Daniel Wolff

KI-Biases im Gesundheitswesen

– Teil 1 (Terminologie und Typologie)

Der Einsatz von künstlicher Intelligenz im Gesundheitswesen soll dort zu mehr Gleichheit führen, bewirkt bisweilen aber genau das Gegenteil. Dieser Aufsatz behandelt das damit angesprochene Problem von KI-Biases. Während der vorliegende erste Teil eine Bias-Typologie herausarbeitet, widmet sich der in in Heft 01/2023 erscheinende zweite Teil der unionsrechtlichen Adressierung der Bias-Problematik.

1 Einleitung: (Un-)Gleichheit im Gesundheitswesen durch KI

Mit dem Einsatz von Künstlicher Intelligenz (KI) im Gesundheitswesen (KIG) ist die große Hoffnung verbunden, die gegenwärtigen und zukünftigen Herausforderungen der Medizin (besser) bewältigen zu können.[1] Bereits heute hilft der Einsatz von KI medizinischem Personal etwa bei der Analyse von Tumorbildern, der Entscheidung zwischen verschiedenen Behandlungsoptionen und jüngst bei der Bekämpfung der COVID-19-Pandemie.[2] Insbesondere in besonders „musteraffinen" Teilgebieten der Medizin wie der Radiologie und der Pathologie[3] wird KI mit guten und sich immer weiter verbessernden Ergebnissen eingesetzt.[4] So sind Computerprogramme in der Krebsdiagnose bereits heute schneller und bisweilen signifikant präziser sowie weniger fehleranfällig als Fachärztinnen und -ärzte.[5] Auch bei der Diagnostik von Herzerkrankungen, Schlaganfällen und Augenkrankheiten wird KIG bereits erfolgreich eingesetzt.[6]

Neben der damit angesprochenen und hier im Fokus stehenden klinischen Versorgung, bei der KIG wohl am intensivsten als (bildanalytische) Unterstützung bei Diagnose-, Prognose- und Behandlungsentscheidungen zum Einsatz kommt,[7] können KI-basierte Systeme auch innerhalb von Gesundheitsorganisationen und -institutionen (z. B. zur Effizienzsteigerung von Planungs- und Managementprozessen), in der medizinischen Forschung oder im öffentlichen Gesundheitswesen eingesetzt werden.[8] Angesichts eines großen Investitionsimpetus, fortschreitenden technischen Entwicklungen und den damit verbundenen Leistungssteigerungen von KIG gilt es als ausgemacht, dass der routinemäßige Einsatz von KIG in allen Bereichen der medizinischen Versorgung keine Frage des „Ob", sondern nur eine des „Wann" ist.[9]

Das technische Ziel des *machine learning* (ML) im Bereich der Medizin besteht darin, die in digitalen Gesundheits- und Behandlungsdaten hunderter Millionen Menschen enthaltenen Informationen[10] für die Behandlung einzelner Patientinnen und Patienten fruchtbar zu machen und dem medizinischen Personal damit einen Rückgriff auf Informationen zu ermöglichen, die weit über die Praxiserfahrung der jeweils behandelnden Person(en) hinausgehen.[11] Diese Masse an Informationen kann allein mithilfe von KIG in ihrer Gänze erfasst und verarbeitet werden, wodurch etwa bislang unbekannte Krankheitsprädiktoren, Behandlungsoptionen oder unerwünschte Arzneimittelnebenwirkungen ermittelt werden können.[12] Konkret werden in den Daten Muster zu Tage gefördert, die Menschen nicht nur nicht gefunden, sondern noch nicht einmal gesucht hätten.[13]

1 Vgl. *Sonar/Weber*, in: dies. (Hrsg.), Künstliche Intelligenz und Gesundheit, 2022, 155 (155). S. zu „Artificial Intelligence in Medicine" als Forschungszweig *Combi*, Artif Intell Med 76 (2017), 37 (38).

2 Vgl. statt vieler *Hoffman,* Hastings Cent Rep 51 (2021), 8 (8).

3 Vgl. *Topol,* Deep Medicine, 2020, 21. Die klinische Forschung zum Einsatz von KI betrifft zu 40 % den Bereich der Radiologie. Weitere Forschungsschwerpunkte sind (in absteigender Reihenfolge) Pathologie, Ophthalmologie, Kardiologie und Onkologie; s. dazu *Celi et al.,* PLOS Digit Health 1 (2022):0000022, 1 (2).

4 Vgl. statt vieler *Terry,* YJHPLE 18 (2019), 133 (145); *Jorzig/Sarangi,* Digitalisierung im Gesundheitswesen, 2020, 111 f., *Chan,* in: Bielicki (Hrsg.), Regulating Artificial Intelligence in Industry, 2021, 66 (66 f.).

5 Beispiele sind etwa der „Watson Oncology Advisor" und „DeepMind" von Google; s. dazu *Winn,* Cancer Lett 44 (2018):43; *Buelens,* in: De Bruyne/Vanleenhove (Hrsg.), Artificial Intelligence and the Law, 2021, 487 (489).

6 S. dazu *Topol* (Fn. 3), 18 f.

7 S. für Beispiele *Schneider,* ZfmE 67 (2021), 327 (330); s. zu zukünftigen Potenzialen *Sonar/Weber* (Fn. 1), 162.

8 S. dazu ebd., 159 f.; *Daelman,* in: De Bruyne/Vanleenhove (Hrsg.), Artificial Intelligence and the Law, 2021, 123 (134 ff.); s. zu Einsatzbereichen von KI in der Pharmaindustrie *Heil,* PharmR 2022, 473 (474).

9 So zu Recht *Reddy et al.,* JAMIA 27 (2020), 491 (493); i. Erg. auch *Shackelford/Dockery,* Cornell J Law Public Policy 30 (2020), 279 (294). In Deutschland hat diese Entwicklung im Jahr 2019 durch § 33a SGB V krankenversicherungsrechtliche Anerkennung gefunden; s. dazu *Steege,* GuP 2021, 125 (129).

