Djeffal DuD 2021, 529 (530).

[104] Ebd.

Teil 3

Der Blick der Praxis und Schluss

Kapitel 8

Entscheide Du, KI! – wie uns künstliche Intelligenz in der Anwaltsberatung helfen kann und wo die Grenzen sind

Martin Gerecke

A. Einleitung

Die nachstehende Untersuchung des bekannten amerikanischen Psychologen Larry Richard, von der mir nicht bekannt ist, inwieweit sie wirklich auf empirischen Daten beruht, zeigt, dass wir Anwälte für den Einsatz von Künstlicher Intelligenz ("KI") an sich nicht geschaffen sind. Wir sind im Vergleich zum Rest der (hier: US-amerikanischen) Bevölkerung übermäßig skeptisch ("Skepticism"), vertrauen also einem (teil-)autonomen System möglicherweise nicht. Zudem entscheiden wir auch lieber selbst ("Autonomy") und lassen uns dadurch u.U. nicht gern von einer anderen Intelligenz leiten.

Lawyer psychology

Trait	US Population	Lawyers
Skepticism**	50% (by definition of what the "norm" is)	93rd %-ile
Autonomy**	50%	89%
Abstract reasoning**	50%	81%
Urgency*	50%	71%
Resilience*	50%	30%
Sociability**	50%	7% (12% including rainmakers)

* indicates one standard deviation from norm
** indicates two standard deviations from norm

Source: Dr. Larry Richard (using the "Caliper" instrument)

AdamSmithEsq.com

Ist der Einsatz von KI für uns Anwälte trotzdem sinnvoll? Ja, unbedingt! Besteht umgekehrt die Gefahr, dass der Einsatz von KI unsere anwaltliche Beratung determiniert? Ja, hoffentlich sogar. Wird sie uns deshalb ersetzen und überflüssig machen? Nein, wie dieser Beitrag zeigen soll.

B. Der Hype um KI

Der diesjährige Hype um KI wurde maßgeblich initiiert vom amerikanischen Unternehmen OpenAI und ihrem statistischen Sprachmodell ChatGPT. Das gab es auch zuvor schon, aber jetzt wurde es erstmals für die Allgemeinheit zugänglich gemacht. Das Programm liefert verblüffend detaillierte Antworten auf alle möglichen Fragen, wenngleich nicht immer die richtigen (wir nennen diese Fehler „Halluzinationen"). Der kostenlose Zugang führte dazu, dass viele Menschen, denen das Thema "Künstliche Intelligenz" lange Zeit suspekt, weil eben wenig greifbar, war, die Möglichkeit nutzten, das Programm auszuprobieren und KI erstmals wirklich zu erleben.

Die Einführung von ChatGPT war ein Schock für die Konkurrenz. Google rief den sog. Code Red aus und sah seine Marktanteile bedroht. Sogar die Co-Gründer Larry Page und Sergey Brin wurden zur Krisensitzung einberufen.[1] Kurze Zeit später – für viele überhastet – stellte Google seinen eigenen experimentellen KI-Dienst, der auf generativer KI basiert, vor: Google Bard.
Schaut man sich in den einschlägigen Suchmaschinen die Anzahl der Suchanfragen für den Begriff "ChatGPT" an, übertrifft dieser im Januar 2023 erstmals den lange Zeit führenden Spitzenreiter "Covid". Der "SPIEGEL" titelte sogar „Die neue Weltmacht – Wie ChatGPT und Co. unser Leben verändern", übrigens mit einem geprompteten KI-generierten Bild (eine Art Cyborg).[2] Das Titelbild illustriert aus meiner Sicht sehr gut, wessen Arbeitsplätze durch den Einsatz künstlicher Intelligenz tatsächlich ernsthaft bedroht sind: Es sind neben den Journalisten die Fotografen. Die Frage ist eine Provokation, aber sie liegt nahe: Wer braucht noch professionelle Fotografen, wenn mir die KI derart perfekte Bilder erstellen kann wie den Aufmacher des "SPIEGEL"?
Die neueste Version von ChatGPT – GPT-4 – zeigte auch erstaunliche Resultate beim Uniform Bar Exam (dem amerikanischen Jura-Examen, vergleichbar dem 2. Staatsexamen in Deutschland). Die KI bestand nicht nur das Examen; mit einem Score von 75% war sie deutlich besser als der 68%-Durchschnitt aller Examinierten.[3] Dies zeigt: KI-Sprachmodelle können auch komplexe juristische Antworten geben, sogar Fälle lösen.

