AEQUITAS

D2.1 Bias-Bericht

Kofinanziert von der Europäischen Union

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Inhaltsverzeichnis

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Kofinanziert von der Europäischen Union

AEQUITAS

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Kofinanziert von der Europäischen Union

AEQUITAS

Einleitung

Im Rahmen des AEQUITAS-Projekts wird eine Datenbank zu gender- und rassenspezifischen Bias in medizinischen Anwendungen der Künstlichen Intelligenz (KI) erstellt, wobei der Fokus auf drei Krankheiten liegt: Herz-Kreislauf-Erkrankungen, Diabetes und Depressionen.

Um diese Aufgabe zu erfüllen, haben die Konsortialpartner zunächst verschiedene bibliographische Quellen zu den oben genannten Bias-Arten erfasst. Das Universitätsklinikum Köln (UKK) hat als Task Leader und Fachexperte die Informationsbeschaffung organisiert und eine Vorlage zur Zuordnung (Mapping) bereitgestellt, mit der die Projektpartner die bibliographischen Quellen zugeordnet haben, um sicherzustellen, dass die relevanten Informationen problemlos in die Datenbank übertragen werden können.

Der vorliegende Bericht präsentiert die theoretischen und wissenschaftlichen Grundlagen, die die Auswahl der Erfassungsaktivität und des Mapping-Templates geleitet haben, begleitet von Fallstudien, die die verschiedenen Arten von Bias aufzeigen; die policy-relevanten Implikationen, wonach durch biomedizinische KI induziertes Bias die durch die Charta der Grundrechte der Europäischen Union geschützten Rechte beeinträchtigen; eine Beschreibung der Datenerfassungsaktivität; das Mapping-Template; eine Liste der von den AEQUITAS-Partnern gesammelten bibliographischen Quellen; sowie weiteres Begleitmaterial. Der Rest des Berichts ist wie folgt gegliedert:

Zuerst stellen wir in der Einführung die theoretischen Grundlagen unserer Arbeit vor, die sich mit den Anwendungen der KI in der Medizin sowie mit dem Konzept des Bias in Computersystemen und in der Medizin befasst. Wir legen zunächst den Fokus auf die Medizin und zeigen, wie sich Bias in Bezug auf rasse- und genderbezogene Diskriminierung in der medizinischen Versorgung manifestiert, und anschließend, wie

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

ethische Fragen, die sich in der medizinischen Praxis und der biomedizinischen Forschung ergeben, durch die Bioethik behandelt werden. Dabei geben wir eine kurze Einführung in die vier Prinzipien der Bioethik: Autonomie, Nichtschaden, Wohltun und Gerechtigkeit.

Als Nächstes wenden wir uns dem Bereich der KI zu und stellen die Arten von Bias vor, die in KI-Systemen auftreten können, wie sie sich in der Pipeline für Maschinelles Lernen und Künstliche Intelligenz (ML/KI) manifestieren. Jede Art von Bias wird durch Beispiele und eine Fallstudie zu gender- und rassenspezifischen Bias sowie deren Auswirkungen auf gesellschaftlicher Ebene in Bezug auf die drei Schwerpunktthemen des AEQUITAS-Projekts (Herz-Kreislauf-Erkrankungen, Diabetes und Depressionen) ergänzt, die aus bibliographischen Quellen stammen, die die AEQUITAS-Partnern nach Abschluss von T2.2 gesammelt haben. Wenn dies nicht möglich war, weil das erfasste Material die spezifische Art des betrachteten KI-Bias nicht eindeutig belegte, wurde ein alternativer Fall aus einem anderen medizinischen Bereich vorgestellt, der sich einfach auf die AEQUITAS-Zielkrankheiten übertragen ließ. Die Fallstudienbeschreibungen stützen sich bei Bedarf auf zusätzliche wissenschaftliche Quellen, um sie zu belegen.

Abschließend wird in der Sektion „Policybezogene Implikationen“ aufgezeigt, wie sich die verschiedenen Arten von KI-Bias auf die durch die Charta der Grundrechte der Europäischen Union geschützten Grundrechte auswirken, insbesondere auf die Grundsätze der Menschenwürde, der Gleichheit vor dem Gesetz, der Nichtdiskriminierung sowie auf das Recht auf Unversehrtheit der Person, das Recht auf Gesundheitsschutz, den Datenschutz und das Recht auf einen wirksamen Rechtsbehelf und schließt mit den Schutzmaßnahmen, die bei Konformitätsbewertungen, der Überwachung nach dem Marktzugang und der öffentlichen Beschaffung getroffen werden können.

Der Bericht schließt mit den Referenzen und den folgenden Anhängen:

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Anhang 1: Quellenerfassung und Mappingmethode, die das Mapping-Template enthält und den Prozess der Erfassung, des Mappings und der Informationsbewertung beschreibt, der während der Aufgaben T2.1 und T2.2 durchgeführt wurde.

Anhang 2: Enthält Begleitmaterial für die Aufgaben T2.1 und T2.2, d. h. Folien aus Partnermeetings, in denen der Prozess beschrieben wird, präsentiert von UKK.

Anhang 3: Die Liste der von den AEQUITAS-Partnern gesammelten bibliographischen Quellen.

KI-Anwendungen in der Medizin

Der Aufstieg von KI-Anwendungen in den letzten Jahren hat die Medizin stark beeinflusst, darunter die digitalisierte Datenerfassung, Maschinelles Lernen und die Recheninfrastruktur (Yu et al., 2018). Insbesondere die Einführung von Deep-Learning-Algorithmen in Bereichen wie Computer Vision und Natürlichsprachliche (Natural Language Processing-NLP) Systeme hat Computeranwendungen in der Radiologie, Pathologie, Kardiologie, Diabetologie, Psychiatrie, Onkologie usw., revolutioniert (Esteva et al., 2019; Koteluk et al., 2021; Rajpurkar et al., 2022; Gou et al., 2024). Die Weltgesundheitsorganisation (WHO) listet die folgenden Anwendungsbereiche von KI-Systemen im Healthcare-Bereich auf: Diagnose und prädiktive Diagnostik, klinische Versorgung, Forschung und Arzneimittelentwicklung, Management und Planung von Gesundheitssystemen, öffentliche Gesundheit und Gesundheitsüberwachung, Gesundheitsförderung, Krankheitsprävention, prädiktive Überwachung, Notfallvorsorge und Ausbruchsbekämpfung (World Health Organization, 2021).

Die Einführung von KI-Anwendungen in der Medizin bringt jedoch eine Reihe von Herausforderungen mit sich, darunter Implementierungsprobleme wie Modellvertrauen und Datenbeschränkungen, Fragen der Rechenschaftspflicht, darunter regulatorische Herausforderungen und die ordnungsgemäße Zuweisung von Zuständigkeiten, sowie die Gewährleistung von Gerechtigkeit durch ethischen

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Datengebrauch, gerechte Verteilung der Vorteile und Erkennung und Minderung von Bias (Rajpurkar et al., 2022).

Der Fokus des AEQUITAS-Projekts liegt auf Fällen mit gender- und rassenspezifischem Bias bei Herz-Kreislauf-Erkrankungen, Diabetes und Depressionen. Medizinische KI-Anwendungen unterstützen die kardiovaskuläre Versorgung durch klinische Entscheidungunterstützung, Telemedizin, Risikobewertung, personalisierte Therapie, prädiktive Analytik und Fernüberwachung (Bernstein et al., 2025; Naskar et al., 2025) und verbessern das Diabetes-Management (einschließlich Patientenüberwachung und Selbstmanagement), die Diagnose, Behandlung und Prävention (Contreras & Vehi, 2018; Khalifa & Albadawy, 2024; Naskar et al., 2025; Sheng et al., 2024). In Bezug auf Depressionen sind sie an der Vorsorgeuntersuchung, Diagnose und Behandlung beteiligt (Alhuwaydi, 2024), wobei ein besonderer Fokus auf der Erkennung und Früherkennung unter Verwendung von Large Language Models (LLMs) liegt (Cao et al., 2025; Kumari et al., 2025; Mao et al., 2023; Wang et al., 2025). In allen oben genannten Bereichen gibt es Herausforderungen hinsichtlich Bias, beispielsweise in Bezug auf Herz-Kreislauf-Erkrankungen, Diabetes und Depressionen, siehe (van Assen et al. 2024), (Cronjé et al. 2023), (Dang et al. 2024).

Herausforderungen wie Bias, sei es in der Medizin oder in der KI, werden durch eine Kombination aus Bioethik und KI-Ethik angegangen. In der folgenden Sektion geben wir einen kurzen Überblick über diese beiden Bereiche der angewandten Ethik, die als theoretische und wissenschaftliche Grundlage für die Entwicklung des Bias Mapping Templates dienten.