10 Etwa in Gestalt von Bildern (z.B. Computertomografie-Scans), elektronischen Krankenakten, Rezepten, medizinischer Literatur und Krankenversicherungsdaten; s. *Winn* (Fn. 5).

11 Vgl. *Rajkomar/Dean/Kohane,* N Engl J Med 380 (2019), 1347 (1347 f.).

12 Vgl. statt vieler *Price,* Harv. J Law Technol 33 (2019), 65 (71).

13 Vgl. *Hütten*, Die Große Zukunft, SZ v. 6.12.2019.

Prof. Dr. Daniel Wolff, LL.M. (Yale),

Juniorprofessor für Öffentliches Recht an der Universität Augsburg

E-Mail:
daniel.wolff@jura.uni-augsburg.de

Mit dem Einsatz von KIG sind viele Hoffnungen verbunden. Erkrankungen sollen genauer sowie schneller diagnostiziert[14] und die Entscheidungsfindung mit Blick auf Untersuchungen und Behandlungsempfehlungen von Ärztinnen und Ärzten verbessert sowie beschleunigt werden.[15] Dadurch verspricht man sich nicht nur, Behandlungsfehler mit potenziell lebensbedrohlichen Auswirkungen auf Patientinnen und Patienten zu vermeiden, sondern auch überflüssige ärztliche Besuche, Krankenhausbehandlungen und Operationen, womit wiederum weitreichende finanzielle Einsparpotenziale verbunden sind.[16] KIG soll ferner entscheidende Weiterentwicklungen innerhalb der personalisierten Medizin ermöglichen.[17] Da der Einsatz von KIG das medizinische Personal von Routinetätigkeiten entlastet und auf diese Weise Zeitkapazitäten freisetzt, verspricht man sich für die Zukunft mehr Interaktion zwischen dem Personal und den Patientinnen und Patienten sowie in der Folge bessere Behandlungsergebnisse, eine persönlichere Betreuung sowie einen erfüllenderen Arbeitsalltag für die *health professionals*.[18]

Eine weitere mit dem Einsatz von KIG verbundene Hoffnung ist die Herstellung von mehr Gleichheit im Gesundheitswesen. So soll KIG zum einen zu einer „Demokratisierung medizinischer Expertise" führen.[19] Derzeit bestehen sowohl zwischen als auch innerhalb von Staaten große Unterschiede in der Qualität der medizinischen Versorgung.[20] Diese Unterschiede beruhen unter anderem auf der unterschiedlichen Verfügbarkeit von (hoch-)qualifiziertem und spezialisiertem medizinischem Personal. Der Einsatz von KIG verspricht hingegen weitgehend unabhängig davon, wo man sich in ärztliche Obhut begibt, exzellente Gesundheitsversorgung für alle.[21]

Zum anderen verspricht KIG, menschliche Voreingenommenheit beseitigen oder zumindest reduzieren zu können.[22] Der dahinterstehende Gedanke ist folgender: Maschinen sind anders als Menschen frei von (bewussten oder unbewussten) Vorurteilen, Stereotypen, (Gefühls-)Stimmungen und kognitiven Verzerrungen.[23] Maschinen – so das Argument – haben keine Vorstellung von „Rasse", Geschlecht, Behinderung oder Religion.[24] Insoweit könnte der Einsatz von KIG Wege zu einer weniger diskriminierenden und damit „faireren"[25] Entscheidungsfindung in der Medizin eröffnen.[26]

Bei genauerem Hinsehen stellt sich die (Un-)Gleichheitsthematik im Zusammenhang mit KIG allerdings als zumindest ambivalent heraus. Denn KIG erweist sich auch als Verstärker von Ungleichheiten im Bereich der Gesundheitsversorgung, insoweit bestimmte Ungleichheiten durch maschinelle Lernverfahren aufrechterhalten oder sogar noch verstärkt werden.[27] Vor diesem Hintergrund ist die Gewährleistung von Diskriminierungsfreiheit gar als „[d]ie zentralste Herausforderung bei der KI-Regulierung" bezeichnet worden.[28]

Im Gesundheits- und Medizinwesen hat diese Diskussion erst vergleichsweise spät eingesetzt,[29] obwohl hier die Lösung der aufgeworfenen Probleme besonders dringlich erscheint.[30] Potenzielle Folgen von *biased* KIG sind nämlich Behandlungsfehler und Unter- oder Überbehandlungen,[31] die ihrerseits Schäden an Leben und Gesundheit der betroffenen Personen nach sich ziehen können.[32] KIG gilt im Folgenden dann als *biased*, wenn ihre Anwendung nicht dazu führt, dass – erstens – die KIG für alle Bevölkerungsgruppen gleichermaßen gut funktioniert *(unequal performance)*, oder – zweitens – dass die KIG allen Bevölkerungsgruppen die gleichen medizinischen Ressourcen zuweist *(unequal allocation)* oder – drittens – dass Angehörige aller Bevölkerungsgruppen gleichermaßen in puncto Diagnose- bzw. Behandlungsergebnis profitieren *(unequal outcome)*.[33] Insoweit wird der *Bias*-Begriff hier nicht pejorativ, sondern rein deskriptiv verwendet.[34]

Um Gleichheit im Gesundheitswesen sicherstellen zu können, bedürfen KIG-*Biases* der regulatorischen Adressierung.[35] Im Folgenden wird versucht, die Ungleichheitsproblematik sowie deren regulatorische Adressierung im Unionsrecht differenziert herauszuarbeiten. Dazu müssen zunächst die im Zusammenhang mit KIG verwendeten Begrifflichkeiten und die Funktionsweise von KIG geklärt werden (2). Im nächsten Schritt wird eine Typologie von *Biases* im Kontext von KIG entfaltet (3). Auf dieser Grundlage wird dann die regulatorische Adressierung des *Bias*-Problems durch das Unionsrecht erörtert (4). Am Ende steht ein kurzes Fazit (5).