C. Wie kann uns die KI helfen? Die Stärken

Wir arbeiten schon jetzt mit KI-Anwendungen, insbesondere im Zusammenhang mit Legal Tech-Lösungen. Dazu zählen u.a. eine maschinenlernende Software zur Dokumentenanalyse, die uns dabei hilft, Verträge und sonstige Rechtsdokumente effizient und schnell zu analysieren. Dies ist z.B. bei Due Diligence-Prozessen enorm hilfreich, wenn mit Hilfe von künstlicher Intelligenz

---

[1] Google Calls In Larry Page and Sergey Brin to Tackle ChatGPT and A.I. Chatbots - The New York Times (nytimes.com), abrufbar unter https://www.nytimes.com/2023/01/20/technology/google-chatgpt-artificial-intelligence.html
[2] Nr. 10 v. 04.03.2023.
[3] GPT-4 Passes the Bar Exam | Illinois Institute of Technology (iit.edu), abrufbar unter https://www.iit.edu/news/gpt-4-passes-bar-exam%20.

relevante Bestimmungen (z.B. Change-of-Control-Klauseln) genauer und schneller gefunden, Doubletten herausgefiltert und Formulierungen verglichen werden können.
Ein ganz einfaches Beispiel der KI-Nutzung im Anwaltsbereich, das sicher viele Kollegen schon einmal genutzt haben, ist das Übersetzungsprogramm DeepL, das meines Erachtens sehr gute Übersetzungsleistungen bietet und heute schon zumindest im ersten Schritt der Übersetzung professionelle Fachübersetzer überflüssig macht. Ein kleines Lifehack am Rande: ich lasse häufig von DeepL bereits übersetzte, englische Texte noch einmal durch ChatGPT laufen mit der Aufforderung (*prompt*), den Text zu verfeinern. Das Ergebnis ist wirklich ausgezeichnet.
Auf internationaler Ebene haben wir weitere Legal-Tech KI-Tools, die sich in der Rechtsanwaltsszene inzwischen großer Beliebtheit erfreuen. Ich will hier nur beispielhaft Ironclad, ein System zum Vergleichen von Texten (Track Changes), Lawdroid und Docket Alarm, die beide eine sehr gute Sammlung von Gesetzen bzw. Gerichtsfällen aufweisen und auf diese Weise sogar Prognosen über gerichtliche Entscheidungen treffen können (Docket Alarm), und Lexion, ein Tool, das bei der Vertragserstellung Hilfe leistet, nennen.
Auch ChatGPT kann eine Hilfe bei der anwaltlichen Arbeit sein. Gibt man der KI einen einfachen Sachverhalt an die Hand und stellt ihr eine Frage, z.B.:

„Du bist Inhouse-Counsel eines Zulieferers. Der Zulieferer möchte mit einem Kunden einen Vertrag schließen. Die Haftung des Zulieferers für einfache Fahrlässigkeit soll beschränkt werden. Bitte entwirf eine Klausel für den Vertrag.“,

dann wirft ChatGPT eine durchaus brauchbare – nicht perfekte – Haftungsklausel aus:

"Haftungsbeschränkung für einfache Fahrlässigkeit: Der Zulieferer haftet nicht für Schäden, die durch einfache Fahrlässigkeit verursacht werden. Dies gilt jedoch nicht für (i) Schäden, die aus einer Verletzung wesentlicher Pflichten des Zulieferers resultieren, die für die Erreichung des Vertragszwecks von besonderer Bedeutung sind (Kardinalpflichten) oder (ii) für Schäden aus einer Haftung nach dem Produkthaftungsgesetz."