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Ethik und Bias in der Medizin und KI

Bioethik und Bias in der Medizin

Bias in der Medizin ist gut dokumentiert: Siehe beispielsweise (Hammond et al. 2021) zu kognitivem Bias, das aus systematischen Denkfehlern aufgrund menschlicher Verarbeitungsbeschränkungen oder inadäquater mentaler Modelle besteht, und (FitzGerald und Hurst 2017) zu implizitem Bias, das Assoziationen außerhalb des Bewusstseins umfasst, die zu einer negativen Bewertung einer Person auf der Grundlage irrelevanter Merkmale wie Rasse oder Gender führen.

Rassebezogene Bias in der Medizin sind beispielsweise in den USA gut untersucht. Dort ist dokumentiert, dass Afroamerikaner sowie Angehörige anderer Minderheiten weniger medizinische Behandlungen und eine schlechtere medizinische Versorgung erhalten, da sie weniger aggressive Behandlungen bekommen, seltener operiert werden und seltener an Spezialisten überwiesen werden als Weiße (Bowser, 2001; Williams & Wyatt, 2015).

Genderspezifische Vorurteile lassen sich auf Geschlechterblindheit und stereotype Vorurteile gegenüber Männern und Frauen zurückführen (Hamberg, 2008), hinzu kommt ein allgemeiner Mangel an Wissen über die Funktionsweise des weiblichen Körpers und seine biologischen Unterschiede zum männlichen Körper. Beispielsweise wurden kritisch kranke Frauen ab 50 Jahren seltener als kritisch kranke Männer auf eine Intensivstation aufgenommen (Bierman, 2007), und selbst männliche Mausmodelle sind in der grundlegenden, präklinischen und chirurgischen biomedizinischen Forschung insgesamt stärker vertreten als weibliche Modelle (Yoon et al., 2014).

Es ist auch wichtig zu beachten, dass LGBT+-Personen beim Zugang zur Gesundheitsversorgung diskriminiert werden und Stereotypen ausgesetzt sind, die die

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Kofinanziert von der Europäischen Union

AEQUITAS

heterosexuelle Bevölkerung nicht betreffen. Diese sozialen und kulturellen Faktoren perpetuieren Diskriminierung und wirken sich auf die Gesundheit aus. Eine Studie in den USA, die auf Daten aus der National Health Interview Survey (NHIS) 2013–14 basiert, ergab beispielsweise, dass LGB-Erwachsene im Vergleich zu ihren heterosexuellen Mitmenschen häufiger über einen schlechten Gesundheitszustand, funktionelle Einschränkungen, schwere psychische Belastungen und Schwierigkeiten bei der Finanzierung der Gesundheitsversorgung berichteten. Diese Ungleichheiten sind auf den Stress von Minderheiten und die vielfältige gesellschaftliche Marginalisierung zurückzuführen (Liu et al., 2023).

Andererseits unterliegt die Medizin als Disziplin seit der Antike bis heute einem hohen ethischen Standard (Baker & McCullough, 2008). Seit Jahrhunderten besteht die gesellschaftliche Erwartung, dass Ärzte die ethischen Regeln der beruflichen Verantwortung befolgen, die durch die Standards ihres Berufsstandes festgelegt sind und sich in beruflichen Normen wie dem Hippokratischen Eid aus dem Jahr 400 v. Chr. (Miles, 2005) oder den Deklarationen von Genf und Helsinki (Tröhler, 2008) manifestieren. Wie von (Vevaina et al., 1993) betont, sind Ärzte aufgrund der Investitionen, die die Gesellschaft in ihre Ausbildung tätigt (finanziell und durch die Nutzung ihrer Mitglieder als Lernmaterial während der gesamten Ausbildung und Karriere der Ärzte), und aufgrund des praktischen Monopols, das ihrem Berufsstand durch die Zulassung gewährt wird, verpflichtet, sich an den Ethikkodex ihres Berufsstandes zu halten.

Die biomedizinische Ethik (oder Bioethik) ist ein Bereich der praktischen (oder angewandten) Ethik, der sich mit moralischen Fragen befasst, die sich in der medizinischen Praxis und der biomedizinischen Forschung ergeben (Vevaina et al., 1993). Kernstücke der biomedizinischen Ethik sind die vier Prinzipien, die von Beauchamp und Childress definiert wurden (Beauchamp & Childress, 2019):

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

1. Autonomie: Respektierung der Entscheidungsfähigkeit autonomer Personen. Zwei allgemeine Bedingungen sind für die Autonomie unerlässlich: Freiheit, die sich in Unabhängigkeit von kontrollierenden Einflüssen manifestiert, und Entscheidungsfähigkeit, d. h. die Fähigkeit zu absichtlichem Handeln.

2. Nichtschaden: Vermeidung der Verursachung von Schaden.

3. Wohltun: Ergreifen positiver Maßnahmen, um anderen zu helfen, insbesondere durch Verhinderung von Schaden, Beseitigung von Schaden und Förderung des Wohls.

4. Gerechtigkeit: Faire Verteilung von Vorteilen, Risiken und Kosten. Gerechtigkeit wird als faire, gleichberechtigte und angemessene Behandlung von Einzelpersonen und Gruppen verstanden, angesichts der vielen Ungleichheiten im Healthcare-Bereich und in der Forschung aufgrund von Rasse, ethnischer Zugehörigkeit, Geschlecht und sozialem Status.

KI-Ethik und Bias

Die Einführung von KI und die rasante Entwicklung von KI-Anwendungen haben eine Vielzahl ethischer Fragen aufgeworfen (Christoforaki & Beyan, 2022), darunter insbesondere Bias und Diskriminierung.

Daher wurde die KI-Ethik als Bereich der praktischen (oder angewandten) Ethik entwickelt, der „eine Reihe von Werten, Prinzipien und Techniken umfasst, die allgemein akzeptierte Standards von Recht und Unrecht anwenden, um das moralische Verhalten bei der Entwicklung und Nutzung von KI-Technologien zu leiten“ (Leslie, 2019, S. 3).

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Kofinanziert von der Europäischen Union

AEQUITAS

Die KI-Ethik stützt sich sowohl auf die Bioethik (die vier oben dargestellten Prinzipien) als auch auf den Menschenrechtsdiskurs, der unter anderem das Recht auf Freiheit, Gleichheit und Würde vor dem Gesetz, den Schutz der bürgerlichen, politischen und sozialen Rechte, die universelle Anerkennung der Persönlichkeit und das Recht auf freie und ungehinderte Teilnahme am Leben der Gemeinschaft umfasst (Leslie, 2019).

Die vier mit „Erklärbarkeit" ergänzten bioethischen Prinzipien werden für KI wie folgt dargestellt (Floridi et al., 2018):

1. Autonomie als die Fähigkeit des Menschen, selbst zu entscheiden, ob er eine Entscheidung trifft, wobei das Risiko besteht, zu viel an Maschinen zu delegieren.

2. Nichtschaden, als Verhinderung von Schäden, die entweder durch menschliche Absicht oder durch das nicht vorhersehbare Verhalten von Maschinen entstehen.

3. Wohltun, als Förderung des Wohlergehens, Wahrung der Würde und Nachhaltigkeit des Planeten.

4. Gerechtigkeit, als Verhinderung und Beseitigung bereits bestehender ungerechter Diskriminierungen sowie neuer Schäden und Gewährleistung einer gleichmäßigen Verteilung der Vorteile der KI.

5. Erklärbarkeit, definiert als Verständlichkeit und Rechenschaftspflicht für die Entscheidungsprozesse der KI.

Daher hat sich auch die KI-Ethik auf eine Reihe von Prinzipien geeinigt, die auf den vier klassischen Prinzipien der medizinischen Ethik sowie anderen Ansätzen basieren und in (Christoforaki & Beyan, 2022) zusammengefasst sind. Wie jedoch in (Mittelstadt, 2019) festgestellt wird, mangelt es der KI-Entwicklung im Vergleich zur Medizin an: (1) gemeinsamen Zielen und Treuhänderpflicht, (2) Berufserfahrung und Normen, (3)

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

bewährten Methoden zur Umsetzung von Prinzipien in die Praxis und (4) robusten rechtlichen und beruflichen Mechanismen der Rechenschaftspflicht; dies untergräbt den Erfolg des prinzipienbasierten Ansatzes. Natürlich gibt es auch ein komplexes regulatorisches Umfeld für die Entwicklung und Nutzung von KI in der EU, einschließlich der Antidiskriminierungsgesetze, ein Thema, das jedoch außerhalb des Rahmens dieses Berichts liegt.