14 Vgl. *Daelman* (Fn. 8), 133.

15 S. zum ganzen *Topol* (Fn. 3), 17.

16 S. dazu eingehend *PwC*, What Doctor? Why KI and Robotics Will Define New Health, Juni 2017; s. auch *Morley et al.,* Soc Sci Med 260 (2020):113172, 1 (2).

17 S. dazu *Fröhlich*, BMC Medicine 16.1 (2018): 150.

18 S. zum Ganzen *Crigger et al.,* J Med Syst 46 (2022):12, 1 (1); *Baumgartner,* in: Bauer et al. (Hrsg.), Diskriminierung und Antidiskriminierung, 2021, 149 (151).

19 S. dazu und zum Folgenden *Price* (Fn. 12), 73.

20 Vgl. *Schönberger,* Int J Law Inf Technol 27 (2019), 171 (180). Die sozialmedizinischen Erkenntnisse mit Blick auf Deutschland referiert *Schneider* (Fn. 7), 335 f. Zu Diskriminierungen von bestimmten vulnerablen Gruppen im Kontext des Zugangs zu und der Qualität von medizinischen Dienstleistungen in ausgewählten Mitgliedstaaten der Europäischen Union s. *FRA*, Inequalities and multiple discrimination in access to and quality of healthcare, 2013, passim.

21 Vgl. statt vieler *Desai,* JTLP 21 (2021), 149 (152).

22 Vgl. statt aller *High-Level Expert Group on Artificial Intelligence,* Ethics Guidelines for Trustworthy KI, 2019, 12.

23 Vgl. im hiesigen Kontext *Sonar/Weber* (Fn.1), 161; s. allgemein dazu *Kment/Borchert,* Künstliche Intelligenz und Algorithmen in der Rechtsanwendung, 2022, Rn. 120.

24 Referierend *Tischbirek,* in: Wischmeyer/Rademacher (Hrsg.), Regulating Artificial Intelligence, 2020, 103 (104); *Martini,* Blackbox Algorithmus, 2019, 47.

25 Zur Definition von Fairness als Nichtvorhandensein von Diskriminierung s. *Joos/Meding,* DuD 2022, 376 (376).

26 Vgl. *Xenidis/Senden,* in: Bernitz et al. (Hrsg), General Principles of EU Law and the EU Digital Order, 2019, 151 (153); *Hoffman/Podgurski,* YJHPLE 19 (2020), 1 (14).

27 So spezifisch für den Gesundheitsbereich *Schönberger* (Fn. 20), 181; s. auch *Dizon,* Ateneo Law J 64 (2020), 1127 (1192). S. zu weiteren „ethical challenges of KI in healthcare" *Gerke/Minssen/Cohen,* in: Bohr/Memarzadeh (Hrsg.), Artificial Intelligence in Healthcare, 2020, 295 (295).

28 So *Weber,* EuZ 2022, B1 (B12); ähnlich *Martini* (Fn. 24), 230.

29 Zur damit zusammenhängenden Rückständigkeit der Medizin mit Blick auf den Einsatz von KI s. *Topol* (Fn. 3), 22; zur Überschaubarkeit des ethischen und rechtswissenschaftlichen Schrifttums zu KIG *Schönberger* (Fn. 20), 172 u. 179; *Hoffman/Podgurski* (Fn. 26), 5 f.

30 Vgl. *Terry* (Fn. 4), 186.

31 Vgl. *Sonar/Weber* (Fn.1), 170.

32 Vgl. *Gerke/Minssen/Cohen* (Fn. 27), 304; *Henderson/Flood/Scassa,* CJLT 19 (2022), 475 (484). S. zu weiteren unbeabsichtigten Nebeneffekten des Einsatzes von KIG *Cabitza/Rasoini/Gensini,* JAMA 318 (2017), 517 (517).

33 Vgl. *Rajkoma et al.,* Ann Intern Med 169 (2018), 866 (868 f.); *Hoffman/Podgurski* (Fn. 26), 7, die diese Kriterien als Fairness-Maßstäbe heranziehen.

34 S. zu verschiedenen Verwendungen des *Bias*-Begriffs *Danks/London,* IJCAI 17 (2017), 4691 (4691).

35 Vgl. *Rajkoma et al.* (Fn. 33), 866.

2 Terminologie und Funktionsweise von KIG

Der Begriff KI wird äußerst disparat verwendet. Nach dem für den hiesigen unionsrechtlichen Kontext besonders maßgeblichen, von der Kommission in ihrem Weißbuch „Zur Künstlichen Intelligenz – Ein europäisches Konzept für Exzellenz und Vertrauen" verwendeten Begriffsverständnis kann KI als ein „Bestand an Technologien, die Daten, Algorithmen und Rechenleistung kombinieren" definiert werden.[36] Weiterführend lässt sich eine starke von einer schwachen KI unterscheiden. Starke KI, die ähnlich dem menschlichen Verstand eigene Aufgaben und Ziele jenseits vorbestimmter Parameter definieren und verfolgen kann, gibt es bislang zumindest in der Medizin nicht[37] und wäre mit Blick auf Art. 22 DSGVO rechtlich als problematisch einzustufen.[38] Die im Folgenden allein im Fokus stehende schwache KI versucht hingegen „lediglich", bestimmte Ziele innerhalb einer Reihe vordefinierter Parameter zu erreichen.[39] Das prominenteste Beispiel für schwache KIG sind *clinical decision support systems.* Diese analysieren Gesundheitsdaten, um auf Auffälligkeiten oder potenzielle Behandlungskomplikationen hinzuweisen und schlagen etwa beobachtende, medikamentöse und (post-)operative Therapieoptionen vor.[40]