Man kann die Antwort weiter verfeinern und ChatGPT z.B. auffordern, die Klausel so auszugestalten, dass die Haftungsbegrenzung beidseitig gilt. Auch hierbei produziert das Programm sehr brauchbare Ergebnisse. Man kann ChatGPT aber nicht nur für Klauselvorschläge nutzen, man kann mit ihm auch Schreiben, z.B. an eine Datenschutzbehörde, verfassen oder Texte konsolidieren.
Ich kann KI-Tools also aktuell schon in der anwaltlichen Beratung (vorausgesetzt, ich nutze keine Mandanteninformationen, dazu sogleich) sehr gut unterstützend einsetzen. Die Programme können dabei helfen, Schriftsätze oder Schreiben vorzubereiten (Fakten sammeln, passende Gerichtsentscheidungen herauszusuchen, mögliche Prozesschancen evaluieren), Texte zu entwerfen (Verträge, Anträge, Schriftsätze, E-Mails etc.), Texte zu verarbeiten (Terminsberichte, Strukturierung von Parteivorbringen, Zusammenfassungen von Vertragsverhandlungen, etc.) oder fremde Texte zu prüfen (auf doppelte Klauseln, wiederholtes Parteivorbringen etc.). Damit kann die KI faktisch auch Rechtsberatungsdienstleistungen erbringen. Sie kann dies aber nicht so gut wie ein Anwalt und das wird auch in naher Zukunft so bleiben.

D. Wie kann uns die KI helfen? Die Schwächen

Die Anwaltsberatung ist (aus guten Gründen) ein streng regulierter Bereich. Zum einen sorgt das Datenschutzrecht dafür, dass ich als Anwalt personenbezogene Daten (d.h. insbesondere den Namen des Mandanten und weitere Kontaktinformationen) besser nicht einfach unbedacht in die Cloud eines amerikanischen KI-Programms eingebe. In seiner „Schrems II"-Entscheidung erklärte der EuGH[4] im Jahr 2020 den damaligen sog. Angemessenheitsbeschluss zum „EU-US Privacy Shield" für ungültig und versagte damit seine Bestätigung, dass die USA ein angemessenes (dem EU-Standard vergleichbares) Datenschutzniveau gewährleiste. Die Übermittlung von personenbezogenen Daten in Länder außerhalb der EU ist aber nur auf der Grundlage derartiger Datenschutzgarantien bzw. Angemessenheitsbeschlüsse möglich. Aktuell führen die Kommission und die US-Regierung Verhandlungen über einen neuen Angemessenheitsbeschluss.

---

[4] EuGH C-311/18 - Facebook Ireland und Schrems, ECLI:EU:C:2020:559.

Zum anderen hindert das anwaltliche Berufsrecht die uneingeschränkte Nutzung von ausländischen KI-Tools für die Anwaltsberatung. Nach § 43a Abs. 2 BRAO ist der Anwalt grundsätzlich zur Verschwiegenheit verpflichtet. Bei der Hinzuziehung von Dienstleistern (hier: KI-/Legal Tech-Programme) darf der Anwalt dem Dienstleister vertrauliche Daten nur insoweit zukommen lassen, wie dies für die Inanspruchnahme der Dienstleistung erforderlich ist (§ 43e Abs. 1 BRAO). Nach § 43e Abs. 3 Nr. 1 BRAO ist der Dienstleister vom Anwalt mittels Vertrags unter Belehrung über die strafrechtlichen Folgen einer Pflichtverletzung zur Verschwiegenheit zu verpflichten. Kein Anwalt wird dieser Pflicht beim Einsatz von Legal Tech/KI-Programmen aktuell nachkommen können.

Andererseits bin als Anwalt nicht zwingend darauf angewiesen, personenbezogene Daten in die KI einzugeben; die KI liefert häufig auch mit generischen Informationen brauchbare Ergebnisse. Gleichwohl erfordern der Daten- und Geheimnisschutz besondere Vorsicht.