In Bezug auf die Menschenrechte gehören laut einem vom Europarat finanzierten Bericht aus dem Jahr 2018 (Committee of experts on internet intermediaries (MSI-NET), 2018) insbesondere folgende Menschenrechte zu denjenigen, die von Algorithmen und automatisierten Datenverarbeitungstechniken betroffen sind:

- Anspruch auf faires Gerichtsverfahren
- Privatsphäre und Datenschutz
- Meinungsfreiheit
- Wirksame Rechtsmittel
- Versammlungs- und Vereinigungsfreiheit
- Diskriminierungsverbot
- Soziale Rechte und Zugang zu öffentlichen Dienstleistungen
- Allgemeines und gleiches Wahlrecht

Das Bias wird ausdrücklich als möglicher Diskriminierungsfaktor gegenüber gesellschaftlichen Gruppen aufgrund von Alter, sexueller Orientierung, ethnischer Zugehörigkeit, Geschlecht oder sozioökonomischem Status genannt (Committee of experts on internet intermediaries (MSI-NET), 2018, S. 27). Darüber hinaus wird in der Rahmenkonvention des Europarates über künstliche Intelligenz und Menschenrechte,

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Kofinanziert von der Europäischen Union

AEQUITAS

Demokratie und Rechtsstaatlichkeit ausdrücklich erwähnt, dass die Mitgliedstaaten „shall adopt or maintain measures with a view to ensuring that activities within the lifecycle of artificial intelligence systems respect equality, including gender equality, and the prohibition of discrimination, as provided under applicable international and domestic law“ (*Maßnahmen ergreifen oder aufrechterhalten, um sicherzustellen, dass Aktivitäten im Lebenszyklus von Systemen der künstlichen Intelligenz die Gleichstellung, einschließlich der Gleichstellung der Geschlechter, und das Diskriminierungsverbot gemäß den geltenden internationalen und nationalen Rechtsvorschriften achten*), (Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, 2024, p. 4).

Zivilgesellschaftliche Organisationen (CSOs) als Akteure im Healthcare-Bereich (Vayena et al., 2018) können eine wichtige Rolle bei der Identifizierung und Bekämpfung von KI-Bias sowie der KI-Governance im Allgemeinen spielen, indem sie sich für eine ethische KI-Entwicklung einsetzen, Interessengruppen zur Verantwortung ziehen, die Öffentlichkeit aufklären, marginalisierte Gemeinschaften vertreten, politische und regulatorische Rahmenbedingungen gestalten und die Zusammenarbeit zwischen Regierungen, Technologieunternehmen und der Öffentlichkeit fördern (Korir, 2024).

Innerhalb dieses theoretischen Rahmens werden verschiedene technische Lösungen explizit entwickelt, um Bias zu beseitigen. In der folgenden Sektion stellen wir eine Klassifizierung von KI-induziertem Bias vor, die als Grundlage für unser Mapping-Template diente, wobei der Fokus auf deren Auswirkungen auf genderspezifische und rassebezogene Diskriminierung liegt. Menschliches mentales Bias (Hofmann, 2023), zum Beispiel kognitives Bias wie Bestätigungs- oder Verfügbarkeits-Bias, werden zwar in der Medizin als sehr wirkungsvoll angesehen, fallen jedoch nicht in den Rahmen des aktuellen Projekts.

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Bias in KI-Systemen

Bias in Computersystemen wird in (Friedman & Nissenbaum, 1996, S. 332) als ein Begriff definiert, der sich „ [referring] to computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favour of others. A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate", (*auf Computersysteme bezieht, die bestimmte Personen oder Personengruppen systematisch und ungerechtfertigt gegenüber anderen benachteiligen. Ein System diskriminiert ungerechtfertigt, wenn es einer Person oder Personengruppe eine Chance oder einen Vorteil verweigert oder ihr aus unangemessenen oder ungerechtfertigten Gründen ein unerwünschtes Ergebnis zuweist).*

Gemäß (Friedman & Nissenbaum, 1996) lässt sich Bias in Computersystemen in drei Kategorien einteilen: bereits vorbestehendes Bias, technisches Bias und entstehendes Bias. Im folgenden Unterabschnitt untersuchen wir jede Art von Bias und erläutern sie anhand von Fallstudien, wie sie in der wissenschaftlichen Literatur zu finden sind.

Vorbestehendes Bias

Vorbestehende Bias resultiert aus Vorurteilen in sozialen Institutionen, Praktiken und Haltungen, die bereits existieren, unabhängig sind und in der Regel bereits vor der Schaffung des Systems vorhanden waren. Diese Art von Bias wird bewusst oder unbewusst in das System integriert, manchmal sogar dann, wenn die Entwickler des Systems versuchen, dies zu vermeiden.

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Fallstudie: Diagnose von Herz-Kreislauf-Erkrankungen bei Frauen

Herz-Kreislauf-Erkrankungen (HKE) werden häufig als „Männerkrankheit“ wahrgenommen, was zu einer Unterdiagnose und Unterbehandlung bei Frauen beigetragen hat.

Wie in (Al Hamid et al., 2024) gezeigt wurde, wurden kardiovaskuläre Erkrankungen bei Frauen seltener gemeldet, die entweder mildere Symptome als Männer aufwiesen oder deren Symptome fälschlicherweise als gastrointestinale oder angstbedingte Symptome diagnostiziert wurden; daher wurden Frauen weniger diagnostische Tests und Medikamente angeboten und sie wurden seltener an KardiologInnen überwiesen und/oder ins Krankenhaus eingewiesen. Darüber hinaus war es bei Frauen im Falle eines Krankenhausaufenthalts weniger wahrscheinlich, dass sie eine Koronarintervention erhielten. Daher wurden die Risikofaktoren bei Frauen vom medizinischen Personal, insbesondere von Männern, weniger berücksichtigt. Angesichts der Tatsache, dass Frauen im Bereich der Kardiologie nach wie vor unterrepräsentiert sind (Fatunde et al., 2025), kann man zu dem Schluss kommen, dass Frauen aufgrund bestehendes Bias weniger wahrscheinlich eine angemessene Gesundheitsversorgung erhalten.

KI-Systeme werden anhand von Daten trainiert, die aus bestehenden Praktiken stammen, sodass ein KI-basiertes Diagnosesystem für HKE dieses Bias übernimmt und zu einer Diskriminierung von Frauen führt, unabhängig von den Entscheidungen, die bei der technischen Implementierung getroffen werden.

Technisches Bias

Technisches Bias entsteht durch technische Einschränkungen oder Überlegungen, insbesondere wenn Systementwickler versuchen, vom Menschen geschaffene Konstrukte für Computer nutzbar zu machen, beispielsweise durch die Quantifizierung von Qualitativem, die Diskretisierung von Kontinuierlichem oder die Formalisierung

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

von Nicht-Formellem. Darüber hinaus kann die Dekontextualisierung von Algorithmen aus den Bereichen, in denen sie eingesetzt werden, dazu führen, dass sie nicht alle Gruppen unter allen relevanten Bedingungen fair behandeln.

Fallstudie: Vorhersagegenauigkeit von Modellen zur Vorhersage des Schlaganfallrisikos bei schwarzen und weißen Bevölkerungsgruppen

(Hong et al., 2023) führten eine retrospektive Studie zur Vorhersagegenauigkeit des Schlaganfallrisikos durch, in der sie bestehende schlaganfallspezifische Risikovorhersagemodelle und neuartige ML-Techniken verglichen, wobei unter anderem berücksichtigt wurde, ob die PatientInnen schwarz oder weiß waren. Alle Algorithmen zeigten bei schwarzen Personen eine schlechtere Leistung als bei weißen Personen. Diese Situation ist laut den Autoren möglicherweise auf Risikofaktoren zurückzuführen, die in den Daten nicht erfasst sind, wie z. B. die Art der Krankenversicherung, Sprachbarrieren und andere Faktoren, die sich aus dem unterschiedlichen Zugang zum gesundheitlichen Versorgungssystem ergeben, d. h., die Daten sind aus dem sozioökonomischen Kontext, in dem sie entstanden sind, herausgelöst. Gleichzeitig sind alle oben genannten Risikofaktoren Konstrukte, die sich nur schwer in einer für Computer geeigneten Form darstellen lassen. Zu all dem könnte man noch hinzufügen, dass modernste KI-Algorithmen von Natur aus intransparent sind, was die Merkmale betrifft, die sie auswählen, um eine hohe Genauigkeit zu erreichen (Knight, 2017), sodass selbst ihre Entwickler nicht erklären können, wie sie funktionieren, und somit auch nicht kontrollieren können, ob die oben genannten sozioökonomischen Faktoren tatsächlich in den internen Abläufen des KI-Systems berücksichtigt werden.

Entstehendes Bias

Entstehendes Bias tritt im Kontext der Nutzung durch tatsächliche Nutzer auf, typischerweise nach Abschluss der Entwicklung, als Folge von Veränderungen im

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

gesellschaftlichen Wissen, die nicht in die Systementwicklung einfließen können oder wurden, oder aufgrund einer Bevölkerungsgruppe mit anderen Kenntnissen oder kulturellen Werten als denen, von denen bei der Entwicklung ausgegangen wurde.

Fallstudie: Dataset Shifts

Dataset Shifts sind Diskrepanzen zwischen den Verteilungen der Trainings- und Testdatensätze während der Algorithmusentwicklung und kann zu unterschiedlichen Leistungen auf Untergruppenebene führen (Chen et al., 2023).

