

Regulierung künstlicher Intelligenz im öffentlichen Gesundheitswesen zur Vermeidung genderspezifischer und rassenspezifischer Bias

**Ein auf die Charta der Grundrechte der
Europäischen Union abgestimmtes
Regulierungsmodell für biomedizinische KI**

Policy-Dokument



**Kofinanziert von der
Europäischen Union**

Einleitung

Künstliche Intelligenz (KI) wird europaweit zunehmend in Gesundheitssysteme integriert und unterstützt die Diagnose, die klinische Entscheidungsfindung, die Behandlungsplanung sowie die Vorhersage von Patientenrisiken. Diese Technologien bieten erhebliche Möglichkeiten zur Verbesserung der Effizienz und der Ergebnisse im Gesundheitswesen. Gleichzeitig gibt es immer mehr Belege dafür, dass **KI-Systeme bestehende Ungleichheiten im Gesundheitswesen reproduzieren oder verstärken können**.¹

Mehrere Mechanismen können bei der KI-gestützten Entscheidungsfindung im Gesundheitswesen zu verzerrten Ergebnissen führen. Dazu gehören Ungleichgewichte in den Trainingsdaten, Messungsbias bei klinischen Variablen, ungeeignete Proxyindikatoren sowie Unterschiede zwischen Entwicklungsumgebungen und realen Einsatzbedingungen.² Bias kann auch entstehen, wenn KI-Systeme in klinische Arbeitsabläufe eingeführt werden, ohne dass das medizinische Fachpersonal angemessen geschult wurde oder ohne substanzielle menschliche Kontroll- und Rechenschaftsmechanismen zur Überwachung ihrer Leistung bei unterschiedlichen Patientengruppen.³

Darüber hinaus hat empirische Forschung gezeigt, dass solche **Bias im klinischen Kontext schwerwiegende Folgen haben können**. Studien berichten von anhaltenden Ungleichheiten bei der Diagnose und Behandlung schwerwiegender Erkrankungen. Frauen und Patienten aus ethnischen Minderheiten mit Herz-Kreislauf-Erkrankungen erhalten im Vergleich zu Männern und weißen Patienten oft eine verzögerte oder weniger intensive Diagnose und Behandlung,⁴ und ethnische Unterschiede wirken sich auf die Diabetesversorgung und die Behandlungsergebnisse aus,⁵ während psychische Erkrankungen wie Depressionen sowohl bei Frauen als auch in Minderheitengruppen

¹ Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H. und Wang, Y., *Künstliche Intelligenz im Gesundheitswesen: Vergangenheit, Gegenwart und Zukunft*, *Stroke and Vascular Neurology* 2(4): 230–243, 2017.

² Vokinger, K. N., Feuerriegel, S., & Kesselheim, A. S., *Verringerung von Biasen im maschinellen Lernen für die Medizin*, *Communications Medicine* 1, 25, 2021.

³ Panesar, S., Kliot, M., Parrish, R., Fernandez-Miranda, J., Cagle, Y., & Britz, G.W., *Chancen und Risiken künstlicher Intelligenz in der Neurochirurgie*, *Neurosurgery* 87(1): 33–44, 2020.

⁴ Balla, S., Gomez, S.E. und Rodriguez, F., *Ungleichheiten in der kardiovaskulären Versorgung und bei den Behandlungsergebnissen bei Frauen aus ethnischen Minderheiten*, *Current Treatment Options in Cardiovascular Medicine* 22(12):75, 2020.

⁵ Young, C. & Myers, A. K., *Rassen- und ethnische Ungleichheiten in der klinischen Versorgung und Behandlung von Diabetes: Eine narrative Übersicht*, *Endocrine Practice* 29(4): 295–300, 2023.

zu selten diagnostiziert und behandelt werden.⁶ Ein gut dokumentiertes Beispiel ist eine explorative Studie im „*Journal of Medical Internet Research*“, die genderspezifische Leistungsunterschiede bei Algorithmen des maschinellen Lernens zur Vorhersage von Herzerkrankungen feststellte. Insbesondere zeigte sich, dass weibliche Patientinnen in den Trainingsdatensätzen durchweg unterrepräsentiert waren und dass die Modelle bei Frauen höhere Falsch-Negativ-Raten aufwiesen, was verdeutlicht, wie Datenbias zu **einer ungleichen Diagnoseleistung** beitragen kann.⁷ Bezieht man diese Problematik auf die rassische Gleichberechtigung aus, so hat eine weitere Studie ergeben, dass KI-Modelle, die darauf trainiert wurden, den Schweregrad einer Depression anhand der Sprache in sozialen Medien vorherzusagen, bei weißen Teilnehmenden gute Ergebnisse erzielten, bei schwarzen Teilnehmenden jedoch schlechte – was rassistisch bedingte Leistungsunterschiede offenbart, die auf Unterschiede in den Ausdrucksmustern zurückzuführen sind, die von den Modellen nicht erfasst werden.⁸

Um diesen Risiken der Voreingenommenheit zu begegnen, sind **Regulierungsmodelle** erforderlich, **die internationale und europäische ethische und rechtliche Prinzipien in konkrete Steuerungsmechanismen umsetzen**, die Gesundheitseinrichtungen in der Praxis anwenden können. Innerhalb der Europäischen Union (EU) hat sich das regulatorische Umfeld mit der Verabschiedung des „AI Act“ rasch weiterentwickelt, das Verpflichtungen für risikoreiche KI-Systeme und deren Betreiber, wie beispielsweise Kliniken, festlegt.⁹ Die Umsetzung dieser rechtlichen Anforderungen als praktische Verfahren in Gesundheitseinrichtungen stellt jedoch nach wie vor eine erhebliche Herausforderung dar.

Dieses Policy-Dokument schlägt **ein Regulierungsmodell für die Governance biomedizinischer KI** vor, **das mit dem EU-Acquis**, einschließlich der Charta der Grundrechte der Europäischen Union, **im Einklang steht** und für öffentliche Gesundheitssysteme sowie zivilgesellschaftliche Organisationen (CSOs) konzipiert ist. Das vorgeschlagene Modell stützt sich auf europäische Rechtsinstrumente, darunter den „AI Act“, die Medizinprodukteverordnung (MDR), die In-vitro-Diagnostika-Verordnung (IVDR), die Datenschutz-Grundverordnung (DSGVO)

⁶ Sorkin, D.H., Ngo-Metzger, Q., Billimek, J., August, K.J., Greenfield, S., & Kaplan, S.H., [Unterdiagnostizierte und unterbehandelte Depressionen bei Patienten mit Typ-2-Diabetes unterschiedlicher rassistischer und ethnischer Herkunft](#), *Diabetes Care* 34(3): 598–600, 2011.

⁷ Straw, I., Rees, G. und Nachev, P., [Genderspezifische Leistungsunterschiede bei Algorithmen des maschinellen Lernens zur Vorhersage von Herzerkrankungen: Eine explorative Studie](#), *Journal of Medical Internet Research* 26: e46936, 2024.

⁸ Rai, S., Stadel, E. C., Giorgi, S., Francisco, A., Ungar, L. H., Curtis, B., & Guntuku, S. C., [Wichtige sprachliche Indikatoren für Depressionen in sozialen Medien hängen von der ethnischen Zugehörigkeit ab](#), *Proceedings of the National Academy of Sciences* 121(14): e2319837121, 2024.

⁹ Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates [zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz \(„AI Act“\)](#), PE/24/2024/REV/1, Amtsblatt der Europäischen Union, 13. Juni 2024.

und den im Aufbau befindlichen Europäischen Gesundheitsdatenraum (EHDS). Es setzt Grundrechtsprinzipien, insbesondere Würde des Menschen, Gleichheit vor dem Gesetz, Nichtdiskriminierung und Gesundheitsschutz, in praktische Schutzmaßnahmen um, die für den gesamten Lebenszyklus von KI-Systemen gelten. Verweise auf ausgewählte Beispiele aus Nicht-EU-Ländern werden nur dort aufgenommen, wo sie übertragbare Erkenntnisse liefern, und sollen den europäischen rechtlichen und rechtsbasierten Rahmen ergänzen, auf dem das Modell beruht.

Darüber hinaus stützt sich das Regulierungsmodell auf etablierte Regulierungsansätze aus anderen Hochrisikosektoren, darunter das Sicherheitsmanagement in der Luftfahrt, die Governance von Finanzmodellrisiken, die Pharmakovigilanz in der Arzneimittelregulierung und algorithmische Folgenabschätzungen in der öffentlichen Verwaltung. Diese Ansätze liefern wertvolle Erkenntnisse für den Aufbau von Governance-Rahmenwerken, die die komplexen Lebenszyklusrisiken im Zusammenhang mit biomedizinischer KI bewältigen können, und wurden hier angepasst, um Überlegungen zur Gender- und Rassengleichheit explizit in den gesamten KI-Lebenszyklus einzubetten.

Das Modell bietet Gesundheitseinrichtungen und zivilgesellschaftlichen Organisationen einen **praktischen Governance-Rahmen, der die Beschaffung, Validierung, den Einsatz, die Überwachung und die Rechenschaftspflicht** für biomedizinische KI-Systeme **abdeckt**. Durch die Einbettung des Schutzes der Grundrechte in operative Prozesse zielt dieser Rahmen darauf ab, die Risiken gender- und rassenspezifischer Voreingenommenheit bei KI-gestützten Entscheidungen im Gesundheitswesen zu verringern und zu gerechteren Gesundheitsergebnissen in der gesamten EU beizutragen. Darüber hinaus **bietet** das Dokument **eine Reihe praktischer Leitlinien, Tools und Hilfsmittel, die Kliniken und zivilgesellschaftliche Organisationen an unterschiedliche nationale Kontexte anpassen können**, um das Modell in der Praxis konsistent umzusetzen.

Der europäische Rechtsrahmen für KI im Gesundheitswesen

Die EU hat ein mehrschichtiges Regulierungsumfeld für biomedizinische KI geschaffen, das aus vier sich ergänzenden rechtlichen Ebenen besteht: dem „AI Act“, dem Medizinproduktegesetz und den dazugehörigen Leitlinien, dem Datenschutz sowie der Governance von Gesundheitsdaten.

Der „AI Act“ schafft einen risikobasierten Rechtsrahmen für KI-Systeme. Viele Anwendungen im Gesundheitswesen, darunter Tools zur Unterstützung klinischer Entscheidungen und Diagnosealgorithmen, fallen in die Kategorie der hochrisikanten KI-Systeme, die strenge Anforderungen in Bezug auf Risikomanagement, Daten-Governance, Transparenz, menschliche Aufsicht und Genauigkeit erfüllen

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-LIT

müssen. Zu den geltenden Anforderungen gehören auch die Identifizierung und Minderung von Risiken diskriminierender Ergebnisse durch eine disaggregierte Analyse von Bevölkerungsgruppen, wobei Gender, ethnische Herkunft und andere soziale Kategorien als übergreifende Variablen behandelt werden. **Gesundheitseinrichtungen, die KI-Systeme einsetzen, haben gemäß dem „AI Act“ zudem spezifische Verpflichtungen.** Dazu gehören unter anderem die Sicherstellung, dass die Systeme gemäß den Anweisungen des Anbieters verwendet werden, die Überwachung ihrer Leistung, die Führung von Betriebsprotokollen sowie die Meldung schwerwiegender Vorfälle an die zuständigen Behörden, insbesondere wenn solche Vorfälle unterschiedliche Patientengruppen unverhältnismäßig stark betreffen könnten.