Bestehen beim Einsatz von KI-Tools in der Anwaltspraxis Diskriminierungspotentiale? Ja, weil der Prozess der Output-Gewinnung von KI-Anwendungen diskriminierend ist. Diskriminierung im hier verstandenen Sinne meint dem lateinischen Wortsinn nach eine Absonderung, Unterscheidung oder Abgrenzung von Inhalten, man könnte auch sagen, der Umstand, dass ich einem bestimmten Inhalt einen Vorzug gebe und dafür einen anderen benachteilige. Ich gebe einem Inhalt etwas mehr Raum und lasse einen anderen weg. So verstanden, besitzt die gesamte KI-Struktur, angefangen von den Daten, d.h. dem Textkorpus, der zur Auswertung zur Verfügung steht, über den Trainingsprozess bis zum Einsatz des Outputs natürlich Diskriminierungspotential. Die Art und Weise wie der Betreiber das KI-Modell entwickelt hat, ist diskriminierend und auch ihr Trainingsprozess ist diskriminierend. Die KI lernt an der Gesamtheit des Internets, aber wir Menschen haben die Inhalte des Internets erstellt; wir sind die Herren des Internets und wir bestücken das Internet nicht gleich, fair, rational oder ausgewogen. Bei der Verwertung juristischer Texte ist dies besonders eklatant, denn die meisten aktuellen KI-Sprachmodelle lernen nicht mit juristischen Texten. Die Auswahl ist sehr beschränkt und rein selektiv und alles was selektiv ist, ist per se diskriminierend. Auch die Bearbeitung des Outputs der KI durch menschliche Intervention (sog. Labelling) ist diskriminierend, denn Menschen labeln die Ergebnisse nach ihren eigenen subjektiven Vorstellungen.

Die KI selbst bemüht sich übrigens mitunter um Diskriminierungsvermeidung. Ich gebe Ihnen ein schönes Beispiel.
In diesem Sommer findet die Fifa Frauen-WM in Australien und Neuseeland statt. Nehmen wir hypothetisch an, die Fifa vergibt die Rechte an einen Pay-TV-Anbieter, der die Spiele nur verschlüsselt und gegen Entgelt ausstrahlt. Können deutsche TV-Zuschauer trotzdem hoffen, die WM unverschlüsselt im Free-TV zu sehen? Die Antwort richtet sich nach § 13 Medienstaatsvertrag (MStV), der festlegt, dass bestimmte Großereignisse (sportlicher Art) zwingend auch im frei empfangbaren und allgemein zugänglichen Fernsehprogramm in Deutschland gezeigt werden müssen. Die Großereignisse sind abschließend in § 13 Abs. 2 MStV aufgelistet, darunter auch die „Fußball-Europa- und -Weltmeisterschaften" und dort „alle Spiele mit deutscher Beteiligung sowie unabhängig von einer deutschen Beteiligung das Eröffnungsspiel, die Halbfinalspiele und das Endspiel". Es fragt sich, ob mit der Nennung der Fußball-WM in § 13 Abs. 2 Nr. 2 MStV nur die Männer-WM gemeint ist und nicht auch die Frauen-WM. Ausdrücklich bezieht sich die Norm nicht nur auf die Herren-Mannschaft. Die streng rechtliche Antwort ist dennoch, dass mit der Erwähnung der Fußball-WM nicht auch die Frauen-WM gemeint ist. Zum Zeitpunkt der ersten Fassung von § 13 MStV erfüllten nur Männer-Fußballwettbewerbe die hohen Anforderungen für die Einstufung als Großereignis (Ereignis von erheblicher gesellschaftlicher Bedeutung). Auch wenn Frauen-Fußballwettbewerbe inzwischen einen ähnlichen Status erreicht haben (man denke an das Finale der Fußball-Europameisterschaft 2022, bei der fast 18 Millionen Zuschauer die Niederlage der deutschen Frauen gegen England sahen)[5], lässt sich der Katalog des § 13 Abs. 2 MStV nicht einfach um die Frauen-Fußballwettbewerbe ergänzen oder diese in § 13 Abs. 2 MStV "hineinlesen". Eine solche Erweiterung würde eine ausdrückliche Änderung von § 13 MStV erfordern.
Dieses Ergebnis lässt sich aus der Historie der Vorschrift begründen; es gibt dazu aber noch keinerlei Rechtsprechung oder Literaturmeinung. Es lohnte sich also, ChatGPT einmal zu dieser Frage hinzuzuziehen. Das Ergebnis ist verblüffend: Nach Ansicht von ChatGPT zählen natürlich auch Frauen-Fußballwettbewerbe zu den Großereignissen von § 13 Abs. 2 MStV. Es gehe hierbei um Ereignisse von erheblicher gesellschaftlicher Bedeutung und hierzu zähle nicht nur