Bei der Hautkrebserkennung beispielsweise stammen viele Bilddatensätze, die zum Trainieren von KI-Algorithmen zur Erkennung von Hautkrebs verwendet werden, aus Ländern mit hellhäutiger Bevölkerung (Guo et al., 2021), wodurch bestimmte Bevölkerungsgruppen unterrepräsentiert sind. Mit diesen Datensätzen trainierte KI-Algorithmen zeigen bei der Anwendung in Ländern mit einer vielfältigeren Bevölkerung eine unterdurchschnittliche Leistung und diskriminieren dunkelhäutige Personen. Datensätze sind schwierig und kostspielig zu sammeln, zu annotieren und zu validieren, sodass KI-Systeme, die in Ländern mit niedrigem und mittlerem Einkommen entwickelt werden, auf öffentlich zugängliche Datensätze angewiesen sind, die möglicherweise nicht die Bevölkerungsverteilung widerspiegeln, was zu einer Diskrepanz zwischen der Quell- und der Zielpopulation führt. Dasselbe kann auch in Ländern mit hohem Einkommen auftreten, beispielsweise aufgrund von Bevölkerungsverschiebungen durch zunehmende Einwanderung oder aufgrund von Abweichungen bei der selbst angegebenen ethnischen Zugehörigkeit. Wie in (Chen et al., 2023) festgestellt wird, „since it is now accepted that race is a social construct and that there is greater genetic variability within a particular race than there is between races", (*Da es mittlerweile anerkannt ist, dass Rasse ein soziales Konstrukt ist und dass die genetische Variabilität innerhalb einer bestimmten Rasse größer ist als zwischen den Rassen)*, […] "the medical community has begun to realise that the taxonomies of the past do not adequately represent the groups of people that they purport to", (*Die*

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Kofinanziert von der Europäischen Union

AEQUITAS

medizinische Fachwelt hat erkannt, dass die bisherigen Klassifizierungen die Personengruppen, die sie darstellen sollen, nicht angemessen repräsentieren), und "they can obscure culture, history, socioeconomic status and other confounders of fairness", (*sie können Kultur, Geschichte, sozioökonomischen Status und andere Faktoren, die die Fairness beeinträchtigen können, verschleiern)*.

ML/KI-Spezifische Bias-Arten

Während die oben genannten Bias-Arten für alle Computersysteme gelten, haben KI-Anwendungen spezifischere Anforderungen, sodass wir eine detailliertere Taxonomie benötigen. Daher haben wir uns entschieden, der in (Suresh & Guttag, 2020) vorgestellten Klassifizierung von Bias zu folgen, da sie die Bias-Arten in jedem Schritt der ML/KI-Pipeline identifiziert, wie in Abbildung 1 dargestellt.

Eine typische ML/KI-Pipeline lässt sich wie folgt beschreiben:

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Kofinanziert von der Europäischen Union

AEQUITAS

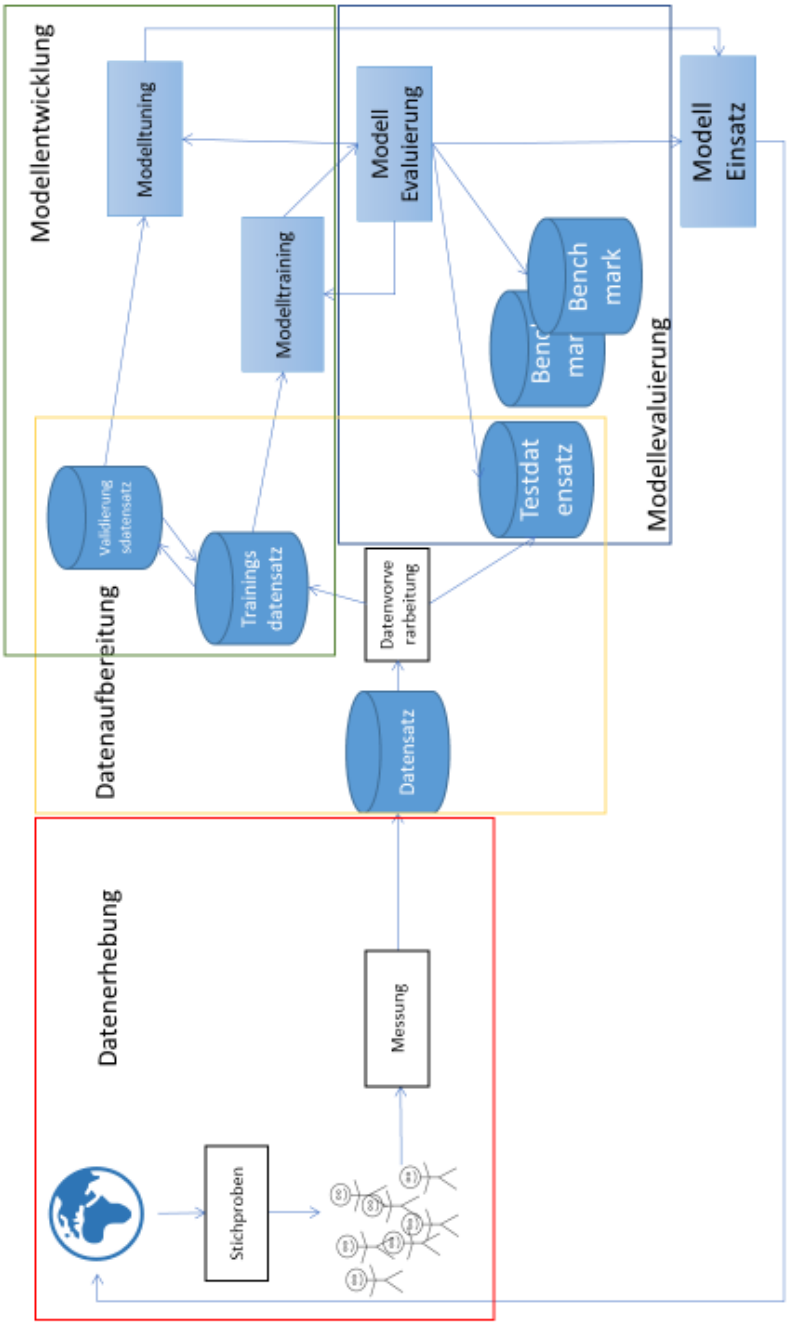

Abbildung 1 ML/KI-Pipeline. Bild adaptiert aus (Suresh & Guttag, 2020)

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Kofinanziert von der Europäischen Union

AEQUITAS

- **Datenerhebung**. Die Erstellung eines ML/KI-Systems beginnt mit der Datenerhebung. Dazu müssen zunächst Daten gesammelt und aufbereitet werden, um einen Datensatz für das KI-System zusammenzustellen. Vorhandene Daten müssen durch die Identifizierung einer Zielpopulation gesammelt werden. Im nächsten Schritt werden die für die zu implementierende Anwendung relevanten Merkmale definiert und gemessen und/oder die Daten mit entsprechenden Labels versehen. Dies ist ein kostenintensiver und langwieriger Prozess, weshalb KI-Spezialisten in den meisten Fällen auf vorhandene Datensätze (entweder öffentliche oder erworbene) zurückgreifen.

- **Datenaufbereitung**. In dieser Phase wird der Datensatz in drei Untersätze aufgeteilt: (i) den Trainingsdatensatz – den tatsächlichen Datensatz, der zum Trainieren des Modells verwendet wird; (ii) den Validierungsdatensatz, eine Auswahl von Daten, die zur Bewertung der Modellanpassung an den Trainingsdatensatz verwendet wird, während die Modell-Hyperparameter (Modellparameter, die nicht aus den Daten gelernt werden können, z. B. die Anzahl der Schichten und Neuronen in einem neuronalen Netzwerkmodell) angepasst werden. In dieser Phase müssen die Daten möglicherweise vorverarbeitet werden (z. B. aufbereitet, normalisiert); und (iii) den Testdatensatz – der Teil der Daten, der zur Bewertung des endgültigen Modells verwendet wird und einen Goldstandard liefert, sobald ein Modell vollständig trainiert ist.

- **Modellentwicklung**. In dieser Phase wird das Modell anhand der Trainingsdaten trainiert und durch Anpassung der Hyperparameter am Validierungsdatensatz feinabgestimmt.

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

- **Modellevaluierung**. Das trainierte Modell wird anhand des Testdatensatzes und manchmal auch anhand von Benchmark-Datensätzen bewertet. Dabei handelt es sich um unabhängig zusammengestellte Datensätze, die dazu dienen, die Robustheit des Modells zu demonstrieren und/oder einen Vergleich mit anderen Methoden zu ermöglichen.

- **Modelleinsatz**. Anwendung des Modells in einer realen Umgebung. Dies kann je nach den Ergebnissen zu Änderungen führen und auch eine Feedbackschleife zum Anfang der Pipeline erzeugen.

Unter Berücksichtigung der oben beschriebenen Phasen der ML/KI-Pipeline übernehmen wir die Klassifizierung von Bias-Arten von (Suresh & Guttag, 2020). Konkret identifizieren sie die folgenden Kategorien: Historisches Bias, Repräsentationsbias, Messungsbias, Aggregationsbias, Bias beim Lernen, Evaluationsbias und Bias im Einsatz. In den folgenden Unterabschnitten definieren wir die oben aufgeführten Verzerrungen und bieten Fallstudien aus den für das Projekt gesammelten Quellen an.