Der vielleicht größte und gegenwärtig – auch und gerade im medizinischen Kontext – wichtigste Teilbereich von KI ist das *machine learning* (ML).[41] Damit werden algorithmische Modelle in die Lage versetzt, Datenmuster zu erkennen und diese dann zur Vorhersage künftiger Daten oder zur Entscheidungsfindung unter Ungewissheit zu nutzen. Anders als bei einer Programmierung mit Regeln zur Ausführung von vorgegebenen Handlungsanweisungen lernt ein *ML*-Modell aus Beispielen.[42]

Der *ML*-Prozess beginnt mit der Definition der Zielvariablen *(target variable).* Dabei handelt es sich um datenbasierte *proxies*, die KI-Anwender:innen anhand des restlichen Datensatzes vorhersagen möchten (z. B. das [Nicht-]Vorhandensein einer bestimmten Erkrankung).[43] Nach der Definition der Zielvariablen gilt es, zu ihrer Vorhersage Beispiele (Trainingsdaten) in Form von Inputdaten *(features)* und Outputdaten *(labels)* bereitzustellen. So werden etwa durch Patholog:innen analysierte und anschließend digitalisierte Gewebeproben in *features* (z. B. Pixel der digitalisierten Gewebeproben) und *labels* (z. B. Informationen, die darauf hinweisen, dass eine Gewebeprobe Anzeichen von Veränderungen enthält, die auf eine Krebserkrankung hindeuten) umgewandelt. Mithilfe eines ersten Algorithmus, des *learner*, wird sodann die Zuordnung von *features* zu *labels* dahingehend bestimmt, dass ein Informationen verallgemeinerndes Modell erstellt wird, das die Testdaten (z. B. neue, noch nicht analysierte Gewebeproben) zutreffend analysiert (etwa auf die Frage hin, ob Gewebeveränderungen vorliegen oder nicht).[44] Aus den *labeled features* (z. B. die kategorisierten Gewebeproben) setzt sich wiederum die *ground truth* zusammen. Das Gebrauchmachen von dem Modell zur Analyse neuer Inputdaten übernimmt schließlich ein weiterer Algorithmus, der *classifier.*[45]

3 Typologie von Biases im Kontext von KIG

Biases im Zusammenhang mit KIG treten insbesondere als *Algorithmic Bias*, *Data Bias* und *Implementation Bias*[46] auf.[47] Der *Algorithmic Bias*[48] umfasst Ungleichbehandlungen, die allein aus dem spezifischen Design des *ML*-Modells folgen, während aus den Trainingsdaten resultierende Ungleichbehandlungen dem *Data Bias* zuzuordnen sind. Die im Folgenden nicht weiter behandelte Kategorie des *Implementation Bias* betrifft schließlich Ungleichbehandlungen, die im Zuge des Gebrauchmachens von KIG durch das Gesundheitspersonal in der Interaktion mit den Patientinnen und Patienten hervorgerufen werden.[49]

3.1 Algorithmic Bias

Ein *Algorithmic Bias* kann seinen Ursprung in verschiedenen Stufen des *ML*-Prozesses haben, wobei insbesondere der *Labeling Bias* vom *Modeling Bias* zu unterscheiden ist.[50]

3.1.1 Labeling Bias

Der *Algorithmic Bias* tritt zunächst im Kontext der Spezifikation der Zielvariablen bzw. des *labeling* auf (*Labeling Bias*).[51] Bei der genauen Spezifikation der Zielvariablen, die eine Übersetzung des Problems in eine von Computern analysierbare Frage beinhaltet, bzw. bei der Wahl der *labels* kommt der betreffenden KI-Entwickler:in ein signifikanter Spielraum zu. In der Folge läuft das Gebrauchmachen von diesem Spielraum Gefahr, zu unbeabsichtigten Benachteiligungen bestimmter Gruppen zu führen.[52]

36 COM (2020) 65 final, 19.2.2020, 2.

37 S. dazu *Morley et al.* (Fn. 16), 2.

38 S. dazu *Schneider* (Fn. 7), 330; *v. Ungern-Sternberg*, in: Mangold/Payandeh (Hrsg.), Hdb, AntidiskrR, 2022, § 28 Rn. 45.

39 Ähnlich *Henderson/Flood/Scassa* (Fn. 32), 477.

40 S. dazu *Chan* (Fn. 4), 67; *Mittelstadt et al.,* Big Data Soc 3 (2016), 1 (3).

41 Vgl. *Gerke/Minssen/Cohen* (Fn. 27), 296; s. zu *Deep Learning* als besonders leistungsstarke *ML*-Form *Rajkomar/Dean/Kohane* (Fn. 11), 1348.

42 Vgl. *Shalev-Shwartz/Ben-David*, Understanding Machine Learning, 2014, S. 2 ff.

43 Vgl. *Barocas/Selbst,* Cal L Rev 104 (2016), 671 (678). Konzepten wie „Gesundheit" müssen für Computersysteme in Gestalt diskreter Variablen formalisiert werden; vgl. *Hildebrandt,* EDPL 7 (2021), 358 (361).

44 So *Rajkomar/Dean/Kohane* (Fn. 11), 1348; s. dazu auch *Barocas/Selbst* (Fn. 43), 680. Insoweit wird der Begriff des Algorithmus hier minimalistisch als „sequence of computational steps that transform the input into the output" verstanden; s. dazu *Cormen et. al.*, Introduction to Algorithms, 3. Aufl. 2009, 5.

45 Zur Unterscheidung zwischen *learner* und *classifier* s. *Burrell*, Big Data & Society 3 (2016):1, 1 (3).

46 Im Mittelpunkt stehen unbeabsichtigte Ungleichbehandlungen. Allerdings können faktisch alle durch den Einsatz von KIG hervorgerufenen Ungleichbehandlungen auch vorsätzlich herbeigeführt werden; s. dazu *Barocas/Selbst* (Fn. 43), 692 f. m.w.N.