---

[5] EM-Finale der Frauen bricht alle Rekorde :: DFB - Deutscher Fußball-Bund e.V., abrufbar unter https://www.dfb.de/news/detail/em-finale-der-frauen-bricht-alle-rekorde-242702/.

die Männer-WM, sondern natürlich auch die Spiele mit weiblicher deutscher Beteiligung. Hier schlägt das Pendel also ins genaue Gegenteil: Die KI, der ich eben noch Diskriminierungspotential vorwarf, bemüht sich übermäßig um Gleichberechtigung und kommt in diesem Bemühen zu einem rechtlich falschen Ergebnis.

E. Wie sieht die Zukunft aus?

Wie sieht also die Zukunft im Anwaltsmarkt aus?

KI-Systeme werden uns viel stärker als bisher in der anwaltlichen Beratung helfen. Sie sind gut darin, große Datenmengen zu verarbeiten und Muster zu erkennen. Sie sind allerdings nicht in der Lage, menschliche Kreativität und das intuitive Verständnis von Situationen und Problemen zu ersetzen. Es gibt anwaltliche Tätigkeiten, die allein darauf hinauslaufen, die Antwort auf eine konkrete Rechtsfrage, die die Rechtsprechung oder das juristische Schrifttum bereits gegeben haben, zu finden. Dann, so hat es *Markus Kaulartz* formuliert[6], entbrennt in der Tat ein Wettstreit um die Frage, wer eine solche Antwort besser geben kann: „Der Anwalt, der auf beck-online recherchiert, oder die KI, die beck-online vollständig gelesen und verstanden hat?" Einen solchen Wettstreit mag die KI zukünftig gewinnen. Viele Entscheidungen des Anwalts betreffen jedoch intuitive, moralische oder ethische Fragen. Eine KI wird die Erfahrung und das Urteilsvermögen eines menschlichen Juristen nicht ersetzen können – auch in Zukunft nicht. Emotionale Intelligenz, Empathie, ein Verständnis für zwischenmenschliche Interaktionen (gerade in Gerichtsprozessen), kommunikative Fähigkeiten, Kreativität und ein Gespür dafür, dass juristisch nicht immer die erstbeste Lösung die richtige ist, sondern manchmal auch die zweitbeste oder eine vermeintlich falsche Lösung, sind wichtige Fähigkeiten eines Juristen, die KI-Programme nicht ersetzen können.
Die KI wird uns jedoch zwingen, die Art und Weise der anwaltlichen Beratung zu ändern. Wir werden unsere Arbeitsprozesse anpassen und einen noch größeren Fokus auf spezialisierte Beratung legen müssen, denn der juristische Laie hat zukünftig einen einfacheren Zugang zum Recht bei alltäglichen Rechtsfragen.

---

[6] Kaulartz NJW-Editorial 01/2023.

Er hat mitunter schlicht eine Alternative bei der Lösung von Rechtsproblemen. Wir Anwälte werden zudem unsere Vergütungsstrukturen anpassen müssen. Wir können nicht mehr wie früher für einfache Commodity-Beratung, Übersetzungen von juristischen Texten, Terminsberichten oder die Zusammenfassung von Mandantengesprächen hohe Vergütungen aufrufen. Hier wird sich die Markterwartung schnell ändern. Es erscheint zudem ratsam, dass Anwälte dort, wo es der Daten- und Geheimnisschutz zulassen, KI-Tools für die eigenen internen Prozesse (Knowledge-Management, Mustersammlungen etc.) trainieren.