Was die Anwendung des „AI Act“ betrifft, so galten bereits im Jahr 2025 die Verbote, und einige Aufsichts- und Governance-Vorschriften traten ebenfalls in Kraft.¹⁰ Der nächste wichtige Meilenstein ist der August 2026, wenn der Großteil der Verpflichtungen aus dem „AI Act“ in Kraft tritt (einschließlich der in Anhang III genannten Hochrisikosysteme und der Transparenzpflichten gemäß Artikel 50). Für hochriskante KI, die in regulierte Produkte integriert ist (darunter fallen auch viele KI-Systeme in Medizinprodukten), soll die Regelung ab August 2027 gelten.

Für Kliniken sind die folgenden Bestimmungen des „AI Act“ von größter praktischer Bedeutung:

- **Anforderungen an hochriskante Systeme.** Anbieter müssen ein Risikomanagementsystem für den gesamten Lebenszyklus (Art. 9) sowie Daten-Governance-Verfahren umsetzen, die Folgendes umfassen: explizite Vorgaben zur Erkennung und Minderung von Bias (Art. 10), einschließlich der Bewertung der Repräsentativität der Daten und potenzieller Leistungsunterschiede zwischen verschiedenen Bevölkerungsgruppen anhand einer nach Gender und ethnischer Herkunft aufgeschlüsselten Analyse; Maßnahmen zur Gewährleistung der Transparenz und zur Bereitstellung von Informationen für die Betreiber (Art. 13); menschliche Überwachung zur Verhinderung von Automatisierungsbias und zur Ermöglichung von Übersteuerungs- oder Stoppverfahren (Art. 14); sowie Kontrollen hinsichtlich Genauigkeit, Robustheit und Cybersicherheit unter ausdrücklicher Berücksichtigung von Rückkopplungsschleifen und KI-spezifischen Angriffen (Art. 15), die andernfalls bestehende Ungleichheiten verschärfen könnten, wenn sie nicht angegangen werden.
- **Pflichten der Betreiber.** Kliniken und andere Betreiber müssen hochriskante KI-Systeme gemäß den Anweisungen des Anbieters einsetzen (Art. 26); eine kompetente menschliche Aufsicht zuweisen (Art. 26); sicherstellen, dass die

¹⁰ Europäische Kommission, [des EU-KI-Gesetzes](#), Brüssel: Europäische Kommission, 2024.

Eingabedaten repräsentativ und vollständig sind, sofern der Betreiber die Dateneingaben kontrolliert (Art. 26), sowie (soweit möglich) die demografische Vielfalt in klinischen Populationen berücksichtigen; den Systembetrieb überwachen (Art. 26); automatisch generierte Protokolle mindestens sechs Monate lang aufbewahren (Art. 26); und schwerwiegende Vorfälle rechtzeitig an die Anbieter und die zuständigen Behörden melden (Art. 26, 73), insbesondere wenn das Risiko ungleicher klinischer Ergebnisse besteht.

- **Grundrechte-Folgenabschätzung (GRFA).** Vor der Inbetriebnahme bestimmter hochrisikanter KI-Systeme müssen öffentlich-rechtliche Einrichtungen (sowie private Einrichtungen, die öffentliche Dienstleistungen erbringen) eine GRFA durchführen, in der Prozesse, betroffene Gruppen, Risiken, Maßnahmen zur menschlichen Aufsicht sowie Mechanismen zur Risikominderung oder zur Bearbeitung von Beschwerden beschrieben werden, und anschließend die Marktüberwachungsbehörde benachrichtigen (Art. 28). Diese Anforderung ist für öffentliche Kliniken in vielen Mitgliedstaaten von unmittelbarer Relevanz, da sie einen Mechanismus bereitstellt, um vor dem Einsatz Risiken von Diskriminierung und ungleichen Auswirkungen zu identifizieren und anzugehen, die durch eine nach Geschlecht und Gender sowie ethnischer Zugehörigkeit aufgeschlüsselte Analyse ermittelt werden.

Über den „AI Act“ hinaus **unterliegen KI-Systeme, die im medizinischen Kontext eingesetzt werden, auch dem Medizinprodukterecht und den damit verbundenen Leitlinien**, einschließlich der *Medizinprodukteverordnung (MDR)*¹¹ und der *In-vitro-Diagnostika-Verordnung (IVDR)*,¹² , die die Sicherheit und Leistungsfähigkeit medizinischer Technologien regeln. Diese Verordnungen verlangen eine klinische Bewertung und eine Überwachung nach dem Inverkehrbringen für Medizinproduktesoftware, einschließlich KI-basierter Tools, wobei auf eine gerechte Leistungsfähigkeit über disaggregierte Patientengruppen hinweg geachtet werden muss, sofern die Daten dies erlauben. Darüber hinaus enthält der Leitfaden der Koordinierungsgruppe für Medizinprodukte (MDCG) *Leitlinie zur Qualifizierung und Klassifizierung von Software gemäß MDR/IVDR* Richtlinien dazu, wie Software im Rahmen der MDR und der IVDR zu qualifizieren und zu klassifizieren ist.¹³ Sie betont, dass die Klassifizierung vom Verwendungszweck und nicht vom Einsatzort der Software (z. B. cloudbasiert oder eingebettet) abhängt, und enthält Beispiele aus Klinikinformationssystemen und Software zur Unterstützung klinischer Entscheidungen,

¹¹ Europäische Union, *Verordnung (EU) 2017/745 über Medizinprodukte (MDR)*, Amtsblatt der Europäischen Union, L 117, Brüssel: EU, 2017.

¹² Europäische Union, *Verordnung (EU) 2017/746 über In-vitro-Diagnostika (IVDR)*, Amtsblatt der Europäischen Union, L 117, Brüssel: EU, 2017.

¹³ Koordinierungsgruppe für Medizinprodukte, *Leitlinien zur Qualifizierung und Klassifizierung von Software in der Verordnung (EU) 2017/745 (MDR) und der Verordnung (EU) 2017/746 (IVDR)*, Brüssel: MDCG, 2019.

wobei die Notwendigkeit hervorgehoben wird, unterschiedliche Auswirkungen auf verschiedene Patientengruppen zu berücksichtigen. Der *Leitfaden* der MDCG zur *klinischen und Leistungsbewertung von Medizinprodukt-Software* bietet zusätzliche Unterstützung und bekräftigt die Erwartung, dass die Nachweise in einem angemessenen Verhältnis zur klinischen Rolle der Software stehen sollten, einschließlich des Nachweises, dass die Leistung der KI über Populationen hinweg konsistent ist, die nach biologischem Geschlecht und Gender sowie nach rassischem oder ethnischen Hintergrund differenziert werden.¹⁴

Das Verhältnis zwischen dem „AI Act“ und dem bestehenden Medizinprodukterecht wird in den *Leitlinien zum Zusammenspiel zwischen dem „AI Act“ und der MDR/IVDR* erläutert.¹⁵ Das Dokument erklärt, dass der „AI Act“ parallel zu diesen beiden Verordnungen gilt. In der Praxis bedeutet dies, **dass viele medizinische KI-Systeme beide Rechtsrahmen gleichzeitig erfüllen müssen**. In den Leitlinien wird zudem klargestellt, wann medizinische KI im Sinne des „AI Act“ als hochriskant einzustufen ist. Dies ist in der Regel dann der Fall, wenn das KI-System selbst ein Medizinprodukt (oder ein Sicherheitsbauteil eines solchen) ist und daher einer Konformitätsbewertung durch eine benannte Stelle unterzogen werden muss. Für die Beschaffung in Kliniken hat dies praktische Auswirkungen. Kliniken sollten von Anbietern verlangen, den regulatorischen Status ihrer Systeme nachzuweisen, einschließlich der MDR-/IVDR-Einstufung, der Einbeziehung einer benannten Stelle, sofern zutreffend, und der Übereinstimmung mit den Anforderungen des „AI Act“ für hochrisikante Systeme sowie der Dokumentation der Leistung in aufgeschlüsselten Bevölkerungsgruppen und gegebenenfalls der Maßnahmen zur Verringerung von Bias. Beschaffungsverträge sollten zudem vorsehen, dass Anbieter die Dokumentation, Transparenzinformationen und Überwachungsmaßnahmen bereitstellen, die von den Betreibern gemäß dem „AI Act“ benötigt werden, einschließlich Bestimmungen zur Gewährleistung gerechter Ergebnisse für alle Patienten.

Mehrere **weitere EU-Instrumente sind in der Praxis ebenfalls für die Governance von KI im Gesundheitswesen von Bedeutung**. Die *Datenschutz-Grundverordnung (DSGVO)* bleibt eine zentrale Einschränkung für biomedizinische KI, da Gesundheitsdaten zu den besonderen Kategorien personenbezogener Daten gehören und der Einsatz in Kliniken häufig Datenschutz-Folgenabschätzungen (DPIA) erfordert. Der „AI Act“ (Art. 26 Abs. 9, Art. 27 Abs. 4) verknüpft die Praxis der Betreiber ausdrücklich mit den Verpflichtungen aus der Datenschutz-Folgenabschätzung (Verwendung von Anbieterinformationen zur Unterstützung der Datenschutz-Folgenabschätzung) und positioniert die GRFA als Ergänzung zur Datenschutz-Folgenabschätzung, was die Berücksichtigung unterschiedlicher Auswirkungen auf Bevölkerungsgruppen, differenziert nach

¹⁴ Koordinierungsgruppe für Medizinprodukte, *Leitlinien zur klinischen Bewertung (MDR) und zur Leistungsbewertung (IVDR) von Software in Medizinprodukten*, Brüssel: MDCG, 2020.

¹⁵ Koordinierungsgruppe für Medizinprodukte und Europäischer KI-Rat, *Leitlinien zum Zusammenspiel von MDR/IVDR und „AI Act“*, Brüssel: Europäische Kommission, 2025.

Geschlecht und Gender sowie rassischer/ethnischer Herkunft, einschließen kann. Darüber hinaus zielt der *Europäische Gesundheitsdatenraum (EHDS)* darauf ab, den sicheren Zugang zu Gesundheitsdaten für die Gesundheitsversorgung und die Forschung zu erleichtern und gleichzeitig strenge Datenschutzgarantien aufrechtzuerhalten.¹⁶ Es wird erwartet, dass der EHDS eine Schlüsselrolle dabei spielen wird, repräsentativere Datensätze für die Entwicklung biomedizinischer KI zu ermöglichen, den Fokus auf Gerechtigkeit zu legen und das Risiko einer Unterrepräsentation von Patientengruppen mit bestimmten demografischen Merkmalen zu verringern.¹⁷

Was schließlich die Cybersicherheit betrifft, hat die Agentur der Europäischen Union für Cybersicherheit Leitlinien zu Cybersicherheitspraktiken für KI herausgegeben. Darin wird argumentiert, dass bestehende organisatorische Standards (z. B. ISO-Standards für das Sicherheits- und Qualitätsmanagement) an KI-Kontexte angepasst werden können.¹⁸ Die *Leitlinien der MDCG zur Cybersicherheit für Medizinprodukte* beschreiben, wie Hersteller die Anforderungen der MDR und der IVDR im Bereich Cybersicherheit durch Risikomanagement und Mindestanforderungen an die Cybersicherheit erfüllen sollten.¹⁹ Wichtig ist, dass eine starke Cybersicherheits-Governance den gerechten Einsatz von KI untermauert: Der Schutz der Datenintegrität, des Datenzugriffs und der Systemzuverlässigkeit trägt dazu bei, eine unbeabsichtigte Verstärkung von Ungleichheiten aufgrund des Gender-Prinzips sowie der ethnischen Herkunft zu verhindern.