47 Ähnlich *Norori et al.,* Patter 2 (2021):100347, 1 (1). Eine Übersicht über verschiedene Kategorisierungsversuche bietet *Xenidis/Senden* (Fn. 26), 155.

48 Teilweise wird insoweit auch von *Coding Bias* oder *ML Bias* gesprochen; s. *Henderson/Flood/Scassa* (Fn. 32), 480 f.

49 S. dazu *Rajkoma et al.* (Fn. 33), 868, sowie im Anschluss daran *Giovanola/Tribelli,* Beyond bias and discrimination, KI & Soc 2022, 1 (4), wo zwischen Diskriminierungen bei der Interaktion von KIG mit dem Gesundheitspersonal bzw. mit den Patientinnen und Patienten unterschieden wird.

50 Angesichts dessen, dass jegliche Algorithmen von Menschen programmiert werden, ist zu erwarten, dass die programmierten Algorithmen von den *Biases*, denen Menschen unterliegen nicht unbeeinflusst bleiben; s. dazu statt aller *Macnish*, Ethics Inf Technol 14 (2012), 151 (158). Zudem ist zu berücksichtigen, dass die Gesamtheit der KI-Entwickler:innen weder die Diversität der Weltbevölkerung noch diejenige der Bevölkerungen westlicher Industrienationen, zu denen sie hauptsächlich gehören, abbilden; vgl. *Crigger et al.* (Fn. 18), 4 m.w.N. Vielmehr sind Beschäftigte in Forschung und Entwicklung von KI ganz überwiegend weiße und asiatische Männer; vgl. *Gebru*, in: Dubber/Pasquale/Das (Hrsg.), The Oxford Handbook of Ethics of KI, 2020, 251 ff.; s. dazu auch *Rubenstein,* Fla L Rev 73 (2021), 747 (776).

51 Ähnlich *Wang et al.*, JAMIA 29 (2022), 1323 (1327).

52 Vgl. *Barocas/Selbst* (Fn. 43), 678 u. 715.

Ein Beispiel für den *Labeling Bias* ist der viel diskutierte Fall eines in den USA verwendeten *high-risk care-management*-Programms. Konkret ging es um eine KIG, mit deren Einsatz Patientinnen und Patienten mit besonderen medizinischen Bedarfen identifiziert werden sollten, um diesen zusätzliche medizinische Versorgung zuteil werden zu lassen und dadurch bessere Behandlungsergebnisse zu erzielen.[53] Das eigentliche Prognoseziel war demnach der zukünftige medizinische Versorgungsbedarf der Patientinnen und Patienten.[54] Als konkrete Zielvariable[55] waren hingegen die erwarteten Behandlungskosten definiert worden, die auf Grundlage der in der Vergangenheit bei den jeweiligen Personen angefallenen Ausgaben für Gesundheitsdienstleistungen prognostiziert wurden.[56] Unterstellt wird dabei, dass medizinische Ausgaben mit dem Bedarf an medizinischer Versorgung gleichzusetzen sind, sodass Patientinnen und Patienten mit den höchsten medizinischen Ausgaben von den zusätzlichen Behandlungsangeboten am meisten profitieren würden.[57] Eine in *Science* veröffentlichte Studie zeigte jedoch auf, dass die in das Programm aufgenommenen Afroamerikaner:innen im Vergleich zu ebenfalls aufgenommenen weißen Patientinnen und Patienten signifikant mehr Erkrankungen und höhere Versorgungsbedarfe aufwiesen. Der Algorithmus interpretierte die geringeren monetären Ausgaben von Afroamerikaner:innen fälschlicherweise als Hinweis darauf, dass sie vergleichsweise gesund und deshalb nicht in das Programm aufzunehmen sind. Unberücksichtigt blieb, dass Afroamerikaner:innen in den USA signifikant häufiger mit Hindernissen beim Zugang zur Gesundheitsversorgung konfrontiert sind und sie aufgrund fehlender finanzieller Mittel durchschnittlich deutlich weniger Geld für Gesundheitsdienstleistungen ausgeben als weiße US-Amerikaner:innen. Die Studie kam zu dem Ergebnis, dass der Anteil von Afroamerikaner:innen in dem Programm von 17,7% auf 46,5% steigen würde, wenn man auf den tatsächlichen Gesundheitszustand abstellen würde.[58] Das Kriterium der in der Vergangenheit angefallenen Gesundheitskosten zeigte sich somit nur innerhalb bestimmter Bevölkerungsgruppen als aussagekräftig, wohingegen es bei bevölkerungsgruppenübergreifender Anwendung zu schwerwiegenden Benachteiligungen führte.

3.1.2 Modeling Bias

Modeling Bias bezeichnet ungerechtfertigte Ungleichbehandlungen, die aus dem spezifischen Design eines *ML*-Modells resultieren.[59] Ein in diesem Zusammenhang vielfach beschriebenes Phänomen sind medizinisch ungerechtfertigte Ungleichbehandlungen infolge der *feature selection*, also der Entscheidung darüber, welche Inputdaten konkret in das Modell einbezogen werden. *Feature selection* ist letztlich unumgänglich: Damit KI Prognosen abgeben kann, muss die Welt simplifiziert werden. Hinzu kommt, dass die Abbildung von Komplexität sehr viele Daten erfordert, was unter Umständen datenschutzrechtliche Probleme verursacht[60] und vor allem sehr teuer ist.[61] Vor diesem Hintergrund werden *ML*-Modelle nicht mit unendlich vielen, sondern allein mit solchen Inputdaten erstellt, die nach Auffassung der KI-Entwickler:innen für ein hinreichend aussagekräftiges Modell ausreichen.[62] In der Folge werden Entscheidungen mithilfe von KI auf der Grundlage nicht oder zumindest nur schwer verallgemeinerbarer Kategorisierungen getroffen.[63]