Historisches Bias (Historical Bias)

Historisches Bias entspricht dem von (Friedman und Nissenbaum, 1996) definierten vorbestehendes Bias, der bereits vorhandene Vorurteile und Stereotypen in den Daten berücksichtigt. Ein Beispiel hierfür findet sich bei (Calderone, 1990), der untersucht, ob die Häufigkeit der Verabreichung von Analgetika und Sedativa an PatientInnen nach einer koronaren Bypass-Operation je nach Geschlecht und Alter der PatientInnen unterschiedlich ist. Das Ergebnis zeigte, dass PatientInnen unter 61 Jahren signifikant häufiger Analgetika erhielten als PatientInnen ab 62 Jahren, denen stattdessen signifikant häufiger Sedativa verabreicht wurden. Die Fallstudie zur Vorhersagegenauigkeit von Modellen zur Vorhersage des Schlaganfallrisikos bei schwarzen und weißen Bevölkerungsgruppen wird im Unterabschnitt „Vorbestehendes

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Kofinanziert von der Europäischen Union

Bias” vorgestellt. Wir werden jedoch eine weitere Fallstudie präsentieren, die Historisches Bias hinsichtlich des Einsatzes von KI in der psychischen Gesundheit aufzeigt.

Fallstudie: Künstliche Intelligenz in der psychischen Gesundheit und die Bias von sprachbasierten Modellen

(Straw & Callison-Burch, 2020) präsentieren eine systematische Literaturrecherche zur Verwendung von NLP in der psychischen Gesundheit mit dem Ziel, herauszufinden, wie diese Bias gesundheitliche Ungleichheiten vergrößern können. KI-Modelle, die NLP zur Erstellung von Profilen der psychischen Gesundheit verwenden, sammeln große Datensätze mit Ausdruckssprache, die in der Regel aus sozialen Medien, Online-Foren, Blogs und Chatrooms stammen.

Diese Daten sind jedoch bereits durch den persönlichen Hintergrund und den sozialen Kontext einer Person beeinflusst. Insbesondere in Bezug auf Geschlecht und Sprache gibt es eine umfangreiche Bibliografie (zur englischen Sprache), die in (Pennebaker et al., 2003) zusammengefasst ist und Unterschiede in der Wortwahl von Frauen und Männern aufzeigt.

Beispielsweise verwenden Frauen eine weniger dominante Sprache, was sich in mehr Höflichkeit, weniger Fluchen, mehr Verstärkern (z. B. „wirklich“, „so“) und mehr Heckenausdrücken (d. h. Einschränkungen oder unbestimmte Wörter wie „irgendwie“, „vielleicht“ oder „möglicherweise“) äußert. Männer hingegen wurden als direktiv, präzise und auch weniger emotional in ihrer Sprache beschrieben, die sich durch Verweise auf Quantität, wertende Adjektive (z. B. „gut“, „dumm“), elliptische Sätze („Tolles Bild.“) und „Ich“-Bezüge auszeichnet. Wie die Autoren anmerken, stimmen diese Unterschiede mit einem soziologischen Rahmen für Geschlechtsunterschiede überein, können aber auch auf alternative Erklärungen zurückgeführt werden, wie z. B. das größere soziale Engagement von Frauen.

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

In Bezug auf die psychische Gesundheit schreiben Männer und Frauen Selbstmordbriefe, in denen sie ihre Selbstmordgedanken unterschiedlich zum Ausdruck bringen: Frauen verinnerlichen negative Emotionen, während Männer zunehmende Wut zum Ausdruck bringen (Straw & Callison-Burch, 2020). Ein KI-System, das psychische Probleme für ein Geschlecht untersucht, kann für das andere Geschlecht ungeeignet sein (und dies unter Berücksichtigung des Geschlechts in einem binären Kontext, der einen großen Teil der Bevölkerung ausschließt).

Repräsentationsbias (Representation bias)

Ein Repräsentationsbias tritt auf, wenn die Entwicklungsstichprobe während der Datenerfassungsphase einen Teil der Population unterrepräsentiert. Dies kann auf folgende Weise geschehen: wenn die Zielpopulation so definiert ist, dass sie nicht die Nutzerpopulation widerspiegelt; wenn die Zielpopulation unterrepräsentierte Gruppen enthält; wenn die Stichprobe aus der Zielpopulation entnommen wird und die Stichprobenmethode begrenzt oder ungleichmäßig ist. Ein Repräsentationsbias führt zu einer schlechten Generalisierung für eine Teilgruppe der Nutzerpopulation. Ein typisches Beispiel für ein Repräsentationsbias betrifft die Erkennung von Hautkrebs, da viele Bilddatensätze bestimmte demografische Gruppen unterrepräsentieren, was dazu führt, dass maschinelle Lernmodelle hauptsächlich mit Bildern von hellhäutigen Personen trainiert werden (Guo et al., 2021). Unter Berücksichtigung der Zielkrankheiten des AEQUITAS-Projekts präsentieren wir eine Fallstudie zum Repräsentationsbias in Bezug auf die ethnische Zugehörigkeit bei Typ-2-Diabetes.

Fallstudie: Bewertung von Bias in Algorithmen zur Vorhersage des Risikos für Typ-2-Diabetes

Laut (Cronjé et al., 2023) sind in der Literatur zur Risikoprognose für Diabetes trotz ihres vergleichsweisen geringeren Risikos nicht-hispanische weiße Bevölkerungsgruppen in den USA weiterhin überrepräsentiert.

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

In einer anderen Studie zum Thema ethnisch-rassische Gerechtigkeit in der künstlichen Intelligenz für das Diabetesmanagement war die durchschnittliche Verteilung in den untersuchten Artikeln, die Angaben zur ethnischen Zugehörigkeit enthielten, wie folgt: 69,5 % Weiße, 17,1 % Schwarze und 3,7 % Asiatische.

In einer anderen Übersicht zum Thema „Ethnoracial Equity in Artificial Intelligence for Diabetes Management" (*Ethnische und rassische Gerechtigkeit in der künstlichen Intelligenz für das Diabetes-Management*) betrug in den untersuchten Artikeln, die Angaben zur ethnischen Zugehörigkeit enthielten, die durchschnittliche Verteilung 69,5 % Weiße, 17,1 % Schwarze und 3,7 % Asiaten, während nur zwei Artikel die Einbeziehung von Teilnehmenden mit amerikanischer Ureinwohner-Herkunft angaben (Pham et al., 2021).

Es ist gut dokumentiert, dass die Ungleichheiten bei den Diabetes-Ergebnissen weitgehend auf komplexe, miteinander verbundene soziale Determinanten der Gesundheit zurückzuführen sind, darunter der Zugang zu gesunden Lebensmitteln, eine hochwertige Gesundheitsversorgung, der Versicherungsstatus, Bildungsbarrieren und unterschiedliche Technologieadoptionsraten. Zu diesen Ergebnissen gehören höhere Komplikationsraten und eine schlechtere Blutzuckerkontrolle bei Minderheiten und einkommensschwachen Bevölkerungsgruppen (Alipour & Alipour, 2025).

Infolgedessen würde ein auf bestehenden Datensätzen trainiertes KI-System nur unzureichend generalisieren, was zu voreingenommenen Prognosemodellen führen würde, die beispielsweise bei Präventionsmaßnahmen Personen bestimmter ethnischer Gruppen begünstigen könnten.

Messungsbias

Messungsbias tritt auf, wenn Merkmale und Labels für ein Vorhersageproblem ausgewählt, gesammelt oder berechnet werden, insbesondere bei Verwendung eines Proxys. (einer Approximation eines Konstrukts, das nicht direkt kodiert oder

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

beobachtbar ist). Ein Beispiel dafür findet sich in einer Studie von (Obermeyer et al., 2019), in der Gesundheitskosten als Proxy verwendet wurden, um vorherzusagen und zu bewerten, welche PatientInnen am meisten von einer zusätzlichen Versorgung profitieren würden, was zu einer Diskriminierung aufgrund der ethnischen Zugehörigkeit führte. Gesundheitskosten sind jedoch ein schlechter Proxy für Gesundheitsbedürfnisse, da schwarze PatientInnen, die unverhältnismäßig stark von Armut betroffen sind, oft weniger für den Zugang zu Gesundheitsversorgung ausgeben als weiße PatientInnen. Aufgrund dieses Bias kam der Algorithmus fälschlicherweise zu dem Schluss, dass schwarze PatientInnen gesünder seien als gleich kranke weiße PatientInnen, und stufte sie daher beim Zugang zum medizinischen Versorgungssystem als PatientInnen mit niedrigerer Priorität ein.