Zusammen schaffen diese Rechtsinstrumente ein komplexes regulatorisches Umfeld, in dem KI-Systeme im Gesundheitswesen betrieben werden sollen. Das hier vorgeschlagene Regulierungsmodell für KI in europäischen öffentlichen Kliniken ist ausdrücklich auf diese vier zentralen rechtlichen „Ebenen“ ausgerichtet, die die Praxis der KI im Gesundheitswesen regeln. Dementsprechend sieht das Modell Mechanismen vor, um ungleiche Auswirkungen auf Bevölkerungsgruppen, differenziert nach biologischem und sozialem Gender sowie rassischer oder ethnischer Herkunft, über den gesamten KI-Lebenszyklus hinweg zu überwachen und abzumildern.

¹⁶ Europäische Union, *Verordnung über den Europäischen Gesundheitsdatenraum (EHDS)*, Brüssel: Europäische Union, 2024.

¹⁷ Europäische Kommission, *Europäischer Gesundheitsdatenraum (EHDS)*, Brüssel: Europäische Kommission, 2025.

¹⁸ Agentur der Europäischen Union für Cybersicherheit (ENISA), *Cybersicherheit von KI und Standardisierung*, Athen: ENISA, 2023.

¹⁹ Koordinierungsgruppe für Medizinprodukte, *MDCG 2019-16 Rev. 1 – Leitlinien zur Cybersicherheit für Medizinprodukte*, Brüssel: Europäische Kommission, 2020.

Europäische und internationale Governance-Rahmenwerke für verantwortungsvolle KI im Gesundheitswesen

Es ist wichtig zu beachten, dass die oben erwähnten europäischen Regulierungs- und Standardisierungsrahmen die primäre Rechtsgrundlage für die Regulierung der biomedizinischen KI innerhalb der Union bilden. Neben dem EU-Regulierungsrahmen für biomedizinische KI bieten mehrere nicht bindende internationale Instrumente Leitlinien für den Umgang mit KI-Risiken. Insbesondere **werden die in dieser Sektion vorgestellten zusätzlichen Governance-Rahmenwerke einbezogen, um Umsetzungstechniken aufzuzeigen, die auf den EU-Kontext übertragen werden können, wobei sie den europäischen rechtlichen Anforderungen und den Grundrechtsprinzipien uneingeschränkt untergeordnet bleiben.** Sie werden selektiv als Analogien aus anderen Rechtsordnungen herangezogen, in denen in der Praxis vergleichbare Mechanismen zur Lebenszyklus-Governance geschaffen wurden, insbesondere in Bereichen wie dem Modellrisikomanagement, der Algorithmenprüfung und der institutionellen Folgenabschätzung.

So bietet beispielsweise das *AI Risk Management Framework* (RMF) des National Institute of Standards and Technology (NIST) einen umfassenden Ansatz zum Management von KI-Risiken durch einen Lebenszyklusansatz, der sich um Governance, Risikokartierung, Risikomessung und Risikomanagement dreht.²⁰ Das Rahmenwerk erkennt Fairness und die Bekämpfung schädlicher Bias-Formen ausdrücklich als zentrale Aspekte einer vertrauenswürdigen KI an. Darüber hinaus ermutigt es Organisationen im Rahmen der Risikokartierung und -messung dazu, unterschiedliche Auswirkungen auf verschiedene Patientengruppen zu identifizieren und zu bewerten, wobei auch biologisches Geschlecht und Gender sowie rassische und ethnische Unterschiede berücksichtigt werden. Ziel ist es, Vertrauenswürdigkeit und Risikokontrollen in die gesamte Entwicklung, den Einsatz und die Überwachung zu integrieren und dabei technische Bewertung mit organisatorischer Rechenschaftspflicht zu verbinden. Für Kliniken liegt die Stärke des KI-RMF in seiner Umsetzbarkeit in konkrete Praktiken, wie beispielsweise **definierte Rollen für die Überwachung der Gleichbehandlung, die methodische Bewertung der Leistung nach aufgeschlüsselten Bevölkerungsgruppen und dokumentierte Strategien zur Risikominderung**, wobei gleichzeitig eine umfassendere, auf Rechten basierende Perspektive gewahrt bleibt.

Auf der Seite der Managementsysteme legt die *Norm ISO/IEC 42001* der Internationalen Organisation für Normung Anforderungen an ein organisatorisches KI-Managementsystem fest, wobei der Fokus auf Richtlinien, Zielen, Prozessen und der

²⁰ National Institute of Standards and Technology, *Rahmenwerk für das Risikomanagement im Bereich künstlicher Intelligenz (AI RMF 1.0)*, Gaithersburg, MD: NIST, 2023.

kontinuierlichen Verbesserung für eine verantwortungsvolle KI liegt.²¹ Obwohl sie keine konkreten Vorgaben zu bestimmten Schäden enthält, unterstützt sie die Integration von Fairness- und Nichtdiskriminierungsprinzipien in organisatorische Richtlinien und Risikokontrollen, wodurch Institutionen in die Lage versetzt werden, potenzielle gender- und rassenspezifische Bias auf koordinierte Weise anzugehen. Ergänzend dazu bietet *die Norm ISO/IEC 23894* Leitlinien zur Integration des Risikomanagements in die Entwicklung und den Einsatz von KI, einschließlich der Identifizierung und Behandlung von Risiken im Zusammenhang mit unbeabsichtigten oder unverhältnismäßigen Auswirkungen auf verschiedene Bevölkerungsgruppen.²² Speziell für klinische KI definiert *die Norm ISO 14971* Risikomanagement-Prinzipien, die in der Regulierung von Medizinprodukten weit verbreitet sind und die systematische Gefahrenidentifizierung, Risikobewertung, Risikokontrolle und Lebenszyklusüberwachung beschreiben.²³ Diese Prozesse lassen sich erweitern, um im Rahmen der Schadensidentifizierung auch ungleiche Leistungen oder Ergebnisse bei nach demografischen Merkmalen aufgeschlüsselten Bevölkerungsgruppen zu berücksichtigen. Die beschriebenen Normen sind wertvoll, da sie **eine Managementsystemstruktur vorschlagen, die Kliniken übernehmen können**. Sie bieten zudem eine gemeinsame Sprache, mit der Anforderungen im Zusammenhang mit Gerechtigkeit – wie die nach demografischen Merkmalen aufgeschlüsselte Bevölkerungsbewertung und die Minderung von Bias – bei der Beschaffung und bei Vertragsabschlüssen mit Lieferanten festgelegt werden können.

Auch ethische Governance-Rahmenwerke bieten wichtige normative Leitlinien. Die *Ethik-Leitlinien* der Europäischen Kommission *für vertrauenswürdige KI* sind ein nützliches, wenn auch nicht verbindliches politisches Instrument.²⁴ Sie definieren sieben Kernanforderungen für verantwortungsvolle KI-Systeme, darunter Vielfalt, Nichtdiskriminierung und Fairness, die ausdrücklich die Vermeidung von Bias und Fairness fordern und die Berücksichtigung von Auswirkungen auf Bevölkerungsgruppen mit unterschiedlichen demografischen Merkmalen wie biologischem Geschlecht und Gender sowie rassischer/ethnischer Herkunft fordern. Die „*Assessment List for Trustworthy Artificial Intelligence*“ (ALTAI), ein freiwilliges Tool für die Selbstbewertung, das von der hochrangigen Expertengruppe der Europäischen Kommission für künstliche Intelligenz entwickelt wurde, setzt diese Leitlinien in Form einer Selbstbewertungsliste

²¹ Internationale Organisation für Normung, *ISO/IEC 42001:2023 – Managementsystem für künstliche Intelligenz*, Genf: ISO, 2023.

²² Internationale Organisation für Normung, *ISO/IEC 23894:2023 – Risikomanagement für künstliche Intelligenz*, Genf: ISO, 2023.

²³ Internationale Organisation für Normung, *ISO 14971:2019 – Medizinprodukte: Anwendung des Risikomanagements auf Medizinprodukte*, Genf: ISO, 2019.

²⁴ Europäische Kommission, *Ethische Leitlinien für vertrauenswürdige KI*, Brüssel: Europäische Kommission, 2019.

um.²⁵ In einem Regulierungsmodell für das Gesundheitswesen liegt der Wert von ALTAI in seinem **fragengesteuerten Format, das in praktische Checklisten für die Beschaffung und den Einsatz angepasst werden kann**. Diese veranlassen die Beteiligten, Aspekte wie die Repräsentativität der Trainingsdaten, das Risiko unterschiedlicher klinischer Ergebnisse und die Angemessenheit von Schutzmaßnahmen gegen diskriminierende Auswirkungen zu berücksichtigen, wodurch Gleichstellungsaspekte in die routinemäßige Entscheidungsfindung von Klinikern, IT-Spezialisten und Klinikmanagern eingebettet werden.

Aus menschenrechtlicher Perspektive ist das *Rahmenübereinkommen* des Europarats *über KI* ein relevantes Rechtsinstrument, das ausdrücklich auf den Prinzipien der Menschenwürde, Transparenz, Rechenschaftspflicht, Gleichheit und Nichtdiskriminierung sowie des Datenschutzes verankert ist und damit eine Rechtsgrundlage für die Bekämpfung diskriminierender Auswirkungen bietet, einschließlich solcher, die mit genderspezifischen und rassistischen Bias in KI-Systemen zusammenhängen.²⁶ Ebenso bieten *die Prinzipien* der Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) *zur künstlichen Intelligenz*²⁷ und *die Empfehlung* der UNESCO *zur Ethik der KI*²⁸ internationale normative Leitlinien, die die Identifizierung und Minderung ungerechter oder ungleicher Auswirkungen auf bestimmte Bevölkerungsgruppen fördern und damit die Erwartung bekräftigen, dass KI-Systeme strukturelle Ungleichheiten nicht reproduzieren oder verstärken sollten. Für eine gesundheitspezifische ethische Steuerung hat die Weltgesundheitsorganisation sechs Kernprinzipien für KI im Gesundheitswesen formuliert: Autonomie, Wohlbefinden, Erklärbarkeit, Verantwortung, Gerechtigkeit und Reaktionsfähigkeit.²⁹ Das Prinzip der **Gerechtigkeit erfordert ausdrücklich die Berücksichtigung von Unterschieden beim Zugang, bei der Leistung und bei den Ergebnissen in verschiedenen Patientengruppen, wobei auch genderspezifische sowie rassische oder ethnische Unterschiede zu berücksichtigen sind**.

Diese internationalen Governance-Rahmenwerke stehen in direktem Zusammenhang mit den an der Charta ausgerichteten Anforderungen hinsichtlich Würde, Gleichheit, Nichtdiskriminierung und dem Recht auf Gesundheit und legen einen gemeinsamen Schwerpunkt auf die Lebenszyklus-Governance – Prinzipien, die sich direkt in praktische Kontrollmaßnahmen für Gesundheitseinrichtungen umsetzen lassen.

²⁵ Europäische Kommission, *Bewertungsliste für vertrauenswürdige künstliche Intelligenz (ALTAI) – Selbstbewertung*, Brüssel: Europäische Kommission, 2020.

²⁶ Europarat, *Rahmenübereinkommen über künstliche Intelligenz (CAI)*, Straßburg: Europarat, 2023.

²⁷ Organisation für wirtschaftliche Zusammenarbeit und Entwicklung, *OECD-Prinzipien zur künstlichen Intelligenz*, Paris: OECD, 2019.

²⁸ UNESCO, *Empfehlung zur Ethik der künstlichen Intelligenz*, Paris: UNESCO, 2021.

²⁹ Weltgesundheitsorganisation, *Ethik und Governance der künstlichen Intelligenz im Gesundheitswesen*, Genf: WHO, 2021.