Abschließend: Vielleicht kennen Sie den bekannten Zeit-Podcast „Alles gesagt". Dort diskutieren zwei Zeit-Journalisten mit ihrem Interview-Gast so lange, bis wirklich alles gesagt ist, d.h. genau genommen so lange, bis der Gast ein bestimmtes, vorher vereinbartes, Code-Wort sagt. Vor einiger Zeit war Armin Wolf beim Podcast zu Gast. Armin Wolf ist einer der bekanntesten österreichischen Journalisten. Er ist Moderator des Nachrichten-Journals ZIB 2 im ORF. Auf die Frage, wie künstliche Intelligenz zukünftig den Journalismus-Bereich ändern wird, antwortete er sinngemäß: Die Zusammenfassung von Ticker-Meldungen, die heute einen großen Teil der journalistischen Tätigkeit in Nachrichtensendungen ausmache, werde zukünftig eine KI übernehmen können. Eine KI wird sich aber nie investigativ mit einem Informanten in einer Tiefgarage treffen können, um an geheime Informationen heranzukommen.[7] Diesen Gedanken will ich auch für die Anwaltspraxis nutzbar machen: Die KI-Systeme werden unsere Arbeit in einigen Teilen überflüssig machen. Sie werden aber nie das aus Erfahrung und Intuition gespeiste emotionale Verständnis aufbringen, das es braucht, um Mandanten seriös und erfolgreich beraten zu können. Der Schritt zum Transhumanismus ist noch weit.

---

[7] Interviewpodcast: Armin Wolf, verstehen Sie Österreich? | ZEIT ONLINE, abrufbar unter https://www.zeit.de/gesellschaft/2023-04/armin-wolf-interviewpodcast-alles-gesagt.

Kapitel 9

Schluss – Empfehlungen zur Vermeidung von Diskriminierungen beim KI-Einsatz

Susanne Lilian Gössl

Die Vermeidung von Diskriminierung durch den Einsatz von KI-Systemen ist, wie unter anderem die Beiträge dieses Bandes gezeigt haben, eine komplexe Fragestellung mit noch vielen ungelösten Problemen.
In der einführend erwähnten Studie wurde eine Reihe von Möglichkeiten identifiziert, wie Diskriminierungen beim KI-Einsatz vermieden oder zumindest verringert werden können.
Diese lauten, zusammengefasst:[1]

A. Allgemeine Empfehlungen

1. Nicht alle KI-Systeme sollten einer einheitlichen Regulierung unterworfen werden. Stattdessen ist ein abhängig von den von ihnen ausgehenden Risiken einer möglichen Geschlechterdiskriminierung abgestuftes System zu entwickeln. Es ist auch im (bisher) unregulierten Bereich sinnvoll, dass KI-Verwendende eine solche Abstufung mitdenken und ihre Aktivitä-

[1] Gössl/Yakar Geschlechterneutrale KI. Eine Handreichung, 2023, abrufbar unter https://www.schleswig-holstein.de/DE/fachinhalte/G/gleichstellung/geschlechterneutrale_ki.html, S. 140 ff.

ten zur Vermeidung von Diskriminierungen bei hohem Risiko für betroffene Personen entsprechend intensiv gestalten. Dies ist das System, das die EU aktuell andenkt und das sich auch weltweit gerade durchsetzt.

2. Der Einsatz von KI-Systemen insbesondere im Hochrisikobereich sollte in Zukunft von einem Zertifizierungsverfahren bzw. einem positiven Audit abhängig gemacht werden – entweder als Selbstverpflichtung oder aufgrund zukünftiger gesetzgeberischer Regelung.
3. Intern sollte bei der Entwicklung, dem Training oder der Anwendung von KI-Anwendungen ein Monitoring-Programm etabliert werden, welches Sorge trägt, dass Diskriminierungen soweit möglich verhindert werden. Die Etablierung von KI-Beauftragten – parallel zu Datenschutz- und Gleichstellungsbeauftragten – könnte hier sinnvoll sein.
4. Eine Entscheidung unter Einsatz von KI, die als hochrisikoreich eingestuft wird, sollte immer einer menschlichen Letztkontrolle unterliegen. Um dem Problem des *automation bias* vorzubeugen, sollte die entscheidende Person stets ausführlich i.S. der *explainable AI* das Zustandekommen des Entscheidungsvorschlags eines KI-Systems erhalten und auf Basis auch dieser Erklärung die Entscheidung des KI-Systems übernehmen oder gerade nicht übernehmen.