Weitere Ursachen für Messungsbias können auftreten, wenn die Messmethode zwischen den Gruppen variiert, beispielsweise wenn zwei Gruppen hinsichtlich desselben Verhaltens überwacht werden, eine davon jedoch strenger oder häufiger als die andere. Ebenso kann die Messgenauigkeit zwischen den Gruppen variieren, was in medizinischen Anwendungen zu systematisch höheren Fehldiagnosen oder Unterdiagnosen in bestimmten Gruppen führen kann. Beispielsweise neigen ärztliche Fachkräfte aufgrund falscher Vorstellungen über biologische Unterschiede zwischen Schwarzen und Weißen dazu, die Schmerzen schwarzer PatientInnen im Vergleich zu nicht-schwarzen PatientInnen zu unterschätzen, was dazu führt, dass schwarze PatientInnen seltener Schmerzmittel erhalten und, wenn doch, dann in geringeren Mengen (Hoffman et al., 2016).

Fallstudie: Rassen- und ethnische Unterschiede im Zusammenhang zwischen mittlerem Glukosewert und Hämoglobin A1c

Der A1C-Test misst den durchschnittlichen Glukosegehalt (Zucker) im Blut und wird zur Erkennung von Prädiabetes oder zur Diagnose von Typ-2-Diabetes eingesetzt. A1C ist jedoch nur ein indirekter Messwert und steht in keinem kausalen Zusammenhang mit

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Gesundheitsergebnissen, da es zahlreiche Möglichkeiten gibt, die Beziehung zwischen direkten Messungen der Glykämie (der Glukosekonzentration im Blut) und A1C direkt zu verändern. Es gibt sogar erhebliche Schwankungen in der Beziehung zwischen Glykämie und A1C zwischen verschiedenen Personen und sogar innerhalb einer Person im Laufe der Zeit. Darüber hinaus haben Studien einen signifikant höheren Hämoglobin-A1c-Wert (A1C) bei afroamerikanischen PatientInnen als bei weißen PatientInnen mit dem gleichen durchschnittlichen Glukosewert festgestellt (Karter et al., 2023).

Wenn ein KI-System zur Diagnose von Diabetes darauf trainiert wird, A1C-Testergebnisse als Ersatz für die Glykämie zu verwenden, ohne andere Faktoren wie die ethnische Zugehörigkeit der PatientInnen zu berücksichtigen, kann dies zu voreiligen Diabetesdiagnosen und unangemessenen Behandlungen führen, was zu einem Bias im Versorgungssystem und gesundheitlichen Ungleichheiten im Gesundheitswesen führt. Wie jedoch in der Übersichtsstudie zu Biasen, die die Fairness von KI-/ML-Modellen bei Diabetes beeinträchtigen könnten (einschließlich Messungsbias), von (Alipour & Alipour, 2025) festgestellt wurde, erwähnen die untersuchten Studien zwar ausdrücklich, dass sich Messungsbias durch KI-Modelle ausbreiten kann, wenn es nicht korrigiert wird, aber keine von ihnen hat solche Bias bei der Modellentwicklung berücksichtigt, es ausdrücklich gemildert oder über die Korrektur von Unterschieden in der Messgenauigkeit berichtet.

Aggregationsbias

Ein Aggregationsbias entsteht, wenn ein einheitliches Modell für einen Datensatz verwendet wird, der verschiedene Gruppen von Personen oder Gegenständen umfasst.

Betrachten wir das Beispiel der Zuordnung von Eingabedaten (z. B. das Einkommen einer Person) zu Labels, die diese beschreiben (z. B. niedrig, mittel, hoch), wobei davon

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

ausgegangen wird, dass diese über alle Teilmengen der Daten hinweg konsistent sind. In der Realität können der Hintergrund oder die Kultur einer Person die tatsächliche Bedeutung dieser Zahlen verändern. Beispielsweise kann ein „hohes" Einkommen in einer kleinen ländlichen Stadt oder einem einkommensschwachen Land (low- or middle-income country-LMIC) etwas ganz anderes bedeuten als in einer Großstadt oder einem Land mit hohem Einkommen.

Fallstudie: Digitale Health-Tools zur Passivüberwachung von Depressionen

Die Verwendung digitaler Tools zur Messung physiologischer und verhaltensbezogener Variablen für die passive Überwachung von Depressionen wird in einer systematischen Übersicht zu diesem Thema von (De Angel et al., 2022) behandelt. Die untersuchten Artikel befassten sich mit Zusammenhängen zwischen Depressionen und objektiven Verhaltensdaten, die von Sensoren in Smartphones und tragbaren Geräten erfasst wurden. Diese Daten wurden in Merkmale umgewandelt, die von KI-Modellen für Prognosen verwendet wurden und sich auf Schlaf, körperliche Aktivität, Tagesrhythmus, Geselligkeit, Standort und Telefonnutzung bezogen.

Die Autoren betonen jedoch die Heterogenität, die sich aus der Vielfalt der zur Erstellung dieser Merkmale verwendeten Methoden ergibt. So kann beispielsweise das Merkmal „Schlafqualität" durch die Messung der Anzahl der Aufwachphasen, der Gesamtzahl der Wachminuten oder des Verhältnisses von Wach- zu Schlafphasen in einer Schlafperiode definiert werden, wobei auch die Unterschiede in der Art und Weise berücksichtigt werden müssen, wie Sensoren in verschiedenen Geräten ein Ereignis als „Schlaf" beschreiben. Da alle oben genannten Unterscheidungen nicht berücksichtigt und gemeinsam als „Schlafqualität" zusammengefasst werden und da ein Datensatz von Personen oder Gruppen mit unterschiedlichen Hintergründen, Kulturen oder Normen stammen kann, kann dieses Merkmal für jede dieser Gruppen oder Personen eine andere Bedeutung haben.

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Die Aggregation solcher Daten zu einem einzigen Merkmal kann zu einem System führen, das für keine Gruppe geeignet ist oder das die dominante Bevölkerungsgruppe privilegiert, wenn auch ein Repräsentationsbias vorliegt. Beispielsweise gibt es Hinweise darauf, dass es genderspezifische Unterschiede im Schlafverhalten von Männern und Frauen gibt, wobei Letztere in der Schlafforschung oft unterrepräsentiert sind. Darüber hinaus werden andere Faktoren, die normalerweise bei Schlafmustern und -störungen nicht berücksichtigt werden, nicht zwischen dem sozialen Konstrukt „Geschlecht“ und dem biologischen Geschlecht unterschieden, und es werden keine intersektionalen Identitäten berücksichtigt, die durch Alter, ethnische Zugehörigkeit und sozioökonomische Klasse definiert sind (Lok et al., 2024).

Bias beim Lernen

Bias beim Lernen entsteht, wenn Modellierungsentscheidungen Leistungsunterschiede zwischen verschiedenen Beispielen in den Daten verstärken. Ein Beispiel hierfür ist die differentielle Privatsphäre (Differential Privacy), ein Mechanismus, der in KI-Systemen verwendet wird und sicherstellt, dass anhand der Outputs eines Systems nicht festgestellt werden kann, ob die Daten einer bestimmten Person im ursprünglichen Datensatz enthalten waren. Die differentielle Privatsphäre wird in Datensätzen des Gesundheitswesens eingesetzt, um sensible Patientendaten zu schützen, beispielsweise im Fall seltener Krankheiten, bei denen jeder Patient mehr oder weniger einzigartig in einem begrenzten Bereich ist, der von einer Klinik abgedeckt wird, sodass es selbst bei anonymisierten Daten nicht sehr schwierig ist, die Identität der Person zu ermitteln. Es hat sich jedoch gezeigt, dass die differentielle Privatsphäre den Einfluss unterrepräsentierter Daten auf das Modell verringert. Wenn das KI-System also von vornherein voreingenommen ist, verstärkt die Anwendung einer Maßnahme zum Schutz der Privatsphäre diesen Bias noch mehr (Bagdasaryan & Shmatikov, 2019).

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Fallstudie: Differentielle Privatsphäre und gesundheitliche Ungleichheiten

Im September 2018 gab das US Census Bureau bekannt, dass es differentielle Privatsphäre für Datenprodukte aus den Daten der Volkszählung 2020 implementieren werde. Allerdings untersuchten (Santos-Lozada et al., 2020), wie die Implementierung von differenzieller Privatsphäre das Wissen über gesundheitliche Ungleichheiten in Bezug auf die Sterblichkeit verändern kann, insbesondere für ethnische Minderheiten in ländlichen Gebieten und weniger urbanen Umgebungen. Ihre Ergebnisse deuten darauf hin, dass sich die differentielle Privatsphäre stärker auf die Schätzungen der Sterblichkeitsrate für nicht-hispanische Schwarze und Hispanics auswirken wird als auf die Schätzungen für nicht-hispanische Weiße. Diese Ergebnisse wurden von Kurz et al. (2022) bestätigt, die zeigen, dass die Anwendung der differentiellen Privatsphäre auf dieselben Daten zu einer falschen Darstellung der Medicaid-Teilnahmequoten unter bereits marginalisierten ethnischen Gruppen führen kann.