Erfahrungen aus sektorübergreifenden Regulierungsmodellen

Risikoreiche Sektoren wie Luftfahrt, Bankwesen und Pharmaindustrie haben seit langem ausgefeilte Governance-Systeme zur Bewältigung technologischer Gefahren entwickelt. Diese Sektoren bieten wertvolle Regulierungsansätze, die auf die KI-Governance im Gesundheitswesen und den Umgang mit damit verbundenen Herausforderungen im Zusammenhang mit gendesspezifischen und ethnischen Bias angepasst werden können.

Zwar sind genderspezifische und rassenspezifische Bias im engeren Sinne keine technologischen Gefahren, doch treten sie als soziotechnische Risiken zutage, wenn Daten, Algorithmen und organisatorische Praktiken mit strukturellen Ungleichheiten im Gesundheitswesen interagieren. Hier bezieht sich „Gefahr“ auf die Manifestation von Bias auf Systemebene in Ergebnissen, Entscheidungen und klinischen Resultaten und nicht auf deren Ursprung. Das Gesundheitswesen selbst liefert bereits umfangreiche Belege für solche Effekte, darunter dokumentierte Ungleichheiten bei Diagnose, Behandlung und Risikovorhersage zwischen Bevölkerungsgruppen, die nach biologischem Geschlecht und Gender sowie ethnischer Zugehörigkeit unterschieden werden. Sektorübergreifende Beispiele ergänzen daher die Erkenntnisse aus dem Gesundheitswesen, indem sie ausgereifte Governance-Mechanismen für den Umgang mit komplexen, wirkungsstarken Systemen veranschaulichen – neben empirischen Befunden aus dem Gesundheitswesen, die die Notwendigkeit solcher regulatorischen Kontrollen belegen.

Ein besonders relevantes Beispiel ist das in der Luftfahrt verwendete *Sicherheitsmanagementsystem (SMS)*. Internationale Vorschriften für die Zivilluftfahrt behandeln Sicherheit nicht als einmaligen Zertifizierungsprozess, sondern als kontinuierlichen, organisationsweiten Prozess, der Entscheidungen auf der Grundlage von Gefahrenidentifizierung, Risikomanagement, Sicherheitssicherung (Überwachung und Audit), Veränderungsmanagement und kontinuierlicher Verbesserung ermöglicht.³⁰ Die Anwendung dieser Logik auf KI im Gesundheitswesen würde **die Einrichtung eines „KI-Sicherheitsmanagementsystems“** bedeuten, **das explizit Gleichstellungsaspekte einbezieht**, darunter: (1) einen Meldeweg für Gefahren und Beinahevorfälle für klinisches Personal, über den potenzielle Vorfälle gender- oder rassenspezifischer Bias-Situationen erfasst werden; (2) ein formelles Risikoprotokoll für jedes KI-System, in dem Leistungsunterschiede zwischen aufgeschlüsselten Patientengruppen dokumentiert werden; (3) Auslöser für Änderungskontrollen, die die Auswirkungen von Modellaktualisierungen, Datenverschiebungen oder Änderungen der Arbeitsabläufe auf die Gleichstellungsergebnisse bewerten; und (4) regelmäßige Überprüfungen der Sicherheitsgewährleistung mit Korrekturmaßnahmen, die sowohl

³⁰ Internationale Zivilluftfahrt-Organisation, *Handbuch zum Sicherheitsmanagement (Doc 9859)*, 4. Auflage, Montréal: ICAO, 2018.

technische Risiken als auch festgestellte gender- oder rassenspezifische Ungleichheiten angehen. Die ausdrückliche Betonung des Lebenszyklus-Risikomanagements und der Überwachung nach dem Inverkehrbringen im „AI Act“ (Art. 9, 72) steht in engem Einklang mit dieser SMS-Logik.

Ein zweiter nützlicher Ansatz stammt aus der Finanzregulierung, wo komplexe quantitative Tools als „Modelle“ behandelt werden, die eine solide Entwicklung, eine unabhängige Validierung sowie Aufsicht durch die Governance erfordern. In der Bankenaufsicht wurden in Ländern wie den Vereinigten Staaten vergleichbare Rahmenwerke für das Risikomanagement eingeführt (z. B. die „*Supervisory Guidance on Model Risk Management*“ der US-Notenbank) und liefern ein Beispiel dafür, wie Validierung, Aufsicht und unabhängige Modellüberprüfung umgesetzt werden können.³¹ Ähnliche Ansätze könnten auf biomedizinische KI-Systeme angewendet werden, um sicherzustellen, dass **Tools zur Unterstützung klinischer Entscheidungen (intern oder extern) unabhängig validiert werden, bevor sie in Einrichtungen des Gesundheitswesens eingeführt werden, wobei die Leistung explizit über Populationen hinweg bewertet wird, die nach biologischem Geschlecht und Gender sowie nach rassischen oder ethnischen Merkmalen differenziert sind.** Mindestanforderungen an die Validierungsunterlagen sollten den Verwendungszweck des Systems, seine Betriebsgrenzen, bekannte Ausfallmodi und die aufgeschlüsselte Leistungsfähigkeit in verschiedenen Bevölkerungsgruppen dokumentieren, einschließlich etwaiger Ungleichheiten oder angewandter Abhilfemaßnahmen. Eine erneute Validierung sollte immer dann ausgelöst werden, wenn das System an einem neuen Standort eingesetzt, auf eine andere Patientengruppe angewendet, an neue Geräte oder Protokolle angeschlossen, vom Anbieter aktualisiert wird oder Anzeichen einer Leistungsabweichung zeigt. Diese Vorgehensweisen stehen im Einklang mit dem „AI Act“ (Art. 13, 14, 60), der Tests, Transparenz gegenüber den Betreibern und eine menschliche Aufsicht vorschreibt, um Automatisierungs-Bias zu erkennen und zu verhindern und sicherzustellen, dass die KI-gestützte Versorgung für verschiedene Bevölkerungsgruppen sowohl sicher als auch gerecht ist.

Was die Governance nach dem Einsatz betrifft, bietet die Arzneimittelregulierung ein starkes Vorbild: In der Pharmakovigilanz werden Nebenwirkungen und Sicherheitssignale systematisch erfasst und analysiert, um aufkommende Risiken zu erkennen. Insbesondere *das Modul VI der „Good Pharmacovigilance Practices“ (GVP)* der Europäischen Arzneimittelagentur legt detaillierte Anforderungen für die Erfassung, Verwaltung und Einreichung von Meldungen über vermutete Nebenwirkungen fest und spiegelt damit einen ökosystemorientierten Ansatz für die Meldung und Nachverfolgung von Vorfällen wider.³² Ein vergleichbares

³¹ Verwaltungsrat des Federal Reserve Systems und Amt des Währungsaufsichtsbeamten, [Aufsichtsrichtlinie zum Modellrisikomanagement \(SR 11-7\)](#), Washington, D.C., 4. April 2011.

³² Europäische Arzneimittelagentur, [GVP-Modul VI: Erfassung und Meldung vermuteter unerwünschter Arzneimittelwirkungen](#), London: EMA, 2017

„KI-Vigilanz“-System könnte Kliniken ermöglichen, Vorfälle im Zusammenhang mit KI-gestützten klinischen Entscheidungen zu melden und zu analysieren, wobei besonderes Augenmerk auf Gerechtigkeitsergebnisse gelegt wird. Zu den Schlüsselementen würden gehören: (1) eine systematische Berichterstattung, die alle beobachteten Ungleichheiten in der Versorgung oder bei den Ergebnissen hinsichtlich des biologischen oder sozialen Gender-Profils sowie rassischer/ethnischer Unterschiede erfasst; (2) die verpflichtende Dokumentation schwerwiegender Vorfälle, bei denen Bias zu einer unterschiedlichen Behandlung beigetragen haben könnte; (3) die Verpflichtung, den Leistungserbringer und die zuständigen Behörden über Bedenken hinsichtlich der Gleichbehandlung zu informieren (Benachrichtigung der Behörden, wenn Bias zu schwerwiegendem Schaden führt, wie z. B. Fehldiagnosen oder ungleiche Behandlung, während routinemäßige Ungleichheiten im Rahmen der internen Qualitätsüberwachung behandelt werden); und (4) Funktionen zur Trend- und Signalerfassung, die sich abzeichnende Muster von Bias über Abteilungen, Einrichtungen oder das gesamte Gesundheitsnetzwerk hinweg identifizieren. Eine solche Überwachung würde proaktive Abhilfemaßnahmen und eine kontinuierliche Verbesserung ermöglichen und sicherstellen, dass die KI-gestützte Versorgung nicht nur sicher, sondern auch gerecht ist.

Schließlich bieten Governance-Tools des öffentlichen Sektors wie algorithmische Folgenabschätzungen und Bias-Prüfungen praktische Methoden zur Bewertung der gesellschaftlichen Auswirkungen automatisierter Entscheidungssysteme, einschließlich potenzieller Ungleichheiten hinsichtlich biologischen Geschlechts und Gender oder rassischer und ethnischer Unterschiede. Diese Tools setzen ethische Prinzipien und Menschenrechtsstandards in umsetzbare Formate um, mit denen Organisationen Ungleichheiten vor der Einführung antizipieren und mindern können. So bewertet beispielsweise Kanadas bundesweite „*Algorithmic Impact Assessment*“ die Risiken von KI-Systemen und verknüpft sie mit Maßnahmen zur Risikominderung.³³ Dieses Tool fasst den Einsatz ausdrücklich als ein wirkungsorientiertes Governance-Problem auf. Darüber hinaus sieht der Ansatz der Stadt New York für automatisierte Instrumente zur Personalentscheidung Biasprüfungen und Offenlegungspflichten gegenüber der Öffentlichkeit vor.³⁴ Dementsprechend könnten Kliniken und im Gesundheitswesen tätige zivilgesellschaftliche Organisationen ein **Paket zur „KI-Folgenabschätzung und -Prüfung“** einführen, **das einen strukturierten, mit Fokus auf Gerechtigkeit gestalteten Fragebogen mit regelmäßigen unabhängigen Fairness-Audits kombiniert** und eine veröffentlichungsfähige Zusammenfassung der Leistung über aufgeschlüsselte Bevölkerungsgruppen sowie der Maßnahmen zur Risikominderung liefert. Die Umsetzung eines solchen Pakets würde dazu beitragen, Transparenz, öffentliche

³³ Sekretariat des kanadischen Finanzrats, *Algorithmische Folgenabschätzung*, Ottawa: Regierung von Kanada, 2026.

³⁴ New Yorker Amt für Verbraucher- und Arbeitnehmerschutz, *Automatisierte Tools zur Beschäftigungsentscheidung*, New York: Stadt New York, 2023.

Rechenschaftspflicht und das Vertrauen der Patienten in die KI-gestützte Versorgung zu fördern.

Die Einbeziehung von Regulierungsmodellen aus Bereichen wie der Flugsicherheit, dem Risikomanagement bei Finanzmodellen, der Pharmakovigilanz und der algorithmischen Governance ist als bewusster Transfer etablierter Governance-Prinzipien gedacht. Diese Bereiche wurden ausgewählt, weil sie formalisierte Ansätze zur Risikoerkennung, -überwachung und -minderung in komplexen Systemen, in denen unterschiedliche Auswirkungen auf Einzelpersonen auftreten können, in der Praxis umsetzen. In einigen davon sind Fairness-Praktiken bereits in der Praxis verankert, darunter die demografische Risikostratifizierung in der Pharmakovigilanz, Modellvalidierung und Biaskontrollen in der Finanzaufsicht sowie die systematische Meldung unerwünschter Ereignisse, die bestimmte Bevölkerungsgruppen betreffen. Ihre Relevanz liegt sowohl in der konzeptionellen Ähnlichkeit als auch in ihrem bewährten Einsatz von Lebenszyklus-Governance, Meldepflichten und unabhängigen Aufsichtsmechanismen, die direkt angepasst werden können, um Risiken für die Gender- und Rassengleichheit in biomedizinischen KI-Systemen anzugehen.