B. Empfehlungen für die KI-nutzende Praxis

5. Unternehmen sollten gesetzgeberischen Aktivitäten zuvorkommen und sich Selbstverpflichtungen auferlegen, die u.a. ein *commitment* zu diskriminierungsfreier KI, regelmäßigen Zertifizierungen und internen Monitoring-Programmen enthalten. Ein solches *commitment* könnte auch als Marketing-Strategie genutzt werden.
6. Vorgehensweisen zur Verringerung von KI-Diskriminierungen sollten dokumentiert werden, um Fehler leichter aufzufinden und zukünftigen Haftungsfragen vorzubeugen.
7. Für Personen, die von KI-Entscheidungen betroffen sein können, ist eine transparente und nachvollziehbare Erläuterung i.S.d. *explainable AI* sinnvoll, die u.a. folgende Informationen enthalten:
   - dass und warum KI eingesetzt wird

- Erläuterung möglicher Diskriminierungsrisiken und eventuell weiterer betroffener Rechte.

8. Es sollte stets eine reale Möglichkeit geben, sich für oder gegen den KI-Einsatz zu entscheiden. Durch diese Kommunikation kann das bisher nur mäßig bestehende Vertrauen in die Nutzung von KI-Anwendungen aufgebaut werden. Hier sollte in die Gestaltung der Nutzer*innenoberflächen investiert werden, um eine geeignete und vertrauensbildende Kommunikation zu etablieren.
10. Da einerseits umfangreiche Datensätze das Risiko einer Diskriminierung verringern, andererseits aber das Datenschutzrecht davon ausgeht, dass so wenig Daten wie möglich gesammelt werden sollten, ist beim Sammeln von personenbezogenen Daten besonders darauf zu achten, dass die Personen, die ihre Daten geben, über deren Verwendung u.a. als Trainingsdaten und den dahinterstehenden Zweck aufgeklärt werden.
11. Bei der Entwicklung von KI-Systemen oder ihrem Training und ihrer Überprüfung sollten divers und interdisziplinär zusammengesetzte Gruppen tätig werden. Unternehmen sollten darauf achten, auch gegenüber ihren Angestellten klar ihr *commitment* zu fairen, nicht-diskriminierenden Algorithmen zu kommunizieren.
12. KI-einsetzende oder -entwickelnde Stellen könnten sich zusammenschließen und, sollten kontaminierte Datensätze oder diskriminierende *proxies* gefunden werden, diese Ergebnisse und Proben teilen, um bei anderen Akteuren ähnliche Fehler zu verhindern.
13. Die Praxis sollte mit Personen aus der Forschung und weiteren Stakeholdern zusammenarbeiten, um allgemeine Standards und Vorgehensweisen zu entwickeln, wie zukünftig Diskriminierungen aufgefunden und verhindert bzw. minimiert werden können.

C. Empfehlungen für die Forschung

14. Rechtswissenschaften, Informatik und verwandte Wissenschaften und evtl. Sozialwissenschaften sollten stärker zusammenarbeiten und einerseits mit der Praxis und weiteren Stakeholdern allgemeine Standards und

Vorgehensweisen entwickeln, wie zukünftig Diskriminierungen aufgefunden und verhindert bzw. minimiert werden können. Andererseits sollten bereits die Studierenden früh mit den anderen Disziplinen in Berührung kommen, um für die Probleme des KI-Einsatzes jeweils aus der Sicht der anderen Disziplin(en) sensibilisiert zu werden.

15. Die Forschung könnte nach dem Vorbild anderer Stellen in Deutschland daran mitwirken, vertrauenswürdige Zertifizierungsstellen zu entwickeln oder entsprechendes Personal aus- oder fortzubilden.

digital | recht

Schriften zum Immaterialgüter-, IT-,
Medien-, Daten- und Wettbewerbsrecht

Susanne Lilian Gössl (Hrsg.)

Diskriminierungsfreie KI

In diesem Tagungsband werden Fragen der Regulierung(sdefizite) zur Vermeidung von Diskriminierungen durch KI-Einsatz vertieft. Es fließen neben juristischen auch Erkenntnisse der Informatik und der Soziologie ein. Ein Fokus liegt auf KI-VO (Entwurf), DMA und DSA, Staatshaftungsrecht, der proxy-Diskriminierung i.R.d. AGG und dem Datenschutzrecht.