Feature selection kann Ungleichbehandlungen zulasten von Minderheiten und besonders vulnerablen Gruppen zur Folge haben, wenn die ausgewählten *features* die Besonderheiten besagter Gruppen aufgrund ihrer fehlenden Spezifität nicht abbilden.[64] Ein klassisches Beispiel außerhalb des medizinischen Kontexts betrifft die Auswahl bzw. Bevorzugung von Bewerber:innen im Rahmen von Einstellungsverfahren, die einen Abschluss von einer prestigeträchtigen Bildungseinrichtung erworben haben. Die Fokussierung auf das Inputdatum „Bildungseinrichtung" ist eine leicht zugängliche und kostengünstige Information. Überdies weist dieser Parameter statistisch gesehen durchaus eine gewisse Aussagekraft hinsichtlich der schwer vorherzusagenden Jobperformance einer entsprechenden Bewerber:in auf, auch wenn eine holistische Berücksichtigung verschiedener Parameter, die auf das Vorhandensein der für die jeweilige Stelle notwendigen Fähigkeiten schließen lassen, eine noch deutlich größere Aussagekraft hätte. Da der Besuch entsprechender Bildungseinrichtungen in vielen Staaten mit hohen Kosten verbunden ist, besuchen Angehörige sozial schwächerer Bevölkerungsgruppen diese Institutionen jedoch kaum und werden insoweit durch die vorgenommene *feature selection* benachteiligt.[65]

3.2 Data Bias

Je höher Qualität und Quantität der Trainingsdaten sind, desto besser ist die Funktionalität der KI.[66] Für die *Bias*-Thematik bedeutet dies, dass ein in den Trainingsdaten enthaltener *Bias* direkt in die Funktionsweise des *ML*-Modells eingespeist wird, das sodann seinerseits als Grundlage für zukünftige menschliche Entscheidungen zulasten der betroffenen Bevölkerungsgruppen fungiert.[67] In Abwandlung des informationstechnischen *GIGO*-Prinzips („garbage in, garbage out") kann man insofern auch vom *BIBO*-Grundsatz sprechen („bias in, bias out").[68]

Konkret lassen sich vier Formen des *Data Bias* unterscheiden. Dies sind auf der einen Seite der *Measurement Bias* und der *Population Bias* sowie auf der anderen Seite der *Historical Bias* und der *Statistical Bias*.

53 S. *Obermeyer et al.,* Science 366 (2019), 447 (447).
54 Vgl. *Wang et al.* (Fn. 51), 1327.
55 In der Literatur finden sich insoweit unterschiedliche Zuordnungen. Während *Wang et al.* (Fn. 51), 1327 und *Orwat,* Diskriminierungsrisiken durch Verwendung von Algorithmen, 2021, 53, wie hier ein Problem der Wahl bzw. Spezifikation der Zielvariable erkennen, diskutiert u.a. die Studie den Fall als Ungleichbehandlung infolge von „label choice"; s. *Obermeyer et al.* (Fn. 53), 452 f.; so auch *Omar,* Denver Law Rev 98 (2021), 807 (810); *Wiens/Price II/Sjoding,* Nat Med 26 (2020), 25 (25).
56 S. dazu *Schneider* (Fn. 7), 334.
57 Vgl. *Wiens/Price II/Sjoding* (Fn. 55), 25.
58 S. dazu und zum Folgenden *Obermeyer et al.* (Fn. 53), 447 ff.
59 So auch *Wang et al.* (Fn. 51), 1327.
60 Vgl. dazu auch *v. Ungern-Sternberg* (Fn. 38), § 28 Rn. 75.
61 Vgl. *Xenidis/Senden* (Fn. 26), 160.
62 Vgl. *Daelman* (Fn. 8), 140.
63 Vgl. *Wang et al.* (Fn. 51), 1327.
64 Vgl. *Barocas/Selbst* (Fn. 43), 688.
65 Vgl. *Xenidis/Senden* (Fn. 26), 160.
66 Vgl. *Daelman* (Fn. 8), 138.
67 Vgl. *Barocas/Selbst* (Fn. 43), 671, 683 u. 687 f..
68 So der Titel des Aufsatzes von *Mayson*, Yale Law Journal 128 (2018), 2218; s. auch *Fukuda-Parr,* Glob Policy 12 (2021), 32 (35 f.).

3.2.1 Measurement Bias

Der *Measurement Bias* ist die Folge von Qualitätsunterschieden der Trainingsdaten für verschiedene Bevölkerungsgruppen sowie von Unterschieden in der Auswahl und Berechnung der Trainingsdaten für diese Gruppen.[69] So sind etwa Gesundheitsdaten von sozial schwächeren und marginalisierten Bevölkerungsgruppen, die klassischerweise weniger bzw. weniger gute medizinische Versorgung erhalten, regelmäßig besonders fehler- und lückenhaft.[70] In der Folge wird die Funktionalität der KIG mit Blick auf diese Gruppen schlechter ausfallen.[71]

3.2.2 Population Bias

Der *Population Bias*[72] entsteht, wenn die KI deshalb eine schlechtere Performance für bestimmte Bevölkerungsgruppen erbringt, weil die für das Training des *ML*-Modells verwendeten Daten die für den Einsatz der KI vorgesehenen Bevölkerungsgruppen, die im Gesundheitsbereich regelmäßig die gesamte Bevölkerung umfassen, nicht repräsentativ abbilden.[73] Dieser *Bias* spielt in der Praxis eine besonders große Rolle,[74] sind doch bestimmte Bevölkerungsgruppen in den bislang existierenden und genutzten Trainingsdaten regelmäßig deutlich überrepräsentiert (insbesondere weiße Männer) und andere, insbesondere vulnerable und marginalisierte Gruppen deutlich unterrepräsentiert.[75] In der Folge erstellt KIG, die auf entsprechenden Datensätzen beruht, z. B. erheblich schlechtere Vorhersagen für schwarze Frauen.[76]

Neben den vielfältigen (sprachlichen, sozioökonomischen usw.) Hindernissen, die bestimmte Minderheiten für die Inanspruchnahme von Gesundheitsdienstleistungen überwinden müssen[77] und die in der Folge zu einem Mangel an Gesundheitsdaten von diesen Gruppen führen,[78] besteht ein weiterer Grund für die Unterrepräsentation von bestimmten (Minderheiten-) Gruppen in den existierenden Trainingsdaten darin, dass umfassende, für das Training von *ML*-Modellen vielfach eingesetzte Datensätze oftmals aus medizinischen Einrichtungen stammen, deren Patientinnen und Patienten die Gesamtbevölkerung nicht abbilden. In der Literatur wird etwa darauf hingewiesen, dass „MIMIC-III", eine große, allgemein genutzte und öffentlich zugängliche Datenbank für die Entwicklung von KI-Anwendungen in der Intensivmedizin ausschließlich aus Daten besteht, die aus der Intensivmedizin eines großen Krankenhauses in Boston, Massachusetts stammen, das eine ganz überwiegend aus wohlhabenden weißen US-Amerikaner:innen zusammengesetzte Patientenpopulation aufweist. Werden KIG-Anwendungen demnach mit diesen Daten „trainiert", werden sie für andere Patientenpopulationen eine schlechtere Funktionalität aufweisen.[79]

Weitere Beispiele für ungleiche KIG-Performances aufgrund nicht repräsentativer Trainingsdaten sind Legion. Angefangen vom Einsatz einer KIG zur Bildanalyse in der Dermatologie, die mit vornehmlich von weißen Patientinnen und Patienten stammenden Daten trainiert wurde und deren Vorhersagen für schwarze Patientinnen und Patienten signifikant weniger präzise waren als für weiße[80], über eine KIG zur Analyse von Thoraxbildern, die infolge der Unterrepräsentation von Frauen in den Trainingsdaten schlechtere Analysen für Frauen als für Männer vorlegte[81], bis hin zur KIG eines kanadischen Unternehmens, das anhand von Sprachproblemen eine Alzheimererkrankung diagnostizieren sollte, dies mit hinreichender Genauigkeit aber nur für englische Muttersprachler:innen eines kanadischen Dialekts leisten konnte, da bei der KI-Entwicklung allein Trainingsdaten von dieser Bevölkerungsgruppe verwendet worden waren.[82]

3.2.3 Historical Bias

Der dritte Subtyp des *Data Bias* ist der *Historical Bias*, der auftritt, wenn die KIG deshalb unterschiedlich gute Prognosen für verschiedene Bevölkerungsgruppen abgibt, weil die Trainingsdaten von den *Folgen* überkommener Diskriminierungen bestimmter Bevölkerungsgruppen geprägt sind.[83] Auf diese Weise werden besagte Diskriminierungen und ihre Folgen in Gestalt der von der KIG in der Vergangenheit abgegebenen Prognosen in die Zukunft fortgeschrieben und so perpetuiert.[84] Zu beachten ist allerdings, dass die Trainingsdaten zwar von historischen Diskriminierungen geprägt sind, aber die gegenwärtigen sozialen Realitäten durchaus widerspiegeln[85] und daher die Funktionalität der KIG nicht zwingend beeinträchtigen. Ein illustratives Beispiel dafür ist eine in den USA zur effizienten Terminorganisation in Arztpraxen eingesetzte KIG.[86] Diese soll u. a. diejenigen Patientinnen und Patienten identifizieren, die mit erhöhter Wahrscheinlichkeit ihren vereinbarten Arzttermin versäumen werden. Um entsprechende Prognosen anzustellen, wird entscheidend auf die Versäumnisrate der Patientinnen und Patienten in der Vergangenheit abgestellt. Die entsprechend identifizierten Personen werden in der Folge bei der Terminplanung gezielt überbucht, um im Falle einer Terminversäumung keine Behandlungspausen entstehen zu lassen. Da Afroamerikaner:innen aufgrund ihrer tendenziell schlechteren sozioökonomischen Lage etwa mit mehr Problemen bei der Anfahrt oder Kinderbetreuung zu kämpfen haben sowie weniger berufliche Flexibilität aufweisen, verpassen sie statistisch signifikant häufiger Arzttermine. Da sich die soziale Situation von Afroamerikaner:innen aller Voraussicht nach in absehbarer Zeit nicht fundamental verändern wird und deshalb auch in Zukunft mit deutlich häufigeren Terminversäumnissen von Afroamerikaner:innen zu rechnen ist, stellt sich die KIG als sehr funktional dar. Ihr Einsatz führt in der Konsequenz allerdings dazu, dass Afroamerikaner:innen wesentlich häufiger als