Blaupause für ein an der Charta ausgerichtetes KI-Regulierungsmodell für öffentliche Kliniken und zivilgesellschaftliche Organisationen

Diese Sektion beschreibt eine Architektur für ein Regulierungsmodell, das **in den Prinzipien der EU-Charta verankert ist, jedoch in Form praktischer Verfahren dargestellt wird, die Kliniken und zivilgesellschaftliche Organisationen in der Praxis umsetzen können**. Die Charta bildet die normative Grundlage für die Governance von Gesundheitstechnologien in Europa, insbesondere durch die Prinzipien der Würde des Menschen, der Gleichheit, der Nichtdiskriminierung und des Rechts auf Gesundheitsversorgung.³⁵ Diese Rechte sind besonders relevant, wenn automatisierte Systeme klinische Entscheidungen beeinflussen, die sich auf die Patientenversorgung auswirken, einschließlich potenzieller Ungleichheiten zwischen Bevölkerungsgruppen, die nach biologischem Geschlecht und Gender unterschieden werden.

Auf der grundlegendsten Ebene stützt Artikel 1 der Charta, der festlegt, dass die Würde des Menschen unantastbar ist, die Anforderungen an die Governance, die sicherstellen, dass KI-Tools so eingesetzt werden, dass menschliches Urteilsvermögen gewahrt bleibt und die Autonomie der Patienten respektiert wird. In der Praxis erfordert dieses Prinzip **Schutzmaßnahmen gegen eine übermäßige Abhängigkeit von automatisierten Empfehlungen, Mechanismen zur Anfechtung KI-gestützter Entscheidungen sowie**

³⁵ Europäische Union, *Charta der Grundrechte der Europäischen Union*, 2012/C 326/02, Straßburg: EU.

klinische Aufsichtsstrukturen, die Automatisierungs-Bias verhindern – also die Tendenz menschlicher Entscheidungsträger, übermäßig auf algorithmische Ergebnisse zu vertrauen, was Ungleichheiten verstärken kann, wenn KI-Empfehlungen bestimmte untergrupperte Bevölkerungsgruppen systematisch benachteiligen.

Ebenso bilden die Artikel 20 und 21, die die Gleichheit vor dem Gesetz verankern und Diskriminierung verbieten, die rechtliche Grundlage für ausdrückliche Verpflichtungen zur Fairness und zur Verringerung von Bias in der biomedizinischen KI. Krankenhäuser, die KI-Systeme einsetzen, sollten daher **Verfahren zur Überwachung der Systemleistung über verschiedene Patientengruppen hinweg** einführen, einschließlich biologischen Geschlechts, Gender und Rasse oder ethnischer Zugehörigkeit sowie intersektionaler Kombinationen dieser Merkmale. Werden Ungleichheiten bei der Leistung oder den Ergebnissen festgestellt, sollten Korrekturmaßnahmen durch Modellanpassungen, Änderungen im klinischen Arbeitsablauf oder alternative Entscheidungswege umgesetzt werden, um sicherzustellen, dass der Einsatz von KI keine strukturellen Ungleichheiten verfestigt.

Darüber hinaus garantiert Artikel 35 der Charta das Recht auf ein hohes Schutzniveau für die menschliche Gesundheit und untermauert damit das Prinzip, dass KI-Systeme nur dann eingesetzt werden dürfen, wenn sie nachweislich zu einer sicheren, wirksamen und gerechten Versorgung im Gesundheitswesen beitragen. KI-Technologien sollten keine neuen Hindernisse für die Versorgung schaffen oder bestehende Ungleichheiten verschärfen, von denen historisch unterversorgte oder anderweitig benachteiligte Patientengruppen betroffen sind, und ihre Auswirkungen auf solche Patientengruppen sollten kontinuierlich bewertet werden.

Um diese Rechte in der Praxis zu gewährleisten, verfolgt das in diesem Policy-Dokument vorgeschlagene Regulierungsmodell einen **Lebenszyklus-Governance-Ansatz, der die KI-Aufsicht anhand einer Reihe verbindlicher Entscheidungspunkte oder „Gates“ strukturiert**. Diese Kontrollpunkte spiegeln die im „AI Act“ und den damit verbundenen Gesundheitsvorschriften verankerte Regulierungslogik wider und bieten Kliniken gleichzeitig einen klaren und nachvollziehbaren Governance-Pfad. Das Modell geht davon aus, dass die Klinik in der Regel ein „Betreiber“ (im Sinne des „AI Act“) und ein Gesundheitsdienstleister nach nationalem Gesundheitsrecht ist, berücksichtigt jedoch auch den Fall, dass die Klinik zum Anbieter/Hersteller wird (wie im Beispiel der hauseigenen KI).

Kliniken sind die Hauptverantwortlichen für den Betrieb und die Kontrolle von KI-Systemen, während CSOs als unabhängige Aufsichts-, Prüf- und rechtsbasierte Überwachungsgremien fungieren, ohne die tägliche Kontrolle über klinische Einsatzentscheidungen zu haben.

Die Umsetzung dieses Regulierungsmodells folgt dem Prinzip der Verhältnismäßigkeit und trägt dabei der Tatsache Rechnung, dass **Kliniken und CSOs je nach Mitgliedstaat**

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-LIT

und Gesundheitssystem über unterschiedliche technische, organisatorische und analytische Kapazitäten verfügen können. Verbindliche Mindestanforderungen gelten einheitlich, um grundlegende Sicherheits-, Transparenz- und Gerechtigkeitsgarantien zu gewährleisten. Darüber hinaus können fortgeschrittene Verfahren wie erweiterte, disaggregierte Bevölkerungsanalysen, externe Audits und ausgefeilte Tools zur Überwachung von Bias je nach institutioneller Kapazität schrittweise eingeführt werden. Der technische Anhang folgt ebenfalls diesem Ansatz, indem er zwischen Mindest- und empfohlenen Anforderungen unterscheidet.

1. **Gate 1 – Anwendungsbereich und Klassifizierungskriterien.** Vor der Beschaffung oder Entwicklung müssen Einrichtungen den regulatorischen Status eines bestimmten KI-Tools bestimmen. Dazu gehört die Beurteilung, ob das System als Medizinprodukt-Software gemäß der MDR oder IVDR einzustufen ist, ob es ein KI-System mit hohem Risiko im Sinne des „AI Act“ darstellt oder ob es in eine Kategorie mit geringerem Risiko fällt, die dennoch einer behördlichen Aufsicht unterliegt. In dieser frühen Phase können Kliniken Tools kennzeichnen, deren Verwendungszweck möglicherweise verstärkte Auswirkungen auf die Chancengleichheit hat oder verschiedene Patientengruppen betrifft, wobei zu beachten ist, dass diese Tools in späteren Phasen eine detailliertere Bewertung hinsichtlich genderspezifischer und rassistischer Voreingenommenheit erfordern werden. Die von der MDCG erstellten Leitlinien zur Softwarequalifizierung und -klassifizierung enthalten praktische Beispiele, die dazu beitragen können, eine falsche regulatorische Einstufung zu vermeiden und die Grundlage für Bewertungen hinsichtlich Bias und Chancengleichheit zu schaffen.³⁶ Kliniken leiten die Klassifizierung und regulatorische Einstufung. CSOs liefern unabhängige Beratungsbeiträge zu Gerechtigkeitsrisiken, haben in dieser Phase jedoch keine Entscheidungsbefugnis.
2. **Gate 2 – Phase vor der Beschaffung und dem Vertragsabschluss.** Kliniken üben häufig während der Beschaffung ihren größten Einfluss auf die KI-Governance aus. In dieser Phase sollten Einrichtungen sicherstellen, dass sich Anbieter vertraglich verpflichten, Gerechtigkeitsrisiken – einschließlich genderspezifischer und rassistischer Bias – während des gesamten KI-Lebenszyklus zu berücksichtigen. Anbieter sollten verpflichtet werden, Unterlagen vorzulegen, die die Einhaltung regulatorischer Vorschriften belegen, einschließlich Nachweisen über Konformitätsbewertungen, sofern relevant,

³⁶ Koordinierungsgruppe für Medizinprodukte, [Leitlinien zur Qualifizierung und Klassifizierung von Software in der Verordnung \(EU\) 2017/745 \(MDR\) und der Verordnung \(EU\) 2017/746 \(IVDR\)](#), Brüssel: MDCG, 2019.

Ergebnisse der Leistungsvalidierung sowie Informationen, die für die Einsatzstellen erforderlich sind, um ihre Verpflichtungen gemäß dem „AI Act“ zu erfüllen. Beschaffungsverträge sollten die Verantwortlichkeiten zwischen Anbieter und Betreiber ausdrücklich zuweisen, einschließlich Anforderungen in Bezug auf Transparenz, Dokumentation, Protokollierung, menschliche Aufsicht und Unterstützung nach der Einführung. Das Ziel in dieser Phase besteht darin, das Bewusstsein für Bias und die Rechenschaftspflicht in vertragliche Verpflichtungen zu verankern und damit eine Grundlage für eine sichere und gerechte Einsatzplanung zu schaffen. Kliniken sind für Beschaffungsentscheidungen und vertragliche Vereinbarungen verantwortlich. CSOs können die Beschaffungsunterlagen im Hinblick auf Gleichstellungs- und Rechtsaspekte prüfen und unverbindliche Empfehlungen aussprechen.

3. **Gate 3 – Validierungsphase.** Die dritte Phase ist eine Validierungsphase, die sicherstellt, dass KI-Tools vor ihrer Einbindung in klinische Arbeitsabläufe unabhängig bewertet werden. Bei der Validierung sollten sowohl die Gesamtleistung als auch die Ergebnisse über disaggregierte Patientengruppen hinweg bewertet werden, wobei besonderes Augenmerk auf biologisches Geschlecht und Gender sowie ethnischen Hintergrund zu legen ist. Dieser Schritt ist entscheidend, da biomedizinische KI-Systeme unterschiedliche Leistungen bei verschiedenen Bevölkerungsgruppen aufweisen können, wenn die Trainingsdaten nicht repräsentativ sind oder wenn kontextuelle Unterschiede zwischen Entwicklungs- und Einsatzumgebungen übersehen werden.³⁷ Die Bewertung sollte zudem alle angewandten Risikominderungsmaßnahmen dokumentieren und verbleibende Risiken für die Chancengleichheit aufzeigen, um Kliniken die Grundlage für fundierte Entscheidungen zum Einsatz zu liefern. Kliniken sind dafür verantwortlich, die Validierung durchzuführen oder in Auftrag zu geben und die Ergebnisse für Entscheidungen zum Einsatz zu interpretieren. CSOs bieten eine unabhängige Überprüfung der aufgeschlüsselten Leistung und der Risiken durch Bias, ohne Einfluss auf die Validierungsergebnisse zu haben.
4. **Gate 4 – Einsatz.** Sobald ein System die Validierung bestanden hat, folgt in der vierten Phase eine Einsatzprüfung. Kliniken müssen sicherstellen, dass die Arbeitsbedingungen einen sicheren, gerechten und verantwortungsvollen Einsatz unterstützen. Dazu gehören die Einsetzung von geschultem Aufsichtspersonal, die Einführung von Mechanismen, die es den medizinischen Fachkräften erlauben, algorithmische Empfehlungen zu überschreiben oder zu deaktivieren, sowie die Sicherstellung, dass die Nutzenden den beabsichtigten Zweck des Systems, seine Grenzen und etwaige aufgeschlüsselte Leistungsaspekte verstehen. Das medizinische Fachpersonal sollte über

³⁷ Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G. & Chin, M. H., „*Ensuring Fairness in Machine Learning to Advance Health Equity*“, *Annals of Internal Medicine*, 169(12), 866–872, 2018.