Insbesondere unterschieden sich diese Raten für bestimmte Kombinationen von Landkreis, Rasse und ethnischer Zugehörigkeit zwischen den Ergebnissen der differentiellen Datenschutzdaten und den Originaldaten, wobei sie manchmal 10 % überstiegen. Darüber hinaus waren nicht-hispanische Weiße die einzige ethnische und rassische Untergruppe, für die der Algorithmus zur differentiellen Privatsphäre die Medicaid-Teilnahmequoten genau erfasste. Diese Erkenntnis könnte wichtige Auswirkungen auf die Gesundheitspolitik haben, da Volkszählungsdaten zur Planung von Regierungsprogrammen, zur Zuweisung von Ressourcen sowie zur Bewertung und Verfolgung von politischen Maßnahmen verwendet werden.

Evaluationsbias

Evaluationsbias tritt auf, wenn die für eine bestimmte Aufgabe verwendeten Benchmark-Daten nicht repräsentativ für die Nutzerpopulation sind. Benchmarks sind standardisierte Datensätze, die zur Messung der Qualität eines Modells verwendet werden und einen quantitativen Vergleich von Modellen ermöglichen. In der Folge

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

besteht die Gefahr, dass die Entwicklung und der Einsatz von Modellen gefördert werden, die nur für den Teil der Daten, der in der Benchmark vertreten ist, gut funktionieren. Daher kann es zu einer Diskriminierung schutzbedürftiger Untergruppen oder Personen kommen, wenn die Benchmark historischen, repräsentativen oder Messungsbias unterliegt.

Im Healthcare-Bereich kann die Unterrepräsentation bestimmter Bevölkerungsgruppen in Datensätzen entweder darauf zurückzuführen sein, dass Personen oder Gruppen in den Datensätzen fehlen (z. B. schwangere Frauen aufgrund ethischer Beschränkungen) oder dass Personen falsch oder unangemessen in Gruppen eingeteilt werden (z. B. Kategorien wie „gemischte ethnische Zugehörigkeit“ oder „Sonstige“). Die Ursachen hierfür können soziale und technische oder rechtliche/ethische Gründe sein, wie z. B. strukturelle Hindernisse für den Zugang zur Gesundheitsversorgung, technische Hindernisse für die Erfassung oder Digitalisierung relevanter Gesundheitsdaten, individuelle und strukturelle Einschränkungen hinsichtlich der Zustimmung zur Datenweitergabe sowie rechtliche oder ethische Beschränkungen der Datenweitergabe, die den Zugang zu Daten verhindern (Arora et al., 2023). Das Ergebnis ist, dass KI-Systeme, die auf solche Benchmarks kalibriert sind, bei der Anwendung auf Personen aus einer unterrepräsentierten Gruppe unterdurchschnittliche Leistungen erbringen können. Es ist jedoch wichtig zu beachten, dass die Gültigkeit von Benchmarks ein allgemeineres Problem ist und sich nicht auf Bias beschränkt (Brooks, 2025).

Fallstudie: Hautbild-Datensätze

Hautbild-Datensätze repräsentieren bestimmte Bevölkerungsgruppen nur unzureichend, da die meisten Bilder in diesen Datensätzen aus Nordamerika oder Europa stammen und überwiegend hellhäutige Personen zeigen (Guo et al., 2021). Aufgrund der hohen Kosten und der Schwierigkeit, diese Datensätze zu erstellen,

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

können sie neben dem Training von Modellen auch als Benchmarks verwendet werden.

Die Fallstudie, die Entstehendes Bias veranschaulicht, d. h. die Hautkrebs-Bilddatensätze, die zum Trainieren von Vorhersagemodellen verwendet werden, ist ein Beispiel für einen ungeeigneten Benchmark, wenn die Nutzerpopulation aus unterrepräsentierten Gruppen stammt (Guo et al., 2021). Ein ähnlicher Fall, der zwar nicht mit KI zu tun hat, aber die Allgemeingültigkeit des Problems verdeutlicht, betrifft Pulsoximeter (Geräte zur Messung der Blutsauerstoffsättigung, die beispielsweise bei Herzinfarkten oder Herzversagen eingesetzt werden), die nachweislich bei hell pigmentierter Haut genauer funktionieren (Sjoding et al., 2020).

Repräsentations-, Messungs-, Aggregations-, Lern- und Evaluationsbias lassen sich auf die von Friedman und Nissenbaum (1996) definierte Technisches Bias zurückführen.

Bias im Einsatz

Ein Bias im Einsatz entsteht, wenn eine Diskrepanz zwischen dem Problem, das ein Modell lösen soll, und der Art und Weise, wie es tatsächlich eingesetzt wird, besteht. Dies kann insbesondere in Kombination mit kognitiven Verzerrungen wie Bestätigungs- und Automatisierungsverzerrungen zu Problemen führen. Der Bias im Einsatz entspricht dem von (Friedman und Nissenbaum 1996) definierten Entstehendes Bias

Fallstudie: Domänenverschiebung

Der Fall der Datenverschiebung ist im Unterabschnitt Entstehendes Bias über die Erkennung von Hautkrebs dokumentiert. Darüber hinaus können wir den Fall der Domänenverschiebung definieren, der auftritt, wenn ein System implementiert wurde, die behördliche Zulassung erhalten hat und in der klinischen Praxis eingesetzt wird, jedoch auf eine andere Patientengruppe angewendet wird als die, für die es trainiert wurde.

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Kofinanziert von der Europäischen Union

AEQUITAS

Ein Beispiel hierfür wäre ein System, das für eine Klinik in einem Land mit hohem Einkommen entwickelt und in einem Land mit niedrigem oder mittlerem Einkommen (LMIC) eingesetzt wird, ohne dass Faktoren wie die soziodemografischen Merkmale der Patienten oder die Frage berücksichtigt werden, ob die Patienten im Vergleich zu denen, die in den Trainingsdaten enthalten sind, das gleiche Gesamtrisikoniveau aufweisen (Vokinger et al., 2021).

Policybezogene Implikationen

Die in Deliverable D2.1 dargelegten Nachweise zeigen, dass gender- und rassenspezifische Bias in der biomedizinischen KI keine zufälligen oder isolierten technischen Mängel sind, sondern systemische Risiken, die während des gesamten Lebenszyklus von KI-Systemen im Gesundheitswesen auftreten. Bei Herz-Kreislauf-Erkrankungen, Depressionen und Diabetes entstehen Bias durch historisch verzerrte klinische Datensätze, ungleiche Diagnosepraktiken, Proxy-Variablen, die strukturelle Ungleichheiten kodieren, und Einsatzkontexte, die sowohl Nutzen als auch Schaden ungleich verteilen. Diese Ergebnisse bestätigen, dass biomedizinische KI direkt mehrere Rechte und Prinzipien betrifft, die durch die Charta der Grundrechte der Europäischen Union geschützt sind, insbesondere die Prinzipien der Menschenwürde, der Gleichheit vor dem Gesetz und der Nichtdiskriminierung sowie das Recht auf Unversehrtheit der Person, das Recht auf Gesundheitsversorgung, den Datenschutz und das Recht auf einen wirksamen Rechtsbehelf.

Vor diesem Hintergrund müssen die politischen Rahmenbedingungen der EU und der Mitgliedstaaten für KI im Gesundheitswesen die Verringerung von Verzerrungen nicht als freiwillige ethische Ergänzung, sondern als verbindlichen Bestandteil eines rechtmäßigen und rechtskonformen Einsatzes von KI behandeln. Die europäischen und nationalen Regulierungsbemühungen im Bereich der KI im Gesundheitswesen sollten als Teil des umfassenderen Grundrechtsrahmens für KI betrachtet werden (siehe Novossiolova, 2025; Novossiolova et al., 2025; Kasapi, 2025). Der EU AI Act bietet die

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

notwendige regulatorische Grundlage, indem er die meisten biomedizinischen KI-Systeme als Hochrisikosysteme einstuft. Seine Wirksamkeit in der Praxis wird jedoch davon abhängen, wie die Grundrechtsschutzmaßnahmen bei Konformitätsbewertungen, der Überwachung nach dem Inverkehrbringen und der Beschaffung im öffentlichen Sektor umgesetzt werden.

Erstens müssen Garantien für eine sinnvolle menschliche Aufsicht gestärkt und für biomedizinische KI-Systeme während ihres gesamten Lebenszyklus konkretisiert werden. Klinische KI-Tools, die für Diagnosen, Risikostratifizierung, Screening oder Behandlungsunterstützung eingesetzt werden, sollten in keinem Fall als de facto autonome Entscheidungsträger fungieren. Die von Menschen durchgeführte Überprüfung muss nicht nur die Möglichkeit der Übersteuerung durch medizinisches Fachpersonal umfassen, sondern auch eine klare institutionelle Verantwortung für das Verstehen der Systemgrenzen, bekannter Verzerrungsrisiken und Leistungslücken in Untergruppen. Im Einklang mit dem Schutz der Menschenwürde und -integrität durch die Charta sollten medizinische Fachkräfte geschult und institutionell unterstützt werden, um KI-Ergebnisse kritisch zu hinterfragen, anstatt sich ihnen zu unterwerfen. Dies erfordert die Einbettung von KI-Kompetenz, Bewusstsein für Bias und Grundrechtsschulungen in die medizinische Ausbildung und die kontinuierliche berufliche Weiterbildung.