69 Vgl. *Wang et al.* (Fn. 51), 1327.
70 Vgl. *Hoffman* (Fn. 2), 8; *Daelman* (Fn. 8), 139.
71 Ähnlich ebd., 138.
72 Andere Bezeichnungen sind *Data Selection* oder *Representation Bias*.
73 Vgl. statt aller *Wang et al.* (Fn. 51), 1327; *Henderson/Flood/Scassa* (Fn. 32), 481.
74 Vgl. *Joos/Meding* (Fn. 25), 378.
75 Vgl. *Schneider* (Fn. 7), 332; s. auch *Mehrabi et al.,* ArXiv 2019, 4 (5); *Schönberger* (Fn. 20), 180.
76 Vgl. *Vayena/Blasimme/Cohen*, PLoS Med 15 (2018):e1002689.
77 Vgl. *Wójcik,* Health Hum Rights 24 (2022), 93 (98).
78 Vgl. *Schönberger* (Fn. 20), 181 m.w.N.
79 S. *Henderson/Flood/Scassa* (Fn. 32), 482 m.w.N.
80 In der Folge droht eine weitere Vergrößerung der Unterschiede in den Überlebensraten bei Hautkrebserkrankungen; vgl. ebd., 482.
81 Vgl. ebd., 482.
82 S. dazu *Schneider* (Fn. 7), 337; *Hoffman/Podgurski* (Fn. 26), 17 f. Ein weiteres Beispiel ist der Einsatz Bilddiagnose-KIG „Watson for Oncology" in China; s. dazu *Liu et al.*, J Med Internet Res 20 (2018):e11087.
83 Vgl. *Barocas/Selbst* (Fn. 43), 671.
84 Vgl. *Datenethikkommission,* Gutachten, 2019, 167.
85 Deshalb wird teilweise auch von *Social Bias* gesprochen. So etwa *Wachter/Mittelstadt/Russell,* West VA Law Rev 123 (2021), 735 (742).
86 S. zum Folgenden *Samorani et al.,* Manuf Serv Oper Manag 2021, 1 (1).

andere Gruppen überbucht werden, weshalb sie ca. 30 % längere Wartezeiten in den Arztpraxen auf sich nehmen müssen und auf diese Weise durch die Anwendung der KIG ungleich behandelt werden.[87]

3.2.4 Statistical Bias

Der *Statistical Bias* liegt dann vor, wenn Ungleichbehandlungen hervorgerufen werden, die von der ungleichen Verteilung von Fähigkeiten, Risiken und anderen Zielvariablen zwischen verschiedenen Bevölkerungsgruppen herrühren.[88] Der *Statistical Bias* weist Ähnlichkeiten zum *Historical Bias* auf, insofern hier auch Ungleichbehandlungen in Rede stehen, die durch ein *ML*-Modell hervorgerufen werden, das mit repräsentativen und fehlerfreien Datensätzen trainiert wurde und letztlich so funktioniert wie es funktionieren soll.[89] Während der *Historical Bias* allerdings Ergebnis existenter Unterschiede zwischen Bevölkerungsgruppen ist, die Folgen historischer Diskriminierungen und damit historisch kontingent sind, rühren die Unterschiede in den Trainingsdaten im Fall des *Statistical Bias* von natürlichen bzw. naturwissenschaftlich feststellbaren Unterschieden zwischen den jeweiligen Bevölkerungsgruppen her.

Für *Statistical Bias* gilt bisweilen: „it's not a bug, it's a feature!", nämlich wenn bestimmte Kriterien wie Geschlecht oder Ethnizität in der Diagnose oder Therapie einer Krankheit eine Rolle spielen.[90] Ein Beispiel ist der systemische Lupus erythematodes (SLE), eine schwere rheumatologische Autoimmunerkrankung, an der signifikant mehr Frauen als Männer erkranken und die eine deutlich höhere Prävalenz bei Menschen afrikanischer, asiatischer oder hispanischer Abstammung aufweist.[91] Vor diesem Hintergrund müssen Informationen über Geschlecht und ethnische Zugehörigkeit in die Trainingsdaten für ein Diagnoseprogramm für SLE aufgenommen werden, um dessen größtmögliche Genauigkeit sicherzustellen.[92] Die Ungleichbehandlung bestimmter Bevölkerungsgruppen ist demnach für die Funktionalität der entsprechenden KIG sowie für eine optimale Gesundheitsversorgung aller Patientinnen und Patienten notwendig.

In anderen Fällen kann ein *Statistical Bias* hingegen sehr wohl als „bug" angesehen werden. *Tischbirek* weist in diesem Zusammenhang auf die Entscheidung des EuGH in der Rechtssache *Association beige des Consommateurs Test-Achats* hin, in der die Luxemburger Richter:innen Geschlechtsdifferenzierungen in Versicherungsverträgen als unionsrechtswidrig einstuften, laufe eine entsprechende Differenzierung doch der Verwirklichung des mit der Richtlinie 2004/113/EG verfolgten Ziels der Gleichbehandlung von Frauen und Männern zuwider und sei überdies mit Art. 21, 23 GRC unvereinbar.[93] Obwohl Versicherungen aus statistischer Perspektive allen Grund haben, geschlechtsspezifische Tarife anzubieten,[94] zeigt dieser Fall, dass (vermeintlich) rationale Ungleichbehandlungen politisch und in der Folge auch rechtlich abgelehnt werden können.[95]

Fortsetzung des Beitrags („Die regulatorische Adressierung des Bias-Problems von KIG durch das Unionsrecht") in Heft 01/2023

87 S. *Samorani/Blount,* AJPH 110 (2020), 440 (440).

88 Ähnlich *Hacker,* CMLRev 55 (2018), 1143 (1148) und *Harisimiuk/Braun,* Regulating Artificial Intelligence, 2021, 80, die diesen *Bias* terminologisch als *unequal ground truth* erfassen.

89 Vgl. *Schwarcz,* Hous J Health L & Pol'y 21 (2021), 95 (109).

90 Ähnlich *Starke/De Clercq/Elger,* Med Health Care Philos 24 (2021), 341 (343).

91 S. dazu *Lewis/Jawad*, Rheumatology 56 (2017), i67 ff.

92 Vgl. *Starke/De Clercq/Elger* (Fn. 90), 342.

93 S. *EuGH, C-236/09*, ECLI:EU:C:2011:100**,** Rn. 32.

94 So haben Frauen eine höhere Lebenserwartung und eine niedrigere Autounfallwahrscheinlichkeit als Männer, sodass Lebens- und Autoversicherungspolicen für Männer höher anzusetzen wären.

95 S. dazu *Tischbirek* (Fn. 24), 107.