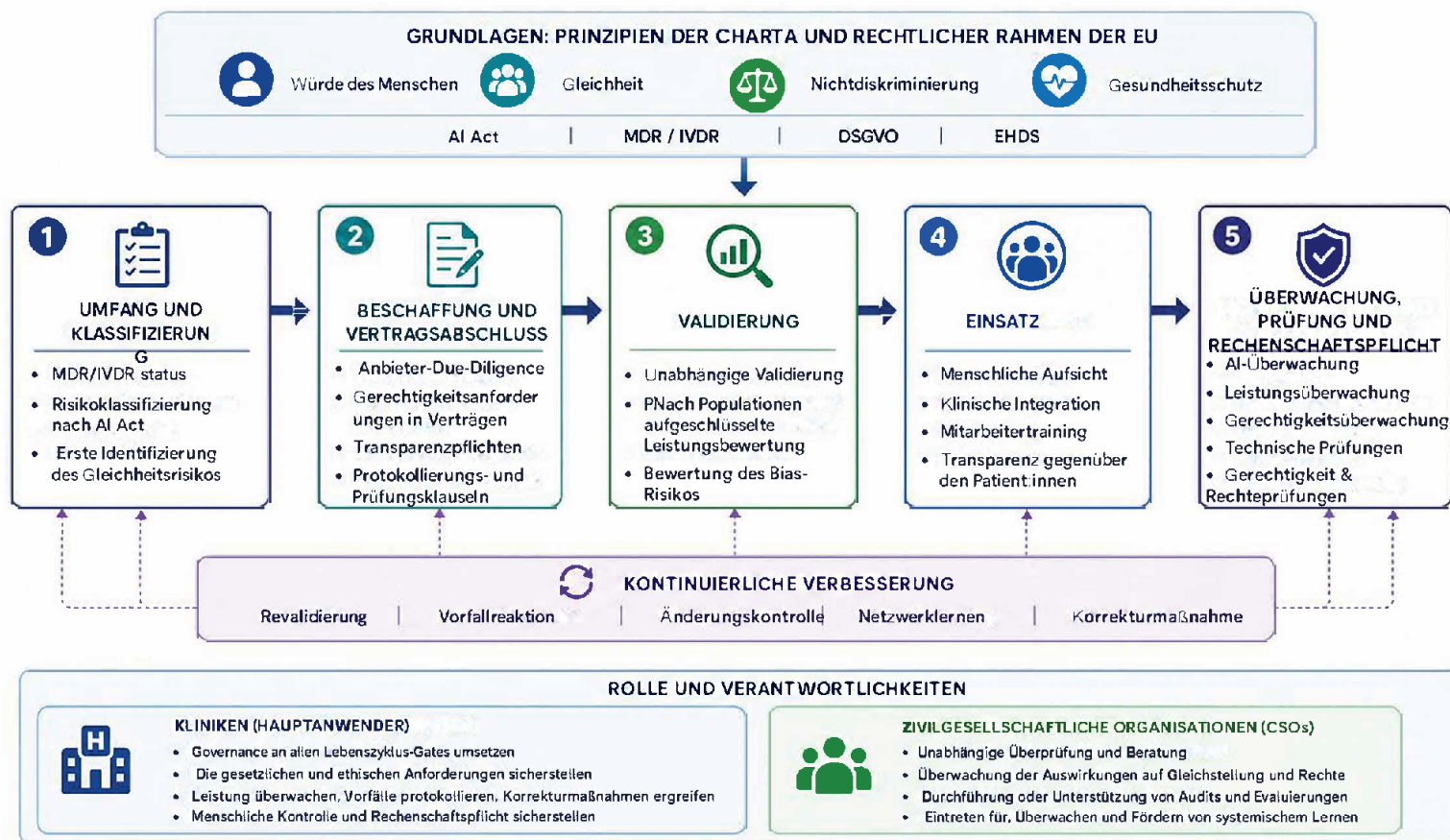
potenzielle ungerechtfertigte Abweichungen in den Ergebnissen informiert werden, wenn Unterschiede hinsichtlich des biologischen Geschlechts, des Gender oder der ethnischen Zugehörigkeit berücksichtigt werden, und Patienten sollten darauf hingewiesen werden, wenn KI-gestützte Tools bei ihrer Versorgung zum Einsatz kommen. Kliniken tragen die volle Verantwortung für Einführungsentscheidungen, die klinische Integration und die Überwachung der Umsetzung. CSOs spielen beim Einsatz keine Rolle, können jedoch die Dokumentation zur Einsatzbereitschaft unter dem Gesichtspunkt der Gerechtigkeit und der Rechte überprüfen.

5. **Gate 5 – Kontroll-, Prüf- und Rechenschaftsphase.** Dieser letzte Schritt regelt die Phase nach dem Einsatz im Lebenszyklus der KI. KI-Systeme sollten einer kontinuierlichen Überwachung hinsichtlich ihrer Gesamtleistung, etwaiger Ungleichheiten zwischen einzelnen Bevölkerungsgruppen sowie sicherheits- oder gleichstellungsrelevanter Vorfälle unterzogen werden. Kliniken sollten Protokolle über den Systembetrieb führen, standardisierte Verfahren zur Meldung von Vorfällen einführen, die potenziellen gender- oder rassenspezifischen Bias erfassen und die Systemleistung regelmäßig durch Revalidierungsmaßnahmen neu bewerten. Wenn Ungleichheiten oder Risiken von Bias auftreten, die nicht wirksam gemindert werden können, sollte das System modifiziert, ausgesetzt oder außer Betrieb genommen werden. Die Ergebnisse sollten sowohl in die interne Governance als auch – sofern angemessen – in einen breiteren Lernprozess auf Netzwerkebene einfließen, um wiederkehrende Bias zu verhindern. Kliniken sind für die fortlaufende Überwachung, Protokollierung, Meldung von Vorfällen und die Umsetzung von Abhilfemaßnahmen verantwortlich. Zivilgesellschaftliche Organisationen führen unabhängige Audits durch, bewerten systemische Gerechtigkeitsrisiken standortübergreifend, ohne jedoch operative Kontrolle über Systemänderungen zu haben.

Zusammen wandeln diese Lebenszyklus-Kontrollpunkte Grundrechte-Prinzipien in konkrete Governance-Prozesse um, mit denen Risiken von gender- und rassenspezifischen Biasen in klinischen KI-Systemen angegangen werden können.

Abbildung 1: An der AEQUITAS-Charta ausgerichteter KI-Governance-Pfad

AEQUITAS: MIT DER EU-CHARTA IM EINKLANG STEHENDER AI GOVERNANCE PATHWAY FÜR BIOMEDIZINISCHE KI IM ÖFFENTLICHEN GESUNDHEITSWESEN



Praktische Richtlinien, operative Tools und Governance-Hilfsmittel für Fachkräfte in Kliniken und zivilgesellschaftlichen Organisationen

Um in realen Gesundheitsumgebungen effektiv zu funktionieren, sieht das in diesem Policy-Dokument vorgeschlagene Regulierungsmodell die Schaffung praktischer Instrumente vor, die institutionsübergreifend wiederverwendet und an unterschiedliche nationale Kontexte angepasst werden können. Es betont die Notwendigkeit standardisierter Verfahren, Dokumentationen und Vorlagen, die es sowohl Kliniken als auch zivilgesellschaftlichen Organisationen ermöglichen, KI-Systeme einheitlich zu verwalten. Die folgenden Instrumente stehen im Einklang mit den Anforderungen des „AI Act“, den Leitlinien zur MDR und zur IVDR sowie sektorübergreifenden Governance-Praktiken.

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-LIT

Beschaffungs- und Vertragsphase

- **Paket für Beschaffung und Vertragsabschluss.** Die EU-Arbeitsgruppe für KI-Beschaffung hat Mustervertragsklauseln (Versionen für Hochrisiko- und Nicht-Hochrisiko-Anwendungen) veröffentlicht und diese an den „AI Act“ angepasst, wobei sie öffentliche Behörden ausdrücklich zur Verwendung dieser Klauseln ermutigt und Erläuterungen zur Umsetzung bereitstellt.³⁸ Für KI im Gesundheitswesen können diese Klauseln im Hinblick auf die Bewertung und Minderung von Risiken gender- und rassenspezifischer Bias in einen klinikspezifischen Anhang als eigenständiges Dokument angepasst werden, das (mindestens) Folgendes vorschreibt:
 - (1) eine standardisierte Berichterstattung über die Zusammensetzung der Trainingsdaten, einschließlich der Darstellung von Gender und ethnischer Zugehörigkeit, um die Bewertung der Repräsentativität und des Risikos von Bias zu unterstützen;
 - (2) Offenlegung der Datenherkunft, einschließlich der Angabe, ob es sich um proprietäre oder nicht-proprietäre Datensätze handelt, ihrer Zugangsbedingungen sowie der Auswirkungen auf Transparenz, unabhängige Validierung und die Bewertung von Gender- und Rassebias;
 - (3) Offenlegung der Leistungsdaten für aufgeschlüsselte Bevölkerungsgruppen sowie der Validierungsnachweise, einschließlich nach Gender und ethnischem Hintergrund stratifizierter Ergebnisse, der Dokumentation bekannter Ungleichheiten, der während der Entwicklung angewandten Maßnahmen zur Verringerung von Verzerrungen sowie der unterstützenden Validierungstests (einschließlich standortspezifischer Überlegungen hinsichtlich Gerechtigkeit und Fairness);
 - (4) Anforderungen an die Protokollierung und Überprüfbarkeit, um sicherzustellen, dass das System in der Lage ist, Nutzungsprotokolle zu erstellen, die unabhängige Prüfungen auf Bias oder ungleiche Auswirkungen unterstützen können;
 - (5) Verpflichtungen zur Aktualisierung und Änderungskontrolle, wonach der Anbieter verpflichtet ist, die Klinik über Modellaktualisierungen, Nachschulungen oder Änderungen der Arbeitsabläufe zu informieren und deren potenzielle Auswirkungen auf die Leistung in Bezug auf differenzierte Patientengruppen zu dokumentieren;
 - (6) Verpflichtungen zur Reaktion auf Vorfälle im Zusammenhang mit Bias, die den Anbieter verpflichten, festgestellte Fehler oder

³⁸ Europäische Kommission, [Aktualisierte EU-Mustervertragsklauseln für KI jetzt verfügbar](#), Brüssel: Europäische Kommission, 2025.

Ungleichheiten zu melden, die sich auf nach biologischem Geschlecht oder Gender differenzierte Patientengruppen auswirken, und Maßnahmen zur Behebung darzulegen;

- (7) eine klare Zuweisung von Verantwortlichkeiten, in der die Pflichten des Anbieters und des Einsatzbetreibers hinsichtlich der Einhaltung der Anforderungen des „AI Act“ sowie des Risikomanagements im Zusammenhang mit Gleichbehandlung und Bias festgelegt sind.

Validierungs- und Vorbereitungsphasen

- **Policy zur Governance von Gesundheitsdaten und zur Repräsentativität.** Um gender- und rassenspezifische Bias zu mindern, sollten alle Kliniken und zivilgesellschaftlichen Organisationen, die KI einsetzen, eine formelle Policy einführen, die sicherstellt, dass die für Trainings-, Feinabstimmungs- oder Implementierungszwecke verwendeten Daten die betreuten Bevölkerungsgruppen repräsentativ, vollständig und genau abbilden. Dazu gehören die Bewertung der demografischen Zusammensetzung, die Dokumentation unterrepräsentierter Patientengruppen und gegebenenfalls die Festlegung von Maßnahmen zur Risikominderung. Die Policy sollte sowohl interne als auch extern bezogene Datensätze abdecken, einschließlich proprietärer und frei zugänglicher Daten, und die Überprüfung der disaggregierten Leistungsdaten sowie aller von den Anbietern angewandten Strategien zur Verringerung von Bias vorschreiben. Eine solche Governance steht im Einklang mit dem „AI Act“ (Art. 10), der die Untersuchung und Minderung von Bias-Risiken vorschreibt, und spiegelt bewährte Verfahren aus der Literatur zur Verringerung von Bias in der Gesundheits-KI wider, die den gesamten Lebenszyklus von der Datenerhebung bis zur Überwachung nach der Einführung abdeckt.³⁹ Diese Policy ist kein Artefakt an sich, sondern eine Policyebene, die anderen Artefakten als Leitfaden dient. Sie fungiert als Prinzip vor der Bereitstellung und fließt direkt in die kontinuierliche Überwachung, Prüfung und Rechenschaftspflicht ein.
- **Grundlagen der Cybersicherheit und KI-spezifisches Bedrohungsmodell.** KI-Systeme in Kliniken können nicht nur anfällig für herkömmliche Angriffe wie Datenvergiftung, adversarische Beispiele oder Modellumgehung sein, sondern auch für Manipulationen oder Fehler, die Gender- oder Rassebias verschärfen, beispielsweise durch eine unverhältnismäßige Beeinträchtigung unterrepräsentierter Bevölkerungsgruppen in den Trainings- oder Implementierungsdaten. Der „AI Act“ (Art. 15) verlangt Widerstandsfähigkeit gegen unbefugte Veränderungen von Ergebnissen oder Leistung und

³⁹ Char, D.S., Shah, N.H., & Magnus, D., „Implementierung von maschinellem Lernen im Healthcare-Bereich – Bewältigung ethischer Herausforderungen“, *New England Journal of Medicine* 378: 981–983 (2018).

unterstreicht damit die Notwendigkeit KI-spezifischer Schutzmaßnahmen. Kliniken sollten daher eine Cybersicherheits-Grundlage und ein KI-spezifisches Bedrohungsmodell etablieren, das neben technischen Bedrohungen auch Biasrisiken ausdrücklich berücksichtigt. Dieses „lebende“ Dokument fungiert sowohl als Richtlinie als auch als praktisches Tool und dient als Checkliste vor der Inbetriebnahme, um die Umsetzung von Cybersicherheitskontrollen im Einklang mit der MDR und der IVDR, der organisatorischen Informationssicherheits-Governance (z. B. ISO/IEC 27001) sowie von vorurteilssensitiven Bedrohungsmodellen und Testplänen sicherzustellen.⁴⁰ Es unterstützt zudem die kontinuierliche Überwachung nach der Inbetriebnahme, um neu auftretende technische und Bias-bezogene Risiken zu erkennen und proaktiv darauf zu reagieren. Durch die Umsetzung dieser Richtlinie können Kliniken die Systemresilienz stärken, die Gerechtigkeit bei KI-gestützten Entscheidungsprozessen gewährleisten und die Einhaltung der Anforderungen des „AI Act“ (Art. 9–15) sowie der MDR/IVDR sicherstellen.

- **Ein „KI-Systemdossier“ pro Tool.** Bei diesem Dokument handelt es sich um ein standardisiertes Dokument, das für jedes KI-System vor dessen Einsatz in einer Gesundheitseinrichtung geführt wird. Das Dossier sollte den Verwendungszweck des Systems, die Integration in den klinischen Arbeitsablauf, Validierungsergebnisse und Leistungskennzahlen zusammenfassen, wobei besonderes Augenmerk auf die Leistung in verschiedenen Patientengruppen zu legen ist, die nach biologischem und sozialem Gender oder rassischer/ethnischer Herkunft differenziert werden, sowie auf bekannte Ungleichheiten oder Maßnahmen zu deren Minderung. Es sollte die Regelungen zur menschlichen Aufsicht dokumentieren, klarstellen, welche Entscheidungen ausschließlich unter der Kontrolle des klinischen Personals bleiben, und etwaige Kontraindikationen vermerken, die für die Patientensicherheit und -gerechtigkeit relevant sind. Durch die Bereitstellung eines formalisierten, auf Bias eingehenden Überblicks über das KI-System fördert das Dossier Transparenz, interne Governance und behördliche Aufsicht und stellt sicher, dass bei Entscheidungen vor der Inbetriebnahme Risiken ungleicher Ergebnisse ausdrücklich berücksichtigt werden.
- **Ein „kombiniertes GRFA- und DPIA-Paket“.** Kliniken, die risikoreiche KI-Systeme einsetzen, sind zunehmend verpflichtet, sowohl Risiken für Grundrechte als auch Risiken für den Schutz personenbezogener Daten zu bewerten. Die Integration dieser Bewertungen in eine einzige Vorlage, die den Bestimmungen des „AI Act“ (Art. 27), der DSGVO und der DPIA entspricht, kann den

⁴⁰ Koordinierungsgruppe für Medizinprodukte, *MDCG 2019-16 Rev. 1 – Leitlinien zur Cybersicherheit für Medizinprodukte*, Brüssel: Europäische Kommission, 2020.

Verwaltungsaufwand verringern und gleichzeitig sicherstellen, dass zentrale Fragen zu den betroffenen Bevölkerungsgruppen, potenziellen Schäden, Strategien zur Risikominderung und Beschwerdemechanismen systematisch behandelt werden, wobei besonderes Augenmerk auf potenzielle Ungleichheiten gelegt wird, die Patientengruppen betreffen, die nach biologischem Geschlecht und Gender sowie Rasse/Ethnizität differenziert werden.

Das „kombinierte GRFA- und DPIA-Paket“ dient in erster Linie als Ex-ante-Prüfungsinstrument, das regelmäßig aktualisiert wird und in der Phase vor der Einführung sowie im Rahmen der Gleichstellungs- und Rechteprüfung zum Einsatz kommt. Bei diesen Bewertungen handelt es sich um strukturierte, dokumentierte Einschätzungen von Risiken für Grundrechte und den Datenschutz, die eher als Kontrollpunkte der Governance denn als Tools zur kontinuierlichen Überwachung fungieren. Zwar sollten sie aktualisiert werden, wenn sich Risiken oder Systembedingungen ändern, doch besteht ihre Kernaufgabe darin, vor und während der Einführung eine systematische und überprüfbare Bewertung potenzieller Auswirkungen zu liefern, die Einhaltung sowohl des „AI Act“ als auch der DSGVO zu unterstützen und dabei zu helfen, genderspezifische oder ethnisch bedingte Bias zu identifizieren und zu mindern.

Um den Menschenrechtsaspekt zu stärken, sollte bei der Konzeption der Folgenabschätzung eine sinnvolle Konsultation mit zivilgesellschaftlichen Organisationen vorgesehen werden, um deren Fachwissen im Bereich der Menschenrechte einzubeziehen. So fordert beispielsweise das Europäische Netzwerk nationaler Menschenrechtsinstitutionen (European Network of National Human Rights Institutions, ENNHRI) die Europäische Kommission nachdrücklich auf, die Vorlagen durch Leitlinien und öffentliche Konsultationen zu ergänzen, um wirksame GRFA zu gewährleisten, die die Gleichstellung in Bezug auf Gender und rassische/ethnische Dimensionen berücksichtigen.⁴¹

- **Eine „Policy zur Veränderungskontrolle“.** Diese Policy regelt Aktualisierungen von KI-Systemen, einschließlich des erneuten Trainings von Modellen, Software-Upgrades oder Änderungen an klinischen Arbeitsabläufen. Organisationen sollten klar definieren, was eine „wesentliche Veränderung“ darstellt, die eine erneute Validierung oder Genehmigung erfordert, um sicherzustellen, dass Aktualisierungen keine unvorhergesehenen Risiken für die Patientenversorgung mit sich bringen. Insbesondere müssen Veränderungen

⁴¹ Europäisches Netzwerk nationaler Menschenrechtsinstitutionen (ENNHRI), *Stellungnahme zur Gewährleistung wirksamer Grundrechtsfolgenabschätzungen (GRFAs) im Rahmen des „AI Act“*, Brüssel: ENNHRI, 2025.

auf potenzielle Auswirkungen auf die Gender- und ethnische Gleichstellung hin bewertet werden, einschließlich Verschiebungen bei der nach demografischen Merkmalen aufgeschlüsselten Leistung oder unterschiedlicher Ergebnisse bei verschiedenen Bevölkerungsgruppen. Kliniken sollten zudem sicherstellen, dass das medizinische Personal und die betroffenen Mitarbeiter über Veränderungen informiert werden, die sich auf die Behandlungsergebnisse verschiedener Patientengruppen auswirken könnten, um so eine angemessene Überwachung und die Ergreifung geeigneter Maßnahmen zur Risikominderung zu ermöglichen. Durch die Einbettung von Gleichstellungsaspekten in das Veränderungsmanagement gewährleisten Kliniken sowohl die Sicherheit als auch die Fairness der KI-gestützten Entscheidungsfindung in der klinischen Praxis.

- **Ein „Qualitätsmanagementsystem“ (QMS) für Betreiber, die gleichzeitig Anbieter sind.** Eine weitere wichtige Dimension des Modells betrifft KI, die intern von Kliniken oder Forschungseinrichtungen entwickelt wird. Öffentliche Gesundheitsorganisationen entwickeln oder passen zunehmend ihre eigenen KI-Tools an, insbesondere in akademischen medizinischen Zentren. In solchen Fällen kann die Einrichtung faktisch als Hersteller oder Anbieter fungieren, was zusätzliche regulatorische Verantwortlichkeiten mit sich bringt. Europäische Leitlinien zu intern entwickelten Medizinprodukten stellen klar, dass Gesundheitseinrichtungen Qualitätsmanagementsysteme betreiben und technische Dokumentationen führen sollten, aus denen hervorgeht, dass intern entwickelte Tools spezifische Patientenbedürfnisse erfüllen, die durch kommerzielle Produkte nicht abgedeckt werden können.⁴² Das QMS sollte sicherstellen, dass KI-Ergebnisse hinsichtlich unterschiedlicher Leistungsmerkmale bei verschiedenen Patientengruppen – differenziert nach biologischem und sozialem Gender sowie rassischem/ethnischem Hintergrund – bewertet werden, dass Risikominderungsstrategien dokumentiert werden und dass Vorkehrungen zur menschlichen Aufsicht die Patientensicherheit und Fairness ausdrücklich gewährleisten. Selbst dort, wo regulatorische Ausnahmeregelungen für selbst entwickelte Geräte gelten, bleiben die Verpflichtungen im Rahmen der Grundrechte und umfassenderer KI-Governance-Rahmenwerke uneingeschränkt relevant, wodurch das QMS zu einem entscheidenden Tool wird, um Gerechtigkeit in die Konzeption, Validierung und den Einsatz biomedizinischer KI-Systeme zu integrieren.

Phase der Überwachung, Prüfung und Rechenschaftspflicht

⁴² Koordinierungsgruppe für Medizinprodukte, *MDCG 2023–1 – Leitlinien zur Ausnahmeregelung für Gesundheitseinrichtungen (MDR/IVDR)*, Brüssel: Europäische Kommission, 2023.

- **Ein „nach Bevölkerungsgruppen aufgeschlüsselter Leistungs- und Gerechtigkeitsbericht“.** Da Schwankungen in der algorithmischen Leistung zu ungleichen Gesundheitsergebnissen beitragen können, sollten Kliniken nicht nur von Anbietern verlangen, vor der Beschaffung relevante Informationen bereitzustellen, sondern auch interne Analysen erstellen, aus denen hervorgeht, wie sich die Systeme bei Patienten mit unterschiedlichen demografischen Merkmalen verhalten. Diese Analysen sollen den Fokus auf die nach biologischem Geschlecht und Gender sowie – soweit rechtmäßig und durchführbar – nach relevanten rassischen oder ethnischen Unterschieden stratifizierte Leistung legen und eine Begründung für etwaige unvermeidbare Leistungsunterschiede enthalten. Wo Leistungslücken bestehen, sollten Organisationen Strategien zur Risikominderung dokumentieren und die Wirksamkeit dieser Maßnahmen im Zeitverlauf überwachen. Dieser Ansatz spiegelt die sich abzeichnenden Erwartungen im Rahmen der europäischen KI-Regulierung wider, dass Organisationen Risiken des Bias in Datensätzen und Modellergebnissen aktiv bewerten und mindern. Der „AI Act“ (Art. 10) verlangt ausdrücklich die Untersuchung und Minderung von Risiken des Bias in Datensätzen und erlaubt unter strengen Schutzvorkehrungen eine begrenzte Verarbeitung von Daten besonderer Kategorien zur Erkennung von Voreingenommenheit.

Der „nach Bevölkerungsgruppen aufgeschlüsselte Leistungs- und Gerechtigkeitsbericht“ dient in erster Linie als Tool zur kontinuierlichen Überwachung innerhalb des Governance-Rahmens. Er erfordert eine fortlaufende Datenerhebung, eine regelmäßige Nachverfolgung der Systemleistung über relevante demografische Gruppen hinweg sowie eine kontinuierliche Aktualisierung von Abhilfemaßnahmen, sobald unterschiedliche Ergebnisse festgestellt werden. Dies ermöglicht es den Gesundheitsorganisationen, Risiken von Bias im Laufe der Zeit zu erkennen und zu beheben. Gleichzeitig sollten die wichtigsten Ergebnisse dieses Berichts, wie beispielsweise festgestellte Ungleichheiten und die Wirksamkeit von Abhilfemaßnahmen, zusammengefasst und in das „Gerechtigkeits- und Rechte-Audit“ einbezogen werden, um Transparenz und öffentliche Rechenschaftspflicht zu gewährleisten.

- **Ein zweistufiges Audit-System, bestehend aus einem technischen Audit und einem Gleichstellungs- und Rechte-Audit.**

Ein praktikables Audit-System für Kliniken sollte ein Gleichgewicht zwischen Strenge und Praktikabilität herstellen und sowohl für das Personal als auch für die Öffentlichkeit verständlich sein. In diesem Zusammenhang ist ein Audit eine formalisierte, regelmäßige Überprüfung der Konformität, Leistung und

Auswirkungen eines KI-Systems auf Rechte, die darauf ausgelegt ist, die laufende Überwachung zu ergänzen, anstatt sie zu ersetzen.

Technische Prüfung: Diese Ebene umfasst eine vertrauliche operative Überprüfung und beinhaltet detaillierte, intern verbleibende Kontrollen, darunter Sicherheitstests, Robustheitsbewertungen, die Validierung der klinischen Leistung und die Einhaltung der Daten-Governance-Compliance. Besondere Aufmerksamkeit sollte der Leistung über biologische Geschlechts- und Gender- sowie rassische oder ethnische Unterschiede hinweg gewidmet werden, wobei zu bewerten ist, ob die Modellergebnisse bestimmte Bevölkerungsgruppen systematisch benachteiligen. Diese Aktivitäten sollten mit den einschlägigen Bestimmungen des „AI Act“ (Art. 9–15), der MDR und der IVDR, soweit anwendbar, im Einklang stehen.

Prüfung von Gerechtigkeit und Rechten: Diese zweite Ebene hat einen Fokus auf Transparenz, Rechenschaftspflicht und die Minderung von Bias. Sie sollte Ungleichheiten bei den Ergebnissen, die in die Tests einbezogenen Bevölkerungsgruppen, bekannte Einschränkungen und Beschwerdemechanismen dokumentieren, einschließlich einer expliziten Analyse der Auswirkungen auf Gender und ethnische Zugehörigkeit. Es sollten auch Belege dafür erfasst werden, dass die menschliche Überwachung sinnvoll ist und voreingenommene Ergebnisse korrigieren kann. In Anlehnung an den in der arbeitsrechtlichen Regulierung verwendeten Ansatz der „Bias-Prüfung und Offenlegung“⁴³ kann eine für die breite Öffentlichkeit bestimmte Zusammenfassung wichtige Ergebnisse, Strategien zur Risikominderung und umgesetzte Verbesserungen vermitteln, ohne geschützte oder vertrauliche Informationen preiszugeben. Das „kombinierte GRFA- und DPIA-Paket“ ist ebenfalls Teil der Ebene der Gleichstellungs- und Rechteprüfung und verstärkt die strenge Bewertung von Grundrechten, Datenschutz und demografischer Gerechtigkeit.

- **Ein „KI-Überwachungsprotokoll“.** Die Überwachung nach der Einführung wird zudem durch die Erstellung eines KI-Überwachungsprotokolls unterstützt, das sich an den in der Arzneimittelregulierung verwendeten Pharmakovigilanz-Systemen orientiert.⁴⁴ Fachkräfte im Gesundheitswesen sollten in der Lage sein, Vorfälle, Beinahevorfälle oder unerwartete Ergebnisse im Zusammenhang mit KI-gestützten Entscheidungen zu melden, wobei

⁴³ New Yorker Amt für Verbraucher- und Arbeitnehmerschutz, *Automatisierte Tools zur Beschäftigungsentscheidung (AEDT)*, New York: Stadt New York, 2025.

⁴⁴ Europäische Arzneimittelagentur, *GVP-Modul VI – Erfassung und Meldung vermuteter unerwünschter Wirkungen (Rev. 2)*, London: EMA, 2017.

explizite Felder zur Erfassung potenzieller beobachteter Ungleichheiten hinsichtlich biologischen Geschlechts und Gender oder ethnischer Zugehörigkeit vorgesehen sind. Die Berichte sollten systematisch auf sich abzeichnende Muster analysiert werden, einschließlich unterschiedlicher Auswirkungen auf bestimmte demografische Gruppen, damit Organisationen Sicherheitssignale, Bedenken hinsichtlich der Gerechtigkeit oder Tendenzen zu Verzerrungen erkennen können. Falls erforderlich, sollten Vorfälle, insbesondere solche, die auf eine ungerechte Behandlung oder ein Risiko für geschützte Gruppen hindeuten, unverzüglich an Systemanbieter, zuständige Klinik-Leitungsgremien und Aufsichtsbehörden weitergeleitet werden. Das Protokoll sollte zudem in regelmäßige Gleichstellungs- und Rechteauidits einfließen und so die kontinuierliche Verbesserung sowie den Abbau von Bias im Einsatz klinischer KI unterstützen.

- **Zusammenarbeit auf Netzwerkebene und Austausch von Vorfällen.** Das im Rahmen von AEQUITAS entwickelte europäische Netzwerk aus zivilgesellschaftlichen Organisationen und öffentlichen Kliniken fungiert als gemeinsame Governance-Ebene und stellt ein einheitliches Vokabular sowie Berichtsstandards für Vorfälle und Befunde im Zusammenhang mit gender- und rassenspezifischem Bias in KI-Systemen bereit. Dieses Netzwerk unterstützt standortübergreifendes Lernen und die Verbreitung bewährter Verfahren zur Erkennung, Minderung und Beseitigung von Ungleichheiten, ähnlich wie in Sektoren mit hohen Sicherheitsanforderungen und in der Pharmakovigilanz. Durch die Verknüpfung von Daten zu Vorfällen mit Bias, aufgeschlüsselten Leistungsdaten und Strategien zur Risikominderung ermöglicht das Netzwerk Kliniken und zivilgesellschaftlichen Organisationen, systemische Muster von Ungleichheit zu identifizieren und gezielte Maßnahmen zu koordinieren. Da das EHDS sowohl die Primärnutzung (Versorgungsleistung) als auch die Sekundärnutzung (Forschung und Policy) unter einem regulierten Rahmen erleichtert, kann das Netzwerk seine Praktiken des Datenaustauschs und des Lernens an die Zeitpläne und Durchführungsrechtsakte des EHDS anpassen, sobald diese ausgereift sind, und so einen gerechteren und transparenteren Einsatz von KI über alle Einrichtungen hinweg unterstützen.
- **Schulungsmodul für klinisches Personal und Experten von zivilgesellschaftlichen Organisationen.** Um Automatisierungsbias und Ungleichheiten in der klinischen KI zu mindern, umfasst das AEQUITAS-Regulierungsmodell obligatorische Trainings, die darauf abzielen, Fachkräfte im Gesundheitswesen mit den Fähigkeiten auszustatten, die Grenzen von KI-Systemen zu verstehen, Ergebnisse kritisch zu interpretieren und Systeme zu übersteuern oder anzuhalten, wenn die Ergebnisse zu verzerrten oder ungerechten Ergebnissen führen könnten. Das Training legt den Schwerpunkt auf das Bewusstsein für Risiken gender- und rassenspezifischer

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-LIT

Bias, Grundrechte-Prinzipien und ethische Überlegungen, einschließlich Gerechtigkeit, Nichtdiskriminierung und Patientenbeteiligung.

Dieses Modul kann als unterstützendes Instrument betrachtet werden, das die Anwendung des Regulierungsmodells aktiv fördert, indem es sicherstellt, dass das für die KI-Aufsicht zuständige Personal Überwachungs-, Validierungs- und Rechenschaftsprozesse unter Berücksichtigung von Gerechtigkeitsaspekten kompetent umsetzen kann. Das Training soll fallbasierte Übungen zur Erkennung und Minderung von genderspezifischen sowie ethnisch-rassischen Ungleichheiten umfassen.

Fazit

KI bietet erhebliche Chancen zur Verbesserung der Gesundheitsergebnisse, doch ohne sorgfältige Governance kann sie Ungleichheiten in Bezug auf biologisches Geschlecht und Gender sowie ethnische Zugehörigkeit ebenso wie andere Formen systemischer Bias reproduzieren oder verschärfen. Das in diesem Policy-Dokument vorgestellte Regulierungsmodell setzt Grundrechte-Prinzipien in konkrete Schutzmaßnahmen über den gesamten KI-Lebenszyklus hinweg um – von der Beschaffung und Validierung bis hin zum Einsatz, zur Überwachung und zur Rechenschaftspflicht –, wobei ein ausdrücklicher Schwerpunkt auf der Erkennung und Minderung von Ergebnisunterschieden zwischen verschiedenen Patientengruppen liegt. Durch die Umsetzung einer KI-Governance-Architektur, die auf einem klaren Satz von Lebenszyklus-Meilensteinen, Richtlinien, praktischen Tools und Artefakten basiert, können Kliniken und zivilgesellschaftliche Organisationen Risiken systematisch bewerten, mindern und dokumentieren, einschließlich ungleicher Auswirkungen auf marginalisierte Patientengruppen. Durch die Einbeziehung sektorübergreifender Erkenntnisse, EU-Rechtsvorschriften und internationaler Governance-Rahmenwerke unterstützt das Regulierungsmodell Fachkräfte im Gesundheitswesen dabei, KI-Systeme verantwortungsvoll einzusetzen und eine gerechte, sichere und transparente Gesundheitsversorgung für alle Patienten zu gewährleisten, unabhängig von Gender, ethnischer Herkunft oder anderen Merkmalen.