Transparenzpflichten sollten im Gesundheitswesen weit ausgelegt werden. PatientInnen und NutzerInnen von Gesundheitsdienstleistungen müssen informiert werden, wenn KI-Systeme bei klinischen Entscheidungen eingesetzt werden, die sie betreffen, einschließlich bei Screenings, Priorisierungen oder Risikobewertungen. Wenn KI-generierte Ergebnisse in öffentliche Gesundheitsdienste einfließen, sollten diese Ergebnisse eindeutig als solche erkennbar sein und mit verständlichen Erläuterungen zu ihrer Rolle, ihren Einschränkungen und bekannten Risiken des Bias versehen sein. Personen sollten auch informiert werden, wenn ihre personenbezogenen Daten für KI-Training, -Tests oder kontinuierliches Lernen

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

verwendet werden, insbesondere wenn es sich um sensible Gesundheitsdaten handelt. Diese Transparenzmaßnahmen sind unerlässlich, um die in der Charta verankerten Rechte auf Datenschutz und wirksame Rechtsbehelfe zu wahren und es Einzelpersonen zu ermöglichen, Entscheidungen, die sich nachteilig auf sie auswirken können, sinnvoll anzufechten.

Zweitens muss die Auswirkungsanalyse auf die Grundrechte zu einer routinemäßigen, durchsetzbaren Anforderung für biomedizinische KI-Systeme werden, die über die Überprüfungen vor der Markteinführung hinausgeht und eine kontinuierliche Bewertung während des Einsatzes umfasst. Die empirischen Belege in D2.1 zeigen, dass viele Bias-Effekte erst dann sichtbar werden, wenn KI-Systeme mit realen Bevölkerungsgruppen und klinischen Arbeitsabläufen interagieren, insbesondere durch intersektionale Effekte, die Geschlecht, ethnische Zugehörigkeit, Alter und sozioökonomischen Status betreffen. Rechtebasierte Folgenabschätzungen, wie sie beispielsweise von der HUDERIA (Human Rights, Democracy and the Rule of Law Impact Assessment)-Methode (Methode zur Risiko- und Wirkungsanalyse von KI-Systemen unter dem Aspekt der Menschenrechte, Demokratie und Rechtsstaatlichkeit) des Europarates inspiriert sind, sollten daher für risikoreiche medizinische KI-Systeme obligatorisch sein und ausdrücklich die unterschiedlichen Leistungen und Ergebnisse in den verschiedenen geschützten Gruppen untersuchen. Diese Bewertungen müssen eine sinnvolle Beteiligung der Interessengruppen, einschließlich zivilgesellschaftlicher Organisationen, Patientenvertreter und Gleichstellungsstellen, beinhalten, um Schäden aufzudecken, die aus rein technischer oder klinischer Sicht möglicherweise nicht sichtbar sind.

Es sollten regelmäßige Audits biomedizinischer KI-Systeme vorgeschrieben werden, um die fortdauernde Einhaltung der Grundrechtsstandards zu überprüfen, wobei besonderes Augenmerk auf Bias, Verschiebungen in den Datensätzen und Veränderungen in der klinischen Anwendung im Laufe der Zeit zu legen ist. Wenn Audits anhaltende oder nicht zu mildernde diskriminierende Auswirkungen aufdecken,

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

muss es klare rechtliche und institutionelle Wege geben, um die Nutzung des Systems einzuschränken, auszusetzen oder zu beenden. Das Recht auf Gesundheitsversorgung kann nicht die fortgesetzte Nutzung von KI-Tools rechtfertigen, die bestimmte Gruppen systematisch benachteiligen, selbst wenn die aggregierten Leistungskennzahlen günstig erscheinen.

Drittens müssen die EU- und nationalen Behörden das Risiko des Missbrauchs und sekundärer Schäden im Zusammenhang mit biomedizinischer KI angehen. Dazu gehören Cybersicherheitslücken, die die Systemintegrität gefährden oder eine böswillige Manipulation klinischer Ergebnisse ermöglichen könnten, sowie die Umnutzung von Gesundheits-KI für Überwachungs-, Profiling- oder Ausgrenzungspraktiken. Biomedizinische KI-Systeme sollten regelmäßigen Sicherheitsbewertungen und strengen Meldepflichten für Vorfälle unterliegen, mit klaren Mechanismen zur Rechenschaftspflicht für Fälle, in denen voreingenommene oder kompromittierte Systeme zu Rechtsverletzungen führen. Haftungsrahmen sollten sicherstellen, dass die Verantwortung nicht allein auf einzelne Kliniker abgewälzt werden kann, wenn Schäden strukturell in KI-Design- oder Einsatzentscheidungen eingebettet sind.

Viertens muss die Förderung ethischer und verantwortungsbewusster Praktiken in die gesamte Wertschöpfungskette der biomedizinischen KI eingebettet werden. Entwickler sollten verpflichtet werden, Risiken des Bias proaktiv anzugehen, indem sie repräsentative Daten sammeln, Ziele und Proxies sorgfältig auswählen, untergruppenspezifische Validierungen durchführen und die Leistung über Geschlechter- und Rassengruppen hinweg transparent berichten. Wichtig ist, dass die in D2.1 untersuchten Belege zeigen, dass „Fairness durch Unwissenheit“ und rein technische Strategien zur Beseitigung des Bias im Gesundheitswesen oft unzureichend sind. Regulatorische Leitlinien und Standards sollten daher über abstrakte Fairness-Metriken hinausgehen und von Entwicklern verlangen, klinisch bedeutsame Ergebnisse

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

in Bezug auf Gleichbehandlung nachzuweisen, die im Zusammenhang mit realen Gesundheitspfaden und Zugangsstrukturen bewertet werden.

Die öffentliche Beschaffungs- und Förderpolitik spielt eine entscheidende Rolle bei der Gestaltung von Anreizen für Entwickler. Gesundheitsbehörden und öffentliche Kliniken sollten Grundrechte und Kriterien zur Vermeidung von Bias in ihre Beschaffungsentscheidungen für KI-Systeme einbeziehen und Lösungen bevorzugen, die robuste, transparente und unabhängig überprüfte Verfahren zur Minderung von Bias aufweisen. EU-Finanzierungsinstrumente, einschließlich künftiger Forschungs- und Innovationsprogramme, sollten weiterhin Projekten Vorrang einräumen, die technische Innovation mit rechtsbasierter Governance, Einbeziehung von Interessengruppen (Stakeholders) und Kapazitätsaufbau im Einklang mit dem AEQUITAS-Modell verbinden.

Schließlich erfordert die Stärkung der gesellschaftlichen Resilienz gegenüber voreingenommener biomedizinischer KI nachhaltige Investitionen in die Sensibilisierung der Öffentlichkeit, das Engagement der Zivilgesellschaft und die sektorübergreifende Zusammenarbeit. Einzelpersonen müssen gestärkt werden, damit sie ihre Rechte in der KI-gestützten Gesundheitsversorgung und die ihnen zum Schutz zur Verfügung stehenden Mechanismen verstehen. Zivilgesellschaftliche Organisationen, Gleichstellungsstellen und Patientengruppen sollten als wesentliche Akteure bei der Überwachung der Auswirkungen von KI, der Unterstützung betroffener Personen und der Information über die Politikentwicklung anerkannt werden. Die Zusammenarbeit zwischen Regierungen, Anbietern, ForscherInnen, Industrie und Zivilgesellschaft ist notwendig, um sicherzustellen, dass die Vorteile der biomedizinischen KI gerecht verteilt werden und bestehende Ungleichheiten im Gesundheitswesen nicht verstärkt werden.

Insgesamt stützen die Ergebnisse von Deliverable D2.1 eine klare policybezogene Schlussfolgerung: Biomedizinische KI kann in der EU nur dann als vertrauenswürdig

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

und legitim angesehen werden, wenn ihre Gestaltung, ihr Einsatz und ihre Governance fest im Schutz der Grundrechte verankert sind.

Der EU AI Act, der im Lichte der Charta der Grundrechte der Europäischen Union ausgelegt und durch konkrete Aufsichts-, Folgenabschätzungs- und Rechenschaftsmechanismen umgesetzt wird, bietet eine entscheidende Gelegenheit, um sicherzustellen, dass Innovationen im Gesundheitswesen die Gerechtigkeit fördern, anstatt historische Diskriminierungsmuster zu reproduzieren.

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Kofinanziert von der Europäischen Union

AEQUITAS

Von der Europäischen Union finanziert. Die geäußerten Ansichten und Meinungen entsprechen jedoch ausschließlich denen des Autors bzw. der Autoren und spiegeln nicht zwingend die der Europäischen Union oder der Europäischen Exekutivagentur für Bildung und Kultur (EACEA) wider. Weder die Europäische Union noch die EACEA können dafür verantwortlich gemacht werden. Projektcode: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI