Ethik in den Biowissenschaften – Sachstandsberichte des DRZE | 29
Herausgegeben von Dirk Lanzerath | Roman Wagner

Wagner | Caspers | Eickhoff | Langner
Steinborn | Dokmanac | Günther

KI in der radiologischen Diagnostik

Medizinische, psychologische, rechtliche und ethische Aspekte

drze VERLAG KARL ALBER

Ethik in den Biowissenschaften – Sachstandsberichte des DRZE

Im Auftrag des
Deutschen Referenzzentrums für Ethik
in den Biowissenschaften

Herausgegeben von
Dirk Lanzerath und Roman Wagner

www.drze.de

Band 29

Roman Wagner I Julian Caspers I Simon Eickhoff
Robert Langner I Michael Steinborn
Sara Dokmanac I Joana Günther

KI in der radiologischen Diagnostik

Medizinische, psychologische, rechtliche und ethische Aspekte

drze
Deutsches Referenzzentrum für Ethik in den Biowissenschaften

VERLAG KARL ALBER

Diese Publikation wird als Vorhaben der Nordrhein-Westfälischen Akademie der Wissenschaften und der Künste im Rahmen des Akademienprogramms von der Bundesrepublik Deutschland und dem Land Nordrhein-Westfalen gefördert.

Redaktion: Roman Wagner

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

1. Auflage 2026

Publiziert von
Verlag Karl Alber – ein Verlag in der
Nomos Verlagsgesellschaft mbH & Co. KG
Waldseestraße 3–5 | 76530 Baden-Baden
www.**verlag-alber.de**

Gesamtherstellung:
Nomos Verlagsgesellschaft mbH & Co. KG
Waldseestraße 3–5 | 76530 Baden-Baden

ISBN (Print): 978-3-495-98824-4
ISBN (ePDF): 978-3-495-98825-1

DOI: https://doi.org/10.5771/9783495988251

Onlineversion
Inlibra

Inhalt

Vorwort 9

I. Künstliche Intelligenz in der radiologischen Diagnostik: Medizinische Aspekte 13

1. Einführung und begriffliche Einordnung 13
2. Historische Einordnung 16
3. Aktuelle Entwicklungen und Herausforderungen 20
4. Perspektiven für KI in der Radiologie 24

Literaturverzeichnis 26

II. Psychologische Aspekte der Nutzung Künstlicher Intelligenz (KI) in der radiologischen Diagnostik . . . 33

1. Kognitive Konsequenzen KI-basierter Automatisierung . . 35
   - 1.1 Theoretischer Rahmen 35
   - 1.2 Kurzfristige kognitive Auswirkungen 36
   - 1.3 Langfristige kognitive Auswirkungen 38
   - 1.4 Maßnahmen gegen negative Auswirkungen der Automatisierung 40
2. Motivationale und emotionale Aspekte KI-basierter Automatisierung 44
   - 2.1 Identität, Autonomie und Bedeutung der Arbeit 44
   - 2.2 Emotionale Reaktionen: Unsicherheit, Sorgen und Erleichterung 47
   - 2.3 Angst vor Fehlern und moralischer Distress 50
3. Verantwortung und Erklärbarkeit 52
   - 3.1 Verantwortung und Rechenschaft bei hybriden Entscheidungsprozessen 52
   - 3.2 Von der Erklärbarkeit zur Interpretierbarkeit 54

4. Vertrauen in KI-basierte Technologien in der Radiologie . 56
4.1 Vertrauen und Zuverlässigkeit 56
4.2 (Fehl-)Kalibrierung des Vertrauens: Folgen, Dynamik und Modifikation 58
5. Patientenperspektiven zum Einsatz von KI in der Radiologie 61
5.1 Erwartungen und wahrgenommene Risiken 61
5.2 KI-Nutzung und Vertrauen in asymmetrischen Arzt-Patient-Beziehungen 63
6. Technologieakzeptanz und organisatorische Bedingungen nachhaltiger KI-Nutzung in der Radiologie 65
6.1 Einstellungen und Akzeptanz gegenüber KI-basierten Technologien 65
6.2 Organisationale Bedingungen einer nachhaltigen KI-Adoption in der Radiologie 68
Literaturverzeichnis 71

III. Künstliche Intelligenz in der radiologischen Diagnostik: Rechtliche Aspekte 77

1. Allgemeine regulatorische Einbettung 77
2. KI-Verordnung 78
2.1 Vorab: Forschungsprivileg 78
2.2 Ziele der Verordnung 80
2.3 Risikobasierter Ansatz 81
2.3.1 Grundkonzept 81
2.3.2 Anforderungen an Hochrisiko-KI-Systeme . . . 83
2.4 Beteiligte und Verantwortliche 84
2.4.1 Pflichten gemäß KI-VO 88
a) Pflichten der Anbieter 88
b) Pflichten der Betreiber 90
c) Einbeziehung in den Pflichtenkreis des Anbieters 92
2.4.2 Exkurs: Pflichten gemäß MP-VO 93
a) Pflichten der Hersteller 93
b) Pflichten der Betreiber 96
2.5 Sanktionsmechanismen 98

2.6 Kritik und Herausforderungen 100
3. Haftungsrichtlinie und Produkthaftungsrecht 103
3.1 Ursprünglicher Regulierungsansatz 104
3.2 Status quo 106
4. European Health Data Space 106
5. Schlussbetrachtung und Ausblick 109
Literaturverzeichnis 111

IV. Künstliche Intelligenz in der radiologischen Diagnostik: Ethische Aspekte 113

1. Problemstellung und technische Hintergründe 113
1.1 Zwei Formen von KI 114
1.1.1 Das symbolische Paradigma und klassische KI 115
1.1.2 Das konnektionistische Paradigma und maschinelles Lernen 116
2. Ethische Analyse 121
2.1 Benefizienz 122
2.2 Gerechtigkeit 124
2.3 Vertrauen 129
2.3.1 Begriffliche Grundlagen des Vertrauensbegriffs 129
2.3.2 Ethische Analyse: Vertrauen und Verlässlichkeit 134
2.4 Transparenz 139
2.5 Verantwortung 141
2.6 Achtung der Autonomie 144
2.6.1 Konzeptionelle Grundlagen des Autonomiebegriffs 144
2.6.2 Autonomie der Ärzt*innen 147
2.6.3 Autonomie der Patient*innen und informierte Einwilligung 152
Literaturverzeichnis 156

Vorwort

Künstliche Intelligenz (KI) hat sich in den vergangenen Jahren zu einem der dynamischsten Innovationsfelder der Medizin entwickelt. Insbesondere in der radiologischen Diagnostik, die in hohem Maße auf die Verarbeitung komplexer Bilddaten angewiesen ist, werden KI-basierte Verfahren als vielversprechende Werkzeuge zur Unterstützung ärztlicher Arbeit diskutiert. Anwendungen reichen von der Optimierung bildgebender Verfahren über die Detektion und Klassifikation von Auffälligkeiten bis hin zur Unterstützung bei Triage- und Priorisierungsentscheidungen. Zugleich wirft der Einsatz solcher Systeme grundlegende medizinische, ethische und rechtliche Fragen auf, die weit über technische Leistungsparameter hinausgehen.

Medizinisch verbinden sich mit KI in der Radiologie Erwartungen an eine höhere Effizienz, eine verbesserte Bildqualität und potenziell auch an eine gesteigerte diagnostische Präzision. Gleichzeitig zeigt sich jedoch, dass die tatsächliche Integration KI-basierter Systeme in den klinischen Alltag bislang hinter diesen Erwartungen zurückbleibt. Fragen der Evidenz, der Interoperabilität mit bestehenden Workflows, der Kosten-Nutzen-Relation sowie der ärztlichen Verantwortlichkeit sind bislang nur teilweise geklärt. KI tritt damit weniger als autonome Entscheidungsträgerin in Erscheinung, denn als komplexes Werkzeug, dessen Nutzen wesentlich von seiner Einbettung in medizinische Praxis, Organisation und Verantwortung abhängt.

Gerade diese Einbettung macht den Einsatz von KI in der radiologischen Diagnostik zu einem genuin ethisch-rechtlichen Thema. Ethisch stellen sich Fragen nach Verantwortung und Haftung, nach Transparenz und Erklärbarkeit algorithmischer Entscheidungen, nach dem angemessenen Maß an Vertrauen in technische Systeme sowie nach möglichen Veränderungen ärztlicher Kompetenzen und Rollenbilder. Hinzu kommen psychologische Effekte der Automatisierung, etwa das Risiko eines übermäßigen Vertrauens in

KI-Vorschläge oder langfristige Veränderungen ärztlicher Expertise. Rechtlich wiederum ist der Einsatz medizinischer KI eingebettet in einen sich rasch entwickelnden Regulierungsrahmen, der den Schutz von Patient*innen, die Sicherheit medizinischer Produkte und das Vertrauen in neue Technologien gewährleisten soll, dabei jedoch weiterhin mit Unsicherheiten und Auslegungsfragen konfrontiert ist.

Vor diesem Hintergrund verfolgt dieser im Rahmen der interdisziplinären Arbeit des durch das Bundesministerium für Forschung, Technologie und Raumfahrt (BMFTR) geförderten (FKZ 01GP2113) Forschungsprojekts »FRAIM« entstandene Sachstandsbericht das Ziel, den Einsatz Künstlicher Intelligenz in der radiologischen Diagnostik interdisziplinär zu beleuchten und eine differenzierte Grundlage für die weitere fachliche, ethische und rechtspolitische Diskussion bereitzustellen. Die Beiträge verbinden medizinische, psychologische, ethische und rechtliche Perspektiven und machen deutlich, dass die Bewertung von KI in der Radiologie weder technikoptimistisch noch technikskeptisch erfolgen sollte, sondern einer sorgfältigen Analyse von Chancen, Grenzen und Voraussetzungen verantwortungsvollen Handelns bedarf.

Den medizinischen Ausgangspunkt bildet der Beitrag von **Julian Caspers und Simon Eickhoff**, die den aktuellen Stellenwert KI-basierter Verfahren in der Radiologie einordnen. Sie erläutern die technischen Grundlagen neuronaler Netzwerke, ordnen deren Einsatz historisch ein und zeigen, dass KI in der radiologischen Praxis bislang vor allem unterstützende Hintergrundprozesse übernimmt, etwa bei der Bildrekonstruktion oder -optimierung. Zugleich verweisen sie auf bestehende Hürden der klinischen Implementierung und betonen die Notwendigkeit einer belastbaren Evidenzbasis, um den tatsächlichen Nutzen dieser Systeme für die Versorgung der Patient*innen valide beurteilen zu können.

An diese medizinische Perspektive schließen **Robert Langner und Michael B. Steinborn** mit einer Analyse der psychologischen Aspekte des KI-Einsatzes in der Radiologie an. Sie untersuchen kognitive Automatisierungseffekte bei Radiolog*innen, beleuchten Risiken wie den sogenannten »automation bias« sowie mögliche langfristige Effekte des »deskilling« ebenso wie Potenziale eines »upskilling« hinsichtlich der ärztlichen Kompetenzen. Darüber hinaus nehmen sie die Perspektive der Patient*innen in den Blick

und zeigen, welche Erwartungen, Hoffnungen und Sorgen mit dem Einsatz von KI verbunden sind. Ihr Beitrag macht deutlich, dass Akzeptanz und verantwortungsvoller Einsatz von KI wesentlich von einer angemessenen Kalibrierung von Vertrauen, Transparenz und Kommunikation abhängen.

Die rechtlichen Rahmenbedingungen werden im Beitrag von **Joana Günther und Sara Dokmanac** analysiert, die die europäische KI-Verordnung (KI-VO) in ihren Grundzügen darstellen und auf ihre Bedeutung für medizinische KI-Systeme eingehen. Sie erläutern den risikobasierten Ansatz der Verordnung, die unterschiedlichen Pflichten für Anbieter*innen, Betreiber*innen und Hersteller*innen sowie die besonderen Regelungen für Forschung. Zugleich zeigen sie auf, dass trotz der neuen Regulierung weiterhin rechtliche Unsicherheiten bestehen, insbesondere aufgrund offener Begriffe und bislang ungeklärter Abgrenzungsfragen.

Eine systematische ethische Einordnung des Einsatzes von KI in der Radiologie liefert schließlich der Beitrag von **Roman Wagner**. Ausgehend von den technischen Besonderheiten selbstlernender, konnektionistischer KI-Systeme identifiziert er zentrale ethisch relevante Merkmale wie den Black-Box-Charakter, die Abhängigkeit von Trainingsdaten und die Gefahr seltsamer Fehler. In seiner prinzipienethischen Analyse arbeitet Wagner heraus, dass KI-Systeme als Werkzeuge zu verstehen sind, deren Einsatz stets in der Verantwortung menschlicher Akteur*innen verbleibt. Fragen nach Verantwortung, Vertrauen und Autonomie erweisen sich damit nicht als grundsätzlich neu, gewinnen jedoch unter den Bedingungen algorithmisch vermittelter Diagnostik eine neue Dringlichkeit.

Der Sachstandsbericht macht insgesamt deutlich, dass Künstliche Intelligenz in der radiologischen Diagnostik weder als normativ neutrales technisches Instrument noch als autonome Akteurin begriffen werden kann. Ihr verantwortungsvoller Einsatz setzt vielmehr eine enge Verzahnung medizinischer Expertise, ethischer Reflexion und rechtlicher Rahmensetzung voraus. In diesem Sinne versteht sich der vorliegende Band als Beitrag zu einer sachlichen, interdisziplinären und zukunftsorientierten Debatte über die Rolle von KI in der Medizin – mit dem Ziel, Orientierung zu bieten, ohne die Komplexität der damit verbundenen Fragen zu verkürzen.

Dirk Lanzerath und Roman Wagner

Julian Caspers, Simon Eickhoff

I. Künstliche Intelligenz in der radiologischen Diagnostik: Medizinische Aspekte

1. Einführung und begriffliche Einordnung

Der Begriff »künstliche Intelligenz« (KI) beschreibt datenverarbeitende Systeme, die in der Lage sind, aus großen und komplexen Datenbeständen zu lernen, Muster zu erkennen und Entscheidungen basierend auf diesen Informationen abzuleiten. Hierdurch wird intelligentes Verhalten nachgebildet, wobei die hierfür verwendeten Ansätze historisch vielfältig waren. Auch wenn das Feld der künstlichen Intelligenz weitaus mehr umfasst, werden in Hinsicht auf die heutige und absehbar zukünftige Anwendung in der Radiologie mit dem Begriff künstliche Intelligenz vor allem Methoden des maschinellen Lernens, insbesondere des tiefen maschinellen Lernens mit neuronalen Netzwerken (*deep learning*), in Verbindung gebracht (Moawad, 2022). Letztere zeigen beeindruckende Leistungen insbesondere in der Bildanalyse, aber auch in der Sprachverarbeitung. Eine weitere, momentan viel Beachtung findende Variante der KI sind generative Verfahren, welche nach einer Aufforderung selbstständig Texte oder Bilder basierend auf gelernten Assoziationen zwischen Begriffen erstellen (Koohi-Moghadam, 2023). Gemein ist diesen KI-Verfahren, dass sie Muster in hochdimensionalen Daten gelernt haben und basierend auf diesem »Wissen« (welches in den entsprechenden Modellparametern verankert, jedoch nicht explizit abrufbar ist oder erklärt werden kann) Entscheidungen auf Grundlage neuer Daten treffen können. Das heißt, die zentrale Fähigkeit solcher KI-Modelle ist ein selbstständiger Lernprozess auf Basis vorhandener Trainingsdaten, wobei die hierdurch gelernten Regeln auch auf neue Situationen, Fälle oder Konstellationen generalisieren (Chen, 2023). Dementsprechend wird eine Reihe von Verfahren, auch wenn sie sich für Nutzer*innen oft ähnlich präsentieren, nicht

dem maschinellen Lernen zugerechnet und daher nicht unter dem Begriff KI im engeren Sinne verstanden. Dies gilt für einfache Automatisierungen, regelbasierte Entscheidungsbäume oder deterministische Maschinen, die lediglich vorprogrammierte Abläufe abarbeiten, ohne selbstständig dazuzulernen. Beispiele hierfür wären unter anderem Segmentierung von Bilddaten in der Radiologie anhand von Schwellwerten oder, vielleicht am eindrücklichsten, klassische Ansätze zur Routenplanung per Navigationssystem, welche über automatisierte (Optimierungs-)Prozesse verfügen, jedoch nicht über KI-Qualitäten im Sinne eines Lernens aus vorhandenen Daten mit Generalisierung auf neue Situationen.

Während die öffentliche Diskussion zu »KI in der Medizin« oft von der Idee geprägt ist, dass ein Computer mehr oder weniger selbstständig Entscheidungen über Diagnosen und damit Therapien trifft (Heinrichs, 2020), d. h., analog zu einem selbstfahrenden Auto agiert, muss angemerkt werden, dass die Nutzung von KI-Verfahren im medizinischen Bereich deutlich komplexer ist.

Versteckte KI arbeitet oft im Hintergrund in medizinischen Geräten und verbessert beispielsweise Bildqualität bei Magnetresonanztomographien (MRT), Computertomographien (CT) und Ultraschall. Sie rekonstruiert Aufnahmen, entfernt Störungen oder optimiert Kontraste, ohne dass Anwendende dies explizit bemerken (Mazurowski, 2019). Auch in der Verarbeitung und Aufbereitung von Daten spielt KI schon heute eine relevante Rolle. Wie in der Optimierung von Bilddaten steht sie hier jedoch im Hintergrund, da die entsprechenden Verfahren nur genutzt werden, um Ärzt*innen möglichst optimale Informationen und Daten bereitzustellen – nicht jedoch, um diese selbst zu bewerten. In einer einfachen Analogie können diese KI-Ansätze daher mit dem Autofokus einer Smartphone-Kamera verglichen werden: Sie übernehmen Hintergrundprozesse, vereinfachen damit den Nutzer*innen die Arbeit und verbessern das Ergebnis.

Als ein Sonderfall können »*closed-loop* KI-Systeme« angesehen werden, welche Geräte wie Beatmungsmaschinen, Infusionssysteme oder Überwachungsmonitore selbstständig steuern und anhand von Patientendaten Parameter in Echtzeit anpassen (Wang, 2025). Die Abgrenzung zu nicht-intelligenten, regelbasierten Systemen ist dabei in der Regel schwierig.

Autonome Systeme zur Diagnose, Therapie und Prognose, welche Ärzt*innen nicht nur unterstützen, sondern sie potenziell ersetzen, erscheinen auch mittelfristig eher unwahrscheinlich (Langlotz, 2019). Große Hoffnungen werden jedoch in unterstützende Systeme gelegt, welche Daten wie Laborergebnisse oder bildgebende Befunde nicht nur verarbeiten, sondern auch auswerten (Hedderich, 2021). Solche Anwendungen haben anhand einer großen Anzahl von vorherigen Fällen Zusammenhänge gelernt, welche es erlauben, auch bei neuen Einzelfällen Hinweise auf mögliche Befunde oder Diagnosen zu geben. Diese (Vorab-)Informationen sollten dann zu einer Entlastung bzw. Effizienzsteigerung bei Ärzt*innen führen. Dies gilt insbesondere für Disziplinen wie die Radiologie, in der sehr ausgereifte Algorithmen zur Bildverarbeitung (»*computer vision*«) nicht nur eine erste Beurteilung der Bilder erlauben, sondern insbesondere auch unerwartete Nebenbefunde für die ärztliche Beurteilung hervorheben können.

Eine Variante solcher Ansätze sind in Assistenzsystemen zu sehen, welche Diagnostik und Therapieplanung unterstützen, indem sie klinisches Wissen, Leitlinien und aktuelle Studien systematisch auswerten und Ärzt*innen basierend auf diesen Informationen konkrete Vorschläge machen, etwa für Risikoeinschätzungen oder Therapieoptionen (Bizzo, 2019). Da solche Systeme nur medizinisches Erfahrungswissen auswerten und verknüpfen, jedoch nicht auf Primärdaten trainiert werden, können sie als eine Art indirekte medizinische KI angesehen werden.

Zusammenfassend muss festgehalten werden, dass »KI in der Medizin« ein sehr heterogenes Feld umfasst, was nicht zuletzt dem Umstand geschuldet ist, dass der Begriff KI oft sehr unscharf gebraucht wird. So nehmen ethische und rechtliche Debatten in diesem Zusammenhang oft die momentan noch nicht in greifbarer Nähe erscheinenden autonomen Systeme in den Fokus (Eickhoff, 2021). Für den praktischen Alltag heute und in der absehbaren Zukunft dürften aber vor allem KI-Anwendungen im Vordergrund stehen, welche ärztliche Entscheidungen unterstützen, indem z. B. Daten aufbereitet oder analysiert werden. Aufgrund der Vorreiterrolle, die der Radiologie in Hinblick auf die Einbindung von KI in der Medizin zukommt, fokussieren die folgenden Abschnitte vor allem die Anwendung von KI in diesem Bereich.

2. Historische Einordnung

Auch wenn die aktuelle mediale Präsenz des Themas und öffentliche Debatte etwas anderes vermuten lassen, ist künstliche Intelligenz keine neuartige Entwicklung der letzten Jahre und weder an sich noch für die Anwendung in der Medizin ein gänzlich neues Thema. Tatsächlich reichen die Anfänge und konzeptionellen Grundlagen von künstlicher Intelligenz bis in die 1950er Jahre zurück. Als einer der Wegbereiter der künstlichen Intelligenz gilt der britische Mathematiker Alan Turing, der in seinen Arbeiten die relevanten theoretischen Grundsteine künstlicher Intelligenz legte. In seiner wegweisenden, im Jahr 1950 veröffentlichten Arbeit »Computing Machinery and Intelligence« (Turing, 1950) stellte er die Frage »Can machines think?« und präsentierte das Konzept des sogenannten »imitation game«. Bei diesem später als »Turing-Test« bekannt gewordenen Gedankenspiel handelt es sich um ein theoretisches Experiment, das zur Beantwortung der Frage herangezogen wird, ob eine Maschine ein dem Menschen gleichwertiges Denkvermögen aufweist. Hierbei kommuniziert ein*e menschliche*Fragesteller*in ohne Sicht- und Hörkontakt über schriftliche Nachrichten zum einen mit einer menschlichen Person und zum anderen mit einer Maschine. Ist der oder die Fragesteller*in nach intensiver Befragung nicht dazu in der Lage, zu differenzieren, bei welchem Gesprächspartner es sich um die menschliche Person und bei welchem es sich um die Maschine handelt, hat die Maschine den Turing-Test bestanden und gilt als (künstlich) intelligent. Trotz diverser Kritik findet der Turing-Test bis heute Beachtung und bietet zwar keine umfassende Definition, aber zumindest einen gut verständlichen Zugang, um den abstrakten Begriff der künstlichen Intelligenz besser zu fassen. Obwohl das Bestehen des Turing-Tests lange als unerreicht galt, geht man heute davon aus, dass moderne sprachbasierte KI-Systeme (*large language models*) dazu in der Lage sind, den Turing-Test zumindest bestimmten Interpretationen zufolge zu bestehen (Jones 2025). Nach den theoretischen Vorarbeiten von Turing gilt die im Jahr 1956 veranstaltete »Dartmouth Summer Research Project on Artificial Intelligence«-Konferenz als eigentliche Geburtsstunde der künstlichen Intelligenz, bei der erstmals der Begriff »künstliche Intelligenz« geprägt wurde und von einigen führenden Mathematikern und Wissenschaftlern relevante Konzepte dieses neuen Feldes kon-

stituiert wurden. Die weitere Entwicklung der künstlichen Intelligenz in den kommenden Jahren war vor allem durch frühe Entwicklungen geprägt, wie beispielsweise einfache Entscheidungssysteme auf Basis von Entscheidungsbäumen, oder den Prototyp eines ersten Chat-Computers mit dem Namen ELIZA Mitte der 1960er Jahre (Weizenbaum, 1966). In diese Zeit fallen auch erste Versuche, KI auf medizinische Bilddaten anzuwenden. So nutzte etwa Gwilym S. Lodwick ein auf dem Bayes-Theorem beruhendes probabilistisches System, um Knochentumoren anhand bestimmter Bildeigenschaften in Röntgen-Bildern zu diagnostizieren (1965). Während die theoretischen Grundlagen der heute gebräuchlichen künstlichen neuronalen Netzwerke bereits in den 1940er Jahren durch die Neurowissenschaftler Warren S. McCulloch und Walter Pitts geprägt wurden, wurde mit dem von Frank Rosenblatt 1958 vorgestellten *Perzeptron*, einem vereinfachten künstlichen neuronalen Netzwerk, welches für einfache Klassifikationsaufgaben geeignet ist (1958), zudem eine wichtige Grundlage für die Entwicklung neuronaler Netzwerke in dieser Zeit gelegt. 1969 führten Marvin Minsky und Seymour Papert jedoch den Beweis, dass nicht-lineare Probleme durch das Perzeptron nicht lösbar sind (1969). Diese und weitere Erkenntnisse hinsichtlich der Limitationen dieser frühen KI-Systeme führten Mitte der 1970er Jahre zu einer großen Enttäuschung der bis dahin stark überzogenen Erwartungen an künstliche Intelligenz, weshalb diese Zeit heute gemeinhin als erster »KI-Winter« bezeichnet wird.

Einen erneuten Aufschwung der künstlichen Intelligenz brachten in den 1980er Jahren die sogenannten »Experten-Systeme«, die eine computer-basierte Entscheidungsunterstützung für komplexe Probleme durch eine Menge definierter »Wenn-dann-Regeln« sowie eine zugrundeliegende Wissensdatenbank liefern sollten. Diese Experten-Systeme fanden insbesondere auch im Kontext medizinischer Problemstellungen Anklang. Eines der ersten solcher Systeme war CASNET (Weiss, 1978), ein Experten-System, welches die Diagnose und Therapie des Glaukoms unterstützen sollte. Ein weiteres Beispiel war MYCIN (Shortliffe, 1977), welches mit einem enorm umfangreichen Regelsystem Krankheiten anhand von eingegebenen Patientendaten erkennen und Therapievorschläge für Antiobiotika-Regime vorschlagen sollte und aus dem später das noch umfangreichere System INTERNIST-1 entstand (Miller, 1985). Letztendlich erreichten die Experten-Systeme insbesondere bei komplexen Pro-

blemstellungen nicht die erhoffte Leistungsfähigkeit und enttäuschten die hochgesteckten Erwartungen an diese Technologie, was Anfang der 1990er den Beginn des zweiten KI-Winters markierte.

Ab den 1990er Jahren traten an die Stelle der zuvor regelbasierten Ansätze der Experten-Systeme vor allem stärker datengetriebene Ansätze. Methoden wie die *support vector machines* (Cortes, 1995), Bayes'sche Netzwerke (Castillo, 2012) oder Ensemble-Methoden wie *random forest* (Breiman, 2001) wurden in dieser Zeit etabliert und fanden zunehmend Anwendung auf medizinischen Daten. Auf dieser Basis entwickelten sich in den 1990er Jahren auch zunehmend kommerziell verfügbare »CAD-Systeme« (*computer aided detection*), die Radiolog*innen bei der Erkennung bestimmter Pathologien unterstützen sollten, beispielsweise bei der Brustkrebs-Erkennung in Mammographien (Freer, 2001) oder der Detektion von Lungenrundherden in Computertomographien der Lunge (Das, 2006). Teilweise konnte mit dem Einsatz dieser Systeme bereits eine Steigerung der Detektionsrate entsprechender Läsionen demonstriert werden; insgesamt blieb die klinische Akzeptanz dieser Systeme jedoch verhalten.

Einige CAD-Systeme beruhten bereits auf künstlichen neuronalen Netzwerken. Bereits in den 1980er Jahren gab es gegenüber dem Perzeptron relevante theoretische Weiterentwicklungen im Bereich künstlicher neuronaler Netzwerke, unter anderem die architekturelle Entwicklung der später im *deep learning* vornehmlich verwendeten *convolutional neural networks* (CNNs) und die Anwendung des Backpropagation Algorithmus auf solche mehrschichtigen neuronalen Netzwerke (Fukushima, 1980; LeCun, 1989). Obwohl es in der Folgezeit kontinuierliche Weiterentwicklungen im Bereich künstlicher neuronaler Netzwerke gab, blieb der große Durchbruch jedoch zunächst aus – vor allem aufgrund der noch begrenzten Rechenleistung damaliger Computersysteme. Erst mit dem deutlichen Anstieg der verfügbaren Rechenkapazitäten und dem Einsatz von Grafikprozessoren ab den späten 2000er und insbesondere ab den 2010er-Jahren konnten tiefe neuronale Netzwerke ihre volle Leistungsfähigkeit entfalten. Als Meilenstein dieses Durchbruchs gilt dabei das Jahr 2012, als das Netzwerk *AlexNet* im *ImageNet Large Scale Visual Recognition Challenge* Wettbewerb einen bis dahin unerreichten Leistungssprung erzielte (Krizhevsky, 2012) und damit den Beginn der modernen *deep learning* Ära markierte. Der losgetretene Trend

übertrug sich unmittelbar auf die Anwendung in der Medizin. Prominente Beispiele für solche Anwendungen waren unter anderem ein *convolutional neural network*, das Hautkrebs in Fotografien und dermatoskopischen Bildern auf fachärztlichem Niveau klassifizieren konnte (Esteva, 2017) oder ein *deep learning* Netzwerk, welches diabetische Retinopathie anhand von Fotografien des Augenfundus erkennt (Gulshan, 2016). Aufgrund der besonderen Eignung von *deep learning* Verfahren für Mustererkennung in Bilddaten rückte jedoch insbesondere die medizinische Bildgebung und Radiologie in den Fokus des wissenschaftlichen Interesses. Die Zahl wissenschaftlicher Publikationen zu künstlicher Intelligenz in der Radiologie stieg in dieser Zeit exponentiell von zuvor ca. 300 Publikationen pro Jahr im Jahr 2010 auf über 2000 Publikationen im Jahr 2020 an (Decuyper, 2021). Nahezu jede Bildgebungsmodalität und Krankheitsentität wurde wissenschaftlich mithilfe von *deep learning* Netzwerken analysiert, vor allem, um Pathologien in Bildern zu erkennen oder Zusammenhänge zwischen medizinischen Bilddaten und Krankheits- oder Therapieverläufen herzustellen. Beispiele hierfür waren Systeme basierend auf *convolutional neural networks* zur Brustkrebserkennung in Mammographien, die hinsichtlich ihrer Leistungsfähigkeit klassische CAD-Systeme übertraf und eine Erkennungsrate auf dem Niveau von Radiolog*innen aufwiesen (Kooi, 2017), Algorithmen zur zuverlässigen Erkennung von Pathologien wie Tuberkulose oder Pneumonien in Röntgen-Thorax Bildern (Lakhani, 2017; Rajpurkar, 2017) oder Systeme zur Segmentierung von Hirntumoren in Magnetresonanztomographien des Schädels (Havaei, 2017). Während der *deep learning* Hype in der medizinischen Bildgebung noch nicht abgeklungen war/ist, kam es Anfang der 2020er Jahre zu einem darauf aufsetzenden Hype um sogenannte generative künstliche Intelligenz. Dieser Hype wurde durch relevante Weiterentwicklungen sogenannter *Transformer-Modelle* ausgelöst (Vaswani, 2017) und erlangte spätestens durch die zu dieser Zeit verblüffenden Fähigkeiten des auf der *Transformer-Architektur* beruhenden und im November 2022 von der Firma OpenAI veröffentlichten Chat-Systems ChatGPT (OpenAI, San Francisco, USA) eine breite öffentliche Wahrnehmung. Neben großen Sprachmodellen erlangten zu dieser Zeit auch weitere generative KI-Systeme zur Erzeugung von Bildern, gesprochener Sprache, Musik oder sogar Videos aus Texteingaben große Beachtung. Schnell wurden die Potentiale generativer KI

auch für die Medizin erkannt. Neben der Generierung täuschend echter synthetischer medizinischer Bilder wie Röntgen-Thorax-Bilder (Bluethgen, 2025), denen das Potential zugeschrieben wird, die Wahrung des Datenschutzes beim Training von KI Modellen mit großen Datenmengen im medizinischen Kontext zu erleichtern, konnten generative Modelle vor allem auch eingesetzt werden, um aus stark verrauschten medizinischen Bilddaten Bilder von diagnostischer Qualität zu erzeugen und so z. B. die Strahlen-Exposition in der Computertomographie (Sadia, 2024) oder die Scan-Dauer in Magnetresonanztomographien relevant zu senken (Kadavigere, 2024). Auch verschiedene Anwendungsmöglichkeiten großer generativer Sprachmodelle im medizinischen Kontext wurden bereits in zahlreichen Arbeiten erforscht, unter anderem, um Fehler in radiologischen Befunden zu erkennen (Gertz, 2024), strukturierte Daten aus Freitextbefunden zu Qualitätssicherungszwecken zu extrahieren (Lehnen, 2024), Fachwissen aus medizinischen Leitlinien abzuleiten (Vach, 2025) oder Bildgebungsprotokolle für Magnetresonanztomographien anhand der klinischen Fragestellung automatisch zu generieren (Boschenriedter, 2025). Auch die automatische Generierung radiologischer Befunde direkt aus medizinischen Bildern auf dem Niveau von Radiolog*innen konnte erfolgreich demonstriert werden (Zhou, 2025). Zuletzt wurde in wissenschaftlichen Arbeiten auch vermehrt sogenannte agentische künstliche Intelligenz fokussiert – generative KI-Systeme, die durch Zugriff auf Werkzeuge, Wissensdatenbasen und Interaktionsmöglichkeiten umfassende Möglichkeiten haben, Aufgaben durch Planung und Erweiterung ihres Kontextes zu lösen – und ihre Anwendbarkeit im medizinischen Kontext demonstriert (Ferber, 2025).

3. Aktuelle Entwicklungen und Herausforderungen

Nach den Entwicklungen der letzten Jahre zeigt sich aktuell ein zunehmender Einzug von künstlicher Intelligenz in die klinische Anwendung. In diesem Zuge zeigte sich in den letzten zehn Jahren ein enormer Zuwachs von Zulassungen KI-basierter Software-Systeme als Medizinprodukte. Nach der ersten Zulassung eines KI-basierten Software-Produktes durch die für die Regulierung von Medizinprodukten in den USA zuständige Food and Drug Administration

(FDA) im Jahr 1995 kam es in den folgenden Jahren lediglich sporadisch zu weiteren Zulassungen. So wurden im Jahr 2015 lediglich sechs Produkte durch die FDA neu zugelassen. Seither zeigt sich jedoch ein rapider Anstieg neu zugelassener Medizinprodukte mit beispielsweise 130 Neuzulassungen im Jahr 2021 und zuletzt 235 Neuzulassungen im Jahr 2024 (FDA, 2025). Bemerkenswert ist hierbei, dass der Radiologie eine deutliche Vorreiterrolle für die Adaptation von KI in der Medizin zukommt: Von den insgesamt 1247 zugelassenen KI-basierten Software-Medizinprodukten, welche im August 2025 von der FDA geführt werden, entfallen 956 Produkte, also 76,7 %, auf den Bereich Radiologie (FDA, 2025). Bei diesen Software-Produkten handelt es sich überwiegend um Software-basierte Assistenzsysteme, die Radiolog*innen bei der Interpretation von Bildern oder bei der Priorisierung der Befundung unterstützen sollen.

Trotz dieses immensen Angebots an KI-basierten Software-Produkten in der Radiologie bleibt die tatsächliche Anwendung in der klinischen Radiologie bisher vergleichsweise zurückhaltend. In der täglichen Praxis haben sich bislang kaum Systeme flächendeckend etabliert, und ihr Beitrag zur Routineversorgung ist bislang begrenzt. Ein wesentlicher Grund hierfür liegt in der Frage der Finanzierung (van Leeuwen, 2024). Viele der derzeit verfügbaren KI-Anwendungen sind kostenintensiv in Anschaffung und Wartung, während belastbare Nachweise für einen klaren ökonomischen Nutzen fehlen. Solange nicht belegt ist, dass KI-Systeme die Effizienz radiologischer Prozesse messbar steigern oder Kosten im Gesundheitssystem senken, fällt es schwer, entsprechende Investitionen im ohnehin stark budgetierten klinischen Umfeld zu rechtfertigen.

Hinzu kommen erhebliche Hürden bei der Integration in bestehende klinische Workflows. Radiologische Abteilungen arbeiten mit komplexen IT-Landschaften, die Bildarchivierungs- und Kommunikationssysteme (engl. Picture Archiving and Communication System, PACS), Radiologie-Informationssysteme (RIS) und Krankenhaus-Informationssysteme (KIS) umfassen. KI-Anwendungen, die nicht nahtlos in diese Strukturen eingebettet werden können, führen im Arbeitsalltag eher zu Mehraufwand als zu Entlastung. Es ist naheliegend, dass Systeme, die nur isoliert funktionieren oder zusätzliche Arbeitsschritte erfordern, nur begrenzten Nutzen entfalten und häufig nicht akzeptiert und rasch wieder verworfen werden. Daher

ist es relevant, dass sich KI-Anwendungen nahtlos in bestehende Arbeitsabläufe und -infrastrukturen eingliedern und z. B. Arbeitslisten priorisieren, auffällige Befunde markieren oder standardisierte Messungen automatisiert durchführen, ohne den klinischen Workflow zu stören.

Eng mit diesen Aspekten verknüpft ist die noch immer unzureichende Evidenzlage in Bezug auf den tatsächlichen klinischen und ökonomischen Nutzen von KI. Zwar gibt es zahlreiche wissenschaftliche Arbeiten zu KI in der Radiologie, die teils beeindruckende Leistungswerte einzelner Modelle berichten, doch handelt es sich hierbei ganz überwiegend um retrospektive, monozentrische Fallsammlungen, während groß angelegte, randomisiert kontrollierte Studien, die die Effektivität von KI-Systemen unter Alltagsbedingungen belegen, aktuell kaum vorhanden sind (Gorenstein, 2023). Ohne diese Evidenz bleibt unklar, ob der Einsatz von KI tatsächlich die diagnostische Genauigkeit, die Versorgungsqualität oder die Sicherheit der Patient*innen verbessert und ob er einen messbaren Mehrwert für das Gesundheitssystem generiert.

Über die genannten praktischen Hürden hinaus bestehen grundsätzliche methodische und wissenschaftliche Herausforderungen. Ein zentrales Problem stellt die sogenannte Domänen-Verschiebung (*domain shift*) dar. Modelle, die in einer bestimmten Institution, Population oder mit Daten eines bestimmten Herstellers trainiert wurden, verlieren häufig an Genauigkeit, wenn sie auf neue Patientenkollektive, Scanner oder Bildgebungsprotokolle angewandt werden (Guan & Liu, 2021). Diese mangelnde Generalisierbarkeit begrenzt die klinische Nutzbarkeit und erfordert Verfahren zur Domänenanpassung sowie standardisierte externe Validierungen.

Ein weiteres Problemfeld betrifft Verzerrungen (*biases*) in KI-Systemen. Wenn Trainingsdaten nicht repräsentativ für die gesamte Patientenpopulation sind, entstehen Verzerrungen, die zu systematisch ungleichen Ergebnissen führen können. In der Radiologie betrifft dies beispielsweise Unterschiede zwischen verschiedenen Altersgruppen, Geschlechtern oder ethnischen Hintergründen (Banerjee, 2023). Solche Verzerrungen können die diagnostische Qualität gefährden und werfen zudem grundlegende ethische Fragen nach Gerechtigkeit und Fairness auf (Lara, 2022).

Ein bislang nur wenig beachteter Aspekt ist zudem die Mensch-Maschine-Interaktion. Während die meisten wissenschaftlichen Ar-

beiten vor allem die technische Validität von KI-Systemen bei der Anwendung auf Bildgebungsdaten fokussieren, wird das Zusammenspiel der Anwender*innen mit den KI-Systemen und hieraus möglicherweise resultierender Effekte wenig beachtet. Bei der Verwendung automatisierter Systeme kann es zu relevanten Beeinflussungen durch kognitive Verzerrungen der Anwender*innen durch übermäßiges Vertrauen in die Empfehlungen einer Maschine kommen, selbst wenn diese fehlerhaft sind (Parasuraman & Manzey, 2010). Dieses als *automation bias* bezeichnete Phänomen konnte unter anderem eindrucksvoll in einem Laborexperiment demonstriert werden, bei dem die Genauigkeit von Radiolog*innen bei der KI-unterstützten Befundung von Mammografien abhängig vom Erfahrungsgrad von ca. 80 % auf bis zu ca. 20 % sank, wenn die KI falsche Entscheidungen suggerierte (Dratsch, 2023). Strategien zur Kommunikation von Unsicherheit von KI-Systemen und zur Vermeidung solcher Verzerrungen sowie zur Förderung eines kritischen Umgangs mit KI-Ausgaben sind daher entscheidend.

Ethische und regulatorische Fragen bilden einen weiteren zentralen Bereich. Der Einsatz von KI in der Radiologie betrifft hochsensible Patientendaten und Entscheidungen von erheblicher Tragweite. Entsprechend hoch sind die Anforderungen an Datenschutz, Transparenz und Nachvollziehbarkeit. Gleichzeitig müssen regulatorische Systeme wie der europäische AI Act und Institutionen wie die US-amerikanische FDA sicherstellen, dass KI-Produkte nicht nur bei der Zulassung, sondern auch während ihres gesamten Lebenszyklus überwacht und regelmäßig evaluiert werden. Eine besondere Herausforderung ergibt sich hier bei adaptiven Systemen, die sich nach der Zulassung durch kontinuierliches Lernen verändern können. Hier gilt es, einen regulatorischen Rahmen zu schaffen, der Innovation ermöglicht, aber zugleich Sicherheit und Verlässlichkeit garantiert.

Zusammenfassend lässt sich feststellen, dass die aktuelle Nutzung von KI in der Radiologie noch weit hinter den Erwartungen zurückbleibt, die die technologische Entwicklung in den letzten Jahren geweckt hat. Neben praktischen Hürden wie Finanzierung, Integration und Evidenzlage sind es auch grundlegende wissenschaftliche, ethische und regulatorische Fragen, die eine breite Anwendung hemmen. Zugleich verdeutlichen diese Herausforderungen, dass die Weiterentwicklung der KI in der Radiologie nicht nur eine technische,

sondern auch eine gesellschaftliche Aufgabe ist, die einen interdisziplinären Diskurs und klare Rahmenbedingungen erfordert.

4. Perspektiven für KI in der Radiologie

In Anbetracht der enormen Dynamik, die die Entwicklung der künstlichen Intelligenz in den vergangenen Jahren entfaltet hat, lassen sich belastbare Prognosen über zukünftige Einsatzformen in der Radiologie nur schwer treffen. Gleichwohl zeichnen sich zentrale Handlungsfelder ab, die über die erfolgreiche und verantwortungsvolle Integration dieser Technologien entscheiden werden. Eine wesentliche Aufgabe der kommenden Jahre besteht darin, die bestehenden und oben erläuterten Herausforderungen gezielt zu adressieren, um einen vertrauensvollen Umgang mit dieser potenten, zugleich aber auch hochsensiblen Technologie im Bereich der Medizin zu ermöglichen. Hierzu gehört insbesondere die Schaffung einer ausreichenden wissenschaftlichen Evidenzbasis, die nicht nur technische Leistungskennzahlen, sondern auch den tatsächlichen Nutzen von KI für die Versorgung der Patient*innen und das Gesundheitssystem belegt.

Eng damit verbunden ist die Frage, wie sich ein hinreichendes Maß an Kontrolle über KI-Systeme bewahren lässt, um deren sicheren und gewinnbringenden Einsatz gewährleisten zu können. Von besonderer Bedeutung wird dabei die Gestaltung der Mensch-Maschine-Interaktion sein. Nur wenn klar definiert bleibt, welche Aufgaben von KI übernommen werden können und wo ärztliche Verantwortung unverzichtbar ist, kann das Vertrauen in diese Systeme gestärkt werden. Auch regulatorische Vorgaben wie der europäische AI Act, der für Hochrisiko-KI-Systeme ausdrücklich eine menschliche Überwachungspflicht vorsieht, unterstreichen die Relevanz dieses Aspekts.

Auf technologischer Ebene eröffnet insbesondere die Entwicklung sogenannter *foundation models* weitreichende Perspektiven. Diese Modelle werden auf sehr großen, multimodalen Datensätzen trainiert, die unterschiedliche Datenquellen wie radiologische Bilddaten, Textdaten aus Arztbriefen oder auch medizinisches Fachwissen aus der Literatur umfassen können. Dadurch entsteht ein breites, abstrahiertes medizinisches Wissen, das in vielfältige Anwendungs-

aufgaben – sogenannte *downstream tasks* – überführt werden kann. Denkbar sind auf dieser Basis sowohl Assistenzsysteme zur Erkennung ganz unterschiedlicher Pathologien in Bilddaten als auch Systeme, die Radiolog*innen beim Erstellen von Befunden oder Arztbriefen unterstützen, oder auch Chatbots, die administrative Tätigkeiten wie etwa die Terminorganisation übernehmen (D'Antonoli, 2025). Im Vergleich zu den gegenwärtig vorherrschenden Insellösungen, die jeweils nur eine eng umgrenzte Aufgabe erfüllen – etwa die Detektion von Lungenrundherden in der Computertomographie, ohne in der Lage zu sein, andere Pathologien wie beispielsweise Frakturen zu erkennen oder andere klinische Dokumente zu verarbeiten – erscheint die Integration eines einzigen, generalistischen Systems auf Basis eines *foundation models* für die klinische Anwendung mittelfristig deutlich realistischer.

Eine weitere Entwicklungsperspektive ergibt sich durch sogenannte *KI-Agenten*. Diese Systeme verfügen über die Möglichkeit, auf Wissensbasen wie lokale Datenbanken von Gesundheitseinrichtungen oder externe Quellen aus dem Internet sowie auf Software- und Kommunikationstools zuzugreifen, um hiermit ihren Kontext eigenständig zu erweitern und komplexe Aufgaben weitgehend autonom zu bearbeiten. Das Potenzial solcher Agenten reicht von der medizinischen Assistenz – etwa bei der Befundung oder Diagnoseunterstützung – bis hin zur administrativen Entlastung in radiologischen Abteilungen (Ferber, 2025). Gerade bei dieser Technologie ist es jedoch essenziell, sorgfältig zu regulieren, in welchem Umfang der Zugriff auf Daten und Werkzeuge im medizinischen Kontext erlaubt werden kann, um die notwendige Kontrolle über die Systeme zu gewährleisten und Missbrauch vorzubeugen.

Auch wenn die Leistungsfähigkeit von KI-Systemen in einzelnen Aufgabenfeldern bereits das Niveau erfahrener Radiolog*innen erreichen kann, ist ein vollständiger Ersatz ärztlicher Expertise in absehbarer Zeit nicht zu erwarten (Langlotz, 2019). Wahrscheinlicher ist eine Verschiebung des beruflichen Rollenbildes: Durch die Automatisierung repetitiver Tätigkeiten wie der Mustererkennung in Bilddaten oder bestimmter administrativer Arbeiten werden Radiolog*innen zunehmend in die Rolle von »Orchestrator*innen« und Supervisor*innen treten, die den Einsatz von KI-Systemen überwachen, Ergebnisse kritisch reflektieren und in den klinischen Gesamt-

kontext einordnen. Damit wird ihre Tätigkeit zwar nicht weniger zentral, wohl aber grundlegend anders strukturiert sein.

Vor diesem Hintergrund gewinnt auch die Aus- und Weiterbildung in der Radiologie neue Bedeutung. Es wird darauf ankommen, Konzepte zu entwickeln, die Studierende und Nachwuchsärztinnen und -ärzte gezielt auf die Arbeit an der Schnittstelle zwischen Medizin und Informationstechnologie vorbereiten. Neben der klassischen radiologischen Expertise müssen Kompetenzen im Umgang mit KI, in der Bewertung algorithmischer Ergebnisse und in der verantwortungsvollen Nutzung digitaler Werkzeuge vermittelt werden. Nur so lässt sich sicherstellen, dass die Radiologie von den Chancen der KI profitiert, ohne dass dabei die Sicherheit der Patient*innen oder die ärztliche Verantwortung kompromittiert werden.

Nicht zuletzt bestimmen regulatorische Rahmenbedingungen, ethische Grundsätze und ökonomische Überlegungen die zukünftige Rolle der KI in der Radiologie. Es stellt sich die Frage, wie die Kosten für Entwicklung, Implementierung und Wartung von KI-Systemen in ein Gesundheitssystem integriert werden können, das bereits erheblich unter ökonomischem Druck steht. Akzeptanz seitens der Radiolog*innen wird nur dann entstehen, wenn die Systeme nachweislich die Versorgungsqualität verbessern, gleichzeitig aber auch die Arbeitsbelastung reduzieren.

Literaturverzeichnis

Banerjee, I., Bhattacharjee, K., Burns, J. L., Trivedi, H., Purkayastha, S., Seyyed-Kalantari, L., Patel, B. N., Shiradkar, R., & Gichoya, J. (2023). »Shortcuts« causing bias in radiology artificial intelligence: Causes, evaluation, and mitigation. *Journal of the American College of Radiology*, *20*(9), 842–851. https://doi.org/10.1016/j.jacr.2023.06.025

Bizzo, B. C., Almeida, R. R., Michalski, M. H., & Alkasab, T. K. (2019). Artificial intelligence and clinical decision support for radiologists and referring providers. *Journal of the American College of Radiology*, *16*(9), 1351–1356. https://doi.org/10.1016/j.jacr.2019.06.010

Bluethgen, C., Chambon, P., Delbrouck, J. B., van der Sluijs, R., Połacin, M., Zambrano Chaves, J. M., Abraham, T. M., Purohit, S., Langlotz C. P., & Chaudhari, A. S. (2025). A vision–language foundation model for the generation of realistic chest X-ray images. *Nature Biomedical Engineering*, *9*(4), 494–506. https://doi.org/10.1038/s41551-024-01246-y

Boschenriedter, C., Rubbert, C., Vach, M., & Caspers, J. (2025). Automated protocol suggestions for cranial MRI examinations using locally fine-tuned BERT models. *Clinical Neuroradiology*, epub ahead of print. https://doi.org/10.1007/s00062-025-01554-z

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Castillo, E., Gutierrez, J. M., & Hadi, A. S. (2012). *Expert systems and probabilistic network models*. Springer Science & Business Media.

Chen, J., Patil, K. R., Yeo, B. T., & Eickhoff, S. B. (2023). Leveraging machine learning for gaining neurobiological and nosological insights in psychiatric research. *Biological Psychiatry*, *93*(1), 18–28. https://doi.org/10.1016/j.biopsych.2022.07.025

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

D'Antonoli, T. A., Bluethgen, C., Cuocolo, R., Klontzas, M. E., Ponsiglione, A., & Kocak, B. (2025). Foundation models for radiology: fundamentals, applications, opportunities, challenges, risks, and prospects. *Diagnostic and Interventional Radiology*. epub ahead of print. https://doi.org/10.4274/dir.2025.253445

Das, M., Mühlenbruch, G., Mahnken, A. H., Flohr, T. G., Gündel, L., Stanzel, S., Kraus, T., Günther, R. W., & Wildberger, J. E. (2006). Small pulmonary nodules: Effect of two computer-aided detection systems on radiologist performance. *Radiology*, *241*(2), 564–571. https://doi.org/10.1148/radiol.2412051139

Decuyper, M., Maebe, J., Van Holen, R., & Vandenberghe, S. (2021). Artificial intelligence with deep learning in nuclear medicine and radiology. *EJNMMI Physics*, *8*(1), 81. https://doi.org/10.1186/s40658-021-00426-y

Dratsch, T., Chen, X., Rezazade Mehrizi, M., Kloeckner, R., Mähringer-Kunz, A., Püsken, M., Baeßler, B., Sauer, S., Maintz, D., & Pinto dos Santos, D. (2023). Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology*, *307*(4), e222176. https://doi.org/10.1148/radiol.222176

Eickhoff, S. B., & Heinrichs, B. (2021). Der vorhersagbare Mensch. *Der Nervenarzt*, *92*(11), 1140–1148. https://doi.org/10.1007/s00115-021-01197-8

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115–118. https://doi.org/10.1038/nature21056

Ferber, D., El Nahhas, O. S., Wölflein, G., Wiest, I. C., Clusmann, J., Leßmann, M. E., Foersch, S., Lammert, J., Tschochohei, M., Jäger, D., Salto-Tellez, M., Schultz, N., Thrun, D., & Kather, J. N. (2025). Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nature Cancer, 6,* 1337–1349. https://doi.org/10.1038/s43018-025-00991-6

Food and Drug Administration (FDA) (2025). *Artificial intelligence and machine learning (AI/ML)– enabled medical devices*. USA Department of Health and Human Services. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices (Zugriff 18.08.2025).

Freer, T. W., & Ulissey, M. J. (2001). Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center. *Radiology*, *220*(3), 781–786. https://doi.org/10.1148/radiol.2203001282

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202. https://doi.org/10.1007/BF00344251

Gertz, R. J., Dratsch, T., Bunck, A. C., Lennartz, S., Iuga, A. I., Hellmich, M. G., Persigehl, T., Pennig, L., Gietzen, C. H., Fervers, P., Maintz, D., Hahnfeldt, R., & Kottlors, J. (2024). Potential of GPT-4 for detecting errors in radiology reports: Implications for reporting accuracy. *Radiology*, *311*(1), e232714. https://doi.org/10.1148/radiol.232714

Gorenstein, L., Soffer, S., Apter, S., Konen, E., & Klang, E. (2023). AI in radiology: is it the time for randomized controlled trials? *European Radiology*, *33*(6), 4223–4225. https://doi.org/10.1007/s00330-022-09381-3

Guan, H., & Liu, M. (2021). Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering*, *69*(3), 1173–1185. https://doi.org/10.1109/TBME.2021.3117407

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA, 316*(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P., & Larochelle, H. (2017). Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, *35*, 18–31. https://doi.org/10.1016/j.media.2016.05.004

Hedderich, D. M., & Eickhoff, S. B. (2021). Machine learning for psychiatry: getting doctors at the black box? *Molecular Psychiatry*, *26*(1), 23–25. https://doi.org/10.1038/s41380-020-00931-z

Heinrichs, B., & Eickhoff, S. B. (2020). Your evidence? Machine learning algorithms for medical diagnosis and prediction. *Human Brain Mapping*, *41*(6), 1435–1444. https://doi.org/10.1002/hbm.24886

Jones, C.R., & Bergen, B.K. (2025). Large language models pass the turing test. *arXiv*, Artikel 2503.23674. https://doi.org/10.48550/arXiv.2503.23674

Priyanka, Kadavigere, R., Nayak S, S., Chandran M, O., Shirlal, A., Pires, T., & Pendem, S. (2024). Impact of artificial intelligence assisted compressed sensing technique on scan time and image quality in musculoskeletal MRI–A systematic review. *Radiography*, *30*(6), 1704–1712. https://doi.org/10.1016/j.radi.2024.08.012

Koohi-Moghadam, M., & Bae, K. T. (2023). Generative AI in medical imaging: Applications, challenges, and ethics. *Journal of Medical Systems*, *47*(1), Artikel 94. https://doi.org/10.1007/s10916-023-01987-4

Kooi, T., Litjens, G., Van Ginneken, B., Gubern-Mérida, A., Sánchez, C. I., Mann, R., den Heeten, A., & Karssemeijer, N. (2017). Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis*, *35*, 303–312. https://doi.org/10.1016/j.media.2016.07.007

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Pereira, F., Burges, C.J., Bottou, L., Weinberger, K.Q. (Hrsg.). *Advances in Neural Information Processing Systems*, *25*.

Lakhani, P., & Sundaram, B. (2017). Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, *284*(2), 574–582. https://doi.org/10.1148/radiol.2017162326

Langlotz, C. P. (2019). Will artificial intelligence replace radiologists? *Radiology: Artificial Intelligence*, *1*(3), Artikel e190058. https://doi.org/10.1148/ryai.2019190058

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

Lehnen, N. C., Dorn, F., Wiest, I. C., Zimmermann, H., Radbruch, A., Kather, J. N., & Paech, D. (2024). Data extraction from free-text reports on mechanical thrombectomy in acute ischemic stroke using ChatGPT: A retrospective analysis. *Radiology*, *311*(1), Artikel e232741. https://doi.org/10.1148/radiol.232741

Lodwick, G. S. (1965). A probabilistic approach to the diagnosis of bone tumors. *Radiologic Clinics of North America*, *3*(3), 487–497. https://doi.org/10.1016/S0033-8389(22)02844-5

Mazurowski, M. A., Buda, M., Saha, A., & Bashir, M. R. (2019). Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. *Journal of Magnetic Resonance Imaging*, *49*(4), 939–954. https://doi.org/10.1002/jmri.26534

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. https://doi.org/10.1007/BF02478259

Minsky, M., & Papert, S. (1969). *Perceptrons. An introduction to computational geometry.* MIT Press.

Miller, R. A., Pople Jr, H. E., & Myers, J. D. (1985). INTERNIST-I, An experimental computer-based diagnostic consultant for general internal medicine. In B. I. Blum, J. A. Reggia, & S. Tuhrim (Hrsg.), *Computer-Assisted Medical Decision Making* (S. 139–158). Springer New York. https://doi.org/10.1007/978-1-4612-5108-8_8

Moawad, A. W., Fuentes, D. T., ElBanan, M. G., Shalaby, A. S., Guccione, J., Kamel, S., Jensen, C. T., & Elsayes, K. M. (2022). Artificial intelligence in diagnostic radiology: where do we stand, challenges, and opportunities. *Journal of Computer Assisted Tomography, 46*(1), 78–90. https://doi.org/10.1097/RCT.0000000000001247

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv*, Artikel 1711.05225. https://doi.org/10.48550/arXiv.1711.05225

Ricci Lara, M. A., Echeveste, R., & Ferrante, E. (2022). Addressing fairness in artificial intelligence for medical imaging. *Nature Communications*, *13*(1), Artikel 4581. https://doi.org/10.1038/s41467-022-32186-3

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Sadia, R. T., Chen, J., & Zhang, J. (2024). CT image denoising methods for image quality improvement and radiation dose reduction. *Journal of Applied Clinical Medical Physics*, *25*(2), Artikel e14270. https://doi.org/10.1002/acm2.14270

Shortliffe, E. H. (1977). Mycin: A knowledge-based computer program applied to infectious diseases. In *Proceedings of the Annual Symposium on Computer Application in Medical Care* (S. 66).

Turing, A.M. (1950). Computing machinery and intelligence. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Vach, M., Gliem, M., Weiss, D., Ivan, V.L., Hauke, F., Boschenriedter, C., Rubbert, C., & Caspers, J. (2025). Evaluating retrieval augmented generation-enhanced large language models for question answering on german neurovascular guidelines. *Clinical Neuroradiology*, epub ahead of print. https://doi.org/10.1007/s00062-025-01562-z

van Leeuwen, K. G., de Rooij, M., Schalekamp, S., van Ginneken, B., & Rutten, M. J. (2024). Clinical use of artificial intelligence products for radiology in the Netherlands between 2020 and 2022. *European Radiology*, 34(1), 348–354. https://doi.org/10.1007/s00330-023-09991-5

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Hrsg.), *31st Conference on Neural Information Processing Systems (NIPS). Advances in Neural Information Processing Systems, 30.*

Wang, X., Si, J., Li, Y., Tse, P., Zhang, G., Wang, X., Ren, J., Xu, J., Sun, J., & Yao, X. (2025). Effectiveness and safety of AI-driven closed-loop systems in diabetes management: a systematic review and meta-analysis. *Diabetology & Metabolic Syndrome, 17*(1), Artikel 238. https://doi.org/10.1186/s13098-025-01819-0

Weiss, S., Kulikowski, C. A., & Safir, A. (1978). Glaucoma consultation by computer. *Computers in Biology and Medicine*, *8*(1), 25–40. https://doi.org/10.1016/0010-4825(78)90011-2

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Zhou, H. Y., Acosta, J. N., Adithan, S., Datta, S., Topol, E. J., & Rajpurkar, P. (2024). MedVersa: A generalist foundation model for medical image interpretation. *arXiv,* Artikel *2405.07988*. https://doi.org/10.48550/arXiv.2405.07988

Robert Langner, Michael B. Steinborn

II. Psychologische Aspekte der Nutzung Künstlicher Intelligenz (KI) in der radiologischen Diagnostik

In der klinischen Radiologie werden heute zahlreiche KI-basierte Anwendungen entwickelt, erprobt oder bereits genutzt, um Arbeitsschritte und -abläufe im diagnostischen Prozess zu automatisieren. Solche radiologischen KI-Anwendungen zielen beispielsweise darauf ab, (i) Bildgebungsdaten zu strukturieren und relevante Bildbereiche visuell zu akzentuieren, (ii) Bildmuster diagnostisch einzuordnen und daraus (vorläufige) Befunde abzuleiten, (iii) Untersuchungen nach Dringlichkeit zu ordnen und damit den klinischen Arbeitsfluss zu steuern oder (iv) visuelle und quantitative Befunde sprachlich darzustellen. Der vielfältige Einsatz von KI verändert dabei den diagnostischen Arbeitsprozess nicht nur auf technischer Ebene, sondern greift in kognitive Prozesse, das ärztliche Selbstbild, die Verteilung von Verantwortung, berufliche Beziehungen sowie die Organisation der Arbeit ein. Leistungsfähigkeit und Sicherheit KI-gestützter radiologischer Diagnostik ergeben sich also nicht allein aus der Genauigkeit und Korrektheit der eingesetzten KI-Systeme, sondern auch aus der Art, wie KI-Anwendungen in die diagnostische Arbeit eingebunden sind. Maßgeblich sind dabei psychologische Mechanismen, die durch den Automatisierungsgrad, die Gestaltung der Automatisierung, die funktionale Platzierung des KI-Einsatzes im Entscheidungsablauf sowie durch organisationale und rechtliche Einsatzbedingungen beeinflusst werden (Parasuraman et al., 2000; Parasuraman & Riley, 1997; Sittig & Singh, 2010). Eine an menschlichen Eigenschaften orientierte Analyse dient dazu, diese Mechanismen systematisch zu bestimmen und aufzuzeigen, wie die Nutzung von KI diagnostisches Denken und Handeln in der Radiologie verändern kann.

Radiologische Diagnostik findet unter Bedingungen statt, die Fehlentscheidungen besonders folgenreich machen. Relevante Befunde sind in der Gesamtheit des zu beurteilenden Bildmaterials vergleichsweise selten und visuell oft schwer zu detektieren, während Zeitdruck und hoher Durchsatz die verfügbare Verarbeitungszeit begrenzen. Hinzu kommt eine asymmetrische Kostenstruktur von Fehlern: Das Übersehen klinisch relevanter Veränderungen wiegt meist schwerer als ein falsch-positiver Befund (»Fehlalarm«). Diese Konstellation beeinflusst Such- und Entscheidungsprozesse bereits implizit. Unter hoher Belastung konzentriert sich die Wahrnehmung auf erwartete Bildmerkmale, zugleich verschieben sich die Maßstäbe der Entscheidungsbildung, abhängig davon, ob zügiges Vorgehen oder Absicherung im Vordergrund steht. Frühe Festlegungen begünstigen vorschnelle Urteile, stärker angehobene Urteilsschwellen erhöhen dagegen die Wahrscheinlichkeit des Übersehens. Beide Fehlermodi sind funktional miteinander gekoppelt und entstehen, weil nur ein Teil der verfügbaren diagnostischen Informationen in die Entscheidung einfließt. Vor diesem Hintergrund liegt der Einsatz von KI nahe, doch ihr Beitrag zur diagnostischen Qualität hängt davon ab, wie zusätzliche Informationen in Wahrnehmung und Bewertung eingebunden sind. Förderlich wirken sie nur dann, wenn sie diagnostische Suchprozesse strukturieren, ohne diese zu überlagern oder zu stören (Drew et al., 2013). Entsprechend beeinflusst KI diagnostische Leistung auch über ihre Wirkungen auf Aufmerksamkeit, emotionale Sicherheit und Prüfverhalten, da diese Faktoren den Verlauf und die Qualität diagnostischer Entscheidungen mitbestimmen.

Die kognitiven Folgen des KI-Einsatzes lassen sich nur erfassen, wenn seine Funktionen entlang des diagnostischen Arbeitsgangs getrennt betrachtet werden. Eingriffe in klinische Priorisierung, visuelle Orientierung, diagnostische Klassifikation oder Befundformulierung greifen jeweils in unterschiedliche Arbeitsphasen ein und verändern entsprechend verschiedene kognitive Prozesse. Eine undifferenzierte Gesamtbewertung des KI-Einsatzes in der Radiologie griffe daher zu kurz. Da Priorisierung, visuelle Orientierung, Klassifikation und Befundformulierung jeweils andere kognitive Prozesse betreffen, lassen sich die Folgen auch nicht in einer einzelnen Leistungskennzahl abbilden, sondern nur als Kombinationen aus Veränderungen diagnostischer Genauigkeit, charakteristischen Fehlermustern sowie Verschiebungen von Tempo, Belastung und emo-

tionaler Sicherheit im Arbeitsvollzug. Darüber hinaus verändert der KI-Einsatz organisationale Abläufe. Er beeinflusst, wem eine Befundentscheidung zugerechnet wird, welche zusätzlichen Prüf- und Dokumentationsschritte erforderlich sind und wie technische Hinweise in die ärztliche Entscheidung eingehen. Diese Anpassungen betreffen nicht das System isoliert, sondern den gesamten diagnostischen Prozess. Die folgende Analyse verknüpft sie mit konkreten Eingriffspunkten und macht sichtbar, wo KI entlastet und wo neue Belastungen oder Unsicherheiten auftreten.

Die Bewertung des KI-Einsatzes in der Radiologie durch verschiedene Akteursgruppen folgt deren unterschiedlichen Zielsetzungen und Anforderungen. Aus radiologischer Perspektive ist z. B. maßgeblich, wie geeignet eine gegebene KI-Anwendung für bestimmte Aufgaben ist, wie sie in bestehende Abläufe integriert ist, welche zusätzlichen Prüf- und Absicherungsschritte entstehen und wie Verantwortung für Befunde im diagnostischen Prozess zugeordnet bleibt. Ob ein vorgeschlagener Befund als unterstützende Rückversicherung, als Unterbrechung des Arbeitsflusses oder als faktische Entscheidungsvorgabe wirkt, hängt von seiner funktionalen Platzierung und Gewichtung im Ablauf ab. Aus Patientenperspektive stehen demgegenüber Fragen der wahrgenommenen Sicherheit, der Fairness diagnostischer Entscheidungen und der Qualität der ärztlichen Interaktion im Vordergrund. Da Akzeptanz neuer medizinischer Technologien aus unterschiedlichen Bedingungen hervorgeht, ist eine analytische Trennung dieser Perspektiven erforderlich. Ärztliche Akzeptanz einer (neuen) Technologie wird typischerweise über wahrgenommene Nützlichkeit und Nutzerfreundlichkeit beschrieben, patientenseitige Akzeptanz über Vertrauen und Kommunikation. Die Kapitelstruktur folgt dieser Differenzierung und entwickelt die Analyse von kognitiven Automatisierungseffekten über affektive und verantwortungsbezogene Aspekte bis hin zu Fragen der Akzeptanz und organisatorischen Einbettung.

1. Kognitive Konsequenzen KI-basierter Automatisierung

1.1 Theoretischer Rahmen

Eine kognitive Betrachtung des KI-Einsatzes in der Radiologie beginnt damit, festzulegen, welche Aufgaben automatisiert werden und

wie menschliches Handeln daran anschließt. Automatisierung kann die ärztliche Aufmerksamkeit ausrichten, Entscheidungen vorbereiten oder die Reihenfolge und Weiterleitung von zu beurteilenden Fällen bestimmen. Da diese Eingriffe an unterschiedlichen Punkten der Informationsverarbeitung ansetzen, verändern sie die Art, wie Fehler und Abhängigkeiten entstehen (Parasuraman et al., 2000). Abhängig von dieser Ausgestaltung bleibt diagnostisches Entscheiden eine primär menschliche Tätigkeit oder nimmt die Form einer nachgeordneten Prüfung an. Eine solche Verschiebung in Richtung Nachprüfung stellt erhöhte Anforderungen an die kognitive Kontrolle, da geringere aktive Beteiligung mit einer erhöhten Anfälligkeit für Nachlässigkeit einhergehen kann (Parasuraman & Riley, 1997). Dasselbe KI-System kann dadurch die diagnostische Treffgenauigkeit erhöhen und zugleich die Art verändern, wie Fehler zustande kommen. Vor diesem Hintergrund lässt sich zwischen kurzfristigen Veränderungen einzelner Bearbeitungssituationen und längerfristigen Anpassungen diagnostischer Routinen unterscheiden. Die folgenden Abschnitte behandeln die zugrunde liegenden Mechanismen sowie Ansatzpunkte zur Begrenzung vorhersehbarer Fehlentwicklungen.

1.2 Kurzfristige kognitive Auswirkungen

Radiologische Bildbefundung beruht auf visueller Suche unter Bedingungen niedriger Ereignishäufigkeit und hoher Ähnlichkeit zwischen relevanten Signalen und Hintergrundstrukturen. Auffälligkeiten treten selten auf und heben sich oft nur schwach vom übrigen Bildmaterial ab. Unter diesen Bedingungen verändert sich mit zunehmender Suchdauer die statistische Gewichtung möglicher Befunde: Die Wahrscheinlichkeit, weitere Auffälligkeiten zu identifizieren, nimmt ab, und die gezielte Aufrechterhaltung von Aufmerksamkeit über die Zeit wird schwieriger, selbst bei hoher Expertise (Wolfe et al., 2005). KI greift in diesen Suchkontext ein, indem sie Hinweise bereitstellt oder bestimmte Bildregionen vorab markiert. Dadurch verschiebt sich, welche Bereiche bevorzugt geprüft werden, für wie lange die Suche fortgeführt wird und in welchem Maß Exploration offen bleibt. Solche Vorgaben können den Zugang zu relevanten Bildbereichen beschleunigen. Zugleich organisieren sie die Suche stärker entlang vorgegebener Relevanzen. Wird Aufmerksamkeit

wiederholt an markierte Bereiche gebunden, steigt die Wahrscheinlichkeit, dass relevante, aber nicht hervorgehobene Befunde außerhalb des Suchfokus bleiben. Hinzu kommt, dass die visuelle Suche häufig endet, sobald ein plausibler Befund identifiziert ist; alternative Deutungen werden dann seltener weiterverfolgt, auch wenn zusätzliche Hinweise vorhanden wären (Berbaum et al., 1990; Drew et al., 2013).

Der Einsatz diagnostischer KI-Hinweise verschiebt die kognitive Aufgabe der Befundung. Statt eigenständig nach Auffälligkeiten zu suchen, richtet sich die Tätigkeit zunehmend auf die Prüfung eines bereits vorgeformten Vorschlags. Diese Umstellung begünstigt Automatisierungsverzerrungen, bei denen maschinelle Empfehlungen bevorzugt akzeptiert und eigenständige Suchbewegungen verkürzt werden (Parasuraman & Riley, 1997). Fehler entstehen dabei auf zwei Wegen. Zum einen können falsche Hinweise direkt übernommen werden; zum anderen wird das Ausbleiben eines Hinweises selbst als Information interpretiert und kann als stillschweigende Entwarnung wirken. Unter diesen Bedingungen werden schwache oder atypische Auffälligkeiten besonders anfällig fürs Übersehen, da sie weder visuell hervorstechen noch algorithmisch markiert sind. Gleichzeitig können hoch saliente Markierungen die Interpretation in eine bestimmte Richtung lenken und harmlose Strukturen als pathologisch erscheinen lassen. Studien zeigen, dass Zeitdruck, autoritative Ergebnisdarstellung und hohe Erwartungen an die Systemverlässlichkeit diese Effekte verstärken (Goddard et al., 2012; Dratsch et al., 2023). Der Bilddatensatz wird damit weniger als offener Suchraum behandelt, sondern als Grundlage zur Verifikation einer vorgegebenen Annahme. Diese Verschiebung der Verifikationspraxis verändert Art und Wahrscheinlichkeit von Fehlern systematisch und stellt ein zentrales kurzfristiges Sicherheitsrisiko dar.

Der Einsatz von KI verändert die Organisation radiologischer Arbeit entlang des diagnostischen Arbeitsgangs. Automatisierte Funktionen übernehmen vor allem wiederkehrende Prüfschritte und Vorstrukturierungen, wodurch sich die verbleibenden Aufgaben stärker auf Situationen konzentrieren, in denen Unsicherheit, Interpretation und Verantwortung zusammenwirken. Die diagnostische Arbeit verschiebt sich damit in Richtung selektiver Bearbeitung solcher Konstellationen, die eine Einordnung KI-basierter Hinweise erfordern. Besonders relevant sind Fälle, in denen Systemvorschläge

und eigene Einschätzungen auseinanderfallen. In diesen Situationen wird diagnostisches Entscheiden mit der Aufgabe verbunden, die eigene Bewertung nachvollziehbar zu begründen und gegenüber KI-gestützten Hinweisen einzuordnen. Diese Begründungstätigkeit gehört dann zum regulären Arbeitsprozess und prägt die diagnostische Praxis zunehmend. Parallel dazu erfordert der Einsatz von KI eine fortlaufende Aktualisierung der Einschätzung der Leistungsfähigkeit des KI-Systems, da die Aussagekraft KI-basierter Hinweise vom jeweiligen Kontext abhängt. Daraus ergeben sich veränderte Belastungsprofile im Arbeitsablauf. Unterschiedlich anspruchsvolle Falltypen folgen in dichter Abfolge aufeinander und machen wiederholte Übergänge und schnelle Wechsel zwischen Kontroll- und Analyseformen erforderlich. Abgesehen von lange bekannten Effizienzeinbußen durch häufige Aufgabenwechsel (Kiesel et al., 2010), richtet sich die ärztliche Aufmerksamkeit dabei möglicherweise stärker auf die Plausibilität, Konsistenz und Anschlussfähigkeit diagnostischer Ergebnisse. Wie sich diese Verschiebungen auswirken, hängt von der konkreten Einbindung der KI in bestehende Routinen ab (Dzindolet et al., 2003; Parasuraman & Manzey, 2010).

1.3 Langfristige kognitive Auswirkungen

Langfristig verändert der Einsatz von KI in der Radiologie nicht nur kognitive Prozesse im diagnostischen Arbeitsablauf, sondern die Bedingungen, unter denen diagnostische Kompetenz entsteht und stabil bleibt; er verändert also die ärztliche Expertise. Radiologische Expertise beruht wesentlich auf der fortgesetzten aktiven Auseinandersetzung mit Bildmaterial: visuelle Erkennungsleistungen im radiologischen Screening gehören zu den Bereichen, in denen die ärztliche Expertise ihre besondere Stärke entfaltet. Es handelt sich dabei nicht um das bloße Wiedererkennen bekannter Muster, sondern um eine aktive, suchende Exploration des Bildes mit dem Ziel, potenziell relevante Abweichungen überhaupt erst zu identifizieren. Diese Form der Leistung ist offen angelegt und auf Sensitivität gegenüber Unerwartetem ausgerichtet. Darin unterscheidet sie sich grundlegend von nachgelagerten Bewertungsprozessen, bei denen bereits identifizierte Befunde eingeordnet, relativiert und diagnostisch gewichtet werden. Visuelle Differenzierung, Mustererkennung und der Umgang mit

Unsicherheit werden dabei weniger durch abstraktes, explizites Faktenwissen als durch kontinuierliche praktische Anwendung aufrechterhalten. KI-gestützte Automatisierung verschiebt jedoch die funktionale Gliederung der Teiltätigkeiten im diagnostischen Prozess. Aufgaben, die KI zuverlässig übernimmt, fallen damit als regelmäßige Übungsgelegenheiten weg, die für den Erhalt diagnostischer Fertigkeiten erforderlich sind. Über längere Zeiträume kann KI somit zu einem Qualifikationsverlust (»deskilling«: Verlust von Fertigkeiten) führen: Kompetenzen, die nicht fortlaufend aktiviert werden, verlieren an Präzision, Geschwindigkeit und Robustheit. Langfristig entsteht so ein systemisches Risiko, da zunehmende KI-Abhängigkeit bei gleichzeitig abnehmender menschlicher Expertise die diagnostische Robustheit und damit die radiologische Versorgungssicherheit untergraben kann (Bainbridge, 1983; Endsley & Kiris, 1995).

Radiologische KI-Nutzung kann langfristig aber auch neue Formen fachlicher Kompetenz (»upskilling«) hervorbringen – jedoch nur unter der Voraussetzung, dass der Umgang mit der KI selbst zum expliziten Gegenstand des Lernens wird. Kompetenz verlagert sich dann von der isolierten Erkennung einzelner Auffälligkeiten hin zur sachgerechten Steuerung eines KI-gestützten diagnostischen Systems. Diese Form der Expertise besteht in konkreten operativen Fertigkeiten: der kontextsensitiven Einschätzung der Zuverlässigkeit von KI-Ausgaben, dem Wissen um typische Fehlmuster, dem reflektierten Umgang mit Abweichungen zwischen eigener Einschätzung und Systemvorschlag sowie der systematischen Integration von KI-Ergebnissen in weitere klinische Informationen. Diese Kompetenzverschiebung bleibt jedoch aus, wenn sie stillschweigend vorausgesetzt wird. Ohne explizite organisationale und didaktische Rahmung entstehen weder stabile Lernprozesse noch ein belastbarer Erhalt dieser Kompetenzen. Parallel dazu formt langfristige KI-Nutzung diagnostische Strategien selbst: Über die Zeit verändern sich Suchroutinen, Prüftiefen und Abbruchkriterien; Befundung kann sich von offener Exploration zu einer Praxis entwickeln, die primär auf Abgleich und Bestätigung ausgerichtet ist. Solche Anpassungen können in hochstrukturierten Kontexten mit hohem Durchsatz funktional sein. Problematisch werden sie dort, wo dieselben Routinen auf Situationen übertragen werden, in denen KI versagt oder nicht verfügbar ist. Ob diese Verschiebungen in erhöhte Robustheit oder in neue Formen der Fehleranfälligkeit im diagnostischen Prozess mün-

den, entscheidet die Qualität metakognitiver Überwachung (Lee & See, 2004).

Über längere Zeiträume lässt sich der Einsatz von KI als Einflussfaktor auf unterschiedliche Entwicklungspfade fachlicher Praxis analysieren. Im Zentrum steht dabei nicht die Annahme einer einheitlichen oder notwendigen Wirkung, sondern die Frage, unter welchen Bedingungen diagnostische Kontrolle stabilisiert oder geschwächt wird. Analytisch relevant ist dies auf der Ebene unterschiedlicher Praxisformen (verstanden als wiederkehrende Handlungs-, Aufmerksamkeits- und Überwachungsformen im diagnostischen Prozess), also der Weise, wie Urteil, Aufmerksamkeit und Überwachung im Zusammenspiel von Mensch und System organisiert sind. Unter bestimmten Bedingungen entstehen Routinen, in denen KI-Ausgaben fortlaufend auf Plausibilität geprüft und als vorläufige Hinweise behandelt werden. Die ärztliche Bildanalyse bleibt dann aktiv, Abweichungen werden aufgegriffen, und das Vertrauen in das System bleibt an eine kontinuierliche Überprüfung gekoppelt. Unter anderen Bedingungen kann sich hingegen eine Praxis herausbilden, in der hohe Systemleistung, Zeitdruck und seltene Fehlerereignisse die Aktivierung ärztlicher Prüfprozesse zunehmend reduzieren. Die Exploration des Bildmaterials tritt dann in den Hintergrund, während die fachliche Aufmerksamkeit auf Ausnahmen fokussiert wird. Diese Divergenz lässt sich als Ergebnis veränderter Kalibrierungs- und Überwachungsdynamiken rekonstruieren: Mit sinkender Häufigkeit korrektiver Eingriffe werden Überwachungsroutinen seltener aktiviert, während die Vigilanz (d. h. die Aufrechterhaltung der Aufmerksamkeit) unter Bedingungen niedriger Ereignisraten instabil wird. Aus epistemischer Sicht folgt daraus, dass langfristige Sicherheit weniger von der Qualität des KI-Systems allein abhängt als von organisationalen Bedingungen, die eigenständige Praxis, Rückmeldung und Überwachung der Systemperformanz dauerhaft absichern.

1.4 Maßnahmen gegen negative Auswirkungen der Automatisierung

Maßnahmen gegen negative Auswirkungen KI-basierter Automatisierung lassen sich nur dann angemessen bestimmen, wenn das zugrunde liegende Problem korrekt gefasst wird. Die relevanten Effekte

ergeben sich nicht primär aus individuellen Fehlhaltungen, sondern aus stabilen, strukturell bedingten Interaktionsmustern zwischen technischen Systemen, Arbeitsumgebung und Nutzungskontext. Entsprechend sind sie nicht als subjektive Kommunikationsdefizite zu analysieren, sondern als Folgen bestimmter Gestaltungs- und Organisationsentscheidungen. Appellative Hinweise wie »Vertrauen Sie der KI nicht zu sehr« setzen am bewussten Urteil an, während viele problematische Effekte aus Routinisierung, Zeitdruck, Aufgabenstruktur und systemseitigen Voreinstellungen hervorgehen. Zwar kann es situativ notwendig sein, Aufmerksamkeit gezielt zu bündeln, etwa durch Fokus-Instruktionen (Steinborn, Langner & Huestegge, 2017); dies bleibt jedoch punktuell. Wo Handeln überwiegend durch Arbeitsumgebung und Werkzeuglogik geprägt ist, ersetzen instruktionale Eingriffe keine dauerhafte Prävention. Aus arbeitspsychologischer Perspektive sind Fehlgebrauch, Nichtgebrauch und Missbrauch von Technologie zu erwartende Resultate spezifischer Kopplungen von Tätigkeit, Technik und Kontext (Parasuraman & Riley, 1997). Daraus folgt, dass wirksame Gegenmaßnahmen als Systemeigenschaften zu konzipieren sind. Sie müssen in Schnittstellendesign, Befundungsprotokolle, zeitliche Platzierung von KI-Ausgaben, Schulungsformate sowie Rückmelde- und Überwachungsschleifen eingebettet sein und kontinuierlich wirken, unabhängig von individueller Wachsamkeit.

Welche Gegenmaßnahmen angemessen sind, lässt sich nur in Abhängigkeit davon bestimmen, auf welcher Ebene KI in den Arbeitsprozess eingreift. Unterschiedliche Eingriffsebenen verändern unterschiedliche kognitive und organisatorische Prozesse und erzeugen entsprechend verschiedene Risikomuster. Eine einheitliche Behandlung von Automatisierungseffekten wäre daher analytisch nicht haltbar. Systeme, die Aufmerksamkeit lenken, greifen nicht in Urteile selbst ein, sondern in die Such- und Wahrnehmungsdynamik, aus der Urteile hervorgehen. Aufmerksamkeit fungiert dabei als eine Art erkenntnisleitender Vorfilter: Informationen, die nicht in den Fokus gelangen, können später weder geprüft noch korrigiert werden. Vorstrukturierte Aufmerksamkeitslenkung kann daher dazu führen, dass relevante Aspekte systematisch unberücksichtigt bleiben. Der resultierende Fehler entsteht vor dem eigentlichen Urteil und bleibt häufig unbemerkt, da Sucharbeit implizit an das System delegiert wird. Entscheidungsunterstützende Systeme verlagern das Risiko

an eine andere Stelle. Hier steht nicht unterlassene Wahrnehmung im Vordergrund, sondern die Übernahme fehlerhafter Ergebnisse oder das Ausbleiben weiterer Prüfung. Der Fehlerort liegt in der Abschlussphase des Urteilsprozesses. Die Forschung zur Automatisierungsverzerrung zeigt, dass fehlerhafte KI-Empfehlungen handlungsleitend bleiben können, selbst wenn die fachlichen Kompetenzen für eine Korrektur prinzipiell gegeben sind (Goddard et al., 2012; Gaube et al., 2021). Fehleinschätzungen sind damit nicht allein als Mangel an Wissen oder Expertise zu erklären, sondern als Ergebnis spezifischer Prozesskonstellationen, in denen eine menschliche Korrekturleistung bezüglich der KI-Ausgabe nicht aktiviert oder nicht umgesetzt wird. KI-basierte Triage-Systeme greifen schließlich auf organisatorischer Ebene ein, indem sie klinische Priorisierung und Arbeitslast strukturieren. Unter Bedingungen hohen Durchsatzes können sie schleichende Nachlässigkeiten bei wenig salienten Abweichungen begünstigen, wie die Automatisierungsforschung zeigt (Parasuraman et al., 2000).

Aus dieser Differenzierung folgt, dass Gegenmaßnahmen jeweils systemspezifisch anzusetzen sind, da Interventionen ihre Wirksamkeit erst dann entfalten, wenn sie an die jeweilige Automatisierungsform und ihren funktionalen Eingriffspunkt gekoppelt sind. Entsprechend ist es erforderlich, zwischen diversen Vermittlungsmechanismen zu unterscheiden, anstatt Automatisierungseffekte als homogenen Problemtyp zu behandeln. Wirksame Gegenmaßnahmen müssen dabei gezielt an den jeweiligen Vermittlungsprozessen ansetzen, statt allgemein an Wachsamkeit oder Einstellung zu appellieren. Tabelle 1 zeigt eine exemplarische Liste konkreter Gestaltungsmöglichkeiten basierend auf der zuvor entwickelten Analyse; sie macht explizit, wie abstrakte Prinzipien in praktische Maßnahmen zur Reduktion unerwünschter kognitiver Konsequenzen der KI-Nutzung in der Radiologie übertragen werden können. Jede Maßnahme ist dabei einem spezifischen Risiko und einem klar benannten Wirkmechanismus zugeordnet, sodass der jeweilige Eingriffspfad nachvollziehbar bleibt. Auf diese Weise wird deutlich, dass Risikominderung nicht durch allgemeine Vorsicht oder erhöhte Aufmerksamkeit erreicht wird, sondern durch gezielte Eingriffe in kognitive und organisationale Prozesse. Die Tabelle folgt damit demselben Grundprinzip wie die vorausgehende Argumentation: Wirksamkeit entsteht aus der Passung zwischen Automatisierungsform, vermitteltem Pro-

zess und gewählter Intervention. Die ersten Maßnahmen adressieren Verzerrungen im unmittelbaren Umgang mit KI-Ergebnissen, etwa bei Ergebnisübernahme und Prüfungstiefe; weitere Maßnahmen zielen auf Aspekte der Arbeitsgestaltung und auf den langfristigen Erhalt fachlicher Kompetenz.

Tabelle 1: *Maßnahmen zur Risikominderung bei KI-basierter Automatisierung in der Radiologie*

	Maßnahme	**Zielt auf Risiko**	**Mechanismus**
1	Zweitleser-Prinzip: KI als unabhängige Kontrollinstanz bei ausgewählten Aufgaben	Verankerung: vorschnelle Fixierung auf erste Hypothesen	Unabhängigkeit: eigenständige Urteilsbildung vor KI-Konsultation
2	Verifikationsprotokoll: obligatorischer KI-Check vor Befundabschluss	Auslassungsfehler: übersehene oder fälschlich ergänzte Befunde	Aktivierung: bewusste Prüfung statt passiver Übernahme
3	Transparenz: Anzeige aufgabenspezifischer Unsicherheiten und Grenzen	Übervertrauen: überhöhte Glaubwürdigkeit durch scheinbare Objektivität	Kalibrierung: realistisches mentales Modell der KI-Fehlermodi
4	Dezenz: zurückhaltende Gestaltung ohne alarmartige Hervorhebungen	Aufmerksamkeitsbindung: dominante KI-Markierungen lenken ab	Balance: Erhalt der eigenständigen visuellen Exploration
5	Integration: gebündelte Ergebnisdarstellung ohne Systemwechsel	Fragmentierung: kognitive Last durch Multi-Tool-Management	Effizienz: reduzierter Aufwand für Zusammenführung der Information
6	Fehlertraining: Kalibrierung anhand lokaler Fehlerfälle und Audits	Abhängigkeit: schleichende Gewöhnung an KI-Unterstützung	Sensibilisierung: Kenntnis der Randbedingungen und Grenzen
7	Kompetenzerhalt: regelmäßige KI-freie Befundung als Übung	Dequalifizierung: Erosion diagnostischer Fertigkeiten	Praxis: Erhaltung der Kompetenz für seltene Befundmuster

Anmerkung. KI = Künstliche Intelligenz. Die Maßnahmen 1–3 adressieren kognitive Verzerrungen; Maßnahmen 4–7 betreffen strukturell-organisatorische Aspekte sowie den Kompetenzerhalt.

2. Motivationale und emotionale Aspekte KI-basierter Automatisierung

2.1 Identität, Autonomie und Bedeutung der Arbeit

Der Einsatz von KI in der Radiologie verändert nicht nur Abläufe, sondern die Bedeutungsstruktur professioneller Arbeit. Motivational relevant wird diese Veränderung dort, wo sich verschiebt, was als zentrale Leistung gilt und woran professionelle Anerkennung gebunden ist. KI-basierte Automatisierung wirkt damit nicht nur auf Effizienz, sondern auf das Verständnis dessen, was fachliche Leistung ausmacht. Diese Bedeutungsverschiebung lässt sich funktional an der veränderten Tätigkeitslogik festmachen. Arbeit verlagert sich von unmittelbarer Bildauswertung hin zur Steuerung eines diagnostischen Systems. Entsprechend gewinnen Kompetenzen an Gewicht, die auf Integration, Priorisierung und Beurteilung algorithmischer Ergebnisse zielen, während klassische Wahrnehmungs- und Klassifikationsleistungen an Sichtbarkeit verlieren. Der Wandel betrifft damit nicht nur einzelne kognitive Fertigkeiten, sondern auch die symbolische Ordnung professioneller Kompetenz. Diese Neuordnung bleibt nicht äußerlich, sondern wird aktiv in das berufliche Selbstverständnis eingebaut. Identität ist hier nicht als stabile Eigenschaft zu verstehen, sondern als fortlaufender Aushandlungsprozess im Umgang mit sich verändernden Anforderungen. Empirische Arbeiten zeigen, dass diese Anpassung ambivalent ausfallen kann (Perez et al., 2024): Einerseits kann die Reduktion von Routine und die Betonung komplexer Koordination als Aufwertung erlebt werden; andererseits entstehen Spannungen, wenn unabhängige Wahrnehmung und eigenständiges ärztliches Urteil als zentrale Marker professioneller Anerkennung an Bedeutung verlieren. Theorien beruflicher Identitätsbildung verorten solche Spannungen in der Aushandlung von Normen, sozialer Anerkennung und symbolischen Kompetenzmarkern (Cruess et al., 2014; Shonhe & Min, 2025).

Status und Autorität ergeben sich im Arbeitsvollzug nicht allein aus formaler Verantwortung, sondern aus spezifischen Kompetenzmarkern, die im praktischen Handeln sichtbar und erfahrbar werden. KI-basierte Automatisierung kann diese Marker verschieben, indem neue Referenzpunkte für Richtigkeit und Verlässlichkeit etabliert werden. Werden im Arbeitsteam algorithmische Ergebnisse

zur primären Vergleichsfolie, verändert sich die Zuschreibung von Autorität – mitunter auch dann, wenn menschliche Expertise weiterhin funktional erforderlich bleibt. Diese Verschiebung betrifft weniger Kompetenz im engeren Sinn als ihre Wahrnehmbarkeit und soziale Zuschreibung; Expertise wird damit nicht aufgehoben, sondern anders verankert. Automatisierte oder vorlagenbasierte Berichte verstärken diesen Effekt, indem sie die radiologische Tätigkeit als Überwachung, Bestätigung oder Abweichungskontrolle rahmen und sie weniger als eigenständige Problemlösung erscheinen lassen. Dadurch verändert sich die performative Darstellung fachlicher Kompetenz im Arbeitsprozess und damit auch ihre soziale Sichtbarkeit. Qualitative Studien zeigen, dass auf solche Verschiebungen mit diskursiven Strategien reagiert wird, die Kompetenz jenseits reiner Bildinterpretation oder Ergebnisübernahme markieren (Lombi & Rossero, 2024). Autorität wird dabei neu begründet, etwa über Kontextwissen, Koordinationsleistung oder die Fähigkeit, algorithmische Ergebnisse situativ einzuordnen. KI verändert damit nicht nur Arbeitsprozesse, sondern auch die Bedingungen, unter denen fachliche Autorität hergestellt und anerkannt wird.

Diese Verschiebung lässt sich präzisieren, wenn Status und Autorität nicht als formale Zuschreibungen, sondern als relationale und erlebte Größen gefasst werden. Autorität ist dann weniger an Zuständigkeit oder Haftung gebunden als an die Erfahrung, im Handlungsvollzug wirksam beteiligt zu sein. Maßgeblich dabei ist, ob das eigene Handeln als Quelle relevanter Effekte erfahren wird. Ein solches Erleben ist agentisch, insofern es Situationen markiert, in denen Eingriffe nicht nur ausgeführt, sondern als wirksam erlebt werden. Autorität hängt in dieser Perspektive eng mit Selbstwirksamkeit zusammen, verstanden als Erwartung, durch eigenes Handeln kompetent Einfluss nehmen zu können. Sie entsteht nicht aus beruflicher Position und organisationaler Hierarchie, sondern aus der wiederholten Erfahrung, dass eigenes Tun Bedeutung für den Verlauf der Tätigkeit hat. Vor diesem Hintergrund wird deutlich, dass KI-basierte Automatisierung in der Radiologie nicht nur diagnostische Arbeitsabläufe verändert, sondern die Verteilung solcher Wirksamkeitsmarker. Wenn algorithmische Outputs zur maßgeblichen Referenz für Richtigkeit werden, verschiebt sich die Zuschreibung von Wirksamkeit vom handelnden Subjekt zum System, selbst bei fortbestehender menschlicher Verantwortung. Der zentrale Ef-

fekt liegt dabei nicht im Verlust formaler Zuständigkeit, sondern in der Umdeutung der Tätigkeit. Sichtbare Spuren eigener Kompetenz werden seltener, agentische Marker abgeschwächt. Automatisierung wirkt damit direkt auf das Erleben von Handlungsmacht, indem sie die Erfahrungsbedingungen wirksamen Handelns verändert.

Ob jedoch die KI-Nutzung in der Radiologie die wahrgenommene ärztliche Handlungsmacht untergräbt oder stabilisiert, lässt sich nur im Zusammenhang mit ihrem institutionellen Status bestimmen. Entscheidend ist, ob KI als unterstützende Ressource oder als autoritative Instanz in die Organisation eingebettet ist. Diese Unterscheidung ist nicht graduell, sondern betrifft die Art der Zuschreibung von Beiträgen und Kontrolle. Als unterstützende Ressource kann KI Handeln erweitern, ohne die Zuschreibung von Kompetenz grundlegend zu verschieben: Das eigene Urteil bleibt der Referenzpunkt, während algorithmische Ausgaben zusätzliche Orientierung liefern. Wird KI hingegen zur autoritativen Instanz, verändert sich, wem Beiträge zur Richtigkeit der Diagnostik zugerechnet werden. In dieser Konstellation bleibt fachliche Expertise zwar funktional notwendig, verliert jedoch an Sichtbarkeit und Anerkennungswert. Die beobachteten diskursiven Strategien lassen sich vor diesem Hintergrund als aktive Versuche verstehen, unter veränderten Zuschreibungsbedingungen agentische Selbstverortung zu bewahren und Kompetenz jenseits algorithmischer Referenzpunkte sichtbar zu halten. Autonomie zeigt sich dabei nicht in regelkonformer Ausführung, sondern dort, wo eigenes Handeln als wirksam erlebt wird. Konfigurierbare KI-Systeme, die dem klinischen Urteil eindeutig nachgeordnet bleiben, können Handlungsspielräume sichern. Wird der Einsatz von KI dagegen verpflichtend und Workflow-bestimmend (z. B. erforderliche Eingabeaufforderungen, erzwungene Triage, Audits), verschiebt sich das Erleben von Autonomie hin zu externer Kontrolle und Überwachung, was wiederum die intrinsische Motivation verringert und zu konformitätsorientiertem, unkritischem Verhalten im Umgang mit KI-Ergebnissen führen kann. Autonomie erweist sich damit als Ergebnis konkreter Design- und Organisationsentscheidungen bezüglich der Einbettung von KI-Anwendungen, nicht als stabile Eigenschaft professioneller Rollen.

Die Einführung von KI-Anwendungen in der Radiologie kann also die Bedeutung professioneller Arbeit verändern und damit unmittelbar auf die Arbeitsmotivation wirken. Diese Veränderung ist

jedoch über diverse KI-Anwendungen nicht einheitlich, sondern teils sogar durch gegenläufige Effekte gekennzeichnet: Der Wegfall von Routinetätigkeiten kann Arbeit als anspruchsvoller, interessanter und wertiger erscheinen lassen, während zugleich der Verlust von Eigenverantwortung und sichtbarer Urheberschaft dazu führen kann, dass Arbeit an Bedeutung verliert. Motivation entsteht aus dem Zusammenspiel dieser beiden Effekte, nicht aus einem davon isoliert. Die Bedeutungszuschreibung ist dabei handlungsrelevant: Sie beeinflusst, ob zusätzliche Prüfungen vorgenommen, Zweitbegutachtungen eingeholt oder Diskrepanzen aktiv als Lerngelegenheiten genutzt werden. Motivation fungiert hier als vermittelnde Größe zwischen der Wahrnehmung von Wertigkeit und Verantwortung einerseits und konkretem sicherheitsrelevantem Verhalten andererseits. Qualitative Interviews zeigen, dass radiologische KI-Tools bislang selten formale Entscheidungsautonomie entziehen, jedoch die Stellung ärztlicher (menschlicher) Kompetenz und Autorität als Referenz für Richtigkeit in Frage stellen (Lombi & Rossero, 2024). Der Eingriff erfolgt damit nicht über explizite Restriktionen, sondern über eine veränderte symbolische Ordnung der Tätigkeit. In dieser Konstellation können Stress und Bedrohungserleben jene kognitiven Muster (z. B. Automatisierungsverzerrungen) verstärken, die in Kapitel 1 als sicherheitsrelevant beschrieben wurden. Motivation wird so zu einem Faktor, der bestehende Risikodynamiken abschwächen oder verstärken kann, abhängig davon, wie Bedeutung und Verantwortung im Arbeitsvollzug erlebt werden.

2.2 Emotionale Reaktionen: Unsicherheit, Sorgen und Erleichterung

Emotionale Konsequenzen des KI-Einsatzes in der Radiologie speisen sich nicht ausschließlich aus konkreten Nutzungserfahrungen, sondern in erheblichem Maße aus Erzählungen über die ärztliche Ersetzbarkeit durch KI. Solche Narrative entfalten ihre affektive Wirkung auch dann, wenn sie nicht als realistische Prognosen akzeptiert werden, denn sie erzeugen eine anhaltende Unsicherheit darüber, welchen Wert fachliche Expertise künftig besitzt, wie sich Karriereverläufe entwickeln und welche Bedeutung die erfahrene Aus- und Weiterbildung behält (Coppola et al., 2021). Diese Unsicherheit ist

nicht punktuell, sondern strukturell, da sie Erwartungen über zukünftige Handlungs- und Anerkennungsbedingungen betrifft. Die Verteilung dieser Unsicherheit ist dabei sozial und biografisch differenziert: In frühen Karrierestufen richtet sie sich vor allem auf Fragen zur (vermeintlichen) Automatisierbarkeit des radiologischen Tätigkeitsfeldes, während sie in späteren Phasen die Neubewertung bereits aufgebauter Expertise betrifft. Damit verschiebt sich der affektive Fokus von antizipierter Eintrittsunsicherheit zu retrospektiver Sinn- und Wertprüfung. Internationale Umfragen und qualitative Arbeiten zeigen, dass solche Sorgen Einstellungen zur Einführung von KI systematisch prägen (Gong et al., 2019; Huisman et al., 2021; Rony et al., 2024; Dang & Li, 2025). Die Ausprägung dieser Effekte variiert dabei mit radiologischer Erfahrung, Wissen über KI und der wahrgenommenen institutionellen Ausrichtung. Emotionale Reaktionen sind damit nicht hauptsächlich Ausdruck individueller Dispositionen, sondern Ergebnis spezifischer Erwartungs- und Deutungsrahmen, in denen die KI-Nutzung verortet wird.

Emotionale Reaktionen im praktischen Umgang mit KI-Anwendungen entstehen bevorzugt dann, wenn die eigene (ärztliche) Einschätzung mit einem KI-basierten Ergebnis divergiert. Eine solche Abweichung markiert nicht automatisch einen Fehlerhinweis, sondern zunächst einen offenen Bewertungsmoment, der eine Positionierung erforderlich macht. Aus dieser Situation können unterschiedliche affektive Dynamiken hervorgehen: Die Abweichung kann zum Anlass werden, die eigene Begründung zu prüfen, Evidenz erneut zu sichten und den Fall aktiv weiterzubearbeiten, was emotional als positiv erlebt wird. Sie kann jedoch auch als Hinweis auf die Unzulänglichkeit der eigenen Einschätzung interpretiert werden, sodass sich Kompetenzängste herausbilden und sich die weitere Bearbeitung zunehmend an der algorithmischen Vorgabe orientiert. In der ersten, emotional positiven Dynamik bleibt die ärztliche Urteilspraxis aktiv und die KI fungiert als Kontrastfolie für Präzisierung und Lernen; in der zweiten, emotional eher negativen Dynamik zieht sich die handelnde menschliche Instanz aus der initiativen Beurteilung zurück. Welche dieser Dynamiken einsetzt, ergibt sich vor allem aus lokalen Bewertungs- und Rückmeldestrukturen. Entscheidend dabei ist, ob Abweichungen zwischen eigener ärztlicher Einschätzung und KI-Systemausgabe als legitimer Bestandteil professioneller Urteilsbildung behandelt werden oder als erklärungsbedürfti-

ge Fehler. Affektive Reaktionen fungieren in diesem Sinne nicht als Störgrößen, sondern als Umschaltpunkte erhöhter Sensitivität, an denen sich entweder aktive Urteilspraxis und Lernorientierung stabilisieren oder Abhängigkeit und Passivierung ausbilden.

Emotionale Reaktionen im Umgang mit KI lassen sich auch als Indikatoren dafür verstehen, wie gut KI funktional in den radiologischen Arbeitsvollzug integriert ist, also ob die ärztliche Interaktion mit dem jeweiligen KI-Tool als unterstützend und entlastend oder als reibungsreich und anstrengend erlebt wird. Negativer Affekt entsteht im praktischen Umgang nämlich häufig aus dem Erleben operativer Schwierigkeiten im Arbeitsprozess. Wiederholte Fehlalarme, schwer integrierbare Darstellungen oder zusätzlicher Dokumentationsaufwand erhöhen die Interaktionskosten und verkomplizieren den Arbeitsprozess, auch dann, wenn das KI-System in anderen Situationen hilfreiche Beiträge liefert. Daneben lassen sich aber auch klar konturierte positive Affekte bei der KI-Nutzung beobachten: Erleichterung entsteht dort, wo KI monotone, zeitaufwendige und gleichzeitig aufmerksamkeitsfordernde Teiloperationen übernimmt und dadurch im selben Handlungskontext ein spürbarer Übergang zu geringerer Beanspruchung erfahren wird. Die affektive Qualität resultiert dabei aus der wahrgenommenen Diskontinuität zwischen einer Phase hoher Belastung und einem nachfolgenden Zustand erweiterter Handlungsspielräume. Darüber hinaus kann die KI-Nutzung auch affektiv positiv bewertete Selbstzuschreibungen wie Stolz erzeugen, z. B. wenn Abweichungen zwischen eigener (korrekter) Einschätzung und algorithmischem Output begründet und der eigenen Leistungsfähigkeit zugerechnet werden können. In diesem Moment wird KI als begrenztes System erfahrbar, an dessen Grenzen menschliche Kompetenz sichtbar wird. Erleichterung markiert damit gelingende funktionale Entlastung im zeitlichen Vollzug der Arbeit, Stolz eine gelingende Selbstzuschreibung von Kompetenz im Urteilsvollzug.

Aktuelle theoretische Ansätze gehen davon aus, dass emotionale Reaktionen keine epiphänomenalen Begleiterscheinungen sind, sondern integraler Bestandteil kognitiver Regulation mit eigener funktionaler Rolle (Damasio, 1994; Cosmides & Tooby, 2000; Schwarz & Clore, 2007). Sie sind systematisch an Situationen gebunden, in denen Handlungen und Entscheidungen als folgenreich antizipiert werden; in diesem Sinne tragen Emotionen zur Gewichtung noch

nicht vollständig spezifizierter Konsequenzen bei. Sie wirken als regulatorische Signale, die Relevanz strukturieren, Prioritäten setzen und Orientierung unter Unsicherheit ermöglichen (Loewenstein et al., 2001; Lerner et al., 2015). Im diagnostischen Arbeiten mit KI greifen emotionale Reaktionen früh in Wahrnehmung, Aufmerksamkeitsverteilung, Entscheidungsfindung und Handlungssteuerung ein. Sie modulieren, wie breit oder fokussiert die Aufmerksamkeit ausgerichtet wird, wie lange Ambiguität toleriert wird und wie Abweichungen zwischen eigener Einschätzung und algorithmischem Ergebnis gewichtet werden. Diese Wirkungen sind kontextabhängig differenziert: Emotionale Reaktionen entfalten keine einheitliche Wirkungsrichtung, sondern modulieren kognitive Prozesse abhängig von Intensität, Belastung und verfügbarer Handlungsfreiheit. Zustände wie Angst, Unsicherheit oder Überforderung können Aufmerksamkeit verengen und Prüfprozesse verkürzen; moderate Anspannung kann dagegen die Sensitivität für Inkonsistenzen erhöhen und eine sorgfältigere Prüfung unterstützen. Man kann also sagen, Emotionen fungieren als ein Bestandteil der Prozesssicherheit, allein schon deswegen, weil sie mitbestimmen, wie ärztliche Urteile gebildet werden und wie stabil diese Urteile gegenüber den Grenzen, aber auch Verheißungen technischer Systeme bleiben (Croskerry et al., 2013).

2.3 Angst vor Fehlern und moralischer Distress

Die Frage der Verantwortung für diagnostische Entscheidungen (s. Abschnitt 3.1) bildet einen zentralen emotionalen Bezugspunkt, an dem Technologieeinsatz, Vertrauen und wahrgenommene moralische Verpflichtung zusammenlaufen. Nach geltender Gesetzgebung verbleibt die Verantwortung für klinische Entscheidungen bei der Ärzteschaft, aber im praktischen Einsatz kann diese Zuordnung de facto verschoben werden, wenn KI-Empfehlungen zugleich als maßgeblich und in ihrer Entstehung nicht nachvollziehbar wahrgenommen werden. Unter solchen Bedingungen kann Unsicherheit darüber entstehen, welche Handlungsoptionen als legitim gelten. Dies ist insbesondere dann der Fall, wenn davon ausgegangen wird, dass jene ärztlichen Einschätzungen, die vom KI-Ergebnis abweichen, besonders kritisch geprüft werden oder dass die Anpassung

an algorithmische Empfehlungen administrativ sicherer ist als ein begründeter Widerspruch. Die formale Verantwortung bleibt dabei bestehen, während die erlebte und erwünschte Handlungsfähigkeit eingeschränkt wird.

Aus dieser Konstellation heraus kann die Sorge vor Fehlern Verhaltensweisen begünstigen, die primär auf Absicherung und formale Korrektheit zielen, etwa ein übermäßig detailliertes Überprüfen wenig relevanter Befunde, ausgedehnte Dokumentationspraktiken zur haftungsbezogenen Absicherung oder auch ein unverhältnismäßiges Übernehmen technischer Hinweise. Die ärztliche Kontrollfunktion verschiebt sich damit vom patientenfokussierten Prüfen von Ergebnissen hin zum Erfüllen formaler Anforderungen, womit die Technologie ihren operativen Zweck im radiologischen Arbeitsprozess verändert: Was als Unterstützung eingeführt wird, fungiert als Auslöser ärztlicher Verhaltensanpassungen (d. h. Absicherungsstrategien und »Compliance-Rituale«), die die beabsichtigten Effekte abschwächen oder umkehren (Lebovitz et al., 2021). Moralisches Belastungserleben (»moral distress«; Jameton, 1984) entsteht dabei, wenn die verantwortungsorientierte ärztliche Praxis dauerhaft mit institutionellen Erwartungen oder systemseitigen Vorgaben kollidiert, wenn also die moralische Verpflichtung zu patientenbezogener Vorsicht und Umsicht in Konflikt zu Anforderungen wie Durchsatz (also Arbeitstempo), Standardisierung oder Nutzung KI-gestützter Abläufe gerät (Kherbache et al., 2022). Diese Lage wird besonders belastend, wenn Zweifel an der Eignung des KI-Systems für bestimmte Kontexte bestehen. Persistiert die Spannung, kann sie zermürbend wirken und die Arbeitsleistung beeinträchtigen: Arbeitszufriedenheit und -motivation nehmen ab, Frustrations- und Schuldgefühle können entstehen, und professionelle Haltung kann in Zynismus kippen (Dave et al., 2023). Moralischer Distress ist damit weniger ein individuelles Problem als das Resultat anhaltender organisational-struktureller Inkongruenzen mit dem ärztlichen Selbstbild und Arbeitsethos.

Die zuvor beschriebenen Konflikte machen deutlich, dass ärztliche Verantwortung und moralische Integrität unter Bedingungen technischer Intransparenz und organisationaler Spannung ohne strukturelle Absicherung kaum tragfähig bleiben. Schutzfaktoren müssen folglich institutionell, also im soziotechnischen System selbst, verankert sein und dürfen nicht an persönliche Belastbarkeit

oder situative Selbstregulation delegiert werden. Zentrale Elemente einer solchen Absicherung sind klare Verantwortungsnormen, die Zurechnungsunsicherheit reduzieren, sowie psychologisch sichere Wege zur Eskalation KI-bezogener Bedenken, die Zweifel artikulierbar machen, ohne Sanktionen oder auch nur Sanktionserwartungen auszulösen. Partizipative Formen der KI-Implementierung in der Radiologie dabei die Kontrolle über Einsatzbedingungen stärken, während kollektive Diskussionsformate Abweichungen als gemeinsame Bearbeitungsgegenstände rahmen, statt sie zu individualisieren. Auf diese Weise bleiben Konflikte bearbeitbar und die ärztliche Aufsicht wird stabilisiert, anstatt in defensive Routinen abzugleiten. Diese Strukturen verweisen zugleich auf weiterführende Implementierungsfragen, die im weiteren Verlauf systematisch aufgegriffen werden. Zunächst richtet sich der Fokus jedoch auf Verantwortung und Erklärbarkeit als zentrale Bedingungen eines angemessenen Umgangs mit KI.

3. Verantwortung und Erklärbarkeit

3.1 Verantwortung und Rechenschaft bei hybriden Entscheidungsprozessen

Im Kontext klinischer KI-Nutzung ist Verantwortung nicht allein als rechtliche Haftungszuschreibung zu verstehen, sondern als psychologisches und organisatorisches Moment, das Handeln unter Unsicherheit strukturiert. Wenn KI in operative radiologisch-diagnostische Arbeitsschritte wie Fallvorsortierung, Mustererkennung oder Prioritätensetzung eingreift, entsteht ein hybrider Entscheidungsprozess, in dem menschliche und algorithmische Beiträge funktional verschränkt sind. Verantwortung lässt sich in solchen Konstellationen nicht mehr eindeutig lokalisieren, sondern wird entlang unterschiedlicher Ebenen gerahmt. Auf der Ebene des Entscheidungsergebnisses bleibt Verantwortung (bislang) an die ärztliche Befundung gebunden, während sie organisational häufig als verteilt konzipiert wird und die Technologie als unterstützende Instanz erscheint. Diese Gleichzeitigkeit inkonsistenter Rahmungen erzeugt Spannungen, die das Verantwortungserleben im Arbeitsvollzug beeinflussen. Aus der Sozialpsychologie ist seit langem bekannt, dass in Situationen, in

denen mehrere Betroffene oder Instanzen beteiligt sind, die vom Einzelnen wahrgenommene Verantwortung »diffundieren«, sich also ausdünnen, kann (Darley & Latané, 1968). Im Kontext der radiologischen KI-Nutzung kann dabei die Annahme vorherrschen, die Sicherheit der eingesetzten Technologie sei bereits an vorgelagerten Stellen hergestellt worden, etwa durch wissenschaftliche Validierung, technische Zertifizierung, institutionelle Beschaffung oder formale Freigabeprozesse. Diese Vorverlagerung von Zuständigkeit und Diffusion von Verantwortlichkeiten kann dann auf ärztlicher Seite zu nachlassender Wachsamkeit führen, da Risiken als bereits abgefedert oder abgesichert gelten. Zugleich können »moralische Knautschzonen« entstehen, in denen die Verantwortung für KI-gestützte Fehlentscheidungen der Ärzteschaft zugeschrieben wird, obwohl deren tatsächlicher Einfluss auf KI-Systemdesign, Trainingsdaten oder Schnittstellengestaltung begrenzt ist (Elish, 2019).

Die Rechenschaftspflicht im Kontext radiologischer KI-Nutzung ist nicht allein organisationsintern bestimmt, sondern ergibt sich aus regulatorischen Anforderungen an Risikomanagement, diagnostische Transparenz und ärztliche Aufsicht. Diese Anforderungen verleihen einer moralisch verankerten ärztlichen Verantwortung einen rechtlich verbindlichen Charakter. Um diesen Pflichten gerecht zu werden, müssen in der klinischen Praxis Kontrollprozesse konkret definiert und organisatorisch verankert werden, etwa für eine lokale Validierung und Aktualisierung von KI-basierten Anwendungen, die Dokumentation von Abweichungen sowie die Überwachung von Leistungsänderungen des KI-Systems über die Zeit. Die Umsetzung dieser Prozesse erzeugt zusätzlichen Lern- und Koordinationsaufwand und verändert etablierte Routinen, Zeitbudgets und Kompetenzanforderungen. Rechenschaftspflicht fungiert damit nicht als bloße formale Vorgabe, sondern greift unmittelbar in den radiologischen Arbeitsalltag ein. Neben der Berücksichtigung des zusätzlichen Aufwands ist dabei entscheidend, wie klar die Zuständigkeiten geregelt sind: Bleibt unbestimmt, wer welche Kontrollaufgaben trägt, entsteht Unsicherheit darüber, was erwartet wird und welche Handlungen als angemessen gelten. Diese Erwartungsunsicherheit begünstigt wiederum Stress und Abwehrreaktionen. Ob Rechenschaftspflichten bezüglich der Nutzung von KI in der Radiologie zur Stärkung von Sicherheit beitragen oder primär die ärztliche Belastung erhöhen, hängt also wesentlich vom zusätzlichen Arbeits-

aufwand sowie der Klarheit der Prozess- und Zuständigkeitsstruktur ab.

Eine weitere Herausforderung für ärztliche Verantwortlichkeit und Rechenschaft in hybriden KI-gestützten Entscheidungsprozessen ist die Schwierigkeit, Ergebnisse von KI-Systemen nachzuvollziehen und ggf. zu hinterfragen. Ohne die Möglichkeit, KI-Ergebnisse anzufechten, ist eine verantwortliche Nutzung nicht denkbar. Opake (d. h. technisch wenig oder nicht durchschaubare) Systeme, wie es typischerweise KI-Systeme sind (»Black Box«-Charakter), verschieben diese Möglichkeit, indem sie begründeten Widerspruch aufwendig machen. Denn dort, wo Erklärungen für Ergebnisse des technischen Systems nicht zugänglich sind, steigt der Aufwand für Prüfung und Begründung erheblich. Die ärztliche Letztverantwortung bleibt also formal bestehen, wird praktisch jedoch schwer einlösbar. Aus diesem Konflikt ergeben sich soziotechnische Gestaltungsfragen, etwa zur Erklärbarkeit von KI oder zur Zuständigkeit für die Widerspruchsprüfung.

3.2 Von der Erklärbarkeit zur Interpretierbarkeit

Forderungen nach erklärbarer KI beruhen häufig auf der impliziten Annahme, dass mehr Transparenz den klinischen Einsatz grundsätzlich verbessert. Diese Annahme ist nicht grundsätzlich falsch, greift jedoch zu kurz, da sie Kriterien aus der technischen Entwicklung, Überprüfung oder Regulierung unreflektiert auf situative radiologische Entscheidungskontexte überträgt. Im klinischen Alltag erfüllen Erklärungen keinen Selbstzweck, sondern sie sollen der Ärzteschaft helfen, sich *angemessen* auf KI-Tools zu verlassen, auch ohne jedes technische Detail zu kennen. D. h., sie sollen die KI-Nutzung im diagnostischen Urteilsprozess steuern, indem sie zwei symmetrische Fehlerformen begrenzen: die unkritische Übernahme von Ergebnissen wie auch ihre reflexhafte Zurückweisung. Diese Steuerung durch Erklärbarkeit erfüllt dabei in erster Linie die folgenden drei Funktionen: Sie stellt Anfechtbarkeit her (sodass KI-Ergebnisse sinnvoll hinterfragt werden können), bewirkt Lernen (sodass Fehlermodi im Laufe der Zeit besser verstanden werden) und vermittelt Vertrauenskalibrierung (sodass klar wird, wann das KI-System wahrscheinlich richtig oder falsch liegt) (Doshi-Velez & Kim, 2017; Miller, 2019).

Eine weitere wichtige Differenzierung im Kontext der Erklärbarkeit von KI-gestützten Ergebnissen in der Radiologie ist, dass verschiedene Arten von Erklärungen unterschiedlichen Zielgruppen und Zwecken dienen: Einige zielen darauf ab, einen allgemeinen Eindruck davon zu vermitteln, wie das betreffende KI-System insgesamt funktioniert; andere konzentrieren sich darauf, warum das Modell in einem bestimmten Fall ein bestimmtes Ergebnis geliefert hat; und wieder andere vermitteln das Ausmaß der Ergebnisunsicherheit oder »Was-wäre-wenn«-Szenarien. Dementsprechend bemisst sich die Relevanz von KI-Erklärbarkeit an ihrem konkreten Zweck und Nutzen im klinischen Alltag – aber auch an ihren Kosten. Denn diese diversen Erklärungstypen unterscheiden sich in kognitiver Belastung, Zeitbedarf und vorausgesetzter Expertise. Detaillierte technische Begründungen können z. B. eine zeitkritische Befundung belasten, während (über)vereinfachte Darstellungen relevante Einschränkungen der Ergebnisgültigkeit verdecken können. Aus diesen Zielkonflikten folgt, dass nicht maximale (und damit oft überbordend informationsreiche) Transparenz leitend sein kann, sondern zweckmäßige Interpretierbarkeit. Gemeint ist die Passung zwischen Erklärung, Nutzungskontext und Entscheidungsrisiko, d. h. Erklärungen, die gerade so detailliert sind, dass sie das Fehlerrisiko im spezifischen klinischen Kontext verringern oder das spezifische Kommunikationsziel optimal erreichen (Amann et al., 2020; Lipton, 2018, Rudin, 2019). Unter- und Überinformation stellen dabei symmetrische Fehlformen dar, da beide falsche Sicherheit erzeugen und eine angemessene KI-Nutzung unterminieren.

Beachtenswert für die ärztliche Kommunikation aber auch das eigene ärztliche Verständnis von KI-Ergebnissen ist, dass (vermeintliche) Erklärungen epistemische Risiken erzeugen können, wenn sie Verstehen suggerieren, ohne die Fähigkeit zur Bewertung oder Anfechtung von Ergebnissen zu erweitern. Anschauliche Visualisierungen und vereinfachte Darstellungen erzeugen oft subjektive Plausibilität und stärken Vertrauen, ohne eine wirkliche Erklärung (d. h., einen Einblick in das tatsächliche Verhalten des KI-Modells) bereitzustellen (Lipton, 2018; Miller, 2019). Solche oberflächlichen Scheinerklärungen wirken dann nicht klärend, sondern vermindern die Sicherheit, wenn sie als zuverlässige Belege und nicht als fehlbare Entscheidungshilfen betrachtet werden und sich (überhöhtes) Vertrauen vom tatsächlichen Systemverhalten zunehmend entkoppelt

(Ghassemi et al., 2021). Verantwortliche Nutzung verlangt daher weniger fallbezogene Illustration als Wissen darüber, unter welchen Bedingungen KI-Ergebnisse Geltung im radiologisch-diagnostischen Entscheidungszusammenhang insgesamt besitzen oder verlieren, z B. Wissen darüber, auf welchen Populationen und Scannern das KI-Modell trainiert wurde, für welche Indikationen es vorgesehen ist und an welchen Grenzen seine Leistung bekanntermaßen nachlässt (Amann et al., 2020; Rudin, 2019). Entscheidend ist, ob diese Bedingungen und dieses Wissen im klinisch-radiologischen Arbeitsprozess vorhanden bleiben und praktisch berücksichtigt werden. Nur so bleiben verantwortliche diagnostische Entscheidungen in der Radiologie auch bei algorithmischer Unterstützung durch technisch opake KI-Systeme dauerhaft möglich (Ghassemi et al., 2021; Miller, 2019; London, 2019).

4. Vertrauen in KI-basierte Technologien in der Radiologie

4.1 Vertrauen und Zuverlässigkeit

Der Einsatz von KI in der Radiologie setzt ein Mindestmaß an Vertrauen in Leistungsfähigkeit und Sicherheit der jeweiligen Anwendung voraus. Ziel klinischer KI-Nutzung ist jedoch nicht die Maximierung dieses Vertrauens, sondern dessen Kalibrierung: KI-basierte Verfahren sollen dort eingesetzt werden, wo sie im jeweiligen Kontext mit hoher Wahrscheinlichkeit zu besseren Entscheidungen beitragen, und dort unterbleiben, wo systematische Fehlentscheidungen zu erwarten sind. Vertrauen wird damit zu einer steuerungsrelevanten Größe, die in konkrete Einsatz- und Nutzungsentscheidungen übersetzt werden muss. Um diese Steuerungsfunktion analytisch fassen zu können, ist eine begriffliche Trennung zweier häufig vermischter Konstrukte erforderlich. Vertrauen bezeichnet eine Einstellung, verstanden als Erwartung an Kompetenz, Vorhersagbarkeit und Intentionen eines Systems. Es beschreibt, was von einer Anwendung angenommen wird, bevor gehandelt wird. Verlässlichkeit bezieht sich demgegenüber auf tatsächliches Nutzungsverhalten, also darauf, wie KI-Ergebnisse im klinischen Entscheiden und Handeln aufgegriffen, geprüft oder ignoriert werden, wie also mit einer KI-Anwendung faktisch umgegangen wird. Einstellung und Verhalten

können dabei auseinanderfallen: Ein hohes Maß an Vertrauen kann mit zurückhaltender Nutzung einhergehen und umgekehrt. Für die Analyse klinischer KI-Nutzung in der Radiologie ist es daher notwendig, Vertrauen und Verlässlichkeit getrennt zu betrachten und ihre Beziehung nicht vorauszusetzen, sondern empirisch und kontextbezogen zu bestimmen.

Empirische Arbeiten zeigen, dass ein hohes Maß an Vertrauen in KI-basierte Technologien dort problematisch wird, wo es zu einer erwartungsbasierten Reduktion von Überwachung und aktiver Kontrolle der KI-Systeme führt. In solchen Konstellationen werden KI-Ergebnisse weniger geprüft, nicht weil Bedenken fehlen, sondern weil Zuverlässigkeit (Verlässlichkeit) antizipiert wird: Kontrolle unterbleibt, da Fehler als unwahrscheinlich gelten (Parasuraman & Riley, 1997; Lee & See, 2004). Eine gegenläufige Fehlerform entsteht bei zurückhaltendem Vertrauen. In diesem Fall werden Systeme selbst dann nicht genutzt, wenn sie unter gegebenen Bedingungen verlässlich sind und objektiv Leistungsgewinne ermöglichen. Fehlanpassung zeigt sich hier nicht als Übernahme falscher Ergebnisse, sondern als Unterausnutzung verfügbarer Unterstützung. Beide Konstellationen verdeutlichen, dass Vertrauen weder maximal noch minimal sein sollte, sondern aufgabenspezifisch ausgerichtet (d. h. kalibriert) werden muss. Kalibriertes Vertrauen bezeichnet vor diesem Hintergrund den aufgaben- und kontextsensitiven Einsatz KI-basierter Verfahren. Ein System kann für Screening-nahe Erkennungsaufgaben gut geeignet sein, unter artefaktreichen Bildprotokollen, bei selten eingesetzten Geräten oder veränderten Prävalenzen jedoch an Verlässlichkeit verlieren. Zusätzlich ist zu berücksichtigen, dass Verzerrungen in den Trainingsdaten die Aussagekraft für bestimmte Patientengruppen einschränken können, was – bei Nichtbeachtung – zu einer unangemessene Einordnung KI-basierter Hinweise im klinischen Alltag führen kann (Kocak et al., 2025).

Vertrauen in KI-Systeme speisen sich im klinischen Alltag parallel aus mehreren Quellen: aus beobachteter Leistungshistorie, aus systemseitigen Transparenz- und Statushinweisen sowie aus sozialen und institutionellen Signalen, die den Einsatz der Technologie rahmen (Lee & See, 2004; Hoff & Bashir, 2015). Diese Quellen unterscheiden sich funktional und sind nicht gleich verlässlich, wirken jedoch gemeinsam auf die Vertrauensbildung ein. Die tatsächliche Zuverlässigkeit eines Systems ist im klinischen Alltag kaum direkt

erfahrbar. Einschätzungen stützen sich daher auf einzelne Nutzungsepisoden statt auf systematische Leistungsprofile, wobei Ersteindrücke, besonders markante Erfolge oder auffällige Fehler sowie Hinweise auf einen hohen »Automatisierungsstatus« des KI-Systems das Vertrauen überproportional prägen und als heuristische Anker wirken können. Gestalterische Merkmale fungieren in diesem Zusammenhang als Glaubwürdigkeitsindikatoren: Benutzeroberfläche, Sprachstil, Visualisierungsformen oder infrastrukturelle Einbettung beeinflussen, wie kompetent und zuverlässig ein System erscheint, ohne notwendigerweise mit seiner tatsächlichen Leistungsfähigkeit zu korrespondieren. Vertrauenskalibrierung ist daher weniger das Ergebnis individueller Beurteilungsfähigkeit, als das Produkt eines soziotechnischen Arrangements, das bestimmt, welche Leistungsbelege sichtbar sind und wie organisationale Reputationssignale vermittelt werden. Fehlkalibriertes Vertrauen ist in diesem Sinne primär ein Gestaltungs- und Organisationsproblem.

4.2 (Fehl-)Kalibrierung des Vertrauens: Folgen, Dynamik und Modifikation

Fehlkalibriertes Vertrauen in KI-Anwendungen in der Radiologie äußert sich in qualitativ unterschiedlichen problematischen Nutzungsmustern, nämlich Fehlgebrauch, Nichtgebrauch und Missbrauch. Fehlgebrauch wird durch Übervertrauen begünstigt, indem fehlerhafte KI-Systemvorschläge oder -ergebnisse akzeptiert und unabhängige Such- und Prüfaktivitäten eingeschränkt werden. Der zentrale Effekt liegt dabei weniger in der Übernahme falscher KI-basierter Hinweise, als in der Reduktion der ärztlichen Suchbreite, wodurch von der KI unbemerkte und nicht markierte, aber potenziell relevante Befunde häufiger übersehen werden, weil ärztliche Vigilanz und Exploration zugunsten systemseitiger Hinweise zurücktreten. Untervertrauen führt demgegenüber zu Nichtgebrauch: KI-Ergebnisse bleiben trotz nachgewiesenem Nutzen unberücksichtigt, etwa infolge von Skepsis, Frustration oder mangelnder KI-Integration in den Arbeitsablauf. Nichtgebrauch ist damit kein rationales Abwägen im Einzelfall, sondern ein organisations- und erwartungsbedingter Nutzungsausfall. Missbrauch schließlich stellt ein drittes, eigenständiges Muster dar, das den Einsatz von KI außerhalb validier-

ter Bedingungen, etwa bei anderen Modalitäten, Populationen oder Erfassungssettings. Häufig liegt dem keine bewusste Regelverletzung zugrunde, sondern eine inadäquate Übergeneralisierung, also eine spezifische Form des Übervertrauens, oft unter Zeit-, Kosten- oder Leistungsdruck. Fehlgebrauch, Nichtgebrauch und Missbrauch sind daher als unterschiedliche Fehlerformen im Umgang mit KI-Technologien in der Radiologie zu behandeln, die jeweils spezifische Gegenmaßnahmen erfordern und nicht durch eine einheitliche Vertrauensstrategie adressiert werden können (Parasuraman & Riley, 1997).

Die Kalibrierung von Vertrauen in radiologische KI-Anwendungen ist kein einmaliger Akt, sondern ein fortlaufender Anpassungsprozess, denn das Vertrauen verändert sich mit KI-Nutzungserfahrungen, Systemaktualisierungen und Aufgabenwechseln und bleibt damit prinzipiell reversibel. Diese Dynamik ist jedoch unter klinischen Bedingungen systematisch verzerrt, da die Voraussetzungen für verlässliches Lernen nur eingeschränkt gegeben sind. Zentral ist dabei die typische Struktur des Feedbacks in der Radiologie: Rückmeldungen über die tatsächliche Leistung eines KI-Systems sind häufig verzögert, unvollständig oder verrauscht. Diagnostische Konsequenzen werden oft erst nach Wochen sichtbar oder lassen sich nicht eindeutig einer einzelnen KI-Systemausgabe zuordnen. Unter diesen Bedingungen wird die Anpassung bestehender Überzeugungen unsicher, da eine einzelne Erfahrung kein stabiles Lernsignal liefert. Hinzu kommt eine asymmetrische Gewichtung von Ereignissen, wobei einzelne auffällige Fehler überproportional vertrauenssenkend wirken und abrupte Nichtnutzung auslösen können, obwohl die mittlere Systemleistung weiterhin hoch ist. Umgekehrt können wiederholte unauffällige Erfolge oder plausible, aber oberflächliche Fehlererklärungen Vertrauen schleichend und unangemessen normalisieren und dadurch langfristig eine nachlassende Wachsamkeit und Sorglosigkeit begünstigen (Dzindolet et al., 2003; Hoff & Bashir, 2015; Lee & See, 2004). Ein weiterer Mechanismus der Fehlkalibrierung betrifft die inadäquate Übertragung von Vertrauen zwischen verschiedenen Aufgaben einer KI-Anwendung (z. B. von der Triage mittels Brust-CT auf die subtilere onkologische Stadieneinteilung). Diese Generalisierung ist epistemisch unbegründet, da Leistungsprofile aufgaben- und kontextspezifisch sind. Sie stellt damit eine Form des Missbrauchsrisikos dar, wobei einheitliche Benut-

zeroberflächen, konsistenter Sprachstil oder gemeinsame Herstelleridentitäten dieses Risiko noch verstärken können, indem Vertrauen an Design oder Marke gebunden wird statt an nachgewiesene Aufgabenleistung. In der Summe zeigt sich, dass die Dynamik von Vertrauen weniger durch individuelle Urteilskraft bestimmt ist als durch Lernbedingungen, Ereignissalienz und gestalterische Rahmungen. Fehlkalibrierung von Vertrauen in radiologische KI-Anwendungen ist damit kein zufälliges Nebenprodukt, sondern ein Ergebnis der Art und Weise, wie Rückmeldung, Sichtbarkeit und Aufgabenbezug im klinischen Alltag organisiert sind.

Ein ausgewogenes Vertrauen in KI-basierte Systeme erfordert eine fortlaufende Kalibrierung, die als Sicherheitsaufgabe zu verstehen ist. Zentrale Voraussetzungen sind kontinuierliche Rückkopplung, lokale Validierung sowie eine explizite Kommunikation von Anwendungsbereich und typischen Fehlermodi. Kalibrierung ist dabei kein einmaliger Schulungseffekt, sondern ein Prozess, der dauerhaft durch geeignete Strukturen gestützt werden muss. Wirksame Maßnahmen greifen auf mehreren Ebenen: Auf individueller Ebene sollte Nutzungsschulung als Kalibrierungstraining angelegt sein, das mit repräsentativen Grenzfällen arbeitet, typische Fehlermodi sichtbar macht und Diskrepanzen zwischen KI-System- und ärztlichem Fachurteil systematisch reflektiert. Auf der Ebene des Arbeitsablaufs unterstützen Audit-Schleifen die Stabilisierung von Überwachungsstrategien, etwa durch regelmäßige Auswertung von Abweichungen getrennt nach KI-positiven und KI-negativen Fällen. Leistungs-Dashboards ermöglichen zusätzlich die Beobachtung von driftenden Abweichungen, Prävalenzverschiebungen und Untergruppenleistungen. Auf Governance-Ebene beruht Kalibrierung auf operativer Transparenz: Klare Angaben zu Verwendungszweck, Versionierung, bekannten Einschränkungen und Update-Management können unangemessene Vertrauensübertragungen begrenzen und eine kontextangemessene Einordnung von Ergebnissen erleichtern (Lee & See, 2004; Parasuraman & Riley, 1997). Entscheidend ist dabei ein Perspektivwechsel: Nicht die Steigerung von Vertrauen in KI-basierte Automatisierung per se ist relevant, sondern die Reduktion sicherheitsrelevanter Fehlhandlungen und Unterlassungen. Da sich Vertrauensdynamiken zwischen professioneller Praxis und Patientenseite unterscheiden, wird Letztere im folgenden Abschnitt gesondert betrachtet.

5. Patientenperspektiven zum Einsatz von KI in der Radiologie

5.1 Erwartungen und wahrgenommene Risiken

Die patientenseitige Einordnung der KI-Nutzung in der Radiologie erfolgt auf der Grundlage unterschiedlicher mentaler Modelle, die als kognitive Rahmen fungieren und Erwartungen an Rolle, Nutzen und Risiko der Technologie strukturieren. Je nachdem, ob KI als autonomer Diagnostiker, Ko-Detektor oder als untergeordnetes technisches Hilfsmittel verstanden wird, verschieben sich Annahmen über Kompetenz, Kontrollbedarf und Verantwortlichkeit. Mit den mentalen Modellen verändern sich zugleich Nutzen- und Risikozuschreibungen: Erwartungen an höhere Genauigkeit, beschleunigte Abläufe und gleichbleibende Qualität werden mit Befürchtungen wie Fehleranfälligkeit, Entmenschlichung der Versorgung, Verlust ärztlicher Aufmerksamkeit, Datenschutzproblemen, sekundärer Datennutzung oder unklaren Verantwortlichkeiten verknüpft. Akzeptanz entsteht dabei nicht aus einer allgemeinen Technikbefürwortung, sondern unter spezifischen Bedingungen. Empirische Arbeiten zeigen, dass Zustimmung maßgeblich davon abhängt, ob ärztliche Aufsicht und Entscheidungsverantwortung im Behandlungsprozess erkennbar bleiben (Promberger & Baron, 2006; Nadarzynski et al., 2019; Karger, 2026). Der patientenseitige Informationsbedarf in der Radiologie richtet sich entsprechend auf konkrete, handlungsnahe Fragen: Welche Rolle KI im diagnostischen Prozess einnimmt, ob radiologische Bilder unabhängig geprüft werden, wie mit Meinungsverschiedenheiten zwischen Mensch und Maschine umgegangen wird und welche Konsequenzen Fehler haben. Demgegenüber kommunizieren Organisationen den KI-Einsatz häufig in allgemeinen Formeln, die Transparenz signalisieren, ohne Informationsautonomie tatsächlich zu ermöglichen (z. B. »es werden fortschrittliche Tools verwendet«). Mentale Modelle werden so weniger durch formale Offenlegung opaker KI-Systeme als durch die Art geprägt, in der ärztliche Verantwortlichkeit, Aufsicht und Entscheidungsprozesse im Umgang mit solchen Systemen nachvollziehbar gemacht werden.

Darüber hinaus ist die patientenseitige Bewertung von KI-Algorithmen in medizinischen Settings stark kontextabhängig und folgt

den Anforderungen der jeweiligen Entscheidungssituation. Es existiert keine einheitliche Akzeptanzlage, vielmehr variiert die Einschätzung algorithmischer Entscheidungen systematisch mit dem Aufgabencharakter. Dort, wo Genauigkeit, Konsistenz und Objektivität im Vordergrund stehen, werden KI-basierte Verfahren häufig positiv bewertet. In Situationen hingegen, in denen individuelles Urteilsvermögen, Empathie oder ein situationssensitives Eingehen als zentral gelten, stoßen algorithmische Entscheidungen auf Zurückhaltung (Dietvorst et al., 2015; Logg et al., 2019; Longoni et al., 2019). Diese Zurückhaltung ist dabei eben nicht als pauschale Technikablehnung zu verstehen, sondern als Ausdruck unterschiedlicher normativer Erwartungen daran, was in bestimmten medizinischen Kontexten als legitime Entscheidungsgrundlage gilt. Patientenseitig ist Akzeptanz damit funktional gebunden und nicht global zu erzielen. Unter diesen Bedingungen übernimmt Transparenz eine regulierende Rolle. Sie soll Orientierung bieten und informierte Zustimmung ermöglichen, ohne Akzeptanz erzwingen zu wollen oder Vorbehalte zu delegitimieren. Transparenz darf daher nicht als persuasive, ärztliche Kommunikationsstrategie eingesetzt werden, die patientenseitige Zweifel als irrational erscheinen lässt oder implizit nahelegt, Zurückhaltung oder Ablehnung gegenüber dem Einsatz von KI-Tools sei unerwünscht. Nur so kann Transparenz dazu beitragen, kontextangemessene Bewertungen von KI-Einsatz zu unterstützen, anstatt soziale oder moralische Druckeffekte zu erzeugen.

Über individuelle Risiko-Nutzen-Abwägungen hinaus wird der medizinische Einsatz von KI auch nach Fairnesskriterien beurteilt. Besonders in Bevölkerungs-Screenings und Triage-Situationen, in denen KI die Priorisierung, Reihenfolge oder Zugang zur medizinischen Versorgung beeinflusst, rücken Verteilungswirkungen in den Vordergrund. Fairness wird hier zu einer eigenständigen Bewertungsdimension neben Genauigkeit und Effizienz. Das öffentliche Bewusstsein für algorithmische Verzerrungen, verstärkt durch prominente Befunde zur Vorhersage von Gesundheitsrisiken, hat Fairness zu einem zentralen Vertrauensfaktor gemacht, auch in Anwendungsfeldern wie der (radiologischen) Bildinterpretation, in denen systematische Urteilsverzerrungen zunächst weniger offensichtlich erscheinen (Obermeyer et al., 2019; Rajkomar et al., 2018). Wahrgenommene Gerechtigkeit hängt dabei entscheidend davon ab, ob ein System als angemessen für Personen »wie mich« gilt und ob

Fehler als gerecht verteilt erscheinen, statt systematisch bestimmte Gruppen stärker zu betreffen. Hinzu kommt die Bedeutung der Anfechtbarkeit KI-basierter Ergebnisse: Die Möglichkeit, KI-gestützte Entscheidungen ohne substantielle Hürden zu hinterfragen und Eskalationswege zu nutzen, beeinflusst maßgeblich, ob der KI-Einsatz als fair wahrgenommen wird, unabhängig von der formalen Leistungsfähigkeit des KI-Systems. Studien zur algorithmischen Entscheidungsfindung zeigen, dass Wahrnehmungen von Gerechtigkeit stark durch Erklärungstypen, Beteiligungsformate, Rechtsmittel und Ergebniskommunikation geprägt werden (Binns et al., 2018). Diese Faktoren sind direkt anschlussfähig an radiologische Befundungs- und Eskalationsprozesse und zeigen auf, wie die ärztliche KI-Nutzung kommuniziert werden sollte, damit auf Patientenseite angemessenes Vertrauen entsteht.

5.2 KI-Nutzung und Vertrauen in asymmetrischen Arzt-Patient-Beziehungen

Vertrauen in den Einsatz von KI im medizinischen Kontext entsteht unter Bedingungen typischerweise asymmetrischer Arzt-Patient-Beziehungen und wird von diesem Vertrauensverhältnis vermittelt oder moduliert. In den meisten radiologischen Untersuchungsabläufen besteht patientenseitig kein direkter Kontakt mit der KI; ihre Wirkung wird vielmehr über institutionelle Strukturen und professionelles Handeln vermittelt. Vertrauen bezieht sich damit auf die Weise, in der der KI-Einsatz in diagnostische Prozesse eingebettet und nach außen sichtbar gemacht wird. Die Effekte des Systems werden indirekt erfahrbar, etwa über Wartezeiten, die Formulierung von Befunden, nachfolgende klinische Entscheidungen und die kommunikative Haltung der behandelnden Seite. Diese beobachtbaren Konsequenzen fungieren als Anhaltspunkte, anhand derer Vertrauen aufgebaut oder revidiert wird. Aufgrund der ungleichen Verteilung von Wissen und Entscheidungsmacht ist Vertrauen dabei besonders sensibel gegenüber Signalen der Verantwortungsübernahme. Diese Vermittlung wirkt in zwei Richtungen: Vertrauen kann gestärkt werden, wenn KI als überwachte Unterstützung erscheint, die diagnostische Konsistenz erhöht und in das ärztliche Urteil sichtbar integriert wird; es kann jedoch untergraben werden, wenn KI als Ersatz für

ärztliche Einordnung, Entscheidung oder Verantwortung kommuniziert wird oder zumindest so erscheint (Karger, 2026).

Eine spezifische Herausforderung der Radiologie liegt in der begrenzten ärztlichen Sichtbarkeit im klinischen Alltag. Radiologische Expertise ist häufig räumlich, zeitlich und kommunikativ vom unmittelbaren Behandlungskontakt getrennt, sodass die Zuschreibung von Verantwortung nicht selbstverständlich entsteht. Diese strukturelle Distanz kann durch den Einsatz von KI verstärkt werden, denn durch KI tritt eine zusätzliche Vermittlungsebene hinzu, die Zuständigkeiten, diagnostische Entscheidungspfade und Verantwortungszuschreibung weiter verschleiern kann. Erscheint ein Befundbericht als technisches Artefakt einer undurchsichtigen Prozesskette (»Black Box«), etwa als Ergebnis von »KI-System plus Institution«, kann Vertrauen in die Behandlungsbeziehung sinken, selbst wenn sich die diagnostische Leistung objektiv verbessert. Vertrauen reagiert hier weniger auf Leistungskennzahlen als auf Zuschreibbarkeit von Verantwortung und Ansprechbarkeit (Chiou & Lee, 2023). Demgegenüber können strukturierte und dem Zweck angemessene Erklärungsangebote – etwa durch die überweisende Ärzteschaft, sorgfältig kuratierte Patientenportale oder klar zugängliche Rückfragekanäle – die patientenseitige Wahrnehmung ärztlicher Präsenz und Verantwortung in der KI-gestützten Radiologie stabilisieren. Wie bereits erwähnt, wirkt Transparenz in solchen Informationsangeboten kontextabhängig: Hinweise auf Unsicherheit können beruhigen, wenn sie Aufsicht und Sorgfalt signalisieren, oder aber verunsichern, wenn sie primär die Fehlbarkeit betonen. Ziel ist daher eine zweckmäßig kontextualisierte Kommunikation der Unsicherheiten im KI-gestützten diagnostischen Prozess, die klar macht, was bekannt ist, was offen bleibt und welche Sicherungen greifen.

Die Umsetzung dieser Einsichten ist möglich, ohne die radiologische Arbeit in permanente Einzelkonsultationen zu überführen. Vertrauensbildung im Kontext der radiologischen KI-Nutzung erfordert keine kompensatorische Verdichtung persönlicher Arzt-Patient-Interaktionen, sondern eine strukturierte Kalibrierung von patientenseitigen Erwartungen gegenüber hybriden Entscheidungsprozessen. Wie bereits bezüglich der Ärzteschaft erwähnt (s. Abschnitt 4), zielt Vertrauenskalibrierung darauf, Vertrauen in KI-basierte Technologien weder zu maximieren noch zu minimieren, sondern es an deren tatsächliche Zuverlässigkeit und Grenzen anzupassen. Drei Leitlini-

en sind dabei zentral: (1) KI sollte, soweit zutreffend, konsequent als ärztlich überwachte Unterstützung kommuniziert werden, die Priorisierung und Befundung erleichtert, während die diagnostische Verantwortung eindeutig außerhalb des KI-Systems verankert bleibt. (2) Systemseitige Unsicherheiten und Grenzen sollten möglichst *konkret* dargestellt werden, wie z B. der validierte Anwendungsbereich, typische Fehlerquellen sowie das Vorgehen bei mehrdeutigen Ergebnissen. (3) Verantwortlichkeiten, Anfechtungs- und Regressmöglichkeiten bezüglich der KI-basierten Ergebnisse sollten *klar* benannt werden, um sichtbar zu machen, wie mit abweichenden Interpretationen umgegangen und aus etwaigen Fehlern gelernt wird. Diese Prinzipien können helfen, patientenseitige Erwartungen mit hybriden Entscheidungsrealitäten in der Radiologie in Einklang zu bringen und das Risiko unkritischer Überhöhung ebenso wie pauschaler Ablehnung zu senken. Da Vertrauen und Akzeptanz auf Patientenseite die Einführung und Verstetigung neuer Technologien wesentlich mitbestimmen, werden sie im nächsten Abschnitt gemeinsam mit der ärztlichen Perspektive adressiert.

6. Technologieakzeptanz und organisatorische Bedingungen nachhaltiger KI-Nutzung in der Radiologie

6.1 Einstellungen und Akzeptanz gegenüber KI-basierten Technologien

Der großflächige und nachhaltige Einsatz von KI in der Radiologie hängt von dessen Akzeptanz in den beiden zentralen Akteursgruppen, also der Ärzte- und Patientenschaft, ab (Caspers et al., 2025). Aus ärztlicher Perspektive steht die Frage im Vordergrund, ob ein gegebenes KI-Tool als zweckmäßiges Automatisierungselement in bestehende Arbeitsabläufe passt. Forschung zur Technologieakzeptanz zeigt, dass vor allem Nutzenwahrnehmung, Bedienbarkeit und unterstützende/hemmende Rahmenbedingungen die Adoption neuer Technologien prägen (Davis, 1989; Venkatesh et al., 2003). Aus Patientensicht muss KI-gestützte Versorgung vor allem als legitim, angemessen, verlässlich und fair gelten, wie im vorangegangenen Kapitel ausgeführt. Werden diese unterschiedlichen Determinanten

der Akzeptanz nicht beachtet, laufen auch objektive, leistungsfähige und potenziell hilfreiche KI-Systeme Gefahr, im klinischen Alltag nicht nachhaltig Fuß zu fassen. Einstellungen wirken als Schwelle für eine Skalierung: Ohne Zuversicht und Akzeptanz auf ärztlicher Ebene bleibt eine Nutzung begrenzt; ohne Akzeptanz auf Patientenebene entstehen Legitimations- und Reputationsprobleme, die eine Ausweitung verhindern (Greenhalgh et al., 2004; Rogers, 2003).

Die ärztliche Haltung gegenüber KI-basierten Technologien in der Radiologie lässt sich als Ergebnis eines Spannungsverhältnisses zwischen wahrgenommenem klinischem Nutzen und wahrgenommenen beruflichen Kosten rekonstruieren. Akzeptanz entsteht dabei nicht aus der Leistungsfähigkeit der Technologie allein, sondern aus der Relation zwischen erwarteten entlastenden Effekten der Automatisierung und erwartetem zusätzlichen Aufwand im konkreten Arbeitsprozess. Ein zentraler Faktor ist die Nützlichkeit in der lokalen Aufgabenumgebung. KI-Akzeptanz wird gefördert, wenn KI-Systeme Fehler reduzieren, Priorisierung verbessern oder Auswertungszeiten verkürzen, ohne nachgelagerte Prozesse nennenswert zu belasten. Ebenso entscheidend ist die Passung zum Arbeitsablauf. Systeme können auch bei hoher Genauigkeit abgelehnt werden, wenn sie den Arbeitsfluss stören, umständlich zu bedienen sind oder eine Vielzahl manuell zu prüfender Ausnahmen erzeugen. In solchen Fällen ist Mehrbelastung unmittelbar erfahrbar, während der Nutzen unsicher oder zeitlich verzögert bleibt. Hinzu kommt die Frage professioneller Autonomie: Wird KI als Ersatz für Expertise positioniert, entsteht Widerstand; wird sie als Erweiterung bestehender Praxis kommuniziert, ist sie mit beruflichen Normen vereinbar (s. Abschnitt 2). Eine weitere Einflussgröße betrifft die ärztliche Verantwortlichkeit und Haftung: Gelten Abweichungen von KI-Empfehlungen als riskant, begünstigt dies eine Ablehnung solcher Technologien. Empirische Befunde zu Akzeptanzfaktoren in der Ärzteschaft zeigen entsprechend eine Mischung aus Leistungsoptimismus und Sorge um Verantwortung, Qualifikationsverlust und Arbeitsbelastung, was die nachhaltige Adoption von KI-Technologien in der Radiologie als soziotechnisches Integrationsproblem ausweist (Waymel et al., 2019; Huisman et al., 2021). Die aus diesen und anderen Überzeugungen resultierende Einstellung zur KI-Nutzung bestimmt am Ende nicht nur, ob KI überhaupt genutzt wird, sondern auch, ob

sie zur überprüfenden Unterstützung, zu diagnostischen Abkürzungen oder zur rein formalen Absicherung eingesetzt wird.

Die Akzeptanz von radiologischen KI-Anwendungen auf Patientenseite bestimmt sich aus einer Abwägung zwischen wahrgenommenem Nutzen und wahrgenommenen Risiken. Vorteile wie Geschwindigkeit, Konsistenz und zugeschriebene Objektivität stehen dabei Befürchtungen gegenüber, die sich weniger auf technische Leistungsparameter als auf normative und relationale Aspekte beziehen, etwa Fehlerfolgen, Entpersonalisierung, unklare Verantwortlichkeiten und Fairness. Diese Bewertungen und Haltungen haben nicht nur individuelle, sondern auch gesellschaftliche Konsequenzen, denn sie bestimmen die Zustimmung zu, die Zufriedenheit mit und die öffentliche Legitimität von großangelegten medizinischen KI-Programmen, insbesondere in Szenarien, in denen eine KI den Zugang zu oder die Priorisierung von ärztlicher Betreuung steuert, etwa in Screening- und Triage-Situationen. Wie bereits erwähnt, legt die Forschung zur algorithmischen Entscheidungsfindung eine konditionale Akzeptanzdynamik nahe: Algorithmische Urteilsfindung erhält Zustimmung, wenn sie als genauer wahrgenommen wird, löst jedoch Ablehnung aus, sobald sie als Bedrohung von Handlungsfähigkeit, menschlicher Würde oder individueller Versorgung interpretiert wird oder wenn KI-systemseitig Fehler salient werden (Dietvorst et al., 2015; Logg et al., 2019; Longoni et al., 2019; Karger, 2026). Die Akzeptanz von KI in der Radiologie auf Patientenseite ist somit weder stabil noch selbstverständlich, sondern kontextabhängig und revidierbar. Entscheidend für die Akzeptanz wird daher das Vertrauen in die radiologische Einrichtung, in die die Technologie eingebettet ist, sowie die Art und Weise, in der Aufsicht, Unsicherheit und Anfechtungswege von ärztlicher Seite aus kommuniziert werden. Neben gesellschaftlicher Akzeptanz spielen allerdings auch förderliche oder hemmende organisational-strukturelle Bedingungen eine Rolle für die nachhaltige Adoption von KI-Technologien in der Radiologie; psychologische Aspekte dieser Bedingungen werden im Folgenden kurz umrissen.

6.2 Organisationale Bedingungen einer nachhaltigen KI-Adoption in der Radiologie

Die Einführung von KI in der Radiologie ist als sozio-organisatorisches Projekt zu analysieren und nicht als bloße Beschaffungsentscheidung. Ob ein KI-basiertes System zwecks Automatisierung in der radiologischen Diagnostik sicher und wirksam eingesetzt wird, hängt – wie in den vorherigen Abschnitten erörtert – nicht allein von seiner technischen Leistungsfähigkeit ab, sondern von der Weise, in der es in bestehende und ggf. anzupassende Strukturen eingebettet wird. Die Implementierungsforschung zeigt, dass die erfolgreiche Einführung und Adoption von Technologien durch ein Bündel von Faktoren bestimmt werden. Dazu zählen Merkmale der Intervention selbst, etwa ihre Komplexität, ebenso wie Eigenschaften des internen professionellen Umfelds, darunter Kultur, Ressourcen und Routinen. Hinzu kommen externe Einflüsse wie regulatorischer oder reputationsbezogener Druck sowie die organisationale Fähigkeit zur fortlaufenden Bewertung und Anpassung (Damschroder et al., 2009; Greenhalgh et al., 2004). Aus soziotechnischer Perspektive entstehen Risiken nicht aus der Technologie isoliert, sondern aus dem Zusammenspiel von Technik, Aufgabenstruktur und organisationalem Kontext. Daraus folgt, dass Gestaltung und Überwachung radiologischer KI-Anwendungen nicht als einmalige oder bei Einführung fixierte Maßnahmen angelegt sein dürfen. Sicherheit und Wirksamkeit von KI-Systemen im diagnostischen Prozess ergeben sich aus kontinuierlicher Abstimmung, laufender Rückmeldung und wiederholter Anpassung der Arbeitsabläufe (Sittig & Singh, 2010; Carayon et al., 2014). Für einen nachhaltigen und skalierbaren KI-Einsatz lassen sich vor diesem Hintergrund mehrere relevante organisationale Facetten identifizieren, die im Folgenden kurz betrachtet werden.

Vertrauen in KI lässt sich nur dann stabilisieren, wenn Einsatz, Leistung und Abweichungen fortlaufend beobachtbar und überprüfbar gemacht werden. Dies erfordert Praktiken, die wiederholt vor Ort nachprüfen, ob ein System im vorgesehenen Anwendungsbereich zuverlässig arbeitet, die im laufenden Betrieb regelmäßig nachsteuern, die Fehler- und Beinahefehler-Ereignisse systematisch erfassen und die KI-Systemleistung kontinuierlich beobachten. Solche Verfahren bestimmen, ob die breite KI-Nutzung im radiologischen Alltag als nachhaltig tragfähig erscheint und ob KI-gestütz-

te Versorgung als legitim bewertet wird. Wird nachvollziehbar gemacht, wofür ein System eingesetzt wird, wann es aktualisiert wurde und wie mit Abweichungen umgegangen wird, kann Vertrauen an überprüfbare Routinen gebunden werden. Bleiben diese Prozesse undurchsichtig, reagieren Organisationen häufig mit Rückzug aus der Nutzung oder mit defensiven Anpassungen, die letztlich aber einen Fehlgebrauch begünstigen. Damit Kontrolle lernfähig wird, müssen Abweichungen und Beinahefehler-Ereignisse ohne Sorge vor Schuldzuweisung thematisiert werden können. Psychologische Sicherheit ist daher eine operative Voraussetzung belastbarer Rückmeldung und Verbesserung (Edmondson, 1999). Ein »Just Culture«-Ansatz behandelt KI-bezogene Abweichungen als systemische Lernanlässe und verhindert, dass strukturelle Fehler an der operativen Ebene personalisiert abgefangen werden (Reason, 1997).

Die Nutzung von KI in der radiologischen Routine unterliegt auch wirtschaftlichen Gegebenheiten, die sich in Anreizsystemen und Leistungskennzahlen niederschlagen können. Werden dabei Durchsatzziele und Produktivitätsmetriken in den Vordergrund gestellt, kann dies unbeabsichtigt zu reduzierter Überprüfung und erhöhter Abhängigkeit von KI-gestützten Ergebnissen führen, insbesondere wenn KI primär als Instrument zur Beschleunigung radiologischer Arbeitsabläufe positioniert ist. Wird Erfolg überwiegend an Durchlaufzeiten gemessen, verengt sich der Bewertungsraum. Fehlertypen, Fehlerfolgen und nachgelagerte Effekte geraten aus dem Blick, obwohl sie für die Sicherheit und Versorgungsqualität zentral sind. Allgemeiner gilt, dass Kennzahlen, sobald sie zu Zielgrößen werden, Verhalten verzerren und Belastungen verlagern können. Effizienzgewinne an einer Stelle gehen dann mit zusätzlichen Kosten an anderer Stelle einher. So erhöhen falsch-positive Befunde die Zahl von Folgeuntersuchungen und die Belastung auf Versorgungsebene, auch wenn interne Produktivitätskennzahlen steigen (Muller, 2018). Einführung und Skalierung von KI erfordern daher Bewertungssysteme, die Effizienz nicht isoliert und auf wenige Leistungsparameter verengt erfassen, sondern Sicherheit, Arbeitsbelastung und nachgelagerte Nutzung systematisch mitberücksichtigen.

Nachhaltiger Nutzen von KI setzt organisationale Lernmechanismen voraus, die über die einmalige Einführung hinausgehen. Abweichungen und Leistungsänderungen des KI-Systems müssen systematisch in eine veränderte Praxis übersetzt werden, etwa durch

angepasste Arbeitsroutinen, aktualisierte Richtlinien oder gezielte Auslöser für die ärztliche Weiterbildung und Revalidierung des Systems. Studien zu Audits und Feedback zeigen, dass Rückmeldungen Verhalten nur dann verändern, wenn sie zeitnah erfolgen, spezifisch sind und klare Handlungsimplikationen enthalten (Ivers et al., 2012). Diese Bedingungen markieren zugleich Anforderungen an die Interpretierbarkeit der KI-bezogenen Überwachung, denn Überwachung wird erst wirksam, wenn sie nicht nur Abweichungen fortlaufend misst, sondern relevante Abweichungen so aufbereitet, dass Anpassung möglich wird. Organisational verankerte Lernschleifen müssen zudem mit Heterogenität umgehen können, denn Leistungs- und Vertrauensmuster unterscheiden sich zwischen Standorten, eingesetzten Technologien und Prävalenzkontexten, sodass einheitliche Vorgaben hier zu kurz greifen. Die organisationale Steuerung muss daher lokale Anpassung ermöglichen, ohne eine unkontrollierte und potenziell invalide Ausweitung von Einsatzbereichen zuzulassen.

Zusammenfassend zeigt unsere Analyse, dass die psychologischen Wirkungen KI-basierter Automatisierung nicht an der Technologie selbst entschieden werden, sondern an den Bedingungen ihres Einsatzes. Dieselben Systeme können die diagnostische Sicherheit erhöhen oder neue Fehlerquellen erzeugen, abhängig davon, wie sie in Arbeitsabläufe, Verantwortungsstrukturen und Rückmeldemechanismen eingebettet sind. Automatisierungsverzerrung, Fehlkalibrierung von Vertrauen, Kompetenzverschiebungen oder moralischer Distress entstehen nicht zufällig und auch nicht primär aus individuellen Einstellungen, sondern aus stabilen Kopplungen zwischen Systemlogik, Aufgabenstruktur und organisationalen Erwartungen. Damit verschiebt sich der Fokus von der Frage, ob KI leistungsfähig ist, zu der Frage, unter welchen Bedingungen ihre Leistungsfähigkeit wirksam und kontrollierbar wird. Nachhaltiger Nutzen setzt voraus, dass Kontrolle, Lernen und Verantwortung nicht an individuelle Wachsamkeit und Sorgfalt delegiert werden, sondern strukturell abgesichert sind. Transparente Einsatzgrenzen, lernfähige Rückkopplung, ausbalancierte Anreizsysteme und psychologisch sichere Umgangsformen mit Abweichungen bestimmen, ob KI als unterstützende Ressource wirkt oder Abhängigkeit, Passivierung und defensive Routinen begünstigt. Erst auf dieser Grundlage wird eine Implementierung möglich, die über punktuelle Effizienzgewinne hinausreicht

und auf langfristige Robustheit, Sicherheit und Skalierbarkeit ausgerichtet ist.

Literaturverzeichnis

Amann, J., Blasimme, A., Vayena, E., & Frey, D. (2020). Explainability for artificial intelligence in health care: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, *20*(1), 310. https://doi.org/10.1186/s12911-020-01332-6

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

Carayon, P., Wetterneck, T. B., Rivera-Rodriguez, A. J., Hundt, A. S., Hoonakker, P., Holden, R., & Gurses, A. P. (2014). Human factors systems approach to healthcare quality and patient safety. *Applied Ergonomics*, *45*(1), 14–25. https://doi.org/10.1016/j.apergo.2013.04.023

Caspers, J., Karger, C., Langner, R., Weißenfels, S., Günther, J., Spranger, T.M., Wagner, R., Lanzerath, D., Eickhoff, S.B., & Heinrichs, B. (2025). Künstliche Intelligenz in der Radiologie – Von der experimentellen Phase zur produktiven Anwendung. *Deutsches Ärzteblatt International, 122*(Sonderausgabe KI), 24–27.

Chiou, E. K., & Lee, J. D. (2023). Trusting automation: Designing for responsivity and resilience. *Human Factors*, *65*(1), 137–165. https://doi.org/10.1177/00187208211009995

Cook, R. I., & Woods, D. D. (1996). Adapting to new technology in the operating room. *Human Factors*, *38*(4), 593–613. https://doi.org/10.1518/001872096778827224

Cosmides, L., & Tooby, J. (2000). Evolutionary psychology and the emotions. In M. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of emotions* (2nd ed., pp. 91-115). New York, NY: Guilford Press

Croskerry, P., Singhal, G., & Mamede, S. (2013). Cognitive debiasing 1: Origins of bias and theory of debiasing. *BMJ Quality & Safety*, *22*(Suppl 2), ii58–ii64. https://doi.org/10.1136/bmjqs-2012-001712

Cruess, R. L., Cruess, S. R., Boudreau, J. D., Snell, L., & Steinert, Y. (2014). Reframing medical education to support professional identity formation. *Academic Medicine*, *89*(11), 1446–1451. https://doi.org/10.1097/ACM.0000000000000427

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York, NY: Putnam

Damschroder, L. J., Aron, D. C., Keith, R. E., Kirsh, S. R., Alexander, J. A., & Lowery, J. C. (2009). Fostering implementation of health services research findings into practice: A consolidated framework for advancing implementation science. *Implementation Science*, *4*(1), 50. https://doi.org/10.1186/1748-5908-4-50

Dang, Q., & Li, G. (2025). Unveiling trust in AI: The interplay of antecedents, consequences, and cultural dynamics. *AI & Society*. https://doi.org/10.1007/s00146-025-02477-6

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, *8*(4 Pt. 1), 377–383. https://doi.org/10.1037/h0025589

Dratsch, T., Chen, X., Rezazade Mehrizi, M., Kloeckner, R., Mahringer-Kunz, A., Pusken, M., ... Pinto Dos Santos, D. (2023). Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology*, *307*(4), e222176. https://doi.org/10.1148/radiol.222176

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv preprint arXiv:1702.08608. https://arxiv.org/abs/1702.08608

Drew, T., Vo, M. L.-H., & Wolfe, J. M. (2013). The invisible gorilla strikes again: Sustained inattentional blindness in expert observers. *Psychological Science*, *24*(9), 1848–1853. https://doi.org/10.1177/0956797613479386

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, *44*(2), 350–383. https://doi.org/10.2307/2666999

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human–robot interaction. *Engaging Science, Technology, and Society*, *5*, 40–60. https://doi.org/10.17351/ests2019.260

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors*, *37*(2), 381–394. https://doi.org/10.1518/001872095779064555

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Gong, B., Nugent, J. P., Guest, W., Parker, W., Chang, P. J., Khosa, F., & Nicolaou, S. (2019). Influence of artificial intelligence on Canadian medical students' preference for radiology specialty: A national survey study. *Academic Radiology*, *26*(4), 566–577. https://doi.org/10.1016/j.acra.2018.10.007

Greenhalgh, T., Robert, G., Macfarlane, F., Bate, P., & Kyriakidou, O. (2004). Diffusion of innovations in service organizations: Systematic review and recommendations. *Milbank Quarterly*, *82*(4), 581–629. https://doi.org/10.1111/j.0887-378X.2004.00325.x

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Huisman, M., Ranschaert, E., Parker, W., Mastrodicasa, D., Koci, M., Pinto de Santos, D., ... Willemink, M. J. (2021). An international survey on AI in radiology: Fear of replacement, knowledge, and attitude. *European Radiology*, *31*(9), 7058–7066. https://doi.org/10.1007/s00330-021-07781-5

Ivers, N., Jamtvedt, G., Flottorp, S., Young, J. M., Odgaard-Jensen, J., French, S. D., ... Oxman, A. D. (2012). Audit and feedback: Effects on professional practice and healthcare outcomes. *Cochrane Database of Systematic Reviews*, *2012*(6), CD000259. https://doi.org/10.1002/14651858.CD000259.pub3

Jameton, A. (1984). *Nursing practice: The ethical issues*. Englewood Cliffs, NJ: Prentice Hall

Karger, C. R. (2026). Patients' perspectives on the implementation of artificial intelligence in radiological diagnostics: A focus group study. *Journal of Medical Internet Research, 30/03/2026 (im Druck)*, 89178. https://doi.org/10.2196/89178

Kherbache, A., Mertens, E., & Denier, Y. (2022). Moral distress in medicine: An ethical analysis. *Journal of Health Psychology*, *27*(8), 1971–1990. https://doi.org/10.1177/13591053211014586

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching—A review. *Psychological Bulletin*, *136*(5), 849–874. https://doi.org/10.1037/a0019842

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science*, *33*(1), 126–148. https://doi.org/10.1287/orsc.2021.1549

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual Review of Psychology*, *66*, 799–823. https://doi.org/10.1146/annurev-psych-010213-115043

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, *61*(10), 36–43. https://doi.org/10.1145/3233231

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin, 127*(2), 267–286. https://doi.org/10.1037/0033-2909.127.2.267

Lombi, L., & Rossero, E. (2024). How artificial intelligence is reshaping the autonomy and boundary work of radiologists: A qualitative study. *Sociology of Health & Illness*, *46*(2), 200–218. https://doi.org/10.1111/1467-9566.13702

London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report*, *49*(1), 15–21. https://doi.org/10.1002/hast.973

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354

Perez, F., Conway, N., Peterson, J., & Roques, O. (2024). Me, my work and AI: How radiologists craft their work and identity. *Journal of Vocational Behavior*, *155*, 104042. https://doi.org/10.1016/j.jvb.2024.104042

Reason, J. T. (1990). *Human error*. Cambridge, UK: Cambridge University Press.

Reason, J. T. (1998). Achieving a safe culture: Theory and practice. *Work & Stress*, *12*(3), 293–306. https://doi.org/10.1080/02678379808256868

Rony, M. K. K., Parvin, M. R., Wahiduzzaman, M., Debnath, M., Bala, S. D., & Kayesh, I. (2024). »I wonder if my years of training and expertise will be devalued by machines«: Concerns about the replacement of medical professionals by artificial intelligence. *SAGE Open Nursing*, *10*, 23779608241245220. https://doi.org/10.1177/23779608241245220

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Schwarz, N., & Clore, G. L. (2007). Feelings and phenomenal experiences. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (2nd ed., pp. 385–407). New York, NY: The Guilford Press

Shonhe, L., & Min, Q. (2025). Mitigating AI-induced professional identity threat and fostering adoption in the workplace. *AI & Society*, *40*(5), 4079–4092. https://doi.org/10.1007/s00146-024-02170-0

Sittig, D. F., & Singh, H. (2010). A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. *Quality and Safety in Health Care*, *19*(Suppl 3), i68–i74. https://doi.org/10.1136/qshc.2010.042085

Steinborn, M. B., Langner, R., & Huestegge, L. (2017). Mobilizing cognition for speeded action: Try-harder instructions promote motivated readiness in the constant-foreperiod paradigm. *Psychological Research*, *81*, 1135–1151. https://doi.org/10.1007/s00426-016-0810-1

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., ... Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine*, *25*(9), 1337–1340. https://doi.org/10.1038/s41591-019-0548-6

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Cognitive psychology: Rare items often missed in visual searches. *Nature*, *435*(7041), 439–440. https://doi.org/10.1038/435439a

Sara Dokmanac, Joana Günther

III. Künstliche Intelligenz in der radiologischen Diagnostik: Rechtliche Aspekte

1. Allgemeine regulatorische Einbettung

Künstliche Intelligenz (KI) birgt im Gesundheitswesen ein enormes Innovationspotenzial. Sie kann wesentlich dazu beitragen, diagnostische und therapeutische Verfahren zu verbessern, Abläufe effizienter zu gestalten und damit die Versorgungsqualität insgesamt zu erhöhen. Sowohl bei der Auswertung medizinischer Bilddaten, der Analyse großer Datenmengen zur Krankheitsfrüherkennung oder in der Entwicklung personalisierter Behandlungsansätze, etwa personalisierter Therapiekonzepte, kann KI medizinische Prozesse nicht nur beschleunigen, sondern auch qualitativ aufwerten. Gleichzeitig wirft der Einsatz von KI-Systemen grundlegende rechtliche und regulatorische Fragestellungen auf, die sowohl Entwickelnde als auch Anwendende vor erhebliche Herausforderungen stellen. In Deutschland fehlt es bislang an einem einheitlichen, speziell auf den Einsatz von KI im Gesundheitswesen zugeschnittenen Regelwerk. Stattdessen besteht ein komplexes Zusammenspiel unterschiedlicher Normen, darunter insbesondere zivilrechtliche, datenschutzrechtliche und haftungsrechtliche Vorschriften. Relevante Rechtsgrundlagen sind unter anderem die Bundesdatenschutzgesetz (BDSG), das Medizinprodukterecht (MPG), das Produkthaftungsgesetz (ProdHaftG) sowie berufsrechtliche Regelungen. Diese fragmentierte Rechtslage führt in der Praxis häufig zu Unsicherheiten und Auslegungsproblemen.

KI-basierte Anwendungen, die als Medizinprodukte einzuordnen sind, unterliegen bislang der europäischen Medizinprodukteverord-

nung[1] (im Folgenden: MP-VO). Diese definiert grundlegende Anforderungen an Sicherheit, Leistungsfähigkeit und klinische Bewertung, berücksichtigt jedoch nicht die spezifischen Charakteristika von KI-Systemen – etwa ihre Komplexität, den autonomen Entscheidungscharakter oder die fehlende Nachvollziehbarkeit von Entscheidungswegen (Black-Box-Problematik).

Vor diesem Hintergrund legte die Europäische Kommission bereits im April 2021 den Entwurf für eine Verordnung zur Regulierung Künstlicher Intelligenz vor. Die am 13. Juni 2024 verabschiedete Verordnung zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz[2] (KI-Verordnung, im Folgenden: KI-VO) ergänzt das bestehende europäische Rechtsgefüge sowohl sektorübergreifend als auch sektorspezifisch. Ihr Ziel ist es, den rechtlichen Rahmen für den Einsatz von KI in der Europäischen Union zu harmonisieren und insbesondere in sensiblen Bereichen – wie dem Gesundheitswesen – ein hohes Maß an Sicherheit, Transparenz und Grundrechtskonformität zu gewährleisten.

2. KI-Verordnung

2.1 Vorab: Forschungsprivileg

Im gesundheitsbezogenen Kontext, in dem Forschung und Entwicklung eng miteinander verzahnt sind, spielt das sog. Forschungsprivileg eine Rolle. Bevor auf die konkreten Regelungen der Verordnung eingegangen wird, soll dieser Ausnahmebereich daher zunächst grob skizziert werden. Unter einem Forschungsprivileg versteht

1 Verordnung (EU) 2017/745 des Europäischen Parlaments und des Rates vom 5. April 2017 über Medizinprodukte, zur Änderung der Richtlinie 2001/83/EG, der Verordnung (EG) Nr. 178/2002 und der Verordnung (EG) Nr. 1223/2009 und zur Aufhebung der Richtlinien 90/385/EWG und 93/42/EWG des Rates, (ABl. L 117 S. 1, 2019 L 117 S. 9, L 334 S. 165; 2021 ABl. L 241 S. 7).

2 Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz und zur Änderung der Verordnungen (EG) Nr. 300/2008, (EU) Nr. 167/2013, (EU) Nr. 168/2013, (EU) 2018/858, (EU) 2018/1139 und (EU) 2019/2144 sowie der Richtlinien 2014/90/EU, (EU) 2016/797 und (EU) 2020/1828 (Verordnung über künstliche Intelligenz), (ABl. L 2024/1689, 12.7.2024).

man eine gesetzlich verankerte Ausnahme oder Sonderregelung, die wissenschaftliche Forschungstätigkeiten ganz oder teilweise von bestimmten rechtlichen Anforderungen ausnimmt (Martini & Wendehorst/*Wendehorst*, 2024, Art. 2 Rn. 83 f.). Ziel solcher Regelungen ist es, die wissenschaftliche Freiheit zu wahren und Innovationen nicht durch übermäßige regulatorische Hürden zu hemmen. Auch die KI-VO enthält entsprechende Ausnahmen zugunsten der wissenschaftlichen Forschung. Nach Art. 2 Abs. 6 KI-VO findet die Verordnung keine Anwendung auf KI-Systeme, die ausschließlich zum Zweck der wissenschaftlichen Forschung und Entwicklung in Betrieb genommen werden. Sobald ein KI-System für andere – insbesondere kommerzielle oder operative – Zwecke eingesetzt wird, greift das Privileg nicht mehr, und das System unterfällt in vollem Umfang den Anforderungen der KI-VO. Darüber hinaus stellt Art. 2 Abs. 8 KI-VO klar, dass Forschungs-, Test- und Entwicklungstätigkeiten zu KI-Systemen oder KI-Modellen vor Inverkehrbringen oder Inbetriebnahme grundsätzlich nicht unter die Verordnung fallen, vorausgesetzt es handelt sich um Tests unter Realbedingungen. Ziel ist es, frühe Entwicklungsphasen von KI-Systemen nicht durch die umfangreichen Vorgaben der Verordnung zu erschweren.

Die Privilegierung wissenschaftlicher Forschung ist weder im europäischen noch im nationalen Recht ein neues Konzept. So sieht etwa § 2 Abs. 2 Nr. 1 Gendiagnostikgesetz (im Folgenden: GenDG) vor, dass das Gesetz nicht für genetische Untersuchungen und Analysen sowie den Umgang mit genetischen Proben und Daten zu Forschungszwecken gilt. Auch das Datenschutzrecht enthält entsprechende Öffnungsklauseln: Art. 89 DSGVO eröffnet den Mitgliedstaaten die Möglichkeit, für wissenschaftliche, historische oder statistische Zwecke besondere Regelungen vorzusehen. Von dieser Öffnungsklausel hat der deutsche Gesetzgeber mit § 27 BDSG Gebrauch gemacht. Die Norm konkretisiert damit die Anforderungen und Voraussetzungen für die Verarbeitung insbesondere besonderer Kategorien personenbezogener Daten im Rahmen wissenschaftlicher Forschung und ergänzt so den unionsrechtlichen Rahmen durch spezifische nationale Vorgaben. Abweichend von Art. 9 Abs. 1 DSGVO kann eine solche Verarbeitung auch ohne Einwilligung der betroffenen Person zulässig sein, soweit sie erforderlich ist und das Interesse der verantwortlichen Stelle an der Verarbeitung das schutzwürdige Interesse der betroffenen Person erheblich überwiegt.

Gemäß Erwägungsgrund Nr. 159 DSGVO ist der Forschungsbegriff dabei weit auszulegen, um dem großen Erkenntnispotential der Auswertung und Erforschung gesundheitsbezogener Daten im Hinblick auf Krankheitsverläufe, Therapien und die Versorgungspraxis Rechnung zu tragen (Spitz, et. al, 2021). Zwar enthält die KI-VO – anders als die DSGVO in Erwägungsgrund Nr. 159 – keinen ausdrücklichen Verweis auf die in Art. 179 AEUV (Vertrag über die Arbeitsweise der Europäischen Union) formulierte Idee eines europäischen Forschungsraums, dennoch ist Art. 2 Abs. 6 KI-VO gleichermaßen in diesem Sinne auszulegen (Spitz, 2025).

Die Privilegierung der Forschung im Rahmen der KI-VO zielt darauf ab, die wissenschaftliche Freiheit zu schützen und Europa als innovationsfreundlichen Standort für digitale Technologien zu stärken (Martini & Wendehorst/*Wendehorst*, 2024, Art. 2 Rn. 73).

2.2 Ziele der Verordnung

Ausgangspunkt der folgenden Betrachtung ist die Zielsetzung der Verordnung. Bereits in Erwägungsgrund Nr. 1 KI-VO heißt es:

> »Zweck dieser Verordnung ist es, das Funktionieren des Binnenmarkts zu verbessern, indem ein einheitlicher Rechtsrahmen insbesondere für die Entwicklung, das Inverkehrbringen, die Inbetriebnahme und die Verwendung von Systemen künstlicher Intelligenz (KI-Systeme) in der Union im Einklang mit den Werten der Union festgelegt wird […]«.

Zwar existieren für die verschiedenen rechtlichen Teilbereiche – etwa das Medizinprodukterecht, das Berufsrecht oder das Haftungsrecht – bereits einschlägige Rechtsnormen. Diese enthalten jedoch bislang keine spezifischen Regelungen für den Einsatz von KI. Mit der KI-VO wird nun erstmals eine sektorübergreifende und zugleich horizontal wirkende Regulierung für Künstliche Intelligenz in Europa angestrebt. Zu diesem Zweck legt die Verordnung einheitliche harmonisierende Anforderungen, insbesondere im Hinblick auf das Inverkehrbringen und die Inbetriebnahme von KI-Systemen, für alle Sektoren fest. Ein zentrales Leitmotiv ist hierbei die Stärkung des Binnenmarktes, der sowohl Rechtssicherheit als auch fairen Wettbewerb fördern soll.

Darüber hinaus verfolgt die KI-VO gemäß Art. 1 KI-VO das Ziel, das Vertrauen in KI zu stärken und gleichzeitig ein hohes Schutzni-

veau in Bezug auf Gesundheit, Sicherheit und die in der Charta verankerten Grundrechte – einschließlich Demokratie, Rechtsstaatlichkeit und Umweltschutz – vor schädlichen Auswirkungen von KI-Systemen in der Union zu gewährleisten. Dabei sollen Innovationen gezielt gefördert werden, während zugleich betont wird, dass KI stets im Dienste des Menschen eingesetzt werden soll und diesen weder ersetzen noch kontrollieren darf, sog. menschenzentrierte KI (Erwägungsgrund Nr. 1, 6, 8 und 27 KI-VO).

Der Begriff des KI-Systems wird in Art. 3 Nr. 1 KI-VO legal definiert als:

> »ein maschinengestütztes System, das für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt ist und das nach seiner Betriebsaufnahme anpassungsfähig sein kann und das aus den erhaltenen Eingaben für explizite oder implizite Ziele ableitet, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können.«

Angesichts der wachsenden Bedeutung von KI soll sichergestellt werden, dass KI-Systeme verantwortungsvoll, sicher und grundrechtskonform entwickelt und eingesetzt werden. Ein besonderes Augenmerk liegt dabei auf der Vermeidung von Verzerrungen durch fehlerhafte oder einseitige Trainingsdaten (sog. Bias) und daraus resultierenden Diskriminierungspotenzialen. Die menschliche Kontrolle bleibt darüber hinaus auch bei KI-gestützten Entscheidungen substanziell. Insgesamt verfolgt die KI-VO einen Regulierungsansatz, der darauf abzielt, harmonisierende Vorgaben für die Verwendung von KI in der europäischen Union zu schaffen und Europa als zentralen Standort für KI zu etablieren.

2.3 Risikobasierter Ansatz

2.3.1 Grundkonzept

Das zentrale Regelungskonzept der KI-VO ist der risikobasierte Ansatz. Die KI-VO bestimmt die regulatorischen Anforderungen in Abhängigkeit des jeweiligen Gefährdungspotenzials eines KI-Systems. Art und Inhalt der Vorschriften richten sich dabei nach Intensität und Umfang der Risiken, die von dem KI-System ausgehen: Je höher das Risiko für die Gesundheit, Sicherheit oder die Grundrechte von

Personen, desto strenger die Anforderungen. Zu diesem Zweck unterteilt die Verordnung KI-Systeme in vier Risikokategorien: unannehmbares Risiko, hohes Risiko, beschränktes Risiko und minimales Risiko.

KI-Systeme, die nach Einschätzung des europäischen Gesetzgebers ein unannehmbares Risiko für die Sicherheit oder die Grundrechte von Personen darstellen, sind gemäß Art. 5 KI-VO grundsätzlich verboten. Davon werden KI-Systeme umfasst, die Techniken der unterschwelligen Beeinflussung oder manipulative, täuschende Techniken einsetzen, um das Verhalten einer Person wesentlich zu ändern und ihr erheblichen Schaden zuzufügen (Art. 5 Abs. 1 lit. a). Darüber hinaus sind KI-Systeme umfasst, die Schutzbedürftige ausnutzen (Art. 5 Abs. 1 lit. b) sowie soziale Bewertungen durchführen (Art. 5 Abs. 1 lit. c). Hierzu zählen insbesondere Systeme, die auf manipulative Weise Einfluss auf das Verhalten oder die Entscheidungsfreiheit von Personen nehmen und dadurch mit hinreichender Wahrscheinlichkeit erheblichen Schaden verursachen können, beispielsweise das »Social Scoring«. Im medizinischen Bereich finden derartige Systeme in der Regel keine Anwendung. Den normativen Kern der Verordnung bilden die Hochrisiko-KI-Systeme im Sinne von Art. 6 KI-VO, die auch den Großteil der medizinischen KI-Anwendungen ausmachen. Diese Kategorie umfasst Systeme, deren Einsatz in sicherheitskritischen oder grundrechtsrelevanten Kontexten erfolgt – darunter insbesondere kritische Infrastrukturen, der Bildungsbereich, das Personalwesen und das Gesundheitswesen. Für Hochrisiko-KI-Systeme gelten umfangreiche Anforderungen, etwa hinsichtlich Risikomanagement, Sicherheit, Nachvollziehbarkeit, Datenqualität und Überwachung (Abschnitt 2.3.2 [»Anforderungen an Hochrisiko-KI-Systeme«]).

Unter KI-Systeme mit beschränktem Risiko fallen solche, von denen zwar keine unmittelbare Gefährdung ausgeht, die aber dennoch Auswirkungen auf das Verhalten oder die Wahrnehmung von Nutzer*innen haben können. Dazu zählen etwa KI-gestützte Chatbots oder automatisierte Assistenzsysteme. Diese Systeme müssen bestimmten Informations- und Transparenzanforderungen genügen. Im Gesundheitswesen wären hier etwa KI-gestützte Systeme zur Terminvergabe einzuordnen. Die KI-VO verlangt in diesen Fällen, dass für Nutzende klar erkennbar ist, dass sie mittels eines automatisierten Systems interagieren und nicht mit medizinischem Fachper-

sonal. Zudem sollen die bereitgestellten Informationen in einfacher und verständlicher Weise kommuniziert werden.

Schließlich gibt es KI-Systeme mit minimalem Risiko, von denen kein besonderes Gefährdungspotenzial ausgeht. Hierzu zählen etwa rein unterstützende oder administrative KI-Systeme ohne sicherheitsrelevante Wirkung. Sie sind daher von regulatorischen Pflichten, die über das allgemeine EU-Recht hinausgehen, weitgehend ausgenommen.

Im Anschluss soll dargestellt werden, welche konkreten Anforderungen sich aus der Verordnung für die jeweiligen Systeme ergeben (siehe dazu sogleich unter Abschnitt 2.3.2 [»Anforderungen an Hochrisiko-KI-Systeme«]). Dabei ist hervorzuheben, dass sich die Regelungen der KI-VO auf den gesamten Lebenszyklus eines KI-Systems erstrecken – von der Entwicklung und Konzeption über das Inverkehrbringen und den Betrieb bis hin zur Marktüberwachung und Sanktionierung von Verstößen. Sie erfassen dabei sowohl technische Anforderungen als auch organisatorische Pflichten und flankierende Durchsetzungsmechanismen.

2.3.2 Anforderungen an Hochrisiko-KI-Systeme

Ob es sich bei einem KI-System um ein regulierungsbedürftiges Hochrisiko-KI-System handelt, hat die jeweils verpflichtete Person grundsätzlich eigenverantwortlich zu beurteilen. Der objektive Maßstab für die Einschätzung ergibt sich aus Art. 6 KI-VO (Klawonn, 2025, Art. 6 Rn. 1). Gemäß Art. 6 Abs. 1 KI-VO liegt ein Hochrisiko-KI-System vor, wenn zwei kumulative Voraussetzungen erfüllt sind:

1. es handelt sich um ein Sicherheitsbauteil oder Produkt im Sinne einer der in Anhang I genannten Harmonisierungsrechtsvorschriften (lit. a), und
2. das Produkt oder Sicherheitsbauteil unterliegt im Rahmen des Inverkehrbringens oder der Inbetriebnahme einer Konformitätsbewertung durch Dritte nach Maßgabe der jeweiligen sektorspezifischen Vorschriften (lit b).

Zu den in Anhang I Abschnitt A aufgeführten Harmonisierungsrechtsvorschriften zählen unter anderem die Medizinprodukteverordnung sowie die Verordnung (EU) 2017/746 über In-vitro-Diagnostika (ABl. L 117 vom 5.5.2017, S. 176). Daraus folgt, dass KI-

basierte Medizinprodukte regelmäßig als Hochrisiko-KI-Systeme einzustufen sind und den spezifischen Anforderungen der KI-VO unterliegen (siehe dazu unter Abschnitt 2.4 [»Beteiligte und Verantwortliche«]).

Darüber hinaus gelten gemäß Art. 6 Abs. 2 KI-VO auch solche KI-Systeme als hochriskant, die in Anhang III ausdrücklich benannt sind. Dabei handelt es sich um Systeme, bei denen aufgrund ihres Einsatzbereichs eine erhebliche Beeinträchtigung der Gesundheit, Sicherheit oder Grundrechte natürlicher Personen zu befürchten ist. Beispiele hierfür sind KI-Systeme zur biometrischen Identifikation (Nr. 1) sowie Anwendungen im Bereich der Bereitstellung grundlegender privater und öffentlicher Dienste und Leistungen (Nr. 5) – etwa zur Bewertung eines Anspruchs auf Sozialleistungen oder zur Risikobewertung und Preisbildung im Rahmen von Lebens- und Krankenversicherungen. Ein in Art. 6 Abs. 2 i. V. m. Anhang III genanntes System kann gemäß Art. 6 Abs. 3 KI-VO ausnahmsweise nicht als hochriskant eingestuft werden, wenn es im konkreten Fall nachweislich kein erhebliches Risiko für die genannten Rechtsgüter birgt. Soweit weder der Ausnahmetatbestand des Art. 6 Abs. 3 KI-VO greift noch ein Ausschluss nach sonstigen Vorschriften gegeben ist, sind für die Hochrisiko-Systeme die Anforderungen aus Art. 8 ff. KI-VO einzuhalten. Dazu zählen insbesondere die Einrichtung eines Risikomanagementsystems (Art. 9 KI-VO), die Qualitätssicherung von Trainingsdaten (Art. 10 KI-VO), Dokumentations- und Aufzeichnungspflichten (Art. 11 und 12 KI-VO) und die Sicherstellung einer menschlichen Aufsicht (Art. 14 KI-VO). Ergänzt werden diese Pflichten durch Vorgaben zur Durchführung eines Konformitätsbewertungsverfahren (Art. 43 KI-VO) sowie zur Anbindung einer dazugehörigen CE-Kennzeichnung (Art. 48 KI-VO). Eine vertiefte Auseinandersetzung mit den Pflichten der einzelnen Beteiligten erfolgt sogleich unter Abschnitt 2.4.1 (»Pflichten gemäß KI-VO«).

2.4 Beteiligte und Verantwortliche

Der Einsatz von KI-Systemen im medizinischen Bereich wirft eine Vielzahl regulatorischer Fragen auf. Die in diesem Zusammenhang relevanten Wirtschaftsbeteiligten, die im Folgenden näher beleuchtet werden, treffen teils unterschiedliche und teils sich überschneidende

Pflichten. Wie bereits in Abschnitt 2.2 (»Ziele der Verordnung«) erwähnt, ist die KI-VO der bislang einzige KI-spezifische Normtext und insbesondere auch auf KI-Systeme im Gesundheitssektor anwendbar (Abschnitt 2.3.2 [»Anforderungen an Hochrisiko-KI-Systeme«]). Erfüllt KI-basierte Software jedoch eine medizinische Zweckbestimmung, gilt sie gemäß Art. 2 Abs. 1 MP-VO auch als Medizinprodukt:

> »›Medizinprodukt‹ bezeichnet ein Instrument, einen Apparat, ein Gerät, eine Software, ein Implantat, ein Reagenz, ein Material oder einen anderen Gegenstand, das dem Hersteller zufolge für Menschen bestimmt ist und allein oder in Kombination einen oder mehrere der folgenden spezifischen medizinischen Zwecke erfüllen soll:
> – Diagnose, Verhütung, Überwachung, Vorhersage, Prognose, Behandlung oder Linderung von Krankheiten,
> – Diagnose, Überwachung, Behandlung, Linderung von oder Kompensierung von Verletzungen oder Behinderungen,
> – Untersuchung, Ersatz oder Veränderung der Anatomie oder eines physiologischen oder pathologischen Vorgangs oder Zustands,
> – Gewinnung von Informationen durch die In-vitro-Untersuchung von aus dem menschlichen Körper – auch aus Organ-, Blut- und Gewebespenden – stammenden Proben, und dessen bestimmungsgemäße Hauptwirkung im oder am menschlichen Körper weder durch pharmakologische oder immunologische Mittel noch metabolisch erreicht wird, dessen Wirkungsweise aber durch solche Mittel unterstützt werden kann.
> Die folgenden Produkte gelten ebenfalls als Medizinprodukte:
> – Produkte zur Empfängnisverhütung oder -förderung,
> – Produkte, die speziell für die Reinigung, Desinfektion oder Sterilisation der in Artikel 1 Absatz 4 genannten Produkte und der in Absatz 1 dieses Spiegelstrichs genannten Produkte bestimmt sind.«

Die Klassifizierung der KI-Software als eigenständiges Medizinprodukt der verschiedenen Klassen erfolgt gemäß Anhang VIII Regel 11 MP-VO allerdings abhängig vom potenziellen Risiko, das von ihrer Nutzung ausgeht. Dies ähnelt der Vorgehensweise bei der KI-VO (Abschnitt 2.3.1 [»Grundkonzept«]). Die umfangreiche Aufzählung der spezifischen medizinischen Zwecke für die Einordnung als Medizinprodukt verdeutlicht zudem, dass die MP-VO in den meisten Fällen von medizinischer KI parallel zur KI-VO anwendbar sein wird. Ein möglicher Anwendungsfall ist der Einsatz einer KI-gestütz-

ten Software zur automatischen Erkennung von Brustkrebs.[3] In diesem Fall unterliegt die medizinische KI sowohl dem Anwendungsbereich der KI-VO als auch der MP-VO. Dies hat eine komplexe und unübersichtliche Rechtslage zur Folge, die zu Unsicherheiten bei den verantwortlichen Personen führen kann. Der folgende Abschnitt verschafft einen Überblick über die relevanten Wirtschaftsbeteiligten und beleuchtet die Abgrenzungsproblematik im Kontext medizinischer KI-Systeme.

Wie in Abschnitt 2.3.2 (»Anforderungen an Hochrisiko-KI-Systeme«) festgestellt, sind medizinische KI-Systeme als Hochrisiko-KI-Systeme einzustufen, sodass die KI-VO uneingeschränkt Anwendung findet. »Anbieter«[4] wird gemäß Art. 3 Nr. 3 KI-VO definiert als »natürliche oder juristische Person […], die ein KI-System […] entwickelt oder entwickeln lässt und es unter ihrem eigenen Namen oder ihrer Handelsmarke in Verkehr bringt oder […] in Betrieb nimmt, sei es entgeltlich oder unentgeltlich«. »Betreiber« sind gemäß Art. 3 Nr. 4 KI-VO natürliche oder juristische Personen oder Organisationen, die ein KI-System in eigener Verantwortung verwenden, wobei die ausschließlich private Nutzung explizit ausgenommen ist. Exemplarisch können an dieser Stelle Gesundheitseinrichtungen sowie medizinisches Fachpersonal angeführt werden. Eine besondere Herausforderung besteht auch in der Abgrenzung der Rolle des Anbieters von der des Betreibers. Dies ist insbesondere in Konstellationen relevant, in denen Beteiligte – etwa Gesundheitseinrichtungen – beide Rollen gleichzeitig ausüben, beispielsweise wenn ein eigenständig entwickeltes medizinisches KI-System im Rahmen derselben Gesundheitseinrichtung zum Einsatz kommt. Darüber hinaus können auch die in Art. 3 Nr. 8 KI-VO aufgeführten Beteiligten – wie »Einführer« (Art. 3 Nr. 6 KI-VO), »Händler« (Art. 3 Nr. 7 KI-VO) oder »Produkthersteller« (Art. 3 Nr. 8 KI-VO) – von Relevanz sein. Im Folgenden soll der Fokus jedoch auf die in

3 Siehe PRAIM-Studie, in der mehr als 460.000 Frauen im Rahmen des nationalen Mammographie-Screening-Programms (MSP) untersucht wurden. Der Einsatz einer KI-gestützten Doppelbefundung führte dabei zu einer 17,6 % höheren Brustkrebs-Erkennungsrate im Vergleich zur herkömmlichen Befundung (Eisemann et al., 2025).

4 Gesetzlich festgelegte Begriffe (z. B. »Anbieter«, »Betreiber«) werden in der im Gesetz verwendeten Form wiedergegeben. Im Übrigen erfolgt eine genderneutrale Formulierung.

diesem Zusammenhang wichtigsten Anbieter und Betreiber gelegt werden.

In der MP-VO kommen die Begriffe »Anbieter« oder »Betreiber« nicht vor. Stattdessen sollen die in der MP-VO verwendeten Definitionen zu »Hersteller« und »Anwender«, sowie die in der am 14.02.2025 ausgefertigten Verordnung über das Betreiben und Benutzen von Medizinprodukten[5] (Medizinprodukte-Betreiberverordnung, im Folgenden: MPBetreibV) vorkommenden »Betreiber« und »Benutzer« näher beleuchtet werden.

In Art. 2 Nr. 30 MP-VO ist der Begriff des Herstellers definiert als »natürliche oder juristische Person, die ein Produkt herstellt oder als neu aufbereitet bzw. entwickeln, herstellen oder als neu aufbereiten lässt und dieses Produkt unter ihrem eigenen Namen oder ihrer Marke vermarktet«. Der Begriff ähnelt im Wesentlichen dem Begriff des Anbieters i. S. d. KI-VO (Martini & Wendehorst/ *Wendehorst*, 2024, Art. 3 Rn. 64). Als Anwender sind gemäß Art. 2 Nr. 37 MP-VO hingegen alle »Angehörigen der Gesundheitsberufe oder Laien, [die] ein Medizinprodukt anwende[n]« qualifiziert. In diesem Kontext nennt das nationale Recht gemäß § 2 Abs. 3 MPBetreibV den Begriff des Benutzers und definiert ihn als eine Person, die »ein Produkt […] am Patienten einsetzt«. Beide Definitionen beziehen sich dabei auf Personen, die ein Produkt tatsächlich anwenden. Sie sehen dabei die Möglichkeit vor, dass auch weisungsgebundene Personen wie Pflegekräfte in den Kreis der Betroffenen fallen. Das nationale Recht fasst den Begriff lediglich enger, indem es den Einsatz gezielt zur Wirkung bei Patient*innen fordert und damit typischerweise medizinisches Personal betrifft (Rehmann & Wagner/*Rehmann*, 2023, Art. 2 Rn. 52). Die MP-VO hingegen umfasst darüber hinaus auch fachfremde Personen.

In der MP-VO wird der Begriff des Betreibers nicht verwendet. Gemäß § 2 Abs. 2 S. 1 MPBetreibV ist damit »jede natürliche oder juristische Person, die für den Betrieb der Gesundheitseinrichtung verantwortlich ist, in der das Produkt durch dessen Beschäftigte betrieben oder benutzt wird«, gemeint. Darüber hinaus wird das medizinische Personal selbst zum Betreiber, sofern ein eigenes Pro-

5 Medizinprodukte-Betreiberverordnung vom 14. Februar 2025 (BGBl. 2025 I Nr. 38), die durch Artikel 1 der Verordnung vom 14. Februar 2025 (BGBl. 2025 I Nr. 39) geändert worden ist.

dukt in die Gesundheitseinrichtung mitgebracht und dort verwendet oder auch außerhalb der Einrichtung zur Nutzung bereitgestellt wird (§ 2 Abs. 2 S. 2, 3 MPBetreibV). Als Gesundheitseinrichtung ist gemäß § 2 Abs. 4 MPBetreibV jeder Ort zu verstehen, an dem Medizinprodukte von medizinischem Fachpersonal, Pflegekräften oder anderen befugten Personen verwendet oder betrieben werden. Auch hier ist eine gewisse Ähnlichkeit zum Betreiberbegriff i. S. d. KI-VO erkennbar. Während die KI-VO jede Verwendung in eigener Verantwortung – außer ausschließlich private – erfasst, konzentriert sich die vorliegende Definition auf den Betrieb durch Beschäftigte innerhalb einer Gesundheitseinrichtung und legt den Fokus auf die Organisationsverantwortung. Eine private Nutzung ist damit nicht ausgeschlossen.

Medizinische KI-Systeme stellen einen Sonderfall dar, in dem die Regelungen der KI-VO und die der MP-VO in der Regel parallel anwendbar sind. Diese Doppelregulierung kann erweiterte Pflichten für die Betroffenen zur Folge haben und die eindeutige Zuweisung bestehender Pflichten erschweren. Die folgenden Abschnitte geben einen Überblick über die bestehenden Pflichten der wichtigsten beteiligten Personen im Gesundheitswesen.

2.4.1 Pflichten gemäß KI-VO

a) Pflichten der Anbieter

Der Begriff des Anbieters i. S. d. KI-VO wurde in Abschnitt 2.4 (»Beteiligte und Verantwortliche«) bereits definiert. Seine Relevanz zeigt sich insbesondere im Hinblick auf den umfassenden, aber nicht abschließenden Pflichtenkatalog des Art. 16 KI-VO. Gemäß Art. 16 lit. a) KI-VO muss der Anbieter zunächst sicherstellen, dass das infrage stehende Hochrisiko-KI-System den Anforderungen der Art. 8 bis 15 KI-VO genügt (Abschnitt 2.3.2 [»Anforderungen an Hochrisiko-KI-Systeme«]). Damit trägt er die größte Verantwortung für die Qualitätssicherung, die Gefahrenabwehr und auch die Risikovorsorge (Hilgendorf & Roth-Isigkeit/*Linardatos*, 2023, § 7 Rn. 1). Um die Kommunikation mit anderen beteiligten Personen zu erleichtern, ist der Anbieter gemäß Art. 16 lit. b) KI-VO dazu verpflichtet, seine Kontaktdaten so bereitzustellen, dass eine Kontaktaufnahme ohne Weiteres möglich ist, etwa auf dem KI-System, der Verpackung oder

der beigefügten Dokumentation (Martini & Wendehorst/*Eisenberger*, 2024, Art. 16 Rn. 19).

Art. 16 KI-VO verweist darüber hinaus auf verschiedene Pflichten, die sich aus anderen Bestimmungen der Verordnung ergeben. Dazu gehören insbesondere die folgenden Pflichten: die Pflicht zur Einrichtung eines Qualitätsmanagementsystems (Art. 17 KI-VO), die Pflicht zur Aufbewahrung von technischer Dokumentation und automatisch erzeugten Protokollen (Art. 18 und 19 KI-VO), die Registrierungspflicht in einer EU-Datenbank (Art. 49 Abs. 1 i. V. m. Art. 71 KI-VO), sowie die Pflicht, bei Verstößen oder Risiken geeignete Korrekturmaßnahmen zu ergreifen und, soweit notwendig, Betreiber, Bevollmächtigte oder Einführer zu informieren (Art. 20 KI-VO). Vor der Einführung oder Nutzung des Systems muss der Anbieter zudem auch ein Konformitätsbewertungsverfahren gemäß Art. 43 KI-VO durchführen. In diesem Zusammenhang wird die Übereinstimmung mit den gesetzlichen Vorgaben überprüft, sodass »KI-spezifische Risiken« möglichst im Vorhinein vermieden werden können (Gerdemann, 2024). Sobald dieses Verfahren erfolgreich abgeschlossen wurde, muss eine sogenannte EU-Konformitätserklärung gemäß Art. 47 KI-VO erstellt werden. Diese stellt verbindlich die Erfüllung der genannten Anforderungen fest (Art. 47 Abs. 2 S. 1 KI-VO) und ordnet eindeutig zu, für welches KI-System sie gilt (Art. 47 Abs. 1 S. 2 KI-VO).

Mit der Ausstellung übernimmt der Anbieter die volle Verantwortung für die Einhaltung der Anforderungen (Art. 47 Abs. 4 S. 1 KI-VO). Er muss die Erklärung gegebenenfalls aktualisieren (Art. 47 Abs. 4 S. 2 KI-VO) und sie mindestens zehn Jahre lang aufbewahren. Auf Anfrage muss er sie den nationalen Behörden bereitstellen (Art. 47 Abs. 1 S. 1 KI-VO). Daraufhin muss eine CE-Kennzeichnung gemäß Art. 48 KI-VO erfolgen. Sie sollte sich nach den Grundsätzen des Art. 30 der Verordnung (EU) Nr. 765/2008 über die Vorschriften für die Akkreditierung und die Marktüberwachung im Zusammenhang mit der Vermarktung von Produkten richten. Mit ihr wird die Konformität mit den geltenden Anforderungen sowie die Übernahme der Verantwortung durch den Anbieter bescheinigt (Art. 47 Abs. 2, 3 KI-VO). Auf begründete Nachfrage der zuständigen Behörde treffen den Anbieter zudem gemäß Art. 16 lit. k) KI-VO Nachweispflichten im Hinblick auf die Einhaltung der Anforderungen aus Art. 8 bis 15 KI-VO. Er hat außerdem gemäß

Art. 16 lit. l) KI-VO sicherzustellen, dass die Barrierefreiheitsanforderungen[6] erfüllt sind. Die mit dem Einsatz von Hochrisiko-KI-Systemen – insbesondere im Gesundheitswesen – verbundenen Risiken begründen den Bedarf nach verstärkter Kontrolle und machen die Anbieter in hohem Maße zu Verantwortlichen.

b) Pflichten der Betreiber

Ungeachtet des bereits umfangreichen Pflichtenkatalogs des Anbieters, der in Abschnitt 2.4.1.1 (»Pflichten der Anbieter«) näher beleuchtet wurde, ergeben sich für Betreiber im Sinne der KI-VO – bislang im Produktrecht eher ungewöhnlich – darüber hinausgehende Pflichten. Dort stehen regelmäßig »Endnutzer« und »Anwender« im Fokus, für die regelmäßig Schutzmaßnahmen als Verpflichtungen vorgesehen sind (Martini & Wendehorst/*Eisenberger*, 2024, Art. 3 Rn. 80). Gerade im Zusammenhang mit Hochrisiko-KI-Systemen – besonders im Gesundheitswesen – ergibt aber ein gesonderter Pflichtenkatalog für Betreiber, die dem tatsächlichen Einsatz deutlich näher sind als die Anbieter, durchaus Sinn und kann zu einer Minimierung des Risikos führen (Schuh & Witt, 2025).

In diesem Zusammenhang kann die Frage aufgeworfen werden, ob Mitarbeitende als Betreiber qualifiziert werden können, wenn sie das KI-System im Rahmen ihrer beruflichen Tätigkeit nutzen. Ein Blick auf den Wortlaut »in eigener Verantwortung« (Art. 3 Nr. 4 KI-VO) lässt zunächst den Schluss zu, dass eine solche Einordnung nicht vorgesehen ist. Soweit Mitarbeitende im Rahmen von dienstlichen Weisungen innerhalb einer hierarchischen Struktur tätig sind – was in Gesundheitseinrichtungen regelmäßig der Fall sein sollte –, haben sie keine Entscheidungsbefugnisse inne und handeln auch nicht in eigener Verantwortung (Hilgendorf & Roth-Isigkeit/*Gless & Janal*, 2023, § 2 Rn. 45). Es sind allerdings auch Ausnahmefälle denkbar. So kann beispielsweise leitendes medizinisches Perso-

6 Richtlinie (EU) 2019/882 des Europäischen Parlaments und des Rates vom 17. April 2019 über die Barrierefreiheitsanforderungen für Produkte und Dienstleistungen, ABl. L 151 vom 07.6.2019; (EU) 2016/2102 des Europäischen Parlaments und des Rates vom 26. Oktober 2016 über den barrierefreien Zugang zu den Websites und mobilen Anwendungen öffentlicher Stellen, ABl. L 327 vom 02.12.2016.

nal, wie ein Chefarzt oder eine Chefärztin im Einzelfall als Betreiber i. S. d. KI-VO anzusehen sein. Dazu müsste z. B. eigenverantwortlich über die Auswahl, den Einsatz und den Zweck des KI-gestützten Systems entschieden oder die organisatorische und inhaltliche Kontrolle über die Anwendung ausgeübt werden. Entscheidend ist somit immer, ob tatsächlich Entscheidungsbefugnisse und eigene Verantwortung für den Einsatz des KI-Systems bestehen. Nur in diesem Fall sind Mitarbeitende – anders als im Regelfall – als Betreiber im Sinne der KI-VO zu qualifizieren.

Die Pflichten der Betreiber von Hochrisiko-KI-Systemen sind im Wesentlichen in Art. 26 ff. KI-VO geregelt. Betreiber müssen gemäß Art. 26 Abs. 1 i. V. m. Abs. 3 und 6 KI-VO sicherstellen, dass das KI-System nur gemäß Betriebsanleitung, Unionsrecht und nationalem Recht verwendet wird. Diese Pflicht darf die unternehmerische Freiheit allerdings nicht unangemessen beeinträchtigen. Protokolle, die vom KI-System automatisch erzeugt werden, müssen für mindestens sechs Monate aufbewahrt werden. Gemäß Art. 26 Abs. 2 i. V. m. Art. 14 KI-VO muss eine natürliche Person mit entsprechender Fachkompetenz das KI-System überwachen. Diese »menschliche Aufsicht« muss dabei dem Charakter des KI-Systems entsprechen (Martini & Wendehorst/*Eisenberger*, 2024, Art. 26 Rn. 27). Bei medizinischen KI-Systemen ist das in der Regel das medizinische Fachpersonal. Grund für diese Vorkehrung ist die »[…] Verhinderung oder Minimierung der Risiken für Gesundheit, Sicherheit oder Grundrechte […]« (Art. 14 Abs. 2 KI-VO).

Der Betreiber muss darüber hinaus sicherstellen, dass die verwendeten Eingabedaten dem vorgesehenen Zweck entsprechen und repräsentativ genug, d. h. nicht diskriminierend verzerrt sind (Art. 26 Abs. 4 KI-VO). Aus der Betriebsanleitung ergeben sich zudem Überwachungspflichten. Werden Risiken für Gesundheit, Sicherheit oder Grundrechte erkannt, müssen unverzüglich die anderen Beteiligten informiert und der Betrieb gegebenenfalls ausgesetzt werden, beispielsweise wenn eine Gefahr für Patientinnen und Patienten angenommen wird (Art. 26 Abs. 5 KI-VO, Art. 79 Abs. 1 KI-VO). Ist der Betreiber gleichzeitig arbeitgebende Person, muss er gemäß Art. 26 Abs. 7 KI-VO die Arbeitnehmervertretung und die Arbeitnehmenden über die Nutzung informieren und dabei alle einschlägigen arbeitsrechtlichen Vorschriften beachten. Zusätzlich zu den allgemeinen Transparenzpflichten i. S. d. Art. 50 KI-VO müssen vom

KI-System betroffene Personen – hier regelmäßig Patient*innen – informiert werden (Art. 26 Abs. 11 KI-VO). Sollte eine Datenschutz-Folgenabschätzung (DSFA) nach Art. 35 DSGVO erforderlich sein – was gemäß Art. 9 Abs. 1 DSGVO bei Gesundheitsdaten der Fall sein dürfte – muss das KI-System gemäß Art. 25 Abs. 8 KI-VO in einer EU-Datenbank registriert werden. Werden personenbezogene Daten, insbesondere sensible Gesundheitsdaten verarbeitet, ist zudem eine DSFA nach Art. 26 Abs. 9 KI-VO i. V. m. Art. 35 DSGVO oder Art. 27 der Richtlinie (EU) 2016/680 des Europäischen Parlaments und des Rates vom 27. April 2016 (Datenschutzrichtlinie für den Bereich Strafverfolgung) erforderlich. Dafür dürfen nur die in Art. 13 KI-VO aufgeführten Informationen verwendet werden, beispielsweise der Name und die Kontaktdaten des Anbieters (Abs. 3 lit. a) oder die Merkmale, Fähigkeiten und Leistungsgrenzen des Hochrisiko-KI-Systems (Abs. 3 lit. b). Betreiber müssen außerdem aktiv mit den zuständigen Behörden – z. B. den Datenschutzbehörden – zusammenarbeiten (Art. 26 Abs. 12 KI-VO). Zusätzlich zur DSFA ist eine Grundrechte-Folgenabschätzung gemäß Art. 27 KI-VO durchzuführen. Ziel ist es, Risiken für Grundrechte zu identifizieren und vorbeugende Maßnahmen zu planen (Martini & Wendehorst/*Eisenberger*, 2024, Art. 27 Rn. 4). Der Inhalt wird durch Art. 27 Abs. 1 S. 2 KI-VO bestimmt.

Betreiber von Hochrisiko-KI-Systemen unterliegen insgesamt umfassenden Verpflichtungen, die die Pflichten in Abschnitt 2.4.1.1 (»Pflichten der Anbieter«) zum Teil ergänzen. Die Pflichten erstrecken sich auf ein breites Spektrum, das von der Sicherheit über den Datenschutz bis hin zu den Grundrechten reicht. Die Einbeziehung des dem Geschehen meist näheren Betreibers ist aufgrund der hohen Risiken, die insbesondere im Gesundheitssektor bestehen, von großer Bedeutung. Es obliegt seiner Verantwortung, die korrekte Verwendung, Überwachung und Dokumentation des medizinischen KI-Systems sicherzustellen, während der Anbieter die notwendige Unterstützung leistet.

c) Einbeziehung in den Pflichtenkreis des Anbieters

Denkbar sind auch Fallkonstellationen, in denen die Trennung zwischen der Betreiber- und Anbietereigenschaft mit erheblichen

Abgrenzungsschwierigkeiten einhergeht, so dass der jeweilige Pflichtenumfang nicht eindeutig bestimmt werden kann. Dieses Abgrenzungsdefizit dürfte in der praktischen Anwendung der KI-VO erheblich werden, wobei sich – ähnlich wie im Datenschutz – der Gedanke einer gemeinsamen Verantwortung aufdrängt (Art. 26 DSGVO).

Die »Verantwortlichkeiten entlang der Wertschöpfungskette« werden in Art. 25 KI-VO behandelt. Die sogenannte Anbieterfiktion lässt sich in Art. 25 Abs. 1 KI-VO beobachten. Diese Regelung bestimmt, dass auch Händler, Einführer, Betreiber oder sonstige Dritte unter bestimmten Umständen als Anbieter eines Hochrisiko-KI-Systems gelten und damit den entsprechenden in Art. 16 KI-VO aufgelisteten Pflichten (Abschnitt 2.4.1.1 [»Pflichten der Anbieter«]) unterliegen. Dies ist insbesondere dann der Fall, wenn sie das System unter eigenem Namen vertreiben, es wesentlich verändern oder dessen Zweck so anpassen, dass es als Hochrisiko-KI eingestuft wird. Der Anbieter des ursprünglichen KI-Systems wird nun von dem Anbieter des neuen spezifischen KI-Systems abgelöst (Art. 25 Abs. 2 S.1 KI-VO), muss diesen nach Art. 25 Abs. 2 S. 2 KI-VO allerdings angemessen unterstützen, indem er relevante Informationen bereitstellt, technischen Zugang ermöglicht und damit bei der Erfüllung der Pflichten bei der Konformitätsbewertung von Hochrisiko-KI-Systemen hilft. Ausgenommen sind dabei nur Fälle, in denen der Erstanbieter die Umwandlung ausdrücklich untersagt hat (Art. 25 Abs. 2 S. 3 KI-VO).

Im Gesundheitssektor lässt sich ein praxisnaher Anwendungsfall der sogenannten Anbieterfiktion beobachten, wenn z. B. ein zunächst rein kosmetisch eingesetztes KI-System durch einen Gesundheitsdienstleister zweckverändert und für die medizinische Diagnose verwendet wird, so dass dieses nachträglich als Hochrisiko-KI-System i. S. d. Art. 6 KI-VO eingestuft wird.

2.4.2 Exkurs: Pflichten gemäß MP-VO

a) Pflichten der Hersteller

Zusätzlich zu den in Abschnitt 2.4.1.1 (»Pflichten der Anbieter«) genannten Pflichten könnten im Falle eines Medizinprodukts die allgemeinen Herstellerpflichten gemäß Art. 10 MP-VO Anwendung finden. Nach Art. 10 Abs. 1 MP-VO dürfen Produkte nur in den

Verkehr gebracht werden, wenn sie den Anforderungen der MP-VO entsprechend ausgelegt und hergestellt worden sind. Die konkreten Anforderungen ergeben sich vor allem aus Anhang I Kapitel II MP-VO (Rehmann & Wagner/*Wagner*, 2023, Art. 10 Rn. 3).

Die zentralen Pflichten der Hersteller nach Art. 10 MP-VO lassen sich wie folgt zusammenfassen:

- Implementierung eines Risikomanagementsystems (Abs. 2),
- Durchführung einer klinischen Bewertung (Abs. 3),
- Erstellung und Pflege der technischen Dokumentation (Abs. 4),
- Ausstellung einer EU-Konformitätsbewertung gemäß Art. 19 MP-VO und Anbringung der CE-Kennzeichnung gemäß Art. 20 MP-VO (Abs. 6) und
- Aufbewahrung von Dokumentation und Konformitätserklärung für mindestens zehn Jahre (Abs. 8 S. 1, 2).

Darüber hinaus bestehen weitere Pflichten: Hersteller müssen ein Überwachungssystem gemäß Art. 10 Abs. 10 i. V. m. Art. 83 MP-VO sowie ein Qualitätsmanagementsystem gemäß Art. 10 Abs. 9 MP-VO implementieren. Zudem muss die eindeutige Kennzeichnung der Produkte über das UDI-System (Unique Device Identification) nach Art. 10 Abs. 7 i. V. m. Art. 27 MP-VO sowie die Registrierung des Produkts und des Herstellers in den vorgeschriebenen Datenbanken gemäß Art. 10 Abs. 7 MP-VO i. V. m. Art. 29 und Art. 31 MP-VO beachtet werden. Die europäische Datenbank für Medizinprodukte (engl.: European database for medical devices, EUDAMED) ist jedoch noch nicht vollständig betriebsbereit. Sie kann – soweit fertiggestellt – bereits freiwillig verwendet werden, eine verpflichtende Nutzung ist jedoch bis zur vollständigen Fertigstellung nicht vorgesehen (Europäische Kommission, o. D.). Für den Fall, dass EUDAMED nicht funktionsfähig ist, regelt § 97 MPDG entsprechende Übergangsregelungen. Die Bekanntmachungen des Bundesministeriums für Gesundheit konkretisieren diese (§ 97 Abs. 2 MPDG).

Weiterhin müssen Hersteller sicherstellen, dass alle Produktinformationen verständlich, dauerhaft und in der richtigen Amtssprache vorliegen (Art. 10 Abs. 11 MP-VO). Bei festgestellter Nichtkonformität sind unverzüglich Korrekturmaßnahmen zu ergreifen und sowohl die Behörden als auch die Händler, Bevollmächtigten und Importeure zu informieren (Art. 10 Abs. 12 MP-VO). Zudem besteht die Pflicht, ein System zur Aufzeichnung und Meldung schwerwiegen-

der Vorkommnisse zu betreiben (Art. 10 Abs. 13 MP-VO), auf Anfrage vollständige Unterlagen und Proben bereitzustellen sowie bei Risikomaßnahmen mit den Behörden zu kooperieren (Art. 10 Abs. 14 MP-VO). Es ist ein Bevollmächtigter gemäß Art. 11 MP-VO sowie eine verantwortliche Person für die Einhaltung der Regulierungsvorschriften gemäß Art. 15 MP-VO zu benennen. Schließlich müssen Hersteller auch für Produktschäden haften und eine angemessene finanzielle Absicherung garantieren (Art. 10 Abs. 16 MP-VO). Weniger relevant im vorliegenden Kontext sind die Sonderregelungen zu Sonderanfertigungen (Art. 10 Abs. 5 MP-VO) und Fremdentwicklungen (Art. 10 Abs. 15 MP-VO).

Die MP-VO formuliert einen umfangreichen Pflichtenkatalog. Dabei fällt auf, dass sich einige Pflichten mit denen aus Abschnitt 2.4.1 (»Pflichten gemäß KI-VO«) überschneiden. So werden beispielsweise ein systematisches Risikomanagement, eine Konformitätsbewertung, eine CE-Kennzeichnung und eine aktive Überwachung gefordert. Beide Verordnungen fordern zudem eine umfangreiche technische Dokumentation. Die KI-VO legt durch Anhang IV allerdings den Fokus auf KI-spezifischere Aspekte wie Trainings- oder Testdaten. Auch bei den Transparenzpflichten gibt es Übereinstimmungen, jedoch fordert die KI-VO eine spezifischere und umfangreichere Bereitstellung durch den Betreiber. Während die KI-VO Anforderungen an die Qualität der Daten und die Datengovernance stellt, enthält die MP-VO dazu keine detaillierte Regelung. Diese Aspekte werden indirekt über das Risikomanagementsystem und die klinische Bewertung abgedeckt. Auch die menschliche Aufsicht ist in der KI-VO spezifisch geregelt, während die MP-VO diese nicht explizit vorschreibt. Beide Verordnungen fordern zudem Aufzeichnungen, doch während sich die MP-VO auf die Meldung schwerwiegender Vorkommnisse fokussiert, erweitert die KI-VO die Aufzeichnungspflicht, um zu gewährleisten, dass »das Funktionieren des Hochrisiko-KI-Systems in einem der Zweckbestimmung des Systems angemessenen Maße rückverfolgt werden kann« (Art. 12 Abs. 2 S. 1 KI-VO). Sowohl die KI-VO als auch die MP-VO fordern eine Registrierung in einer Datenbank. Die KI-VO verlangt jedoch eine Registrierung in einer EU-Datenbank für Hochrisiko-KI-Systeme, während die MP-VO eine Registrierung bei EUDAMED vorsieht. Demnach müssten sich Anbieter bzw. Hersteller von medizinischen

Hochrisiko-KI-Systemen trotz derselben Zielsetzung derzeit noch in beiden Datenbanken registrieren.

Aufgrund der zahlreichen Überschneidungen könnte der Eindruck entstehen, die MP-VO sei durch die spezifischere KI-VO verdrängt worden. Es ist allerdings hervorzuheben, dass es keine Vorrangregelung gibt und somit beide Verordnungen nebeneinander anwendbar sind. So verbleiben Bereiche, die nach wie vor nur durch die MP-VO geregelt werden, wie z. B. die klinische Bewertung oder die Registrierung bei EUDAMED. Eine sorgfältige rechtliche Abstimmung ist in diesem Zusammenhang unerlässlich.

b) Pflichten der Betreiber

In Abschnitt 2.4.1.2 (»Pflichten der Betreiber«) wurden die Pflichten i. S. d. KI-VO bereits ausführlich behandelt. Die Pflichten des in der MP-VO legal definierten Anwenders sind jedoch nicht so umfangreich formuliert wie die des Herstellers und der anderen Beteiligten. Aus der MP-VO ergeben sich allerdings einige wenige indirekte Pflichten für Anwender. So wird gemäß Art. 16 Abs. 1 MP-VO »eine sonstige natürliche oder juristische Person« in den Pflichtenkreis des Herstellers einbezogen, soweit sie ein Produkt verändert oder unter eigenem Namen in Verkehr bringt. Auch praktische Pflichten wie die Einhaltung der Gebrauchsanweisung oder die Meldung schwerwiegender Vorkommnisse an den Hersteller sind denkbar, werden aber nicht konkret erwähnt. Der Anwender zählt in erster Linie zum geschützten Personenkreis, wohingegen der Betreiber nach der KI-VO spezifische Pflichten zu erfüllen hat (Martini & Wendehorst/*Wendehorst*, 2024, Art. 26 Rn. 2).

In Deutschland gelten indes verbindliche Vorgaben gemäß der MPBetreibV, von denen die wichtigsten im Folgenden aufgeführt werden. Der Betreiber ist gemäß § 3 Abs. 1 MPBetreibV verpflichtet, die ihm zugewiesenen Pflichten zu erfüllen, um den sicheren und ordnungsgemäßen Einsatz der Produkte in der Gesundheitseinrichtung zu gewährleisten. Diese Pflichten werden allerdings nicht weiter konkretisiert. Vielmehr sollen alle Pflichten i. S. d. MPBetreibV Beachtung finden und deren Einhaltung dem Ziel Rechnung tragen, den Schutz und die Sicherheit von Patient*innen zu gewährleisten (Bundesrat, 2016). So sind Betreiber und Benutzer gemäß § 4 Abs. 1

MPBetreibV gleichermaßen dazu verpflichtet, die bestimmungsgemäße Anwendung des Produktes sicherzustellen. Das Produkt darf nur von Personen verwendet werden, die »die dafür erforderliche Ausbildung oder Kenntnis und Erfahrung besitzen« (§ 4 Abs. 2 MPBetreibV) und i. S. d. § 4 Abs. 3 S. 1 MPBetreibV ordnungsgemäß eingewiesen wurden. Für die Einhaltung der Anforderungen aus Abs. 1 und 2 ist der Betreiber verantwortlich (§ 4 Abs. 5 S. 1 MPBetreibV).

Der Benutzer muss sich dagegen vor Verwendung des Produkts vergewissern, dass es funktionsfähig und in einwandfreiem Zustand ist (§ 4 Abs. 6 S. 1 MPBetreibV). Die Gebrauchsanweisung mit allen mitgelieferten sicherheitsrelevanten Informationen und Wartungshinweisen ist gemäß § 4 Abs. 7 S. 1 MPBetreibV so aufzubewahren, dass sie dem Benutzer jederzeit zur Verfügung steht. Auch hier ist denkbar, dass dem Benutzer darüber hinaus praktische Pflichten obliegen, wie beispielsweise die Meldung von Fehlern oder schwerwiegenden Vorkommnissen an die nächsthöhere Person in der Verantwortungskette. Dies wäre insbesondere bei risikobehafteten Produkten im Gesundheitswesen sinnvoll, da Fehler so schneller erkannt und das Risiko für Patient*innen minimiert werden könnte. Der Betreiber hat zudem einen sachkundigen und zuverlässigen »Beauftragte[n]« für Medizinproduktesicherheit zu ernennen, soweit er eine Gesundheitseinrichtung mit mehr als 20 Beschäftigten führt (§ 6 Abs. 1 S. 1 MPBetreibV). Gemäß § 6 Abs. 4 MPBetreibV sollte dessen Funktions-E-Mail-Adresse auf der Internetseite bekannt gegeben werden, sodass er bei Bedarf erreicht werden kann. Nach § 7 Abs. 1 S. 1 MPBetreibV hat der Betreiber die Produkte instand zu halten. Davon umfasst sind insbesondere Wartungen und Inspektionen sowie Instandsetzungen zur Wiederherstellung der Funktionsfähigkeit (§ 7 Abs. 2 S. 1 MPBetreibV). Zuletzt muss der Betreiber bei Produkten aus Anlage 1 gemäß § 12 Abs.1 S. 1 i. V. m. S. 2, 3 MPBetreibV spätestens alle zwei Jahre sicherheitstechnische Kontrollen durchführen. Anlage 1 formuliert einen Katalog von Produkten, wobei KI in verschiedenen Bereichen zunehmend eine Rolle spielt.

Dieser Überblick zeigt in vielerlei Hinsicht Übereinstimmungen mit den Pflichten der Betreiber i. S. d. KI-VO. Letztere stellt jedoch umfangreichere Anforderungen an den Betrieb von KI-Systemen und definiert KI-spezifische Pflichten für Betreiber. Die MPBetreibV deckt die Basispflichten ab, die auch KI umfassen, während die KI-VO diese um spezifische Anforderungen an KI-Systeme ergänzt.

Beide Verordnungen sollten zwar parallel beachtet werden, die KI-VO wird aber in den meisten Aspekten umfangreicher sein.

2.5 Sanktionsmechanismen

Beim Einsatz von KI im Gesundheitswesen gibt es strenge Vorschriften, die in Abschnitt 2.4 (»Beteiligte und Verantwortliche«) näher beleuchtet wurden. So regelt die KI-VO den Einsatz von Hochrisiko-KI-Systemen, während die MP-VO die Anforderungen an Medizinprodukte festlegt. Die KI-VO sieht gemäß Erwägungsgrund Nr. 168 S. 1 KI-VO vor, »die Verhängung von Sanktionen und anderen Durchsetzungsmaßnahmen« zu ergreifen, um die Einhaltung der jeweiligen Anforderungen durchzusetzen. Durch sehr hohe Bußgelder – mit abschreckender Wirkung – soll der regelkonforme Einsatz von KI-Systemen sichergestellt werden (Erwägungsgrund Nr. 168 S. 2 KI-VO). Diese Sanktionen sind insbesondere im Gesundheitswesen von großer Bedeutung, da Patient*innen vor potenziell schwerwiegenden und lebensgefährlichen Fehlern geschützt werden müssen. Die Umsetzung wird dabei den Mitgliedstaaten überlassen. Das MPDG setzt im Hinblick auf Medizinprodukte die MP-VO in nationales Recht um und enthält dort konkrete Bußgeld- und Strafregelungen.

Die Sanktionsregelungen der KI-VO (Art. 99 bis 101 KI-VO) orientieren sich sodann systematisch an den Sanktionsvorschriften der DSGVO (Art. 83 ff. DSGVO). Art. 99 KI-VO regelt hierbei allgemeine Sanktionen bei Verstößen gegen die KI-VO durch die Beteiligten. Bei Verstößen gegen Art. 5 KI-VO (Abschnitt 2.3.1 [»Grundkonzept«]) belaufen sich die Bußgeldhöchstgrenzen auf 35 Millionen Euro oder 7 % des gesamten weltweiten Jahresumsatzes des vorangegangenen Geschäftsjahres, je nachdem welcher Betrag höher ausfällt (Art. 99 Abs. 3 KI-VO).[7] Verstöße gegen die Pflichten der verschiedenen Beteiligten, wie beispielsweise von Anbietern gemäß Art. 16

7 Die Bemessung der Bußgeldhöhe als Prozentsatz des weltweiten Jahresumsatzes des vorangegangenen Geschäftsjahres ist bereits aus anderen Regelwerken bekannt: Verordnung (EG) Nr. 1/2003 des Rates vom 16. Dezember 2002 zur Durchführung der in den Artikeln 81 und 82 EG-Vertrag niedergelegten Wettbewerbsregeln; (EU) 2016/679 vom 27. April 2016 (Datenschutz-Grundverordnung).

KI-VO oder Betreibern gemäß Art. 26 KI-VO, werden ebenfalls mit Bußgeldern geahndet. So formuliert Art. 99 Abs. 4 KI-VO:

> »Für Verstöße gegen folgende für Akteure oder notifizierte Stellen geltende Bestimmungen, mit Ausnahme der in Artikel 5 genannten, werden Geldbußen von bis zu 15 000 000 EUR oder – im Falle von Unternehmen – von bis zu 3 % des gesamten weltweiten Jahresumsatzes des vorangegangenen Geschäftsjahres verhängt, je nachdem, welcher Betrag höher ist:
> a) Pflichten der Anbieter gemäß Artikel 16;
> b) Pflichten der Bevollmächtigten gemäß Artikel 22;
> c) Pflichten der Einführer gemäß Artikel 23;
> d) Pflichten der Händler gemäß Artikel 24;
> e) Pflichten der Betreiber gemäß Artikel 26;
> f) für notifizierte Stellen geltende Anforderungen und Pflichten gemäß Artikel 31, Artikel 33 Abs. 1, 3 und 4 bzw. Artikel 34;
> g) Transparenzpflichten für Anbieter und Betreiber gemäß Artikel 50.«

Diese können eine Höhe von bis zu 15 Millionen Euro oder im Falle von Unternehmen 3 % des Jahresumsatzes des vergangenen Jahres erreichen. Die Bemessung erfolgt gemäß Art. 99 Abs. 4 KI-VO unter Berücksichtigung des jeweils höheren Betrags. Nach Art. 99 Abs. 5 KI-VO werden im Falle einer unrichtigen Auskunftserteilung gegenüber Behörden Geldbußen in Höhe von 7,5 Millionen Euro oder ein Prozent des Jahresumsatzes verhängt. Art. 100 KI-VO behandelt spezifisch die Verhängung von Geldbußen gegen Organe, Einrichtungen und sonstige Stellen der Union durch den Europäischen Datenschutzbeauftragten, während Art. 101 KI-VO die Geldbußen gegenüber Anbietern von KI-Modellen mit allgemeinem Verwendungszweck (General Purpose AI, GPAI) regelt. Für KI-Systeme im Gesundheitswesen wird regelmäßig Art. 99 KI-VO die primäre Sanktionsnorm darstellen.

Die Umsetzung dieser Maßnahmen ist den Mitgliedstaaten – soweit sie wirksam, verhältnismäßig und abschreckend sind – selbst überlassen (Art. 99 Abs. 1 KI-VO).[8] Aufgrund dieser Umsetzung soll dann eine dezentrale Durchsetzung der Regelungen erfolgen (Scheffzig & Kilian/*Schefzig*, 2025, Art. 99 Rn. 19). Deutschland hat bislang keine nationalen Sanktionsvorschriften erlassen, die ausschließlich

8 Siehe zur Festlegung der Merkmale *wirksam*, *verhältnismäßig* und *abschreckend* auch EuGH, Urteil vom 21.09.1989 – Rs. 68/88, NJW 1990, 2245, sowie EuGH, Urteil vom 10.07.1990 – Rs. 326/88, BeckRS 2004, 70816.

Verstöße gegen die KI-VO betreffen. Der erste Referentenentwurf eines Gesetzes zur Durchführung der KI-Verordnung – bislang noch nicht offiziell veröffentlicht – sieht vor, Teile der Sanktionsregelungen über das Gesetz über Ordnungswidrigkeiten (OWiG) abzubilden. Nach diesem Entwurf soll die Bundesnetzagentur die Befugnis erhalten, Bußgelder bei Verstößen gegen die KI-VO zu verhängen (BMWK & BMJ, 2024). Mit diesem Entwurf liegt der erste konkrete Schritt zur nationalen Umsetzung der KI-VO vor. Er bietet eine Grundlage für die weitere politische Diskussion. Das geplante KI-Marktüberwachungsgesetz (KIMÜG) wird dabei als stabiles Fundament gesehen, stößt jedoch auch auf Kritik (Schreiber & Bronner, 2025). Die Umsetzungsfrist läuft bis August 2025. Bisher ist jedoch noch kein entsprechendes Gesetz verabschiedet worden. Eine Verabschiedung wird in naher Zukunft erwartet.

Daneben können bei Verstößen gegen die Anforderungen an Medizinprodukte auch Bußgelder nach §§ 92 ff. MPDG verhängt werden, sofern dieselbe Handlung nicht bereits durch die KI-VO sanktioniert wurde.

Das Verbot der Doppelbestrafung[9] (Art. 50 Charta der Grundrechte der Europäischen Union) steht der Anwendung des MPDG nicht entgegen, sofern unterschiedliche Schutzgüter betroffen sind oder unterschiedliche Rechtsgüter vorliegen.

2.6 Kritik und Herausforderungen

Der Zweck der KI-VO wurde bereits in Abschnitt 2.2 (»Ziele der Verordnung«) thematisiert. Mit anderen Worten soll die Verordnung dafür sorgen, dass in der gesamten EU einheitliche Regeln für die Entwicklung, den Verkauf, den Betrieb und die Nutzung von KI-Systemen geschaffen werden, um verantwortungsvolle KI zu ermöglichen, ohne den technischen Fortschritt zu behindern. Nichtsdestotrotz gibt die KI-VO durchaus Anlass zu Kritik, die im Folgenden näher beleuchtet werden soll.

9 EuGH, Urt. v. 14.9.2023 – C-27/22, EuZW 2023, 1045 zur Möglichkeit der Anwendung des Doppelbestrafungsverbots auf Verwaltungssanktionen; EuGH, Urt. v. 22.3.2022 – C-151/20, NZKart 2022, 203, Rn. 36 f. zum Grundsatz *ne bis in idem* und dem Erfordernis des Vorliegens derselben Tat.

In Abschnitt 2.3 (»Risikobasierter Ansatz«) wurde bereits darauf hingewiesen, dass die KI-VO keine einheitlichen Regeln für alle KI-Systeme festlegt, sondern die Anforderungen nach dem jeweiligen Risiko für Gesundheit, Sicherheit oder Grundrechte der Menschen ausrichtet. Insbesondere die Einordnung als »Hochrisiko-KI-System« ist nicht gelungen (Hacker, 2023), da nahezu jedes KI-System in diese Kategorie fällt, während nur besonders extreme Ausprägungen unter die in Art. 5 KI-VO genannten verbotenen Praktiken fallen. Dadurch verliert der risikobasierte Ansatz in vielen Fällen an Wirkung und führt dazu, dass KI-Systeme in der Regel den strengeren Vorgaben für Hochrisiko-KI-Systeme unterliegen (Spranger & Wenzel, 2023).

Die KI-VO stellt zahlreiche Anforderungen an die Beteiligten, die teilweise sehr abstrakt und offen formuliert sind. Bereits die Einordnung dessen, was genau unter ein »KI-System« fällt oder als »Hochrisiko-KI-System« zu qualifizieren ist, bereitet Probleme und führt in der Folge zu Rechtsunsicherheit (Bomhard & Siglmüller, 2024). Die bisher noch allgemeinen Regeln der KI-VO müssen nach Art. 96 KI-VO durch EU-Leitlinien konkretisiert werden. Diese Leitlinien sind jedoch noch nicht veröffentlicht. Da die Umsetzungsfristen unabhängig von der Ausarbeitung der Leitlinien gelten, wird sich der Erfolg der Maßnahmen erst im Zeitverlauf beurteilen lassen (Art. 113 KI-VO). Nicht zu unterschätzen ist dabei auch der bürokratische Aufwand, den die KI-VO für die jeweiligen Mitgliedstaaten mit sich bringt. Dieser Problematik wird nur teilweise mit den verlängerten Fristen begegnet (Becker & Feuerstack, 2024).

Die parallele Anwendung der KI-VO mit sektorspezifischen Regelungen, wie im medizinischen Bereich mit der MP-VO, kann zu einer sogenannten horizontalen Doppelregulierung führen (Abschnitt 2.5 [»Sanktionsmechanismen«]). Die teilweise überlappenden Pflichten schaffen eine undurchsichtige und komplexe Rechtslage für alle Beteiligten, die sich intensiv mit den ihnen obliegenden Pflichten auseinandersetzen müssen, um potenzielle Sanktionen zu vermeiden. Eine entsprechende Überregulierung könnte zulasten von technologischen Innovationen und Fortschritten im medizinischen Sektor entstehen. Es besteht mithin Bedarf nach einer Harmonisierung der Regulierungen (BVMed, 2021).

Die KI-VO differenziert zwar zwischen zahlreichen Beteiligten (Abschnitt 2.4 [»Akteure und Verantwortliche«]), berücksichtigt je-

doch die Unternehmensgröße weitgehend nicht. Dies kann insbesondere für kleine und mittlere Unternehmen (KMU) zu erheblichen Herausforderungen führen. Ihnen könnten die personellen und finanziellen Ressourcen fehlen, um die umfangreichen Anforderungen der KI-VO vollständig zu prüfen und umzusetzen. Gleichzeitig drohen bei Verstößen hohe Geldbußen (Abschnitt 2.5 [»Sanktionsmechanismen«]), was für KMU ein besonders hohes wirtschaftliches Risiko bedeuten kann (Hacker & Berz, 2023).

Nach Art. 2 der Empfehlung 2003/361/EG der Kommission vom 6. Mai 2003 werden KMU wie folgt eingestuft:

> »Mitarbeiterzahlen und finanzielle Schwellenwerte zur Definition der Unternehmensklassen
> 1. Die Größenklasse der Kleinstunternehmen sowie der kleinen und mittleren Unternehmen (KMU) setzt sich aus Unternehmen zusammen, die weniger als 250 Personen beschäftigen und die entweder einen Jahresumsatz von höchstens 50 Mio. EUR erzielen oder deren Jahresbilanzsumme sich auf höchstens 43 Mio. EUR beläuft.
> 2. Innerhalb der Kategorie der KMU wird ein kleines Unternehmen als ein Unternehmen definiert, das weniger als 50 Personen beschäftigt und dessen Jahresumsatz bzw. Jahresbilanz 10 Mio. EUR nicht übersteigt.
> 3. Innerhalb der Kategorie der KMU wird ein Kleinstunternehmen als ein Unternehmen definiert, das weniger als 10 Personen beschäftigt und dessen Jahresumsatz bzw. Jahresbilanz 2 Mio. EUR nicht überschreitet.«

Als KMU gelten demnach Unternehmen mit weniger als 250 Beschäftigten und einem Jahresumsatz von höchstens 50 Millionen Euro. Diese Definition weicht von der in der Allgemeinheit oft vertretenen Vorstellung ab, wonach KMU überwiegend sehr kleine Betriebe sind. Folglich erscheint eine pauschale Privilegierung dieser Unternehmensgruppe gegenüber großen Unternehmen nicht uneingeschränkt gerechtfertigt. Eine Einzelfallbetrachtung wäre dabei differenzierter. Die KI-VO erkennt in Erwägungsgrund Nr. 143 S. 1 ausdrücklich die Bedeutung des Schutzes von KMU zur Förderung von Innovationen an. Dieses Prinzip wird unter anderem in Art. 99 Abs. 6 KI-VO durch die Wahl des jeweils niedrigeren Sanktionsbetrags umgesetzt. Zudem sieht Art. 11 Abs. 1 S. 4 KI-VO Erleichterungen für kleine und Kleinstunternehmen bei der technischen Dokumentation vor. Gleichwohl bleibt fraglich, ob diese punktuellen Entlastungen ausreichen, um die strukturellen Nachteile für KMU wirksam zu kompensieren.

Diese und weitere Aspekte werden von der KI-VO bislang unberücksichtigt gelassen. Sie lässt damit durchaus Raum für Kritik. Erst die zukünftige Entwicklung wird Aufschluss darüber geben, ob die Realisierung der Ziele ohne negative Auswirkungen auf die medizinische Forschung und Innovationsentwicklung im Gesundheitswesen möglich ist.

3. Haftungsrichtlinie und Produkthaftungsrecht

Die zunehmenden Einsatzmöglichkeiten von KI werfen darüber hinaus grundlegende haftungsrechtliche Fragen auf. Gerade in sicherheitsrelevanten Bereichen, wie etwa dem Gesundheitswesen, bedarf es einer umfassenden Betrachtung. Insbesondere stellt sich die Herausforderung, wie Schadensfälle zu behandeln sind, bei denen die Ursache in einem teilweise autonomen Verhalten eines KI-Systems liegt, das auf selbstlernenden oder datenbasierten Prozessen beruht. Ein wichtiger Baustein für die Schaffung eines kohärenten rechtlichen Rahmens für KI in der EU ist daher das Haftungsrecht. Das bisherige europäische Haftungsrecht, dabei insbesondere das Zivil- und Deliktsrecht, ist stark national geprägt. Die bestehenden zivilrechtlichen Regelwerke – insbesondere die verschuldensunabhängige Produkthaftung nach dem Produkthaftungsgesetz (ProdHaftG) – waren bislang auf klassische, deterministisch funktionierende Produkte zugeschnitten. Die derzeitig bestehenden nationalen Haftungsvorschriften umfassen demnach nicht ausdrücklich Schadensersatzansprüche infolge von durch KI-gestützte Produkte oder Dienstleistungen verursachte Schäden. Die sich daraus ergebenden Lücken und Unsicherheiten haben zu einer politischen Diskussion über eine Anpassung des Haftungsrahmens geführt, insbesondere auf europäischer Ebene im Zuge der Überarbeitung der Produkthaftungsrichtlinie (siehe dazu unter Abschnitt 3.2. [»Status quo«]) und einer geplanten Richtlinie über die Haftung für KI-Systeme (KI-HaftungsRL-E) (siehe dazu unter Abschnitt 3.1 [»Ursprünglicher Regulierungsansatz«]).

3.1 Ursprünglicher Regulierungsansatz

Um Rechtsunsicherheiten zu vermeiden und unter der Prämisse eines stets zu wahrenden hohen Schutzstandards plante die Kommission, Haftungsvorschriften für Schäden, die durch KI-Systeme verursacht werden, unionsrechtlich zu vereinheitlichen. Dazu veröffentlichte sie im Jahr 2022 den Vorschlag für eine Richtlinie zur Anpassung der Vorschriften über außervertragliche zivilrechtliche Haftung an künstliche Intelligenz (Richtlinie über KI-Haftung). Ziel war es, einen harmonisierten Rechtsrahmen in der EU zu schaffen, um der Gefahr der Fragmentierung des Binnenmarktes zu entgegnen. Gleichzeitig sollte sichergestellt werden, dass eine durch KI-Technologie geschädigte Person in gleicher Weise entschädigt wird, wie in vergleichbaren Fällen ohne KI-Beteiligung. Mit Zusammenkunft der neuen Kommission wurde der Vorschlag für die KI-Haftungsrichtlinie in ihrem Arbeitsprogramm für 2025 überraschend zurückgezogen. Stattdessen soll auf bereits bestehende Haftungsvorschriften zurückgegriffen werden.

De lege lata gib es verschiedene rechtliche Vorgaben, die einen etwaigen Schadensersatzanspruch beinhalten: Zunächst kommt eine verschuldensunabhängige Haftung des Herstellers eines KI-Systems in Betracht, wenn sein Produkt fehlerhaft ist und dadurch ein Schaden entsteht. Hierbei wird allerdings der Fall außer Acht gelassen, dass das KI-System selbst nicht fehlerhaft ist, sondern erst nach dem Inverkehrbringen durch maschinelles Lernen von geplantem Output »abweicht«, sodass in diesem Fall eine Regelungslücke besteht.

Im Deliktsrecht kommt eine Haftung des medizinischen Personals oder des Klinikträgers – etwa wegen Verletzung der ärztlichen Sorgfaltspflicht oder eines Organisationsverschuldens – in Betracht, wenn durch ein KI-System ein Schaden entsteht (z. B. eine fehlerhafte medizinische Diagnose). Auch im Rahmen eines vertraglichen Anspruchs (z. B. Behandlungsvertrag) – etwa auf Schadensersatz wegen Schlechterfüllung – muss das Verschulden nachgewiesen werden. Problematisch dabei ist, dass die bestehenden Regelungen ein Verschulden oder einen klaren Produktfehler voraussetzen. Beides ist bei KI-Systemen mit lernenden, teilweise intransparenten Entscheidungen häufig schwierig zu beweisen, wodurch echte Haftungslücken entstehen.

Ziel der geplanten Haftungsrichtlinie war es, solche Haftungslücken zu schließen, indem in Ergänzung zum Produkthaftungsrecht ein Schaden auch in den Konstellationen ersetzt werden sollte, in denen gerade kein Produktfehler vorliegt. Zur Umsetzung sah der Vorschlag einen vereinfachten Zugang zu Informationen vor und enthielt Vorschriften über die Offenlegung von Beweismitteln, um die Beweislast der geschädigten Person zu erleichtern. Neben dem Zugang zu Beweismitteln stellte die vorgesehene Kausalitätsvermutung eine wesentliche Maßnahme dar. Demnach sollte der für Schadensersatzansprüche üblicherweise erforderliche ursächliche Kausalzusammenhang zwischen dem Verschulden des Beklagten und dem Output des KI-Systems (widerlegbar) vermutet werden, wenn die in Art. 4 Abs. 1 KI-HaftungsRL-E benannten Voraussetzungen erfüllt sind: Zum einen muss ein Verschulden auf Seiten des Beklagten in der Weise vorliegen, dass gegen eine im Unionsrecht oder im nationalen Recht festgelegte Sorgfaltspflicht, deren unmittelbarer Zweck darin besteht, den eingetretenen Schaden zu verhindern, verstoßen wurde (Art. 4 lit. a KI-HaftungsRL-E). Darüber hinaus muss der Sorgfaltspflichtverstoß den vom KI-System hervorgebrachten Output auf Grundlage der Umstände des Falls nach vernünftigem Ermessen beeinflusst haben (Art. 4 lit. b KI-HaftungsRL-E). Schließlich muss das vom KI-System hervorgebrachte Ergebnis oder aber die Tatsache, dass das KI-System kein Ergebnis hervorgebracht hat, zu dem Schaden geführt haben.

Handelt es sich um einen Schadensersatzanspruch gegenüber einem »Nutzer« eines Hochrisiko-KI-Systems, beispielsweise einem Arzt oder einer Ärztin, konkretisiert Art. 4 Abs. 3 KI-HaftungsRL-E für den Nachweis einer Sorgfaltsverletzung, dass eine solche vorliegt, wenn das KI-System entgegen der beigefügten Gebrauchsanweisung verwendet oder überwacht wurde und die Eingabedaten der Zweckbestimmung des Systems widersprechen. Mit dem Vorschlag sollten die mit KI-Systemen einhergehenden neuen Fragen geklärt und etwaige Rechtsunsicherheiten behoben werden. Durch die Streichung der Richtlinie vom Arbeitsplan muss nun erneut auf die vorliegenden nationalen Haftungsvorschriften zurückgegriffen werden.

3.2 Status quo

Um dennoch den weiterhin bestehenden Gefahren der Fragmentierung des Binnenmarktes entgegenzuwirken und bestehende Haftungslücken zu schließen, wird in diesem Zuge nun auf Änderungen in der Produkthaftungsrichtlinie verwiesen. Bereits vor dem Rückzug der KI-Haftungsrichtlinie hat die EU-Kommission Ende 2024 die Produkthaftungsrichtlinie (ProdHaft-RL) novelliert und dabei die Haftung für Software, einschließlich KI, erweitert (Richtlinie 2024/2853). Sie wurde am 18.11.2024 im Amtsblatt der Europäischen Union veröffentlicht und muss von den Mitgliedstaaten bis zum Jahr 2026 in nationales Recht umgesetzt werden. Durch die Änderungen wurde der Produktbegriff auf Software und KI-Systeme ausgeweitet, sodass KI-Modelle nun ebenfalls haftungsrechtlich erfasst sind. Hersteller von KI-Anwendungen haften künftig auch für Ausfälle oder Schäden, die durch unzureichende Updates oder Sicherheitslücken entstehen (Art. 11 Abs. 2 ProdHaft-RL). Darüber hinaus gibt es Vorgaben bezüglich der Offenlegung von Beweismitteln und zugleich wurde die Anspruchsdurchsetzung zugunsten der Opfer durch Beweiserleichterungen und Vermutungsregeln für Geschädigte erleichtert.

4. European Health Data Space

Das politische Konzept der europäischen Datenräume überspannt im Bereich des Umgangs mit Gesundheitsdaten spezifische Vorschriften wie etwa die Datenschutz-Grundverordnung oder die KI-Verordnung. Die Kommission plant insoweit ein regelbasiertes Daten-Ökosystem, das unter anderem die Sekundärdatennutzung von Gesundheitsdaten maßgeblich erleichtern soll. Diese Einbettung birgt verschiedene rechtliche Folgen für die gesundheitsbezogene KI-Nutzung, die sich nur in der Gesamtschau zufriedenstellend einordnen und klären lassen.

Sowohl in Bezug auf Therapien als auch das Verständnis von Krankheiten und der Entwicklung von Medikamenten ist die Forschung mit medizinischen Daten essenziell für Innovation und Weiterentwicklung in der Medizin. Das Problem ist dabei häufig der Schutz solcher Daten. Dass die Signifikanz der Forschung grund-

sätzlich mitbedacht wird, sieht man bereits an der expliziten Privilegierung der Forschung sowohl in der KI-VO als auch der DSGVO. Durch die neuen Möglichkeiten, die aus der Anwendung von KI resultieren, eröffnen sich weitreichende neue Chancen. Der European Health Data Space (EHDS) ist dabei ein zentrales Vorhaben der EU zur digitalen Transformation des Gesundheitswesens. Es soll einen sicheren, einheitlichen Rahmen schaffen, um Gesundheitsdaten sowohl für die primäre Versorgung als auch für sekundäre Zwecke – insbesondere Forschung, Innovation und Politikgestaltung – grenzüberschreitend verfügbar zu machen. Konkret wurde im März 2025 die zugehörige Verordnung (EU) 2025/327 über den europäischen Gesundheitsdatenraum (Abl. L 2025/327 v. 05.03.2025) veröffentlicht (im Folgenden EHDS-VO).

Im Verhältnis zur DSGVO fungiert die EHDS-VO nicht als konkurrierendes Regelwerk. Vielmehr stellt sie eine Konkretisierung der datenschutzrechtlichen Anforderungen für den Gesundheitsbereich als sektorenspezifische Ausgestaltung dar. Gleichzeitig ergänzt die EHDS-VO die KI-VO, indem sie den Zugang zu qualitativ hochwertigen Gesundheitsdaten regelt, was von großer Bedeutung für das Training, die Prüfung und Validierung von KI-Systemen in der medizinischen Forschung und Versorgung ist. Im Grunde sollen alle personenbezogenen Gesundheitsdaten für die Sekundärnutzung verfügbar gemacht werden. Art. 51 EHDS-VO legt zu diesem Zweck einen Mindestkatalog elektronischer Gesundheitsdaten fest, die von den »Gesundheitsdateninhabern« für die Sekundärnutzung zur Verfügung gestellt werden müssen, soweit nicht ausdrücklich widersprochen wird (sog. »Opt-out-Modell«). Zu den benannten Gesundheitsdaten gehören beispielsweise elektronische Gesundheitsdaten aus Systemen für elektronische Gesundheitsaufzeichnungen (electronic health records, EHR) (lit. a), Daten zu Faktoren, die sich auf die Gesundheit auswirken, einschließlich sozioökonomischer, umweltbedingter und verhaltensbezogener Gesundheitsfaktoren (lit. b), menschliche genetische, epigenomische und genomische Daten (lit. f), Daten aus Wellness-Anwendungen (lit. i) und Gesundheitsdaten aus Biobanken und zugehörigen Datenbanken (lit. q). Darüber hinaus kann der Katalog nach Art. 51 Abs. 2 EHDS-VO durch die Mitgliedstaaten erweitert werden. Um Zugang zu den nach Art. 51 EHDS-VO erhobenen Sekundärdaten zu erlangen, muss die Verarbeitung der Daten für einen der in Art. 53 EHDS-VO gelisteten

Zwecke erforderlich sein. Dazu gehören etwa öffentliche Interessen im Bereich der öffentlichen Gesundheit oder der Gesundheit am Arbeitsplatz, Politikgestaltung und Regulierungstätigkeiten zur Unterstützung von öffentlichen Stellen oder Organen, Statistiken im Sinne von Art. 3 Nr. 1 der Verordnung (EG) Nr. 223/2009, Bildungs- oder Lehrtätigkeiten im Gesundheitswesen oder im Pflegesektor auf der Ebene der Berufs- oder Hochschulbildung, wissenschaftliche Forschung im Bereich des Gesundheitswesens oder des Pflegesektors und die Verbesserung der Pflege und Gesundheitsversorgung. Im Ergebnis ist nahezu jeder mit dem Gesundheits- oder Pflegesektor zusammenhängender Zweck ausreichend, um von der Zugangsstelle für Gesundheitsdaten (Art. 55 EHDS-VO) die entsprechenden Daten zu erhalten.

Obwohl der EHDS darauf abzielt, einen kohärenten und grenzüberschreitenden Rahmen für die Nutzung von Gesundheitsdaten in der EU zu schaffen, steht der Verordnung in der Fachliteratur und Praxis teils erheblicher Kritik gegenüber. Die Verpflichtung von Gesundheitseinrichtungen, sensible medizinische Informationen an neue staatliche Agenturen in jedem EU-Mitgliedstaat weiterzugeben, kann die Vertraulichkeit der »Arzt-Patienten-Beziehung« negativ beeinträchtigen. Dies könnte zur Folge haben, dass Patient*innen relevante Informationen gegenüber dem Behandlungspersonal zurückhalten (Konopik, 2025). Zudem wird beanstandet, dass ethisch-gesellschaftliche Aspekte, darunter insbesondere Transparenz gegenüber Betroffenen, informierte Einwilligung sowie Schutz vor Diskriminierung, bislang nicht hinreichend ausgearbeitet wurden.[10] Bereits in einer Pressemitteilung vom 05. April 2023 äußerte die Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder (DSK) Bedenken hinsichtlich des ausreichenden Schutzes der Gesundheitsdaten und einer nicht angemessenen Berücksichtigung des Grundrechts auf Datenschutz bzw. des Rechts auf informationelle Selbstbestimmung. Besonders kontrovers diskutiert wird in diesem Zusammenhang das »Opt-out-Modell«, wonach Gesundheitsdaten grundsätzlich für Sekundärzwecke genutzt werden dürfen, solange die betroffene Person dem nicht ausdrücklich widerspricht. Dieses Vorgehen wird kritisch gesehen, da es die informatio-

10 Siehe zur informierten Einwilligung auch Lanzerath, 2025, Zeitschrift für medizinische Ethik, 71(1), 53–69.

nelle Selbstbestimmung der Betroffenen einschränken könnte und insbesondere vulnerable Gruppen, etwa Personen mit Sprachbarrieren oder eingeschränkter digitaler Kompetenz, trotz des in Art. 71 Abs. 2 EHDS-VO vorgesehenen »barrierefreien und leicht verständlichen Mechanismus zum Widerspruch«, faktisch von einer wirksamen Widerspruchsmöglichkeit ausgeschlossen sein könnten.

Ein weiterer Kritikpunkt betrifft die technische Realisierbarkeit der EHDS-Vorgaben. Teilweise wird in dem Zusammenhang eine gemeinsame europäische Auslegung und koordinierte Umsetzung gefordert, um Inkonsistenzen zwischen den nationalen Systemen der Mitgliedstaaten und daraus resultierende Sicherheitslücken zu vermeiden (Konopik, 2025). Die DSK fordert zudem die Streichung der vorgesehenen Regelung zur Bereitstellung von persönlichen Genomdaten, da dies in den intimsten Bereich der betroffenen Personen und ihrer Angehörigen eingreife (DSK, Pressemitteilung v. 05. April 2023). Schließlich wird darauf hingewiesen, dass die Verordnung strukturell eine Bevorteilung großer Technologieunternehmen begünstigen könnte. Diese verfügen über die notwendigen Ressourcen, um die komplexen Anforderungen an Datennutzung, Interoperabilität und Sicherheit zu erfüllen, während kleinere Akteure im Gesundheitswesen Gefahr laufen, abgehängt zu werden. Damit droht eine Machtkonzentration im digitalen Gesundheitssektor, die sowohl wettbewerbsrechtliche als auch innovationspolitische Implikationen hat.

Zusammenfassend lässt sich festhalten, dass der EHDS zwar das Potenzial hat, die Nutzung von Gesundheitsdaten in der EU erheblich zu erleichtern und innovative Forschungsansätze zu fördern, zugleich aber gewichtige Bedenken offenbart. Neben datenschutz- und ethisch-rechtlichen Herausforderungen, insbesondere im Hinblick auf Transparenz, informierte Einwilligung und das »Opt-out-Modell«, bestehen Zweifel an der technischen Umsetzbarkeit und der Wahrung der Vertraulichkeit der »Arzt-Patienten-Beziehung«. Maßgeblich ist, ob ein angemessener Schutz individueller Grundrechte sowie faire Wettbewerbsbedingungen gewährleistet werden können.

5. Schlussbetrachtung und Ausblick

Künstliche Intelligenz entwickelt sich rasant weiter und wird auch künftig – insbesondere im Gesundheitswesen – Fortschritte mit sich

bringen. Die Ausschöpfung dieses Potenzials hängt allerdings maßgeblich von der Klärung rechtlicher Rahmenbedingungen und der Sicherstellung von Datensicherheit ab. Die Privilegierung der Forschung im Rahmen der KI-VO macht die Relevanz von KI sowie die Bedeutung der Schaffung eines innovationsfreundlichen Standorts für digitale Technologien deutlich.

Mit der im Juni 2024 verabschiedeten KI-Verordnung wurde ein erster Versuch unternommen, einheitliche Vorschriften für KI zu schaffen. Sie regelt unter anderem die Risikoklassifizierung von Systemen, die Pflichten der Beteiligten sowie technische Anforderungen. Bei Verstößen drohen hohe Geldbußen. Gleichzeitig bestehen jedoch Überschneidungen mit der MP-VO, was mangels Vorrangregelung zu Rechtsunsicherheiten führt. Diese sogenannte horizontale Doppelregulierung bildet gemeinsam mit zahlreichen anderen Aspekten – wie der Offenheit der Begriffe, der mangelnden Konkretisierung und der angezweifelten Einordnung der Hochrisiko-KI-Systeme – den Grundstein für eine kritische Betrachtung.

Der von der Kommission entwickelte Entwurf einer KI-Haftungsrichtlinie sollte Schadensansprüche infolge von durch KI-gestützte Produkte oder Dienstleistungen verursachten Schäden regeln. Er wurde allerdings im aktuellen Arbeitsprogramm überraschend zurückgezogen und es wurde stattdessen auf bestehende Haftungsvorschriften verwiesen. Zudem sorgt der European Health Data Space mit der Zugangserleichterung zu personenbezogenen Gesundheitsdaten für die Sekundärnutzung im Gesundheitswesen für Kritik hinsichtlich des Datenschutzes.

Die Rechtslage bezüglich der Regulierung von KI-Systemen im Gesundheitswesen ist nach wie vor unsicher. Es gibt Unklarheiten und rechtliche Grauzonen im Hinblick auf die KI-VO, Haftungsfragen sowie bei der Unterscheidung zwischen echten und unechten Haftungsprivilegien. Wie sich dies auswirkt, wird sich in der Zukunft ergeben. Erwartete Änderungen in der KI-VO werden zeigen, ob die aktuellen Probleme gelöst werden können. Unternehmen und Institutionen sollten die bestehenden Vorgaben daher sorgfältig beachten und sich frühzeitig auf absehbare Änderungen einstellen.

Literaturverzeichnis

Becker, D., & Feuerstack, D. (2024). Die EU-KI-Verordnung: Überblick und Bewertung der finalen Fassung der KI-VO. *Künstliche Intelligenz und Recht,* (2), 62–69.

Bundesministerium für Wirtschaft und Klimaschutz & Bundesministerium der Justiz (BMJ). (2024, 4. Dezember). *Entwurf eines Gesetzes zur Durchführung der Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024… (Verordnung über künstliche Intelligenz)* [Nicht offiziell veröffentlichter Referentenentwurf]. https://assets.ctfassets.net/jq1xw71av3v5/6AKrXoeiDtYvFgYjrEwLYI/774c67d2d73ccc94aee94f894a936a87/241204_BMWK-BMJ_RefE_AI-Act-Umsetzung.pdf

Bundesrat. (2016). *Zweite Verordnung zur Änderung medizinprodukterechtlicher Vorschriften* [Drucksache 397/16 v. 04.08.2016]. https://www.bundesrat.de/SharedDocs/drucksachen/2016/0301-0400/397-16.pdf

Bomhard, D., & Siglmüller, J. (2024). AI Act – das Trilogergebnis. *Recht Digital (RDi),* 45–52.

BVMed. (2021). BVMed-Positionen zum Entwurf des »Artificial Intelligence Act« (AIA). *Medizin Produkte Recht (MPR),* (5), 176–182.

Ebers, M., & Streitbörger, C. (2024). Die Regulierung von Hochrisiko-KI-Systemen in der KI-Verordnung. *Recht Digital (RDi), 2024,* 393–400.

Eisemann, N., Bunk, S., Mukama, T., Baltus, H., Elsner, S. A., Gomille, T., Hecht, G., Heywang-Köbrunner, S., Rathmann, R., Siegmann-Luz, K., Töllner, T., Vomweg, T. W., Leibig, C., & Katalinic, A. (2025). Nationwide real-world implementation of AI for cancer detection in population-based mammography screening. *Nature Medicine, 31*, 917–924. https://doi.org/10.1038/s41591-024-03408-6

Europäische Kommission. (o. D.). *Medical devices – EUDAMED: Überblick.* https://health.ec.europa.eu/medical-devices-eudamed/overview_de

Gerdemann, S. (2024). Konformitätsbewertung als Kernpflicht der KI-Verordnung. *Neue Juristische Wochenschrift (NJW),* 2209–2215.

Hacker, P. (2023). Die Regulierung von ChatGPT et al. – ein europäisches Trauerspiel. *Gewerblicher Rechtsschutz und Urheberrecht (GRUR),* (5), 289–290.

Hacker, P., & Berz, A. (2023). Der AI Act der Europäischen Union – Überblick, Kritik und Ausblick. *Zeitschrift für Rechtspolitik,* (8), 226–229.

Hilgendorf, E., & Roth-Isigkeit, D. (2023). *Die neue Verordnung der EU zur Künstlichen Intelligenz. Kommentar* (1. Aufl.). C. H. Beck.

Konopik, J. (2025). Der Europäische Raum für Gesundheitsdaten (EHDS) und seine Bedeutung für das deutsche Gesundheitswesen. *Prävention und Gesundheitsförderung.* https://doi.org/10.1007/s11553-025-01200-z

Lachenmann, M. (2024). EU-Rat stimmt KI-Verordnung zu – neue Pflichten für Unternehmen. *MMR-Aktuell*, (4), 01359.

Lanzerath, D. (2025). Access and Benefit-Sharing: Gesundheitsdaten in der medizinischen Forschung nutzen. *Zeitschrift für medizinische Ethik, 71*(1), 53–69. https://doi.org/10.30965/29498570-20250108.

Martini, M., & Wendehorst, C. (Hrsg.). (2024). *KI-VO: Verordnung über künstliche Intelligenz. Kommentar* (1. Aufl.) C. H. Beck.

Rehmann, W., & Wagner, S. (Hrsg.). (2023). *MP-VO. Verordnung (EU) 2017/745 über Medizinprodukte. Kommentar* (4. Aufl.). C. H. Beck.

Schefzig, J., & Kilian, R. (Hrsg.). (2025). *BeckOK KI-Recht (2. Edition)*. C. H. Beck.

Schreiber, M., & Bronner, P. (2025). Der Referentenentwurf für ein Gesetz zur Durchführung der KI-Verordnung. *jurisPR-ITR, 7,* Anm. 2.

Schuh, M., & Witt, H. (2025). KI-Systeme und ihre Betreiber nach der KI-VO: Pflichten und Abgrenzung zum Endnutzer. *Zeitschrift für Europäisches Daten- und Informationsrecht (EuDIR)*, 142–148.

Spitz, M., Cornelius, K., Jungkunz, M., & Schickhardt, C. (2021). Rechtlicher Rahmen für eine privilegierte Nutzung klinischer Daten zu Forschungszwecken. *Medizinrecht (MedR), 39*, 499–504. https://doi.org/10.1007/s00350-021-5898-7

Spitz, M. (2025). Die KI-Verordnung und die Privilegierung medizinischer Forschung. *Medizinrecht MedR), 43,* 601–604. https://doi.org/10.1007/s00350-025-7102-y

Spranger, T. M., & Wenzel, M. (2023). Künstliche Intelligenz in der Medizin: Anmerkungen zum aktuellen Verordnungsentwurf der Europäischen Kommission. In S. Bohnet-Joschko & K. Pilgrim (Hrsg.), *Handbuch Digitale Gesundheitswirtschaft* (S. 263–266). Springer Gabler. https://doi.org/10.1007/978-3-658-41781-9_55

Roman Wagner

IV. Künstliche Intelligenz in der radiologischen Diagnostik: Ethische Aspekte

1. Problemstellung und technische Hintergründe

Künstliche Intelligenz (KI) ist die aktuell wohl meistdiskutierte Technologie. Im Laufe der vergangenen Jahre wurden im Zuge der Weiterentwicklung von KI-Systemen und -Techniken zunehmend neue Bereiche für den Einsatz von Künstlicher Intelligenz erschlossen (Heinrichs et al., 2022). Eines der ethisch relevantesten und in gleichem Maße vielversprechendsten wie umstrittensten Anwendungsfelder stellt die Medizin dar – insbesondere in der Radiologie werden neuartige KI-Verfahren bereits heute eingesetzt (Adlung et al., 2021). Vielversprechend ist der Einsatz von KI in der Radiologie sowie der Medizin deshalb, weil KI-Systeme unter anderem zeit- und kostenintensive Verwaltungsaufgaben automatisieren, Ärzt*innen bei der Diagnosestellung assistieren oder Probleme, die sich aus dem Fachkräftemangel in der Medizin ergeben, abmildern könnten (Topol, 2019). Zugleich verbinden sich mit dem Einsatz von KI in der Medizin aber auch schwerwiegende ethische Bedenken, die unter anderem mit den Problemkontexten einer dehumanisierten und depersonalisierten Medizin, einer reduktiven Datafizierung der Patient*innen, einer vollständigen Automatisierung klinischer Prozesse, einer vollumfänglichen Überwachung von Patient*innen sowie der Unterminierung von Autonomie in Verbindung stehen (Rubeis, 2024).

Im Folgenden wird es nicht darum gehen, alle ethischen Gesichtspunkte zu besprechen, die der Einsatz von KI in der Medizin aufwirft. Die Untersuchung wird sich vielmehr auf die Darstellung einiger ethischer Kernaspekte des Einsatzes von KI in der Radiologie beschränken, wobei diese Aspekte natürlich auch in weiteren KI-Debatten relevant sein können. Um diese ethisch relevanten Aspekte

von KI-Systemen im Kontext der Medizin beleuchten zu können, ist es dabei zunächst notwendig, bestimmte technische Aspekte der fraglichen Systeme zu erörtern. Zu diesem Zweck wird zunächst dargelegt, welche Formen der KI für die medizinische Praxis von Bedeutung sind, um im Anschluss aus ethischer Sicht vier wesentliche Merkmale dieser Arten von KI in den Blick zu nehmen.

1.1 Zwei Formen von KI

Der Einsatz von KI in der Medizin ist kein gänzlich neues Phänomen. So werden beispielsweise im Rahmen der medizinischen Praxis auf symbolischer KI fußende sogenannte »Decision Support Systems« bzw. »Expert Systems« bereits seit Jahrzehnten verwendet, um die medizinische Entscheidungsfindung zu unterstützen (Bottrighi et al., 2025; Holman & Cookson, 1987; Huang et al., 1993).[1] Gleichwohl hat die ethische Debatte um diesen Einsatz in den vergangenen Jahren dadurch neue Konturen und eine neue Dringlichkeit gewonnen, dass die Formen der KI, die in der Medizin Anwendung finden, technisch weiterentwickelt wurden. Diese Weiterentwicklung fußt auf der Unterscheidung von zwei Paradigmen des KI-Designs: dem *symbolischen Paradigma* sowie dem *konnektionistischen Paradigma* (Goel, 2021). KI-Forschung kann in diesem Zusammenhang als Versuch verstanden werden, Kognition künstlich zu reproduzieren (Flasiński, 2011; Freed, 2020). Während der Begriff der Kognition ebenso umstritten ist wie die Frage, welche Vermögen im Einzelnen als kognitiv verstanden werden sollten – häufig besprochene Vermögen sind etwa Wahrnehmung, Handlung, Sprache, Bewusstsein oder Emotionen[2] (Frankish & Ramsey, 2012) –, ist für die Zwecke dieses Beitrags wesentlich, dass die KI-Forschung kognitive Vermögen

1 Dabei ist wichtig zu beachten, dass zeitgenössische Spielarten dieser Systeme entweder auf neueren Formen des maschinellen Lernens fußen oder verschiedene KI-Technologien verbinden (Kokol et al., 2002).

2 Eine philosophisch gehaltvolle Möglichkeit, den Kognitionsbegriff zu erhellen, besteht darin, unter Kognition das *Erkenntnisvermögen* zu verstehen, also das Vermögen, auf Basis von rechtfertigenden epistemischen Gründen Wahres zu erkennen (Woleński, 2004, S. 3ff.). Über ein solches Vermögen verfügen einer philosophischen Tradition zufolge allein geistige, vernünftige und selbstbewusste Wesen (Kern, 2006; Boyle, 2017). Eine solche Erklärung würde zwar den Kognitionsbegriff erhellen, aber zugleich in die Felder der Philosophie des Geistes,

mithilfe der zwei genannten Paradigmen zu fassen versucht, ohne damit zugleich zu behaupten, diese Paradigmen seien im gegenwärtigen kognitionswissenschaftlichen sowie philosophischen Diskurs alternativlos. Abhängig davon, ob Kognition symbolisch oder konnektionistisch verstanden wird, können KI-Modelle also einem der genannten Paradigmen der Kognitionswissenschaft zugeordnet werden (Goel, 2021).

1.1.1 Das symbolische Paradigma und klassische KI

Dem *symbolischen Paradigma* zufolge sind kognitive Prozesse physisch realisierte Formen der Manipulation von Symbolen. Kognitive Vermögen werden diesem Paradigma zufolge als logische Symbolmanipulationsverfahren gedeutet (Fodor, 1975). Die fraglichen Symbole gewinnen dabei Bedeutung durch ihre repräsentationale Funktion, was heißt, dass Symbole für etwas – ein Objekt, ein Ereignis etc. – in der Außenwelt stehen bzw. es repräsentieren. Dieser repräsentationale Zusammenhang konstituiert die Bedeutung der verwendeten Symbole, transformiert also syntaktisch wohlgeordnete Strukturen in semantisch gehaltvolle Elemente. Indem Symbole in logische Relationen zueinander gesetzt und mithilfe logischer Schlussregeln manipuliert werden, können kognitive Systeme spezifische Aufgaben lösen. So kann etwa im medizinischen Kontext ein Expertensystem dazu genutzt werden, aus bestimmten Ausgangssätzen – etwa der Darstellung bestimmter Krankheitsbilder – auf Behandlungsoptionen zu schließen. Klassische Formen von KI ba-

---

der Epistemologie und der philosophischen Wahrheitstheorie und damit über die Grenzen des in diesem Beitrag besprochenen Themas hinausführen. In kognitionswissenschaftlichen Debatten, um die es im Folgenden gehen wird, wird der Kognitionsbegriff meist ohne expliziten Bezug zu den Begriffen der Erkenntnis oder Wahrheit erklärt. Das äußert sich unter anderem darin, dass in diesen Debatten versucht wird, Kognition mithilfe solcher Konzepte wie *Informationsverarbeitung* oder *Problemlösung* zu erhellen. Diese Vermögen können aus einer Ingenieursperspektive (Clark, 2001, S. 7) als subpersonale Mechanismen verstanden werden, die nicht zwingend ein vernünftiges Erkenntnissubjekt voraussetzen. In diesem Beitrag werden die relevante philosophische Debatte und das philosophische Verständnis von Kognition als wesentlich vernünftiges Erkenntnisvermögen ausgeblendet, ohne damit über die Möglichkeit einer philosophischen Kritik am hier besprochenen Kognitionsbegriff zu urteilen.

sieren auf diesem Verständnis kognitiver Vermögen (Matthias, 2004, S. 178).

Symbolbasierte Formen der KI erfordern die explizite Programmierung der jeweiligen Symbolmanipulation, für die die KI genutzt wird. Hierzu werden bestimmte Programmiersprachen genutzt. Wer die Programmiersprache versteht, kann die Funktionsweise des Systems erklären, die vom System produzierten Ergebnisse antizipieren sowie im Nachhinein die vom System durchlaufenen Schritte rekonstruieren. Diese Merkmale symbolischer KI sind zugleich Stärke und Schwäche solcher Systeme: Transparente Systeme sind kontrollierbar, die mithilfe solcher Systeme getroffenen Entscheidungen sind nachvollziehbar und etwaige Fehler sind zumindest im Prinzip identifizierbar und behebbar.

Symbolische KI setzt aber auch voraus, dass jede einzelne Regel, nach der die Symbole des Systems manipuliert werden, explizit von Personen implementiert werden muss. Dieser Umstand limitiert die Komplexität eines solchen Systems, da Programmierer*innen nicht beliebig viele Regeln programmieren können. In vielen Anwendungsfeldern kommt daher symbolische KI an ihre Grenzen. Exemplarisch lässt sich etwa die Erkennung von handgeschriebenen Buchstaben anführen. Während Personen für gewöhnlich keine Probleme damit haben, die Buchstaben unterschiedlicher Handschriften trotz einer Vielzahl kleiner Unterschiede zu identifizieren, müsste im Fall einer symbolischen KI jede mögliche Buchstabenvariation explizit programmiert werden. Diese Begrenzungen führen zu einem eingeschränkten Nutzen von KI in unterschiedlichen Bereichen.

1.1.2 Das konnektionistische Paradigma und maschinelles Lernen

Das konnektionistische Paradigma geht von der Annahme aus, dass kognitive Vermögen durch Hirnprozesse realisiert werden (Churchland, 2013). Gehirne sind neuronale Strukturen, in denen Informationen nicht wie im symbolischen Paradigma in Form von Symbolen verankert sind. Stattdessen bilden Neuronen komplexe Netze, in denen einzelne Neuronen verknüpft sind und durch Signale unterschiedlicher Stärke miteinander interagieren. Diese Muster werden im konnektionistischen Paradigma als Informationsverarbeitungsprozesse gedeutet, bei denen die fragliche Information nicht diskret, also nicht an einem spezifischen Ort im Gehirn repräsentiert ist,

sondern durch die Gewichtung der Signalstärke im gesamten System realisiert wird (Sridhar et al., 2023). So erklärt etwa Andreas Matthias:

> »While symbolic artificial intelligence presupposes the existence of clear and distinct symbolic representations of objects and the relations between them, connectionism does not. Instead, it attempts to emulate the basic principles of neural operation in living systems. It is based on the observation that biological information processing systems do not seem to represent symbols as discrete entities, but distributed all over the neural net. Information is stored by modifying the architecture of the network and the strength of individual connections between neurons.« (Matthias, 2004, S. 178)

Neuartige Formen von Künstlicher Intelligenz basieren auf dem *konnektionistischen Paradigma*; hierzu zählen unterschiedliche Formen sogenannter »Machine-Learning« (ML) Algorithmen (Jordan & Mitchell, 2015). Solche Systeme gelten als selbstlernend, wobei hier zwischen verschiedenen Spielarten selbstlernender KI unterschieden werden muss, etwa zwischen unüberwachtem oder überwachtem Lernen. Diese Arten selbstlernender KI-Systeme unterscheiden sich darin, wie stark die Entwickler*innen der Systeme in den »Lernprozess« eingebunden sind. Der Lernprozess selbst besteht darin, dass ein künstliches neuronales Netz mit Input-Daten gespeist wird, woraufhin durch interne Verarbeitungsprozesse des künstlichen neuronalen Netzes – basierend auf der Signalstärke und der Verknüpfung der Neuronen untereinander – ein Output generiert wird. Dieser Output kann beispielsweise in der Klassifikation von Bildmaterial bestehen (Brujne, 2016). Indem also in ein neuronales Netz eine bestimmte Art von Bildern – beispielsweise neurologische Aufnahmen – eingespeist werden, »lernt« das System entweder selbstständig oder assistiert, Muster zu erkennen, die etwa auf Erkrankungen schließen lassen.

Für die ethische Beurteilung ist für den Unterschied von überwachtem und unüberwachtem Lernen die Rolle von Personen zentral. Im Fall des überwachten Lernens müssen Personen die Input Daten selbstständig *labeln*, d. h. vorsortieren und annotieren, während dies beim unüberwachten Lernen nicht der Fall ist (Rubel, 2021). Der technische Vorteil dieser Verfahren besteht darin, dass keine potenziell unüberschaubare Menge expliziter Regeln für die Erkennung von Mustern in den Input-Daten programmiert werden

muss; vielmehr ist selbstlernende KI dazu fähig, eigenständig Korrelationen und Muster in den Input-Daten zu registrieren und zu reproduzieren. Hierdurch ist konnektionistische KI nicht nur dazu in der Lage, Aufgaben zu bewältigen, die mit symbolischer KI unmöglich waren, sondern erfordert darüber hinaus auch weniger explizites Wissen aufseiten der Entwickler*innen von KI (Rajula et al., 2020).

Diese Vorteile weisen aber auch problematische Seiten auf: Erstens ist die Art und Weise, wie selbstlernende Algorithmen ihre Outputs generieren, selbst für die Entwickler*innen dieser KI nicht erklärbar. Im Gegensatz zu klassischen symbolischen Formen der KI sind zeitgenössische konnektionistische Netze *Black-Boxes*, also weitestgehend opak, d. h. nicht einsehbar (Burrell, 2016). Dieser *Black-Box*-Charakter selbstlernender Systeme könnte dann problematisch sein, wenn mithilfe von KI gewichtige und rechtfertigungspflichtige Entscheidungen getroffen werden. Wenn grundsätzlich nicht nachvollziehbar ist, wie eine KI zu einem Ergebnis gelangt, ist fraglich, ob und inwiefern die entsprechende Entscheidung gerechtfertigt werden kann. Wenn beispielsweise Ärzt*innen mithilfe von KI Diagnosen stellen, aber nicht erklären können, was im Einzelnen für die Diagnosen spricht, ist dies vor dem Hintergrund der Rechtfertigungsforderung zumindest auf den ersten Blick ein Problem: Wie kann eine Diagnose als gerechtfertigt gelten, wenn nicht erklärt werden kann, wie sie zustande gekommen ist?

Darüber hinaus hängt die Qualität des Outputs – also etwa die Genauigkeit der produzierten Bildklassifikationen – nicht nur von der Qualität des Input-Materials ab, sondern auch davon, welche Muster die KI »erkennt«. Die Relevanz der Input-Daten für eine verlässliche Funktionsweise von KI hat sich in der Debatte in der Phrase »garbage in, garbage out« niedergeschlagen. Wenn beispielsweise die verwendeten Input-Daten zu homogen sind, kann es passieren, dass die KI versagt, wenn sie im Vergleich zu den Trainingsdaten ähnliche, aber hinreichend heterogene Bilder klassifizieren soll. In solchen Fällen funktioniert eine KI zwar unter Laborbedingungen zuverlässig, außerhalb dieser aber nicht. Umgekehrt können zu heterogene Input-Daten dazu führen, dass die KI nicht zwischen relevanten Informationen und bloßem »Rauschen« differenziert und dementsprechend keine verwendbaren Outputs generiert. Eine KI sortiert das Input-Material im Allgemeinen nicht nach für Men-

schen intuitiv plausiblen Kriterien der Relevanz oder des leitenden Erkenntnisinteresses, vielmehr werden kontingente statistische Ähnlichkeiten aufgedeckt, die mehr oder weniger relevant für den durch die KI verfolgten Zweck sein können.

Dieses Merkmal neuronaler Netze hängt mit einer dritten Problematik zusammen, die für die ethische Analyse des Einsatzes von KI in der Radiologie relevant ist: Das Auftreten so genannter »strange errors«. Ein »strange error« ist ein Fehler, der durch eine selbstlernende KI als Output produziert wird und Menschen in dieser Form nicht unterlaufen würde. Charles Rathkopf und Bert Heinrichs erklären:

> »Strange errors are errors that (1) result from perturbations to the input data that are either unnoticeable to humans, or otherwise strike them as irrelevant to the classification task, and (2) would strike humans as radically incorrect, if they knew the ground truth.« (Rathkopf & Heinrichs, 2024, S. 339)

Fehler dieser Art sind also deswegen »seltsam«, weil die Art und Weise, wie eine KI Outputs produziert, nicht identisch ist mit der Art und Weise, wie Menschen oder spezifisch ihre Gehirne Informationen verarbeiten. Der Umstand, dass selbstlernende KI-Systeme auf künstlichen neuronalen Netzen fußen, sollte also nicht zu der Annahme verleiten, dass dadurch bereits ein Mensch oder dessen neuronale Verarbeitungsprozesse in all ihrer Komplexität und Wechselwirkung mit weiteren Körperfunktionen sowie der natürlichen und sozialen Umwelt nachgebildet seien (Newen et al., 2018). Menschen als Organismen und Resultate eines langwierigen, evolutionären Prozesses, sowie als mit Interessen und Bedürfnissen ausgestattete Vernunftwesen, erkennen nicht *beliebige Muster* in ihrer Umwelt, sondern für die jeweilige Lebensform *bedeutsame* Muster (Thompson, 2007; Johnson, 2008). Dabei bestimmen soziale Prägung, individuelle Interessen, ein komplexes soziales und natürliches Hintergrundwissen und unsere normative Beziehung zur Welt, welche Muster Personen als bedeutsam erfahren. Zeitgenössische Formen von KI sind uns also nicht nur deswegen fremd, weil wir ihre Funktionsweise aufgrund ihres *Black-Box*-Charakters nicht erklären können, sondern auch deswegen, weil sie keine Organismen sind, die evaluativ auf ihre Lebenswelt bezogen sind. Aufgrund dieses Umstands ist es grundsätzlich nicht möglich, im Vorfeld zu antizipieren, welche Art »seltsamer« Fehler eine KI produzieren wird, wie

beispielsweise in dem Fall, in dem eine KI einen Schulbus und einen Strauß miteinander »verwechselt« hat (Ajanki, 2025).

Der vierte ethisch relevante Aspekt zeitgenössischer konnektionistischer KI betrifft die Frage der Akteur*innenschaft, der Autonomie und des Subjektstatus dieser Systeme. In diesem Zusammenhang ist die Annahme verbreitet, ML-Systeme seien in dem Maß autonom, in dem es sinnvoll erscheint, ihnen Handlungsfähigkeit zuzuschreiben (Floridi, 2015). Weil diese Systeme strukturelle Merkmale neuronaler Prozesse bei Personen modellieren und deren Fähigkeiten simulieren, wird darüber hinaus oft geschlossen, KI-Systeme seien Subjekte, bei denen es sinnvoll sein könnte, ihnen einen moralischen Status zuzusprechen und mit ihnen in freundschaftlicher Art verbunden zu sein (Munn & Weijers, 2023; Danaher, 2020). Diese Formen der Anthropomorphisierung von KI können weitreichende ethische Konsequenzen haben (Nyholm, 2020). Für eine ausführliche Darstellung dieser problematischen Anthropomorphisierungstendenzen müsste eine Auseinandersetzung mit den Begriffen Handlungsfähigkeit (Gallagher, 2020; Mayr, 2018; Horn & Löhrer, 2010) und der Subjektivität (Zahavi, 2005; Boyle, 2024) erfolgen. Darüber hinaus müssten weitläufige Debatten der Kognitionswissenschaften besprochen werden, die insbesondere die spezifische Natur der biologischen Verkörperung sowie der sozialen Dimensionen geistiger Vermögen beleuchten (Varela et al., 2016; Lakoff & Johnson, 1999; Shapiro & Spaulding, 2024). Im Rahmen dieses Beitrags kann nur darauf verwiesen werden, dass die in den genannten Debatten aufgeworfenen Fragestellungen die Grundlage einer angemessenen Analyse der problematischen Anthropomorphisierungstendenzen von KI und deren Subjektstatus sind, ohne sie hier auszuführen. Zugleich sollte klar sein, dass zumindest die Formen von KI, die im medizinischen Kontext eingesetzt werden, weder als handlungsfähige Subjekte, die in einem ethisch relevanten Sinn autonom sein können, noch als moralisch berücksichtigungswürdige Entitäten betrachtet werden sollten. Dennoch findet sich auch im medizinischen KI-Diskurs die Tendenz, KI zu anthropomorphisieren, indem etwa behauptet wird, KI entscheide, urteile, kooperiere mit Personen und sei mehr oder weniger vertrauenswürdig. Aus diesem Grund wird es nötig sein, zu prüfen, in welchen Hinsichten diese Anthropomorphisierung die ethische KI-Debatte prägt.

Damit sind die für die folgende ethische Analyse wesentlichen technischen Merkmale zeitgenössischer ML-Algorithmen benannt: Erstens der *Black-Box*-Charakter selbstlernender Algorithmen, zweitens die Abhängigkeit des Outputs von der Qualität des Inputs, drittens die grundsätzlich immer bestehende Möglichkeit »seltsamer« Fehler sowie viertens die in der gegenwärtigen Debatte weit verbreitete Meinung, KI sei nicht bloß ein Artefakt, sondern stelle zumindest eine Art Zwischenschritt von Artefakt zum handlungsfähigen Subjekt dar.

2. Ethische Analyse

Die ethische Analyse von KI in der Medizin orientiert sich überwiegend an einer prinzipienethischen Ausrichtung (Beauchamp & Childress, 2024). Die im Rahmen des Einsatzes von KI in der Radiologie zentralen ethischen Prinzipien sind die *des Wohltuns* bzw. der *Benefizienz*, der *Gerechtigkeit*, der *Transparenz*, der *Verantwortung*, des *Vertrauens* und der *Achtung der Autonomie*. Selbstverständlich ist diese Auflistung nicht erschöpfend, sondern stellt eine themenspezifische Schwerpunktsetzung dar, die notwendigerweise relevante Aspekte auslassen muss. Die Betonung der genannten Prinzipien sollte daher als Systematisierungsversuch verstanden werden, der den Zweck verfolgt, besonders gewichtige moralische Probleme zu beleuchten, nicht als das letzte Wort der ethischen Evaluation. Weiterhin gilt zu beachten, dass die genannten Prinzipien nicht nach Wertigkeit bzw. Gewichtung hierarchisiert sind. Welches Prinzip im Fall eines konkreten Konflikts schwerwiegender – und das heißt handlungsleitend – ist, hängt vom Fall selbst ab. Nicht zuletzt muss berücksichtigt werden, dass die Anwendung von Prinzipien auf konkrete Fälle nicht als eine deduktive Ableitung vom allgemeinen Prinzip zum konkreten Fall verstanden werden darf, in der wohlüberlegte Einzelfallurteile und Intuitionen einseitig unter allgemeine Prinzipien fallen und stets durch diese bestimmt werden. Im Prozess der moralischen Entscheidungsfindung müssen vielmehr allgemeine Prinzipien und wohlüberlegte Einzelfallurteile miteinander in ein Reflexionsgleichgewicht gebracht werden (Daniels, 1979; DePaul, 1993).

Der folgenden ethischen Analyse werden jeweils kurze, begriffliche Untersuchungen vorangestellt. Diese fallen unterschiedlich ausführlich aus, je nachdem, wie umstritten oder mehrdeutig eines der genannten Prinzipien ist, und je nachdem, wie schwerwiegend unterschiedliche Begriffsbestimmungen die ethische Analyse beeinflussen. So werden die Begriffe des Vertrauens und der Autonomie beispielsweise ausführlicher besprochen als der mit dem Prinzip der Benefizienz verknüpfte Begriff des Wohlergehens. Der Grund dafür ist nicht, dass der Begriff des Wohlergehens philosophisch uninteressant oder unkontrovers wäre, sondern dass das für die folgenden Fragestellungen relevante Verständnis des Wohlergehens in den hier einschlägigen ethischen Debatten vergleichsweise weniger Probleme aufwirft als etwa der Begriff des Vertrauens. Wie sich zeigen wird, führen dagegen unterschiedliche Deutungen, beispielsweise des Vertrauensbegriffs als auch der Autonomie, zu unterschiedlichen moralischen Bewertungen, sodass in diesen Fällen eine umfangreichere konzeptionelle Analyse vorgenommen werden muss.

2.1 Benefizienz

Das Prinzip der Benefizienz verpflichtet Personen dazu, das Wohlergehen anderer Personen zu achten. Während der Begriff des Wohlergehens eine umfangreiche und bis in die eudaimonistische Ethik der Antike zurückreichende Geschichte aufweist (Vasiliou, 2025) und auch in der zeitgenössischen Philosophie und Psychologie umfassend debattiert wird (Griffin, 1986; Bradley, 2015; Seligman, 2011), reicht es für die Zwecke dieses Beitrags, von einem intuitiven Verständnis des Wohlergehens auszugehen, da die relevanten Wohlergehensaspekte in der ethischen Debatte wenig strittig sind. Im Kontext der Medizinethik stehen hierbei in erster Linie die Gesundheit der Patient*innen, die Arbeitsbelastung des ärztlichen Personals sowie die ökonomische Effizienz des Gesundheitswesens im Mittelpunkt. Mit dem Einsatz von KI in der Medizin bzw. der Radiologie verbindet sich also die Hoffnung, dass KI in unterschiedlichen Hinsichten sowohl das Wohl der Patient*innen fördert als auch dem ärztlichen Fachpersonal und nicht zuletzt der Gesellschaft allgemein nützt (Liua et al, 2020). Grundlage dieser Hoffnung ist die Erwartung, dass selbstlernende KI-Systeme bestimmte Aufgaben,

die zuvor ausschließlich von Personen ausgeführt werden konnten, effizienter bearbeiten werden als diese. Zu solchen Aufgaben zählen etwa die Auswertung von Schriftmaterial, die Organisation von Abläufen in Krankenhäusern oder Kliniken, die Diagnosestellung oder die Analyse großer Datensätze, sowohl im Kontext von Therapie als auch in der medizinischen Forschung (Topol, 2019; Steckmann & Heinrichs, 2023). Dadurch könnten Personalkosten eingespart und Ärzt*innen von unliebsamen Verwaltungspflichten befreit werden. Der Einsatz von KI könnte im Zuge dessen zu einer stärker personenfokussierten Medizin führen, indem durch die Automatisierung organisatorischer Abläufe dem ärztlichen Fachpersonal mehr Zeit für den direkten Kontakt mit Patient*innen zur Verfügung stünde. Durch effizientere Prozesse und Kosteneinsparungen könnte zudem die medizinische Versorgung insgesamt verbessert und kostengünstiger gestaltet werden, wodurch die Gesundheitsversorgung der Bevölkerung den Staatshaushalt weniger belasten würde.

Zugleich sollte nicht unkritisch davon ausgegangen werden, dass der Einsatz von KI in der medizinischen Praxis tatsächlich alle an diese Technologie geknüpften Hoffnungen erfüllen wird. So ist zu bedenken, dass eine KI-gestützte Diagnose, sollte sie zu einem anderen Ergebnis gelangen als eine ärztliche Fachperson, die Notwendigkeit aufwirft, eine weitere Meinung einzuholen oder zu prüfen, ob und inwiefern die KI fehlerhaft ist und damit weitere Arbeitsschritte in den ärztlichen Berufsalltag einführen kann. Gerade vor dem Hintergrund der Möglichkeit »seltsamer« Fehler ist ein solches Szenario nicht auszuschließen. Ohne etablierte Verfahren, die eine rasche Fehlererkennung ermöglichen, kann das Auftreten solcher Fehler dazu führen, dass bestimmte Vorgänge des medizinischen Alltags sich als langwieriger und kostspieliger herausstellen, als sie es ohne den Einsatz von KI wären. Da im Vorfeld des Einsatzes dieser Technologie nur schwer abzuschätzen ist, welches der genannten Szenarien – Zeit- und Kostenersparnis oder Mehrkosten und Zusatzarbeit – wahrscheinlicher ist, ist ungewiss, ob der Einsatz von KI in der Medizin tatsächlich das Wohlergehen von Personen fördern wird.

2.2 Gerechtigkeit

Der zweite ethische Gesichtspunkt des Einsatzes von KI in der Medizin bzw. der Radiologie betrifft Fragestellungen der Gerechtigkeit. Die hierbei einschlägige Form der Gerechtigkeit ist eine Spielart der Verteilungsgerechtigkeit, die primär die Verteilung von Ressourcen zur Sicherstellung einer angemessenen Gesundheitsversorgung berührt. Bei der Anwendung von KI in der Medizin ergibt sich dieses Problem insbesondere aus der Selektion der Input-Daten: So können sich Formen der strukturellen Diskriminierung[3] systematisch in KI-Systemen niederschlagen (Grote & Keeling, 2022; Koçak et al., 2025; Coeckelbergh, 2020a). Werden beispielsweise Trainingsdaten verwendet, die nur einen Teil der Bevölkerung repräsentieren, kann dies dazu führen, dass ein selbstlernendes Programm für diese Personengruppe verlässliche Outputs generiert, im Fall anderer Gruppen aber versagt. Mittlerweile klassische Beispiele im medizinischen Kontext hierfür sind etwa KI-Programme in der Dermatologie, die zur Identifikation von Hautkrebs eingesetzt wurden. Werden solche Programme mit Input-Daten von vorwiegend hellhäutigen Personen trainiert, kann das dazu führen, dass sie auf dunkler Haut unzuverlässige Ergebnisse produzieren (Adamson, 2018). Dadurch können strukturelle Formen der Diskriminierung zugespitzt werden, in denen benachteiligte Personengruppen aufgrund von ungerechtfertigten Ungleichbehandlungen Diskriminierungen erfahren (Heinrichs, 2021).

Ein eingeschränkter Zugang zu gesundheitsrelevanten Technologien bzw. Technologien, deren Verlässlichkeit bei diskriminierten Personengruppen abnimmt, ist eine Form von Diskriminierung.

3 Während im Rahmen dieses Beitrags davon ausgegangen wird, dass Diskriminierung eine Form der Ungleichbehandlung darstellt und aufgrund von bestimmten Faktoren wie biologischem oder sozialem Geschlecht, sexueller Orientierung, Hautfarbe etc. moralisch problematisch ist, sollte nicht unerwähnt bleiben, dass sowohl der Begriff der Diskriminierung wie auch seine moralische Relevanz in der philosophischen Debatte diskutiert wird. Der Punkt ist hierbei weniger die allgemeine Zurückweisung der moralischen Relevanz von Diskriminierung, sondern eine ethische Analyse der Wurzeln dieser Problematik. In dieser Debatte wird etwa darauf aufmerksam gemacht, dass es unterschiedliche Formen der Ungleichbehandlung gibt, die sich in ihrer moralischen Relevanz unterscheiden (Halldenius, 2005). Für eine umfangreiche philosophische Auseinandersetzung mit dem Diskriminierungsbegriff vgl. Lippert-Rasmussen (2013).

Diese Überlegungen zeigen, dass Künstliche Intelligenz als ebenso wenig wertfrei anzusehen ist wie andere Technologien und daher nicht allein unter technologischen Gesichtspunkten betrachtet werden kann und sollte. Das Design und der Einsatz von technologischen Produkten haben stets ethische Implikationen, selbst wenn diese von den Entwickler*innen der jeweiligen Technologie nicht antizipiert oder gar intendiert waren (Poel, 2020). Wenn sich also diskriminierende Strukturen durch den Einsatz bestimmter Technologien vertiefen, heißt das *nicht in jedem Fall*, dass konkrete Individuen hierfür verantwortlich gemacht werden können. Es heißt vielmehr, dass strukturelle Lösungen für strukturelle Probleme gesucht werden müssen. Umgekehrt bedeutet das aber auch, dass gute Intentionen aufseiten der Entwickler*innen von KI nicht notwendigerweise verhindern, dass der Einsatz von KI Personengruppen diskriminieren wird.

Wird eine Personengruppe aufgrund bestimmter Merkmale – etwa biologischem oder sozialem Geschlecht, Hautfarbe, kultureller Zugehörigkeit, sexueller Orientierung u. Ä. – ohne rechtfertigende Gründe ungleich behandelt, ist dies eine Form der *Diskriminierung*, die in der Debatte unter dem Begriff »*bias*« diskutiert wird. Ein besonderes Merkmal dieses Diskurses im Rahmen des Einsatzes von KI in der Medizin betrifft hierbei eine eigentümliche Vermischung deskriptiv-technologischer und normativ-moralischer Erwägungen. So finden sich auf der einen Seite Arbeiten innerhalb der KI-Debatte, in denen der *Bias*-Begriff wie oben dargestellt moralisch verstanden wird. Trishan Panch, Heather Mattie und Rifat Atun etwa definieren

> »algorithmic bias in the context of AI and health systems as: ›the instances when the application of an algorithm compounds existing inequities in socioeconomics status, race, ethnic background, religion, gender, disability or sexual orientation to amplify them and adversly impact inequities in health systems.‹« (Panch et al., 2020, S. 1)

Demgegenüber charakterisieren insbesondere Autor*innen, die aus einer technologischen Perspektive argumentieren, »biased algorithms« als ein System, das »systematically produces outcomes that are not statistically expected.« (Filippi et al., 2023, S. 1242) Ähnlich erklären Koçak et al.:

> »The concept of bias in machine learning (ML) research and more generally in the field of predictive modeling is intrinsically tied to the concept of variance. In this context, bias can be defined as the distance (or error) between the prediction and the actual target variable, whereas variance signifies the dependence of predictions on the randomness in the training data sampling.« (Koçak et al., 2025, S.76)

Während die erste dieser Definitionen die *moralische* Dimension von *biases* ausdrückt und damit normativ ist, versuchen die letztgenannten Charakterisierungen, *biases* rein *deskriptiv* – in diesem Fall durch quantitativ bestimmbare Wahrscheinlichkeiten bzw. Erwartbarkeiten – zu bestimmen. Ein solches Vorgehen ist in technologischen Disziplinen nicht unüblich, da dort die Quantifizierung und Formalisierung von Problemen die Grundlage für technische Lösungen darstellt.[4] Zugleich sollte beachtet werden, dass rein deskriptive Definitionen zwar oft mit moralischen Erwägungen zusammengeführt werden, für sich genommen aber kein moralisches Gewicht haben, da die aus dem obigen Zitat stammende Formulierung »Ergebnisse, die statistisch nicht erwartet wurden« an sich moralisch neutral ist. Nicht alle denkbaren statistisch unerwarteten Ergebnisse sind moralisch problematisch, sodass die normative Arbeit darin besteht, auf Basis normativer Erwägungen jene unerwarteten Ergebnisse zu identifizieren, die moralische Probleme aufwerfen.

Um für moralische Fragestellungen nutzbar gemacht werden zu können, muss also in der Bestimmung des relevanten *Bias*-Begriffs ersichtlich werden, dass die fraglichen, nicht erwarteten Ergebnisse moralisch zu beanstanden sind. Für diesen Zweck ist es notwendig, die inhärent moralischen Begriffe der Gerechtigkeit und der Diskriminierung mit dem des *bias* zu verknüpfen und die Erkenntnisse der normativen Gerechtigkeitsforschung in der Ethik und politischen Philosophie zu beachten (Binns, 2018). Diese Notwendigkeit zur Ergänzung technischer *Bias*-Definitionen um moralische Reflexionen wird dadurch verstärkt, dass auch in weniger technischen Diskursen ein deskriptiver *Bias*-Begriff verwendet wird. So wird etwa in diversen Debatten zwischen »desirable« und »undesirable biases« unterschieden (Cirillo et al., 2020). Die Idee ist hierbei, dass bestimmte Formen von *bias* deshalb wünschenswert sein können, weil die Be-

4 Ein Beispiel für eine solche deskriptive und mathematisierte Annäherung an Gerechtigkeitsfragen findet sich etwa in Favier et al. (2023).

achtung bestimmter Faktoren wie etwa dem biologischen Geschlecht für bestimmte medizinische Fragestellungen von Bedeutung ist. Der hier verwendete *Bias*-Begriff bezeichnet damit keine *moralisch problematische* Ungleichbehandlung, sondern eine moralisch neutrale, aber medizinisch signifikante Beachtung bestimmter Merkmale einer Person oder Personengruppe. Für den Fall, dass bestimmte Personengruppen anfälliger für bestimmte Krankheiten sind, kann es wünschenswert (»desirable«) sein, Merkmale in den Blick zu nehmen – etwa das biologische Geschlecht –, deren Beachtung in anderen Kontexten diskriminierend wäre. Einem solchen Verständnis zufolge besteht ein *bias* also darin, dass in unterschiedlichen Fällen unterschiedliche Personen oder Personengruppen dadurch unterschiedlich behandelt werden, weil verschiedene Merkmale in den Mittelpunkt der Betrachtung rücken. Dieses Verständnis lässt offen, ob eine Ungleichbehandlung moralisch problematisch (*undesirable*) oder wünschenswert (*desirable*) ist, sodass der Verweis darauf, dass hier ein *bias* vorliegt, an sich noch keine moralische Handlungsleitung impliziert. Vielmehr muss im Rahmen einer ethischen Analyse geprüft werden, unter welchen Bedingungen die Ungleichbehandlung moralisch ver- oder geboten ist. Eine rein technische oder deskriptive Definition von *biases* und Diskriminierung kann daher nicht die philosophische Analyse der Diskriminierungsproblematik ersetzen.[5]

Vor dem Hintergrund dieser Überlegungen ist ein KI-System dann in einem moralisch problematischen Sinne als *biased* zu bewerten, wenn die von ihm produzierten, statistisch nicht erwarteten Ergebnisse zur Reproduktion ungerechter Strukturen führen – etwa, indem bestimmte Personengruppen aufgrund medizinisch irrelevanter Merkmale systematisch benachteiligt und dadurch im Zugang zu einer gleichwertigen Gesundheitsversorgung diskriminiert werden. Ansätze in der Entwicklung und dem Einsatz von KI, die darum bemüht sind, solchen moralisch relevanten Formen von *bias* zu begegnen, werden unter dem Stichwort der »Fair AI« zusammengefasst (Feuerriegel et al., 2020). Wenngleich die Diskriminierungsproble-

5 Vgl. auch die Differenzierung von formaler und substanzieller Fairness in Rubel et al. (2021). Rein quantifizierte Formen von Gerechtigkeit erfassen zwar formale Fairness, nicht aber substanzielle. Aus diesem Grund kann eine Anwendung von Algorithmen keine philosophische Grundlagenarbeit an Gerechtigkeitserwägungen ersetzen, sondern diese im besten Fall nur ergänzen.

matik nicht allein durch eine technologiefokussierte Analyse von *biases* gelöst werden kann, muss eine angemessene Lösung dieser Probleme im Kontext von KI-Systemen selbstverständlich technische Mittel anwenden. Aus diesem Grund ist es notwendig zu prüfen, welche Quellen von *biases* im Zusammenhang mit KI-Systemen relevant sind. Im Rahmen einer technologischen Charakterisierung von *bias* werden unterschiedliche Formen von *biases* unterschieden, so beispielsweise »data bias«, »measurement bias« oder »algorithmic bias« (Xu et al., 2022, S. 77).

Diese unterschiedlichen Arten von *bias* haben jeweils unterschiedliche Ursachen und ihre moralische Relevanz hängt von ihrem jeweiligen Einsatzbereich ab. So kann der Ursprung eines *bias in den Trainingsdaten selbst* liegen oder in der *Art und Weise, wie ein KI-System die Daten verarbeitet*, also im *Design der KI* (Hooker, 2021), oder aber, im Fall von überwachtem Lernen, in *von Personen vorgenommenen Annotationen* – also dem Prozess der Kennzeichnung und Markierung – *der Trainingsdaten* (Cross et al., 2024). Um den aus diesen Quellen stammenden Diskriminierungseffekten aus technologischer Perspektive zu begegnen, werden so genannte »Fairness metrics« besprochen, also quantifizierte Evaluationskriterien, die im Design und Einsatz einer KI beachtet werden sollten (Xu et al., 2022).

Vor diesem Hintergrund lässt sich ohne Beachtung der bereichsspezifischen Anforderungen an die gerechte Auswahl von Trainingsdaten, das Design der KI und der von Personen durchgeführten Annotation und händischen Korrektur von KI-Algorithmen nicht ohne Berücksichtigung des Kontextes allgemein festlegen, welche moralisch relevanten *biases* zu beachten sind. Das heißt aber nicht, dass keine allgemeinen Rahmenbedingungen benannt werden könnten, die als Orientierungspunkte dienen, moralisch relevante Aspekte mit Hinblick auf die Gerechtigkeitsproblematik zu identifizieren. So ist etwa zu beachten, dass die Trainingsdaten für den Einsatzbereich hinreichend repräsentativ sind (Reeves, 2024), sodass etwa nicht sozioökonomisch schlechter gestellte Gruppen Nachteile in der Gesundheitsversorgung erleiden müssen (Waite & Scott, 2021). Alle Personengruppen sollten in den Trainingsdaten in einer Weise vertreten sein, die es der KI ermöglicht, für jede dieser Gruppen vergleichbar verlässliche Ergebnisse zu produzieren. Diese Anforderung setzt bereits vor dem Trainingsprozess bei der Auswahl der

Trainingsdaten an und gehört damit zum Entwicklungsabschnitt eines KI-Systems. Im anschließenden Trainingsprozess selbst ist zu beachten, dass keine expliziten oder impliziten Vorurteile seitens der Personen, die für die Annotationen und Vergabe von Labels verantwortlich sind, den Lernprozess der KI beeinflussen (Kamiran & Calders, 2012). Zu diesem Zweck sollte erstens ein Bewusstsein dafür geschaffen werden, dass die *Bias*-Problematik im Kontext von KI *kein rein technologisches*, sondern ein *inhärent moralisches* Problem darstellt, bei dem technologische Erwägungen und Lösungsansätze zwar relevant sind, aber ohne eine ethische Analyse unvollständig bleiben. Zweitens sollten Wege gefunden werden, die ethische Analyse strukturell in die Entwicklung, das Design und – im Fall von überwachtem Lernen – den Annotationsprozess zu implementieren, etwa, indem in diesen Prozessen nicht nur technologisch geschulte Fachkräfte involviert sind, sondern auch Ärzt*innen sowie Personen mit ausgewiesener Expertise in ethischer Analyse.[6] Im Anschluss an die Trainingsphase gilt es schließlich, auch in der Anwendungsphase darauf zu achten, dass sich keine diskriminierenden *biases* niedergeschlagen haben und reproduzieren. Spätestens in dieser Phase ist es unerlässlich, das ärztliche Fachpersonal miteinzubinden (Filippi et al., 2023).

2.3 Vertrauen

2.3.1 Begriffliche Grundlagen des Vertrauensbegriffs

Mit den Arbeiten Onora O'Neills (O'Neill, 2002a; O'Neill, 2002b) zur Rolle des Vertrauens in der medizinischen Praxis hat das Prinzip des Vertrauens bzw. der Vertrauenswürdigkeit Einzug in die medizinethische Debatte gehalten. Das Prinzip des Vertrauens ist aber auch Gegenstand weiterführender Debatten der Ethik (Wolfensberger & Wrigley, 2019; Steinfath, 2016; Baier, 1986; Fabris, 2020; Simon, 2020; Simpson, 2023). Vertrauen stellt eine spezifische Art der Einstellung von Personen zu anderen Personen dar, in der sich

6 Vgl. hierzu etwa Winter & Carusi (2023). Winter und Carusi besprechen in dieser Publikation primär die Einbindung von Ärzt*innen in die Entwicklung einer medizinischen KI; dieser Gedanke kann und sollte aber auch auf Fachkräfte ausgedehnt werden, deren Expertise in der ethischen Analyse liegt.

niederschlägt, dass die vertrauende Person von der Person, der sie vertraut, erwartet, diese sei ihr wohlgesonnen oder zumindest nicht feindlich gestimmt und in Bezug auf den das Vertrauen betreffenden Handlungsvollzug kompetent ist. So erklärt Karen Jones:

> »[T]rust is an attitude of optimism that the goodwill and competence of another will extend to cover the domain of our interaction with her, together with the expectation that the one trusted will be directly and favorably moved by the thought that we are counting on her.« (Jones, 1996, S. 4)

In einem Vertrauensverhältnis besteht also die *Erwartung*, dass die Person, der vertraut wird, in ihren Handlungsvollzügen durch das Wohlergehen der Person, die vertraut, motiviert ist. Das bedeutet nicht zwingend, dass diese Motivation sich auf alle Aspekte des Wohlergehens erstreckt, sehr wohl aber – je nach sozialer Interaktion –, dass das Wohlergehen zumindest nicht insofern irrelevant ist, als dass ein Schaden billigend in Kauf genommen wird. Aus diesem Grund erscheint es unangemessen, einer Person zu vertrauen, von der wir wissen, dass sie uns nicht wohlgesonnen, das heißt uns gegenüber feindselig eingestellt ist. Zugleich ist diese Art der Einstellung zwischen Personen wesentlich für soziale Kooperation: Wenn wir einander nicht in basalen Dingen vertrauen würden, wäre kooperatives Handeln unmöglich. Hierzu ist es nicht zwingend nötig, dass Personen einander kennen. Dass etwa die Teilnehmenden des Straßenverkehrs einander vertrauen, dass sich alle anderen zumindest grundsätzlich an die Verkehrsregeln halten, statt anderen Personen absichtlich zu schaden, ist konstitutiv für die Möglichkeit eines geordneten Verkehrswesens. Vertrauen hat damit zumindest den instrumentellen Wert, dass ohne Vertrauensverhältnisse das gesellschaftliche Zusammenleben unmöglich wird.

Dass Vertrauen normativ rechtfertigungsfähig statt bloß kausal bedingt ist, bedeutet, dass im philosophischen Verständnis nicht die Frage untersucht wird, unter welchen Bedingungen Personen anderen Personen *de facto* vertrauen, d. h., was die kausalen Determinationsfaktoren sind, die dazu führen, dass Personen einander vertrauen, sondern unter welchen Bedingungen Vertrauen *gerechtfertigt* ist, d. h., unter welchen Bedingungen wir einander also vertrauen *sollten*. Der Unterschied zwischen kausaler Verursachung und normativer Rechtfertigung ist wesentlich für die Idee normativer Gründe. Zu sagen, dass es Gründe gibt, einer Person zu vertrauen,

impliziert nicht, dass wir dieser Person notwendigerweise vertrauen werden, sondern nur, dass wir ihr begründeterweise vertrauen sollten. Dass kausale und normative Aspekte auseinanderfallen können, zeigt sich etwa daran, dass Personen häufig auch dann Vertrauen schenken – das ist die deskriptiv-kausale Seite –, wenn die Personen, denen sie vertrauen, nicht vertrauenswürdig sind – das ist die normative Perspektive. So vertrauen Personen etwa häufig Autoritätsfiguren, die durch Manipulationstechniken Vertrauen erschleichen. In Fällen dieser Art wird Vertrauen faktisch entgegengebracht, obwohl es starke normative Gründe gibt, dies nicht zu tun. Der normative Grundgedanke lautet hier also, dass Vertrauen besser oder schlechter begründet werden kann und dass eine deskriptive Analyse, wem Personen faktisch vertrauen, für die philosophische Fragestellung der Vertrauenswürdigkeit allein nicht hinreichend ist.

Zugleich muss betont werden, dass die normativen Gründe für Vertrauen vielfältig sind. So gibt es Fälle, in denen wir Personen vertrauen sollten, die sich in der Vergangenheit nicht als vertrauenswürdig herausgestellt haben. In der Debatte werden in diesem Zusammenhang bestimmte Vertrauensformen diskutiert, etwa das therapeutische Vertrauen oder das korrektive Vertrauen (Scheman, 2020, S. 28). Bei diesen Formen des Vertrauens geht es darum, die Person, der vertraut wird, dazu zu bewegen, sich vertrauenswürdig zu verhalten, und zwar selbst dann bzw. gerade, weil sie in der Vergangenheit nicht vertrauenswürdig agierte. Ein solches Vertrauensverhältnis kann etwa in therapeutischen oder erzieherischen Kontexten notwendig sein. Auch zum Zweck, eine Beziehung zu stärken und als Ausdruck des Respekts vor dem Gegenüber kann es angemessen sein, einer Person Vertrauen zu schenken, die sich in gewissen Hinsichten bisher nicht als vertrauenswürdig erwiesen hat. Wenn etwa ein*e Freund*in, die bislang unzuverlässig war, Besserung gelobt, kann es ein Gebot des freundschaftlichen Respekts sein, ihr das zu glauben und – zumindest, bis klar geworden ist, ob sie ihr Versprechen ernst meint – ihr zu vertrauen.

Von diesen Beispielen abgesehen gilt, dass wir einer Person genau dann vertrauen sollten, wenn die Person vertrauenswürdig ist. Die normative Grundlage von Vertrauenswürdigkeit ist dabei, wie eingangs ausgeführt, der minimale Umstand, dass die Person unser Wohlergehen zumindest nicht untergraben will. Diese Person ist also aufgrund ihrer Natur als praktisch vernünftiges Wesen in der Lage,

zu erkennen, dass unser Wohlergehen einen normativen Grund liefert, und auf Basis dieser Einsicht entsprechend zu handeln. Eine solche Person besitzt bestimmte charakterliche Dispositionen, die in der Sprache der Tugendethik als Tugenden bezeichnet werden (Potter, 2003, S. 1ff.). Eine tugendhafte Person ist dadurch gekennzeichnet, dass sie in spezifischen Situationen erkennt, welche Handlung moralisch geboten ist und dementsprechend agiert (Caro & Vaccarezza, 2021). Eine solche Person ist besorgt um das Wohlergehen ihrer Mitmenschen und darum bemüht, moralische Vorgaben zu achten, indem sie weder lügt noch betrügt, gerecht ist, etc. Tugendhafte Personen handeln weder selten noch ausnahmslos tugendhaft, aber aufgrund ihrer charakterlichen Verfassung gibt es gute evidenzbasierte Gründe, ihnen zu vertrauen, da wir sie nur dann als tugendhaft bezeichnen, wenn sie in der Regel moralisch agieren. Eine tugendhafte Person handelt hierbei vertrauenswürdig, *weil sie einsieht,* dass dieses Handeln richtig ist. Eine tugendhafte Person handelt also nicht *zufälligerweise vertrauenswürdig*, sondern deshalb, weil sie dieses Handeln als Ausdruck einer praktisch vernünftigen Einsicht als richtig versteht. Daraus folgt, dass eine solche Person, wenn sie in Einzelfällen nicht vertrauenswürdig agiert, darum bemüht sein wird, diesen Fehler in Zukunft zu beheben. Zu wissen, dass eine Person einen tugendhaften Charakter besitzt, rechtfertigt damit das Vertrauen in diese Person.

Ein weiterer zentraler Aspekt besteht darin, dass Vertrauen mit *interpersonalen* reaktiven Einstellungen wie Dankbarkeit, Stolz, Scham etc. einhergeht, wobei diese Einstellungen nicht bloß *kausal* verursacht werden, sondern *normatives* Gewicht haben. Wird etwa Vertrauen gebrochen, ist die reaktive Einstellung der Enttäuschung nicht bloß eine kausale Reaktion und damit ein nichtnormativer psychischer Mechanismus, sondern in dem Sinne normativ, dass sie angemessen und gerechtfertigt sein kann. Dieses nicht lediglich kausale, sondern normative Verständnis von reaktiven Einstellungen, das ursprünglich von Peter Strawson entwickelt wurde (Strawson, 2008), ist für die moralische Dimension des Vertrauensbegriffs wesentlich. Reaktive Einstellungen sind nicht beliebige emotionale Reaktionen, sondern stellen *interpersonale* Einstellungen dar, die grundlegend für unsere Praxis der moralischen Verantwortungszuschreibung sind (Wallace, 2022). Dass reaktive Einstellungen nicht beliebige emotionale Reaktionen darstellen, zeigt sich darin, dass be-

reits im Begriff der reaktiven Einstellung der Verweis auf den guten Willen wesentlich ist, wie Strawson betont, wenn er schreibt: »What I have called the participant reactive attitudes are essentially natural human reactions to the good or ill will or indifference of others towards us, as displayed in *their* attitudes and actions.« (Strawson, 2008, S. 10–11) Während Strawson eine Reihe weiterer Subkategorien reaktiver Einstellungen vorstellt (Chaplin, 2023, S. 323), ist für die Zwecke dieses Beitrags lediglich relevant, dass reaktive Einstellungen interpersonaler Natur sind und als normativ gehaltvolle und zugleich emotive Reaktionen evaluierbar sind. Nicht nur ist ein Vertrauensbruch vor diesem Hintergrund moralisch bedeutsam und mit reaktiven Einstellungen verbunden, die als moralische Evaluationen fungieren; darüber hinaus kann auch die Angemessenheit dieser reaktiven Einstellungen selbst zur Debatte stehen – etwa dann, wenn sie dem Ausmaß des Vertrauensbruchs nicht gerecht werden. Wer etwa in einem Vertrauensverhältnis stark empört oder gar wütend reagiert, weil die Person, der vertraut wurde, einen minimalen Vertrauensbruch begangen hat, agiert ebenso wenig angemessen, wie eine Person, die bei einem schweren Vertrauensbruch keinerlei reaktive Einstellungen zeigt oder sogar Dankbarkeit zum Ausdruck bringt, wo tatsächlich Empörung angebracht wäre.

Gemäß den vorgestellten Überlegungen ist ein Vertrauensverhältnis eine *normative Relation einer bestimmten Art* zwischen einer bestimmten Klasse von Subjekten, nämlich *moralischen Akteur*innen*. Die relevante Art der normativen Relation betrifft spezifische, moralische reaktive Einstellungen, nämlich *interpersonale reaktive Einstellungen*. Beispiele solcher Einstellungen sind Dankbarkeit, Vergebung, Entrüstung etc. Nicht nur sind allein gegenüber moralischen Akteur*innen moralische reaktive Einstellungen angemessen, darüber hinaus können auch nur solche Akteur*innen im moralisch relevanten Sinne durch die Erwägung des Wohlergehens ihres Gegenübers motiviert sein. Moralische Akteur*innen sind diejenigen Wesen, gegenüber denen es angemessen ist, ihr Handeln nach moralischen Maßstäben zu bewerten. Wesen dieser Art müssen die Fähigkeit besitzen, moralische Gesichtspunkte einzusehen und ihr eigenes Handeln diesen Gesichtspunkten entsprechend auszurichten. Solche Wesen sind selbstbewusst in dem Sinne, dass sie ihre eigenen Handlungsgründe reflektieren, diese in diskursiven Praktiken auf Basis von Argumenten und Einsichten modifizieren oder, wo nötig,

aufgeben können und deren Handlungsmotivation durch praktisch vernünftige, moralische Erwägungen bestimmt werden kann. Die fraglichen Wesen müssen also vernunftbegabt, selbstbewusst – im Sinne des Bewusstseins der eigenen theoretischen und praktischen Gründe sowie ihrer eigenen Situiertheit in moralischen Praktiken – und affektiv-sensitiv für moralisch relevante Aspekte von Situationen sein. Während kognitiv nicht eingeschränkte erwachsene Personen unter diese Klasse fallen, sind weder Tiere noch sehr kleine Kinder moralische Akteur*innen in diesem Sinne,[7] weswegen wir die Handlungen dieser Subjekte nicht bzw. noch nicht moralisch evaluieren. Nur wenn die Subjekte, die in einer Vertrauensrelation miteinander verbunden sind, moralische Akteur*innen in diesem Sinn darstellen, ist es angemessen, die besprochenen Formen reaktiver Einstellungen einzunehmen. Entsprechend wäre es nicht angemessen, diese Art des Vertrauensverhältnisses gegenüber einem Tier einzunehmen und empört zu reagieren, wenn das Tier das Vertrauen bricht. Das heißt nicht, dass Menschen die fraglichen Einstellungen nicht de facto einnehmen können, sondern nur, dass diese Einstellungen in solchen Fällen nicht gerechtfertigt sind.

2.3.2 Ethische Analyse: Vertrauen und Verlässlichkeit

Vor dem Hintergrund dieser begrifflichen Bestimmung des Vertrauens kann nun geprüft werden, ob und inwiefern dieser Begriff auf KI-Systeme Anwendung finden kann und, sollte das nicht der Fall sein, welches alternative Konzept einschlägig ist. Zumindest die in der Radiologie eingesetzten KI-Systeme können weder als Subjekte noch als Akteur*innen in einem engeren Sinne gelten. Solche KI-Systeme sind weder tugendhafte Akteur*innen noch – und diese Gedanken hängen zusammen – können sie auf Basis praktischer Urteilskraft prüfen, welche normativen Gründe für bestimmte Handlungsweisen sprechen. Außerdem sind sie nicht in der

7 Das heißt nicht, dass sie deswegen moralisch nicht berücksichtigungswürdig wären. Im philosophischen Diskurs hat es sich in diesem Zusammenhang eingebürgert, zwischen moralischen Akteur*innen, also *moral agents* und moralisch relevanten Entitäten, also *moral patients* zu unterscheiden. Sowohl Tiere als auch kleine Kinder sind moralisch berücksichtigungswürdige Wesen und damit *moral patients*, aber sie sind nicht zu moralischen Handlungen befähigt, sie sind also keine *moral agents*.

Lage, zu reflektieren, in welcher sozialen Rolle sie sich gegenüber einer anderen Person befinden und welche Pflichten und Rechte hieraus erwachsen. Ebenso wenig sinnvoll erscheint es, einer KI Vertrauen entgegenzubringen, um etwa das persönliche Verhältnis zu ihr zu stärken. Nicht zuletzt ist unklar, was es heißen soll, dass die fraglichen KI-Systeme um unser Wohlergehen besorgt sind oder uns nicht in feindseliger Haltung gegenüberstehen. Aus diesen Gründen erscheint es unsinnig, gegenüber einer KI eine reaktive Einstellung einzunehmen (Rebera, 2024). Weil einer KI diese für Vertrauensverhältnisse relevanten Merkmale fehlen, sind KI-Systeme nicht die Art von Entität, zu denen ein normatives Vertrauensverhältnis eingenommen werden kann. Dennoch ist der Begriff der »vertrauenswürdigen KI« eines der zentralen Schlagwörter der KI-Debatte (Roberson et al., 2022). Aus philosophischer Perspektive wird diese Begriffsverwendung bisweilen kritisch gesehen, weil die Gefahr besteht, dass die oben aufgeführten Aspekte des Vertrauensverhältnisses dadurch aus dem Blick geraten und ein Kategorienfehler begangen wird (Metzinger, 2019). Diese Problematik wird durch die Meinung verschärft, KI sei ein Subjekt, wodurch suggeriert wird, dass das ausgeführte normative Vertrauensverhältnis auch zwischen Personen und KI-Systemen bestehen könnte (Ryan, 2020).

Das bedeutet jedoch nicht, dass im Kontext des KI-Einsatzes in der Medizin keine ethischen Fragestellungen auftreten, die mit dem Vertrauensbegriff in Verbindung stehen. Insbesondere der Begriff der *Verlässlichkeit*, der eng mit dem des Vertrauens verknüpft ist, rückt hier in den Fokus. Ein Verlässlichkeitsverhältnis hat eine andere Struktur und andere Implikationen als ein Vertrauensverhältnis. Sanford Goldberg charakterisiert das Verlässlichkeitsverhältnis so:

> »[W]e might characterize reliance in terms of a supposition one is prepared to act on: where X is a person, artifact, or natural process, and Φ is an action, behavior or process, to rely on X to Φ is to act on the supposition that X will Φ.« (Goldberg, 2020, S. 97)

Anders als im Falle des Vertrauensverhältnisses ist es also nicht zwingend, dass beide Relata des Verhältnisses Personen bzw. moralische Akteur*innen sind. In dem Fall, in dem eine Person sich etwa darauf verlässt, dass die Brücke, über die sie geht, nicht einstürzen wird, sind die Relata des Verhältnisses eine Person und ein Artefakt. Damit wird gleich ersichtlich, dass in diesem Fall nicht dieselben normativen Implikationen vorliegen – so wäre es nicht angemessen,

der Brücke Vorwürfe zu machen, sollte sie die eigene Erwartung nicht erfüllen. Der Punkt hier ist aber nicht, dass das Vertrauensverhältnis sich vom Verlässlichkeitsverhältnis primär durch die unterschiedlichen Relata unterscheidet; eine Verlässlichkeitsbeziehung kann auch zwischen zwei Personen bestehen. Der Unterschied zwischen beiden ist einer der Beziehungs*form* und der mit dieser verknüpften Normativität, nicht allein der Relata (Holton, 1994, S. 4).

Im Vertrauensverhältnis ist das normative Verhältnis zwischen den Relata eines, in dem reaktive Einstellungen und moralische Wertungen angemessen sind. Das ist im Fall des Verlässlichkeitsverhältnisses anders. Dort mag Überraschung oder sogar Wut als Reaktion nachvollziehbar sein, aber es ist nicht gerechtfertigt, wütend *auf* diejenige Entität zu sein, auf die man sich verlassen hat – unabhängig davon, ob die Entität eine Person, ein Artefakt oder ein Naturereignis ist. Zur Erinnerung: Reaktive Einstellungen sind wesentlich mit dem Gedanken verknüpft, dass sie sich auf Entitäten beziehen, bei denen die Erwartung sinnvoll ist, dass diese uns gegenüber einen guten Willen zum Ausdruck bringen. Wenn nun ein Subjekt gar nichts darüber weiß, dass ich ihm vertraue, sodass sich in dessen Handlungen weder ein guter noch ein schlechter Wille mir gegenüber zeigt oder wenn ich mich auf eine Entität verlasse, die nicht die Art von Entität ist, die überhaupt einen Willen haben kann, können natürlich emotionale Reaktionen auftreten, reaktive Einstellungen hingegen sind in solchen Fällen kategorisch unangemessen.

Ein zweiter Unterschied zwischen Vertrauensverhältnissen und Verlässlichkeitsrelationen besteht in der Frage, unter welchen Bedingungen die fraglichen Einstellungen gerechtfertigt sind. Wie oben erörtert, kann es Gründe geben, einer Person zu vertrauen, selbst wenn diese sich in der Vergangenheit als nicht-vertrauenswürdig herausgestellt hat. Das ist im Fall des Verlässlichkeitsverhältnisses anders. Ob wir uns darauf verlassen sollten, dass eine Entität so agieren wird, wie wir erwarten, hängt allein von epistemischen Gründen darüber ab, wie wahrscheinlich dieses Verhalten ist. Zu diesen Gründen zählen beispielsweise vergangene Verhaltensweisen oder Kenntnisse über die Struktur oder das Design der Entität, auf die wir uns verlassen. Sich mehr auf eine strukturell stabile und nach den besten ingenieurswissenschaftlichen Standards konstruierte Brücke als auf eine Brücke zu verlassen, die mit instabilen Materialien von Personen gebaut wurde, die weder Erfahrungen noch anderweitige

Kenntnisse mit dem Bau von Brücken haben, scheint eine gerechtfertigte Haltung zu sein.

Werden nun diese Erwägungen auf KI-Systeme angewandt, wird deutlich, dass das relevante Verhältnis, das häufig unter dem Vertrauensbegriff besprochen wird, eigentlich eine Art der Verlässlichkeitsrelation ist. Dabei ist es nicht wichtig, welches Wort wir im Einzelnen benutzen, das heißt, es spricht nichts dagegen, von vertrauenswürdiger KI zu sprechen, *insofern die dargelegten Unterschiede zwischen den beiden ausgeführten Konzepten* im Blick behalten werden. In Hinblick auf medizinische KI stellt sich damit die Frage, wie *verlässlich* das fragliche KI-System ist, das heißt, wie erfolgreich es etwa darin ist, auf radiologischen Bildern Erkrankungen zu klassifizieren. Eine solche KI, als Werkzeug verstanden, ist im Sinne der Verlässlichkeitsrelation genau dann »vertrauenswürdig«, wenn sie zuverlässig genau die Outputs produziert, zu dessen Zweck sie entwickelt wurde, wenn sie nicht fehleranfällig ist und wenn es Verfahren gibt, die die Verlässlichkeit der KI als Werkzeug überprüfbar machen. So bemerken etwa Juan Manuel Durán und Karin Rolanda Jongsma: »reliability of algorithms provides reasons for trusting the outcomes of medical artificial intelligence«. (Durán & Jongsma, 2021, S. 329). Gerade vor dem Hintergrund der Möglichkeit von »strange errors« ist es hierbei notwendig, dass KI in der Medizin grundsätzlich durch menschliches Personal auditiert wird.

Diese kritische Einschätzung wird in der Debatte allerdings nicht von allen Autor*innen geteilt. Die Gründe hierfür sind vielfältig. Zanotti et al. (2024) etwa betonen, dass die ethischen Dimensionen des Vertrauensbegriffs in Bezug auf KI über bloße Verlässlichkeit hinausgehen und Fragestellungen der Transparenz, der Autonomie der KI nutzenden Personen oder der Gerechtigkeit umfassen. In eine ähnliche Richtung argumentieren Stake et al. (2022). Während diese Einwände in der Sache richtig sind, erscheint es vor dem Hintergrund der oben ausgeführten Argumentation sowie der Gefahren, die sich mit der Anthropomorphisierung von KI verbinden, sinnvoller zu betonen, dass der Einsatz von KI mit Fragen der Vertrauenswürdigkeit und weiteren ethischen Erwägungen verknüpft ist, das relevante Vertrauensverhältnis aber nicht zwischen Personen und der KI, sondern etwa zwischen den Patient*innen und den Entwickler*innen sowie den die KI nutzenden Ärzt*innen besteht.

Während zu einer KI also im strengen Sinne kein Vertrauensverhältnis bestehen kann, heißt das nicht, dass im Einsatz von KI in der Medizin Vertrauen keine Rolle spielt. Tatsächlich ist Vertrauen und Vertrauenswürdigkeit ein zentrales Merkmal der ethischen Auseinandersetzung mit KI, allerdings geht es hier um ein Vertrauensverhältnis von Patient*innen zu den Ärzt*innen, die KI verwenden, sowie von Patient*innen und Ärzt*innen gegenüber den Personen und Firmen, die die KI entwickeln. Die Rechtfertigungsgrundlagen für diese Vertrauensverhältnisse sind vielfältig. Zwischen Ärzt*innen und Patient*innen besteht ein grundsätzliches und strukturelles Vertrauensverhältnis, das einerseits im ärztlichen Berufsethos gründet, andererseits aber auch durch strukturelle Maßnahmen, wie die der Berufsaufsicht durch die Ärztekammer, die Verpflichtung auf die Einholung informierter Einwilligungen vor dem Einsatz ärztlicher Maßnahmen sowie nicht zuletzt durch die tugendhafte Charakterdisposition individueller Ärzt*innen, gerechtfertigt ist. Dementsprechend ist das Vertrauensverhältnis zwischen Patient*in und Ärzt*in dann gestört, wenn Ärzt*innen ohne Aufsicht agieren, auf das Einholen informierter Einwilligung verzichten und wesentliche Charaktertugenden nicht besitzen.

Im Rahmen des Einsatzes von KI in der Medizin kommt eine weitere Dimension des Vertrauensverhältnisses hinzu: Transparenz bezüglich des Einsatzes von KI sowie der Möglichkeiten und Grenzen, die mit KI verbunden sind. Aufgrund der Neuartigkeit dieser Art von KI in der Medizin und den damit einhergehenden fehlenden Erfahrungswerten ist das Offenlegen dieser Information eine vertrauensbildende Maßnahme. Um die Vertrauenswürdigkeit von Personen und Firmen zu verbessern, die KI herstellen, sollten auch weiterhin unabhängige wissenschaftliche Studien zur Verlässlichkeit und zum ethischen Design der verwendeten KI-Produkte durchgeführt werden. Auch die Verteilung von Forschungsgeldern kann dazu beitragen, dass KI-Design und KI-Entwicklung in vertrauenswürdige Bahnen gelenkt werden (Gardner et al., 2022). Hierbei ist insbesondere darauf zu achten, dass die Entwicklung von KI nicht als wertfreier Prozess verstanden wird, sondern dass beachtet wird, in welchen Hinsichten welche Werte in KI-Designs eingebettet sind. Darüber hinaus kann die Vertrauenswürdigkeit von KI-Entwickelnden dadurch gefördert werden, dass Ärzt*innen bereits in den unterschiedlichen Stadien des Entwicklungsprozesses eingebun-

den werden (Winter & Carusi, 2023). Hierdurch kann auch das Vertrauensverhältnis zwischen Ärzt*innen und KI-Entwickler*innen gestärkt werden.

2.4 Transparenz

Dass Transparenz ein zentrales medizinethisches Prinzip darstellt, zeigt sich unter anderem daran, dass im Rahmen einer informierten Einwilligung bestimmte Informationen offengelegt werden müssen, die für die Entscheidungsfindung der Patient*innen von Bedeutung sind. Transparenz dient in diesem Zusammenhang nicht nur der Vertrauenswürdigkeit und Stärkung des Vertrauensverhältnisses zwischen Ärzt*in und Patient*in (O'Neill, 2002a, S. 134ff.), sondern auch der Achtung der Autonomie der Patient*innen. Gerade beim Einsatz einer neuen Technologie, deren Implikationen sich im Vorfeld nicht antizipieren lassen und bei der weder die Zuverlässigkeit noch die exakte Funktionsweise im Detail bekannt sind, gewinnt die Forderung nach Transparenz ein besonderes Gewicht (Walmsley, 2021, S. 589f.).

Im Fall von KI-Systemen muss die Forderung nach Transparenz allerdings mit Benefizienzerwägungen abgeglichen werden. So gilt es zu beachten, dass es beim momentanen Stand der technischen Machbarkeit dieser Art von Transparenz der KI zu Konflikten in Hinblick auf deren Verlässlichkeit kommen kann, sodass ein solches System unzuverlässiger wird, je transparenter es ist, wenngleich daran gearbeitet wird, für diese Problematik technische Lösungen zu entwickeln (London, 2019; Felzmann et al., 2020). Da eines der Hauptargumente für den Einsatz von KI ihre Verlässlichkeit ist und nur ein verlässliches Instrument, etwa zur Identifizierung von Erkrankungen, dem Wohl der Patient*innen zuträglich ist, ist es nicht selbstverständlich, dass die Transparenz von KI-Systemen unter diesen Bedingungen ethisch gefordert werden muss. Sollte sich ein KI-System als herausragend verlässlich darstellen, könnte schlüssig argumentiert werden, dass wir nicht verstehen müssen, *warum und wie* es funktioniert, wenn diese Erklärbarkeit die Verlässlichkeit des Systems einschränkt. Es ist mit anderen Worten denkbar, dass der Nutzen für Patient*innen in solchen Fällen die Forderung nach Transparenz aufwiegt. Nur unter der Bedingung, dass die Trans-

parenz des Systems nicht seine Verlässlichkeit untergräbt, gilt die ethische Verpflichtung, dieses System transparent zu machen, uneingeschränkt.

Ein weiterer Aspekt der internen Transparenz betrifft die Problematik, dass eine *vollständige Transparenz* bezüglich der Art und Weise, wie KI Ergebnisse produziert, nicht erreicht und damit auch nicht gefordert werden kann. Daraus könnte geschlossen werden, dass diese Art von Transparenz unter keinen Umständen Gegenstand ethischer Verpflichtungen sein kann. Dabei gilt es aber zu beachten, dass das Konzept der Transparenz Abstufungen zulässt. So bemerkt Nicholas Diakopolous etwa:

> »Algorithmic transparency cannot be understood as a simple dichotomy between a system being ›transparent‹ or ›not transparent.‹ Instead, there are many flavors and gradiations of transparency that are possible, which may be driven by particular ethical concerns that warrant monitoring of specific aspects of system behavior.« (Diakopolous, 2020, S. 199)

Selbst wenn wir also keine vollständige und detaillierte Erklärung für die Ergebnisse einer KI geben können, ist es doch möglich, zumindest bestimmte Aspekte der KI und ihrer Funktionsweise offenzulegen. Zu diesen Aspekten gehört die *Verlässlichkeit der von KI produzierten Ergebnisse* – etwa der Identifizierung von Erkrankungen –, die *Fehleranfälligkeit des Systems* im Allgemeinen sowie die Fehleranfälligkeit in Bezug auf bestimmte Personengruppen, die *Struktur und den Umfang der verwendeten Trainingsdaten* – insbesondere unter dem Gesichtspunkt möglicher *biases* – oder die Grundzüge des *KI-Designs*, sodass von neutraler Stelle geprüft werden sollte, ob das von der jeweiligen KI verwendete Verfahren zur Identifikation von Erkrankungen für diesen Zweck geeignet ist und den gängigen medizinischen Standards entspricht.

Ein weiterer ethisch relevanter Aspekt von Transparenz betrifft ihre relationale und kontextuelle Natur (Felzmann et al., 2019). Welche Art und Ausprägung von Transparenz ethisch gefordert werden sollte, hängt davon ab, wem gegenüber etwas transparent gemacht wird. Während beispielsweise Transparenz bezüglich der Verlässlichkeit des Systems für alle Nutzer*innen – Patient*innen wie Ärzt*innen – relevant ist, benötigen Patient*innen im Normalfall keine Informationen bezüglich grundlegender Designentscheidungen eines KI-Systems. Für Ärzt*innen hingegen, die die fraglichen Systeme

verwenden, kann es relevant sein, zu wissen, ob und inwiefern sich medizinische Fachkenntnisse im Design niedergeschlagen haben. Diese Information kann Teil der Entscheidung sein, ein System im medizinischen Alltag einzusetzen oder zwischen unterschiedlichen Systemen zu wählen.

Nicht zuletzt gilt es in diesem Zusammenhang zu beachten, dass die offengelegten Informationen auch *nutzbar* sind. Je nachdem, welches Hintergrundwissen eine Person besitzt, ist es notwendig, zu Zwecken der Transparenz Informationen unterschiedlich aufzuarbeiten und zu präsentieren. Für Patient*innen und Ärzt*innen ohne Informatikfachkenntnisse ist eine technisch anspruchsvolle Darstellung der Verarbeitungsschritte einer KI nicht sonderlich hilfreich. Aus diesem Grund muss die Forderung nach Transparenz relativ zu den sie nutzenden Personen verstanden werden. In diesem Zusammenhang wird in der Debatte zwischen zwei Formen von KI-Transparenz differenziert: *Interpretierbarkeit* (*interpretability*) und *Erklärbarkeit* (*explicability*) (Herzog, 2022). Wo Interpretierbarkeit anspruchsvolle technische Kenntnisse erfordert und kleinteilige kausale und funktionale Erklärungsformen meint, bezeichnet Erklärbarkeit in diesem Kontext eine epistemisch weniger anspruchsvolle Form. Wer in diesem Sinne erklären kann, was eine KI tut, weiß deshalb nicht im technischen Detail, wie sie funktioniert, sondern hat ein für die eigenen Entscheidungen relevantes Wissen.

2.5 Verantwortung

Verantwortung ist nicht allein ein Kernthema der Moralphilosophie, der Handlungstheorie und der Willensfreiheitsdebatte, sie ist in unterschiedlichen Hinsichten auch ein zentraler Aspekt moralisch rechtfertigungsfähiger medizinischer Praxis (Peteet, 2023). Dabei gilt nicht nur, dass Ärzt*innen für ihr Handeln verantwortlich und dazu verpflichtet sind, den Maßstäben ihrer Profession zu genügen, sondern auch, dass selbstbestimmte Patient*innen Verantwortung für ihre eigenen Entscheidungen übernehmen müssen. Der Einsatz von KI hat allerdings spezifische Fragestellungen aufgeworfen, die unter anderem mit den oben ausgeführten Transparenzerwägungen zusammenhängen (Coeckelbergh, 2002b). Diese Fragestellungen verweisen auf ein zentrales Element der Idee moralischer Verantwortlichkeit: Kontrolle über die eigenen Handlungen.

Im philosophischen Diskurs muss in diesem Zusammenhang zwischen unterschiedlichen Spielarten der Verantwortung unterschieden werden: kausal, rechtlich und moralisch. Kausal ist jemand oder etwas für ein Ereignis verantwortlich, wenn es durch eine Bewegung oder Aktivität des Subjekts oder Objekts hervorgebracht wurde. Für diese Art von Verantwortlichkeit ist es weder notwendig, dass das kausal verantwortliche Objekt handlungsfähig ist oder gar Absichten bilden kann. Dementsprechend ist kausale Verantwortlichkeit keine inhärent moralische Kategorie. Ein Stein beispielsweise, der sich löst und im Fallen ein Glas zerbricht, ist kausal für das Brechen des Glases verantwortlich, die Frage der moralischen Verantwortlichkeit stellt sich in diesem Kontext aber nicht. Rechtliche und moralische Verantwortlichkeit hingegen stellen normative Konzeptionen dar. Moralisch und rechtlich verantwortlich kann eine Entität nur sein, wenn sie handlungsfähig, vernunftbegabt und sich ihrer eigenen Handlungen und Überzeugungen bewusst ist sowie, mit Ausnahme von Fahrlässigkeitsdelikten, dass sie das Ereignis, für das sie verantwortlich ist, *absichtlich* hervorbringt (Talbert, 2016, S. 1ff.).

Eine problematische Idee, die einerseits mit dem *Black-Box*-Charakter von KI zusammenhängt und andererseits die Anthropomorphisierungstendenz der zeitgenössischen KI-Debatte betrifft, ist die, dass der Einsatz von KI zu sogenannten *Verantwortungslücken* führt (Matthias, 2004). Diese Verantwortungslücken entstehen, so der Gedanke, weil im Fall zeitgenössischer KI keine Person und keine Personengruppe die notwendige Art von Kontrolle besitzt, die für moralische Verantwortlichkeit grundlegend ist. Der *Black-Box*-Charakter von KI sowie ihr hohes Maß an Autonomie untergraben demnach die Kontrolle von Personen über das System. Dieser Gedanke weist zwei Fehler auf:

Erstens entstehen KI-Systeme nicht zufälligerweise, sondern werden *für bestimmte Zwecke* entworfen. Wie im Fall anderer technologischer Instrumente tragen die Entwickler*innen und Firmen eine Mitverantwortung dafür, dass diese Systeme für Zwecke geschaffen werden, die moralisch gerechtfertigt werden können und verlässlich dafür geeignet sind, besagte Zwecke zu erreichen. Weil KI keine vernunftbegabte Akteurin ist, die sich auf Basis praktischer Deliberation selbstständig Zwecke setzen und diese verfolgen kann, sondern in der Zwecksetzung von menschlichen Designentscheidungen

abhängt, liegt die moralische Verantwortung der KI-Aktivität bei denjenigen Personen, die sie entwickeln und einsetzen.

Zweitens ist der epistemische *Black-Box*-Charakter von Entscheidungen kein Merkmal, das allein KI-Systeme aufweisen, sondern auch eines, durch das Menschen gekennzeichnet sind (Schubbach, 2021; Brandt et al., 2025, S. 549). Wie etwa Suzanne Kawamleh richtig bemerkt, gilt auch für menschliche Ärzt*innen, dass die Gründe für bestimmte Klassifikationen oder Diagnosen oft epistemisch mehr oder weniger opak sind. In vielen Fällen können Ärzt*innen keine Erklärung für die Identifikation von Erkrankungen auf Bildern anbieten, die darüber hinausgeht, dass sie etwas Auffälliges entdeckt haben. Während eine solche Einschätzung aufgrund der Erfahrung und Expertise von Ärzt*innen als Begründung fungieren kann, unterschreitet sie epistemisch anspruchsvolle Begründungsmaßstäbe, die etwa das Offenlegen einer hinreichend vollständigen, kausalen Erklärung umfassen oder zumindest die Identifikation relevanter Regeln und Prinzipien, unter die ein bestimmter Fall subsumiert werden kann. In diesen Hinsichten ist die Entscheidungsfindung der Ärzt*innen undurchsichtig. Das gilt insbesondere für die ärztliche Praxis in der Radiologie, wie Kawamleh schreibt:

> »For example, doctors undergoing residency training in radiology cannot be taught a rule-based system by which to classify medical images. Rather, much like a learning algorithm, they are exposed to many examples and told where to look and what to look for.« (Kawamleh, 2023, S. 913)

Wird nun eine Analogie zu KI-Systemen gezogen, wird klar, dass auch die Ergebnisse einer KI als Rechtfertigungsgrundlage einer Diagnose dienen können, sofern sie verlässlich Ergebnisse produziert. Zwar verhindert der *Black-Box*-Charakter von KI eine lückenlose kausale Erklärung oder die Subsumtion eines Einzelfalls unter klar bestimmte Regeln und Prinzipien, aber genau in diesen Merkmalen besteht eine Parallele zwischen der epistemischen Opazität von Ärzt*innen und KI-Systemen.

Obwohl nun die Entscheidungsfindung menschlicher Ärzt*innen in ähnlicher Weise opak ist wie die Generierung von Outputs einer KI, folgt aus dieser Tatsache nicht, dass wir deshalb Ärzt*innen von der Verantwortung für ihre Entscheidungen freisprechen sollten. Von Ärzt*innen wird erwartet, dass sie einschätzen können, wie überzeugend die Grundlage ihrer Diagnosen ist. Dass sie Diagno-

sen unter Vorzeichen epistemischer Unsicherheit stellen müssen, untergräbt nicht ihre Verantwortung. Verantwortlichkeit setzt im Allgemeinen nicht voraus, dass Personen kausale Erklärungen anbieten können. Daraus folgt, dass die Undurchsichtigkeit der Entscheidungsfindung zumindest im medizinischen Kontext mit Verantwortlichkeit kompatibel ist, unabhängig davon, ob diese Opazität im *Black-Box*-Charakter der für die Diagnose verwendeten KI oder der Entscheidungsfindung der Ärzt*innen selbst verortet werden kann.

Die Debatte zu Verantwortungslücken sollte daher weder dazu verleiten, Personen und Firmen von moralischer Verantwortlichkeit für die Funktionsweise der von ihnen entwickelten oder genutzten KI freizusprechen, noch sollte sie zum Anlass genommen werden, die moralische Verantwortlichkeit der KI-Systeme selbst zu prüfen. Solche Debatten mögen aus theoretischer Perspektive interessante Fragestellungen aufwerfen, sollten jedoch nicht von der Tatsache ablenken, dass KI ein Instrument ist, das besser oder schlechter designt sein und mehr oder weniger verantwortungsvoll genutzt werden kann. In jedem Fall liegt die moralische Verantwortlichkeit des Einsatzes von KI-Systemen bei Personen (Tigard, 2021).

2.6 Achtung der Autonomie

2.6.1 Konzeptionelle Grundlagen des Autonomiebegriffs

Autonomie bezeichnet grundsätzlich eine Form von Freiheit, bei der eine freie Entität *selbstbestimmt* agiert – das heißt, ihre Handlung wird nicht extern bestimmt, sondern von ihr selbst festgelegt. Im philosophischen Diskurs haben sich unterschiedliche Vorstellungen von Autonomie entwickelt, so etwa internalistische und externalistische, individualistische und relationale, deskriptive und normative oder formale und substanzielle (Christman, 2014; Frankfurt, 2009; O'Neill, 2003; Oshana, 2015; Dworkin, 1998). Diese Aufzählung macht ersichtlich, dass Debatten um Autonomie nur dann erfolgreich geführt werden können, wenn zunächst geklärt wird, was genau unter dem Autonomiebegriff im Einzelnen verstanden wird, da die genannten Interpretationen von Autonomie nicht nur konzeptionell unterschiedlich sind, sondern auch verschiedene normative Implikationen aufweisen.

Autonomie wird darüber hinaus oft konzeptionell von bloß negativer Freiheit, also der *Freiheit von* bestimmten Determinationsfaktoren, unterschieden, wenngleich diese Art der Freiheit für gewöhnlich eine notwendige Bedingung von Autonomie ist.[8] In vielen Fällen ist eine Einschränkung negativer Freiheit dementsprechend auch eine Einschränkung von Autonomie. Wird etwa eine Person durch Manipulation oder Nötigung daran gehindert, selbstgesetzte Zwecke zu realisieren, so ist ihre negative Freiheit und zugleich ihre Autonomie dadurch eingeschränkt: Sie ist nicht frei *von* Einflussnahmen auf ihr Handeln und diese Unfreiheit macht es ihr unmöglich, sich selbst und ihre Handlungen nach ihren eigenen Vorgaben zu bestimmen. Während negative Freiheit für gewöhnlich begrifflich von Autonomie unterschieden wird, gelten zumindest bestimmte Formen der positiven Freiheit, also die Freiheit *zu* bestimmten Vollzügen, insbesondere in der kantischen Tradition, als Spielarten der Autonomie (Menke, 2018, S. 52f.), wenngleich die Begriffe der positiven Freiheit und der Autonomie nicht deckungsgleich sind. Wo also in der begrifflichen Analyse eine Differenzierung von negativer und positiver Freiheit sowie Autonomie gemacht werden kann, sind diese Begriffe und die unter sie fallenden Phänomene in vielen Fällen miteinander verknüpft (MacCallum, 1967). Es wird zu prüfen sein, inwiefern die Einschränkungen negativer und positiver Freiheit in moralisch problematischer Weise die Autonomie der in der Medizin tätigen Akteur*innen untergräbt.

Der letzte Teilsatz weist darauf hin, dass nicht jede Autonomieeinschränkung notwendigerweise auch moralisch problematisch ist, weil nicht jede Art von Autonomie gleichermaßen moralisch beachtenswert ist. So betont etwa Onora O'Neill:

> »Most contemporary accounts of autonomy see it as a form of independence. (…) Some independent action is spontaneous, disciplined, altruistic and even heroic; some is self-centred, pig-headed, impulsive,

8 Bisweilen wird bestritten, dass negative Freiheit für Autonomie notwendig ist. So vertritt Harry Frankfurt (1969) eine Form des Kompatibilismus, demzufolge Autonomie und kausale Determination einander nicht ausschließen. Ähnliche Überlegungen finden sich bei Gerald Dworkin (1998). Während in Einzelfällen Autonomie und negative Freiheit auseinanderfallen können, gilt im Normalfall aber, dass die Einschränkung von negativer Freiheit auch eine Einschränkung von Autonomie impliziert.

> random, ignorant, out of control and regrettable or unacceptable for these and many other reasons.« (O'Neill, 2002a, S. 28)

Damit Achtung vor Autonomie in einem moralisch relevanten Sinn das Handeln leiten kann, darf die Form der Autonomie, der Achtung entgegengebracht wird, ihrerseits zumindest nicht gegen moralische Normen verstoßen. Wenn sich etwa die Selbstbestimmung einer sadistischen Person darin ausdrückt, Menschen zu foltern, ist es offensichtlich nicht so, dass wir diese Art der Autonomie aus moralischen Gründen achten sollten. Aus diesem Grund ist es notwendig, nicht nur zu prüfen, was – vor dem Hintergrund der erwähnten Vielfalt der Deutungen – genau unter Autonomie verstanden wird, sondern auch zu erwägen, inwiefern das fragliche Autonomieverständnis moralisches Gewicht hat.

Die Vielfalt von Autonomieverständnissen ist dementsprechend nicht bloß relevant für abstrakte konzeptionelle Debatten in der Philosophie. Da Achtung der Autonomie ein zentrales medizinethisches Prinzip ist, ist die Frage, welche Form von Autonomie wir auf diese Weise schützen, unmittelbar von ethischer Bedeutung. Die Problematik, dass sich konzeptionelle Festlegungen auf ethische Problemstellungen auswirken und daher eine hinreichende Klärung des in den jeweiligen Debatten verwendeten Autonomiebegriffs notwendig ist, ist durch zeitgenössische Formen von KI verschärft worden. In den klassischen medizinethischen und konzeptionellen Debatten der Philosophie um den Autonomiebegriff stand die Autonomie der Person in ihren unterschiedlichen Ausprägungen im Mittelpunkt der Debatte. Mit zeitgenössischen Formen von KI wird nun auch die Autonomie künstlicher Systeme diskutiert. Im Fall selbstlernender KI-Algorithmen liegt es nahe, Autonomie als Freiheit bzw. Unabhängigkeit von menschlichen Inputs zu verstehen. Im Kontext von KI-Systemen wird dieses Verständnis von Autonomie auch als minimale funktionale Autonomie bezeichnet (Laitinen, 2010). Problematisch wird eine solche Verwendungsweise des Autonomiebegriffs aber, wenn die Unterschiede von menschlicher Autonomie und der Autonomie, die KI realisieren kann, dadurch verwischt werden. Der Vorschlag hier ist nicht, dass wir den Begriff der Autonomie im Sinne der Unabhängigkeit von menschlichen Inputs nicht auf KI-Systeme anwenden sollten, sondern dass wir klären müssen, was mit diesem Begriff in verschiedenen Kontexten gesagt wird und was, wenn über-

haupt etwas, ethisch aus der fraglichen Festlegung folgt (Heinrichs & Wagner, 2024).

Um folglich aus ethischer Perspektive zu prüfen, wie das Prinzip der Achtung der Autonomie im Zusammenhang mit dem Einsatz von KI in der Radiologie ethischer Handlungsleitung zugrunde liegen kann, muss zunächst festgelegt werden, welche Form von Autonomie einschlägig ist, und dann, wessen Autonomie respektiert werden sollte. Dabei wird die Autonomie der KI außen vor gelassen, da diese nicht unter das Verständnis der Selbstbestimmung fällt, das im ethischen Prinzip der Achtung der Autonomie zum Ausdruck gebracht wird. KI-Systeme mögen autonom im Sinne der Unabhängigkeit von menschlichen Inputs sein, daraus folgt aber nicht, dass wir diese Form der Autonomie deshalb moralisch achten sollten. Die Autonomie des ärztlichen Fachpersonals und die der Patient*innen auf der anderen Seite kann nicht durch solche grundsätzlichen Erwägungen als moralisch irrelevant gekennzeichnet werden. Offensichtlich sind bestimmte Formen von Selbstbestimmung, die Personen im medizinischen Kontext realisieren, von moralischer Bedeutung. Weniger offensichtlich ist die Antwort auf die Frage, welche Art oder Arten von Autonomie das ärztliche Personal und die Patient*innen realisieren und wie diese Formen von Selbstbestimmung moralisch gewichtet werden sollten. Um dies zu beantworten, muss zunächst geprüft werden, in welchen Hinsichten die Selbstbestimmung dieser Personengruppen durch den Einsatz von KI in der Medizin beeinflusst wird.

*2.6.2 Autonomie der Ärzt*innen*

Im Fall des medizinischen Fachpersonals ist die Form der Autonomie, die für gewöhnlich im Vordergrund steht, die der selbstbestimmten Entscheidungsbefugnis im medizinischen Alltag. In diesem Zusammenhang werden in der Regel drei Bedenken aufgeführt: Erstens besteht die Sorge, dass durch den Einsatz von KI in der medizinischen Praxis der Arbeitsalltag so stark automatisiert wird, dass Ärzt*innen sich nach den von der KI erarbeiteten Vorgaben richten müssen, wodurch zusätzlich zu den bereits bestehenden Regularien weitere externe Einflüsse auf die ärztliche Selbstbestimmung einwirken. Dadurch ist zu befürchten, dass der *workflow* der medizinischen Praxis in problematischer Weise durch KI-Systeme beeinflusst

wird (Lombi & Rossero, 2024). Zweitens wird kritisch bemerkt, dass KI-Systeme, die zur Diagnostik eingesetzt werden, Diagnosen selbstständig stellen und den Ärzt*innen damit vorgeben, wie sie ihre Patient*innen zu behandeln haben (Bergquist & Rolandsson, 2022). Darüber hinaus wird drittens die Befürchtung formuliert, dass Ärzt*innen bestimmte Fähigkeiten durch den Einsatz von KI verlieren, weil die mit diesen Fähigkeiten verbundenen Aufgaben von der KI übernommen werden können. Die letztgenannte Problematik wird unter dem Begriff des *de-skilling* verhandelt (Funer & Wiesing, 2024).

Im Fall der ersten beiden dieser Sorgen ist die relevante Form der Autonomie jene, die mit dem Begriff der Kontrolle über die eigenen Entscheidungen zusammenhängt. Der Grundgedanke ist, dass Ärzt*innen in diesem Sinne dann autonom agieren, wenn sie *frei* oder *unabhängig* von kontrollierenden Einflüssen agieren. Die hier zum Ausdruck gebrachte Form von Autonomie kennzeichnet *negative* Freiheit (Berlin, 2002). Negative Freiheit ist die Freiheit *von* unzulässigen Einflüssen. Während vereinzelt in der politischen Philosophie, insbesondere aus libertärer Perspektive heraus, negativer Freiheit ein hoher Wert zugesprochen wird (Nozick, 2013), ist augenscheinlich, dass diese allgemeine Bestimmung beschränkt werden muss. Gerade in kooperativen sozialen Praktiken wie der Arbeit in einem Krankenhaus oder einer medizinischen Praxis, sind externe Sachzwänge, Regularien und Formen der Automatisierung weder neu noch ethisch problematisch, sofern sie der effizienten Versorgung der Patient*innen dienen, ohne dabei die Ärzt*innen zu überfordern. Selbstverständlich gilt, dass etablierte *workflows* sich durch den Einsatz neuer Technologien wandeln können, allerdings sind Änderungen dieser Art an sich nicht moralisch beanstandenswert. Das bedeutet nicht, dass die Automatisierung von *workflows* durch KI aus moralischer Sicht völlig unproblematisch wäre. KI-Systeme sind wie alle technischen Mittel niemals fehlerfrei, sodass menschliche Aufsicht und Urteilsfähigkeit weiterhin fundamental für die Strukturierung von Arbeitsabläufen bleiben sollten.

Auch die zweite der genannten Sorgen, also die, dass die ärztliche Selbstbestimmung in der medizinischen Entscheidungsfindung – so etwa im Fall von Diagnosen und darauf basierenden Behandlungen – durch den Einsatz von KI unterminiert werden könnte, betrifft ein negatives Freiheitsverständnis. Die Sorge ist somit auch hier, dass die

KI als kontrollierendes Element auftreten könnte, das Ärzt*innen in ihrer negativen Freiheit einschränkt. Diese Sorge basiert allerdings auf dem Gedanken, dass KI-Systeme in einem analogen Sinn zu Personen Entscheidungen treffen und Handlungen ausführen. Gerade jedoch die im medizinischen Kontext verwendete KI, etwa jene, die zur Klassifizierung von radiologischem Bildmaterial eingesetzt wird, kann schwerlich als Akteurin in einem robusten Sinne gelten. Solche KI-Systeme sind Werkzeuge, die zwar in einem hohen Maße unabhängig ihre Funktion verrichten, zugleich kann aber schwerlich behauptet werden, die fraglichen Systeme verfügten über ärztliches Urteilsvermögen. Dieses Urteilsvermögen ist es aber, was der ärztlichen Entscheidungsfindung zugrunde liegt, und es umfasst mehr, als bloß auffällige Muster auf Bildmaterial zu erkennen. Ärzt*innen sind in der Lage, zwischen unterschiedlichen Wissensbereichen Transferleistungen zu erbringen und sie verfügen über ein umfangreiches Netz an Hintergrundinformationen. Außerdem sind sie in der Lage, auf Basis praktischer und theoretischer Vernunftvermögen komplexe Schlüsse zu ziehen, ihre Gedanken mit anderen Expert*innen zu teilen und dadurch die eigenen Überzeugungen kritisch zu prüfen.

KI-Systeme, die Ärzt*innen in der Entscheidungsfindung unterstützen, sollten daher nicht, wie vereinzelt in der Literatur vertreten wird (Nyholm, 2018), als Kooperationspartner verstanden werden, die selbstständig Entscheidungen treffen, sondern als *Instrumente*. Kooperationspartner*innen sind für gewöhnlich Subjekte, die dazu in der Lage sind, selbstgewählte Zwecke zu verfolgen, normative Gründe einzusehen, zu prüfen und das eigene Handeln diesen entsprechend auszurichten sowie auf Basis vernünftiger Reflexion Gründe hervorbringen, die relevant für das Handeln der Kooperationspartner*innen sind. Während Nyholm diese Implikation zu vermeiden sucht, suggeriert die Verwendung des Kooperationsbegriffs im Fall von KI-Systemen dennoch, dass sie die genannten Fähigkeiten haben könnten und trägt damit zur Anthropomorphisierung von KI bei. Darüber hinaus suggeriert der Kooperationsbegriff eine Art von Symmetrie, in der beide Kooperationspartner*innen gemeinsam Entscheidungen treffen, sodass die Entscheidung von Kooperationspartner*in A die Entscheidung von Kooperationspartner*in B trumpfen könnte. In solchen Fällen könnte argumentiert werden, dass die Autonomie von B durch A untergraben werden würde.

Werden KI-Systeme hingegen als Instrumente und nicht als handlungs- und entscheidungsfähige Subjekte aufgefasst, wirkt der Kooperationsbegriff unpassend. Instrumente sind Werkzeuge, deren Zwecke vollständig durch diejenigen Personen festgelegt sind, die sie nutzen und die deshalb den Entscheidungsfindungsprozess von Personen unterstützen, nicht aber an diesem als Subjekte teilnehmen. Ein KI-System, das Muster auf radiologischem Bildmaterial offenlegt, kann in dieser Hinsicht als eine Art von Mikroskop verstanden werden, durch das im besten Fall Personen in die Lage versetzt werden, Dinge zu erkennen, die sie sonst nicht erkennen würden. Ebenso wenig aber, wie das Mikroskop etwas sieht, kann eine KI Entscheidungen treffen. Aus diesem Grund ist es auch nicht überzeugend zu argumentieren, ein KI-System untergrabe die Autonomie der Ärzt*innen. Dies wäre dann der Fall, wenn KI eine Art gleichberechtigte Diskurspartnerin wäre, die Entscheidungen trifft, welche in einem Konkurrenzverhältnis zu den Entscheidungen der Ärzt*innen stehen. Wenn eine ärztliche Fachkraft mit bloßem Auge eine Hauterkrankung nicht erkennt, welche sie durch technische Hilfsmittel identifizieren kann, hat das Hilfsmittel nicht ihre Entscheidungsfindung durch eine eigene Entscheidung untergraben, sondern Informationen offengelegt, die die ärztliche Entscheidungsfindung unterstützen. Dass zeitgenössische KI-Systeme die fraglichen Informationen sprachförmig darstellen können, bedeutet nicht, dass sie auf Gründen fußende Entscheidungen getroffen haben. Da zumindest die für die Radiologie relevanten KI-Systeme nicht als entscheidungsbegabte Wesen im Sinne von Personen verstanden werden können, besteht somit auch keine Gefahr, dass KI-Systeme die Autonomie der Ärzt*innen untergraben könnten.

Das heißt wiederum nicht, dass in diesem Zusammenhang keinerlei moralische Bedenken geäußert werden können. So besteht selbstverständlich die Möglichkeit, dass Ärzt*innen entweder selbst die Verlässlichkeit der Ergebnisse ihrer technologischen Hilfsmittel überschätzen. Diese unter dem Begriff »automation bias« (Abdelwanis et al., 2024) diskutierte Problematik stellt eine ernstzunehmende Gefahr für das Wohlergehen der Patient*innen dar. Eine weitere, mit dieser Problematik verbundene Gefahr erwächst aus der bereits besprochenen Möglichkeit von »strange errors«. Da grundsätzlich nicht ausgeschlossen werden kann, dass ein KI-System solche Fehler produziert und weil diese Fehler weder antizipiert noch nach ihrem

Auftreten leicht erkannt werden können, gilt es, eine umfangreiche Automatisierung sensibler medizinischer Prozesse zu vermeiden. Hierzu muss aufseiten des ärztlichen Personals ein Bewusstsein für die Grenzen von KI-Systemen geschaffen werden.

Das dritte der oben erwähnten Probleme ist die Gefahr des *de-skillings*. *De-skilling* bezeichnet allgemein das Phänomen, dass mit dem Auftreten neuer Technologien mittelfristig oder langfristig menschliche Fähigkeiten dadurch verlorengehen, dass die technischen Hilfsmittel effizienter oder in anderen Hinsichten geeigneter für das Erreichen derjenigen Zwecke sind, die zuvor durch rein menschliche Vermögen realisiert wurden. Mit dem Einsatz technischer Hilfsmittel, insofern diese verlässlich und verfügbar sind, geht einher, dass die fraglichen Vermögen nicht mehr von Personen ein- und ausgeübt werden und somit sukzessive verschwinden. Dieser Problemkontext ist weder in der Medizin noch in anderen Gebieten neu. Technologische Innovationen haben in der Geschichte oft dazu geführt, dass vormals von Menschen ausgeführte Praktiken und die mit ihnen verbundenen Vermögen ersetzt wurden (Wood, 2024). Werden KI-Entscheidungssysteme in der radiologischen Praxis verwendet, besteht die Gefahr, dass Ärzt*innen bestimmte Vermögen – wie das Erkennen von Krankheiten auf radiologischem Bildmaterial – im Laufe der Zeit verlieren, weil sie diese Fähigkeiten nicht weiter einüben oder den nachkommenden Generationen weitergeben.

De-skilling ist deshalb eine mögliche Gefahr für die Autonomie der Ärzt*innen, weil durch das Verlorengehen von Fähigkeiten möglicherweise Formen der *positiven Freiheit* untergraben werden könnten. Im Rahmen der *De-skilling*-Problematik ist somit das relevante Autonomieverständnis nicht wie in den beiden anderen bereits besprochenen Fragestellungen bloß die Unabhängigkeit von kontrollierenden Einflüssen – wenngleich die Abhängigkeit von KI-Systemen durch das *de-skilling* diese negative Freiheitsform ebenfalls betreffen –, sondern die positive Freiheit, bestimmte Handlungsformen ausführen zu können. Der Begriff der positiven Freiheit bezieht sich hier also auf das Vermögen oder die Fähigkeit *zu* bestimmten Handlungstypen, so etwa dem Handlungstyp der Identifizierung von Erkrankungen auf radiologischem Bildmaterial. Die hier relevanten Formen positiver Freiheit sind von offenkundiger moralischer Relevanz, da die Identifizierung von Erkrankungen durch Ärzt*innen der Gesundheit der Patient*innen dient. Fraglich ist allerdings, ob das

mit dem Einsatz von KI einhergehende *de-skilling* tatsächlich die positive Freiheit der Ärzt*innen untergräbt. Wenn davon ausgegangen wird, dass die in der Medizin eingesetzten KI-Systeme als *Artefakte* bzw. als *Instrumente* verstanden werden, wird sogleich ersichtlich, dass die Autonomie der Ärzt*innen durch ihren Einsatz nicht nur nicht untergraben, sondern im Zweifel sogar erweitert wird. Durch technologische Entwicklungen sind Menschen in die Lage versetzt worden, Handlungstypen auszuführen, zu denen sie ohne technische Hilfsmittel nicht befähigt wären. Wir können uns in Fahrzeugen in großen Geschwindigkeiten fortbewegen, wir können mithilfe von Flugzeugen fliegen und wir können mithilfe von Mikroskopen Dinge sehen, die mit bloßem Auge nicht sichtbar wären.

In diesem Zusammenhang werden die Begriffe des *up-skilling* bzw. *re-skilling* diskutiert (Crowston & Bolici, 2025). Wenn neue technische Hilfsmittel entwickelt und eingesetzt werden, müssen Personen, die diese Technologien nutzen, lernen, sie zu verwenden. *Up-skilling* bzw. *re-skilling* bezeichnen diesen Wandel der Anforderungen, mit dem sich Personen konfrontiert sehen. Selbst dann also, wenn die Handlungstypen durch den Wandel der technischen Innovationen identisch bleiben, ändern sich gleichwohl die Anforderungen an die Handelnden. Damit beispielsweise ein Mikroskop die ärztlichen Vermögen dadurch verbessern kann, dass Ärzt*innen mithilfe dieses Instruments Dinge erkennen können, die sie mit bloßem Auge nicht sähen, müssen sie lernen, es zu bedienen. Diese Änderung der Anforderungen schlägt sich für gewöhnlich in den jeweiligen Praxistypen darin nieder, dass Lernprozesse angepasst werden. Wenn also KI-Systeme Eingang in die ärztliche Praxis finden und wenn sie sich in bestimmten Bereichen wie der Identifizierung von Krankheiten auf radiologischem Bildmaterial als so verlässlich herausstellen, dass sie diesen Aufgabenbereich übernehmen können, sollte in der ärztlichen Ausbildung eine Schwerpunktverlagerung stattfinden. Statt wie zuvor zu lernen, Bildmaterial auszuwerten, müssten Ärzt*innen nun lernen, KI-Systeme zu bedienen und die von der KI produzierten Outputs zu evaluieren (Natali et al., 2025).

*2.6.3 Autonomie der Patient*innen und informierte Einwilligung*

Die Autonomie der Patient*innen ist in medizinethischen Debatten unmittelbar mit der Pflicht verknüpft, vor etwaigen medizinischen

Eingriffen eine informierte Einwilligung einzuholen. Die informierte Einwilligung dient folglich als Mittel, durch das die Selbstbestimmung der Patient*innen berücksichtigt wird. Während die oben besprochenen Formen der negativen und positiven Freiheit von Ärzt*innen zwar nicht moralisch irrelevant sind, ihre Einschränkung sich aber nicht als Einschränkung einer moralisch schwerwiegenden Form der Autonomie darstellt, verhält sich die Situation im Fall der Patient*innen anders. Hier geht es um die Selbstbestimmung, die den eigenen Körper betrifft, und damit um ein moralisch hochrangiges Gut. Unabhängig davon, wie diese moralische Relevanz der Entscheidung über den eigenen Körper verstanden wird – etwa durch die libertäre Idee des absoluten Eigentums am eigenen Körper oder durch das kantische Instrumentalisierungsverbot –, herrscht im ethischen Diskurs eine weitestgehende Einigkeit darüber, dass Entscheidungen über Eingriffe in den eigenen Körper moralisch relevant sind. Die Frage, die sich in diesem Zusammenhang stellt, ist die, welche Form der Autonomie in diesem Kontext einschlägig ist. Zwar spielen Formen der Unabhängigkeit und der positiven Freiheit bei der Bestimmung der medizinethisch relevanten Form von Autonomie eine zentrale Rolle, Selbstbestimmung geht in diesen Debatten aber über diese Formen der Freiheit hinaus.

Für die Zwecke dieses Beitrags ist es vor dem Hintergrund der erwähnten Pluralität von Autonomieverständnissen sinnvoll, von einem vergleichsweise intuitiven und minimalen Autonomiebegriff auszugehen. Dieses minimale Autonomieverständnis wurde von Ruth Faden, Tom Beauchamp und James Childress für die Zwecke der Medizinethik entwickelt (Faden & Beauchamp, 1986, S. 235ff.; Beauchamp & Childress, 2024, S. 169ff.). Nach diesem Verständnis sind autonome Handlungen solche, die drei Kriterien erfüllen: Autonome Handlungen werden (1) mit der *richtigen Absicht* ausgeführt, sie (2) werden von der handelnden Person *hinreichend verstanden* und sie sind (3) *frei von kontrollierenden Einflüssen*. Anders als hierarchische Ansätze fordert dieses minimalistische Verständnis keine Reflexionsbewegung im Sinne eines Abgleichs von Wünschen erster Stufe gegenüber höherstufigen Wünschen. Anders als normative Konzeptionen der Autonomie in der Tradition Kants muss im minimalen Verständnis Selbstbestimmung keinen konstitutiven Normen der praktischen Vernunft genügen und anders als bei relationalen Autonomiekonzeptionen spielen strukturelle Machtverhältnisse und

soziale Determinationsfaktoren in der Frage, ob und inwiefern eine Handlung autonom war, keine Rolle. Während jede einzelne dieser Auslassungen kritisch reflektiert werden sollte, gilt zugleich, dass für die Zwecke des medizinischen Alltags und für die Bestimmung von autonomen Handlungen in medizinethisch relevanten Situationen die drei genannten Bedingungen konzeptionelle und moralische Orientierungspunkte bieten, die auch von Vertreter*innen anderer Autonomiekonzeptionen akzeptiert werden können.

Im Rahmen des Einsatzes von KI in der Medizin ist das Verstehenskriterium einschlägig. Deswegen muss kurz erörtert werden, wie dieses Kriterium konkretisiert werden kann. Eine informierte Einwilligung im medizinethisch relevanten Sinn setzt nicht voraus, dass Patient*innen in allen wissenschaftlichen oder technischen Details verstehen, wie die Instrumente funktionieren, die Ärzt*innen zur Diagnosestellung einsetzen. In diesem Zusammenhang lassen sich drei Maßstäbe des Verstehens unterscheiden: Der Expertenmaßstab, der Maßstab der vernünftigen Person sowie der subjektive Maßstab (Faden & Beauchamp, 1986, S. 305ff.). Von Patient*innen kann kein Fachwissen verlangt werden, sodass der Expertenmaßstab nicht relevant ist, um zu prüfen, ob ein*e Patient*in das für eine informierte Einwilligung relevante Wissen besitzt. Stattdessen ist es sinnvoll, eine Verknüpfung des Maßstabs der vernünftigen Person und des subjektiven Maßstabs anzustreben. Der Maßstab der vernünftigen Person setzt voraus, dass eine Person versteht, was jede vernünftige Person in ihrer Situation verstehen sollte, um eine begründete Entscheidung zu treffen. Im medizinischen Kontext sollte eine solche Person etwa verstehen, welche Konsequenzen ein Eingriff haben kann, wie ein*e Ärzt*in zu ihrer Diagnose gefunden hat und wie wahrscheinlich es ist, dass diese Diagnose korrekt ist oder welche alternativen Behandlungsmöglichkeiten es gibt. Zusätzlich hierzu sollte – das ist der subjektive Maßstab – die Person verstehen, was für sie in ihrer Individualität, ihren spezifischen Wertvorstellungen sowie ihrer praktischen Identität ausschlaggebend für eine Entscheidung sein könnte. Wer etwa aus religiösen Gründen Bluttransfusionen ablehnt, sollte wissen, dass ein etwaiger Eingriff eine solche Transfusion beinhaltet.

Die Nutzung von KI in der Medizin wirft nun aufgrund des *Black-Box*-Charakters mögliche Probleme für das Verstehenskriterium einer autonomen informierten Einwilligung auf. Die Sorge

besteht darin, dass, die Autonomie der Patient*innen durch den Einsatz von KI gefährdet ist, da weder die Entwickler*innen der KI noch die Ärzt*innen, die diese KI einsetzen und damit auch nicht die Patient*innen verstehen können, wie und warum eine KI zu einem bestimmten Ergebnis, etwa einer Diagnose gelangt ist. So erklärt etwa Jose Luis Guerrero Quiñones:

> »One of the main problems arising from its implementation in healthcare is the lack of transparency of machine learning (ML) algorithms, which is thought to impede the patient's autonomous choice regarding their medical decisions. If the patient is unable to clearly understand why and how an AI algorithm reached certain medical decisions, their autonomy is being hovered.« (Quiñones, 2025, S. 1917)

Vereinzelt wird argumentiert, dass der *Black-Box*-Charakter das mögliche Verständnis von Personen so stark einschränkt, dass KI im medizinischen Kontext deshalb nur in wenigen Ausnahmefällen eingesetzt werden sollte (Chan, 2023). Eine solche Position erscheint in ihrer Radikalität aber wenig überzeugend, gerade vor dem Hintergrund, dass die epistemischen Maßstäbe, die wir in der medizinischen Praxis im Kontext der Rechtfertigung von Eingriffen für gewöhnlich ansetzen, keine vollständige Transparenz der Entscheidungsfindung voraussetzen. In vielen Fällen sind Verlässlichkeitserwägungen und Heuristiken alles, was uns zur Verfügung steht. Zugleich darf daraus nicht geschlossen werden, dass das Verstehenskriterium keine Rolle spielt und Patient*innen in keiner Weise über den Einsatz und die Funktionsweise von KI aufgeklärt werden sollten. Im Gegenteil gilt es zu prüfen, welche Art der Information in diesem Zusammenhang für die Urteilsfindung der Patient*innen ausschlaggebend ist und daher als Grundlage der autonomen Entscheidung dienen kann.

Die Informationen, die Patient*innen im Zusammenhang der informierten Einwilligung offengelegt werden könnten, umfassen zumindest die folgenden Fragen: Wie genau funktioniert die KI? Anhand welcher Daten wurde die KI trainiert? Wie verlässlich war die KI bisher? Hat eine menschliche Fachkraft die Ergebnisse der KI geprüft? Welche Fehler sind der KI bisher unterlaufen, warum und wie oft sind sie aufgetreten? Wenn Fehler aufgetreten sind, wie wurde auf diese Fehler reagiert und wie wahrscheinlich ist es, dass Fehler dieser Art sich wiederholen? Von Patient*innen kann und sollte nicht verlangt werden, dass sie die technischen Details zeitge-

nössischer KI verstehen. Dies würde den Expertenmaßstab an das Verstehen anlegen und damit über die Kriterien der moralisch geforderten Form informierter Einwilligung hinausgehen. Die Auswahl der Trainingsdaten kann Gegenstand der informierten Einwilligung sein. Je nachdem, zu welcher Personengruppe ein*e Patient*in gehört, sind die oben ausgeführten Fragestellungen der Gerechtigkeit hier in Einzelfällen einschlägig. Zugleich sollte diese Information angemessen aufbereitet und vermittelt werden, sodass ersichtlich wird, ob und inwiefern die Wahl der Input-Daten, mit denen die KI trainiert wurde, tatsächlich diskriminierende Implikationen besitzt. In jedem Fall offengelegt werden muss die Verlässlichkeit der eingesetzten KI. Nur dann, wenn das im medizinischen Alltag eingesetzte Instrumentarium verlässlich ist, kann das Vertrauensverhältnis zu den diese Instrumente nutzenden Ärzt*innen gerechtfertigt werden. Insbesondere in Fällen, bei denen ernsthafte medizinische Eingriffe auf durch KI erarbeiteten Diagnosen fußen, ist es weiterhin sinnvoll, offenzulegen, ob und inwiefern menschliche Fachkräfte die Ergebnisse der KI geprüft haben. Insbesondere vor dem Hintergrund der Möglichkeit von »strange errors« gehört zur Entscheidungsgrundlage vernunftgeleiteter Patient*innen, dass das Fehlerrisiko dadurch verringert wird, dass ärztliche Fachkräfte sich nicht allein auf die automatisierten Prozesse von KI-Systemen verlassen. Auch die Transparenz bezüglich der Fehler, die einer KI bislang unterlaufen sind, gehört zu den Verstehenskriterien, welche vernünftige Personen anlegen sollten, wenn sie eine Entscheidung bezüglich eines medizinischen Eingriffs treffen. Da dieser Aspekt mit der Fragestellung der Verlässlichkeit zusammenhängt, kann er unter diese subsumiert werden. Dabei ist es allerdings nicht notwendig, alle technischen Details offenzulegen oder darzulegen, warum die Fehler passiert sind und wie genau auf diese Fehler reagiert wurde. Hierzu würde erneut Fachwissen aufseiten der Patient*innen vorliegen müssen, das ethisch nicht gefordert werden kann.

Literaturverzeichnis

Ajanki, A. (2021, 17. Juli). Deep learning sometimes makes strange mistakes. *Medium*. https://medium.com/@anttiajanki/deep-learning-sometimes-makes-strange-mistakes-e026d96d00c2

Abdelwanis, M., Alarafati, H. K., Tammam, M. M. S., Simsekler, M. C. E. (2024). Exploring the risks of automation bias in healthcare artificial intelligence applications: A bowtie analysis. *Journal of Safety Science and Resilience 5*, 460–469. https://doi.org/10.1016/j.jnlssr.2024.06.001

Adamson, A. (2018). Machine learning and health care disparities in dermatology. *JAMA Dermatology, 154*(11). https://doi.org/10.1001/jamadermatol.2018.2348

Adlung, L., Cohen, Y., Mor, U., Elinav, E. (2021). Machine learning in clinical decision making. *Med 2*, 642–665. https://doi.org/10.1016/j.medj.2021.04.006

Baier, A. (1986). Trust and antitrust. *Ethics, 96*(2), 231–260. https://doi.org/10.1086/292745

Beauchamp, T. L., & Childress, J. F. (2024). *Prinzipien der Bioethik* (J. Pelger, Übers.; D. Lanzerath & A. Halsband, Hrsg.). Verlag Karl Alber. https://doi.org/10.5771/9783495998045

Bergquist, M., Rolandsson, B. (2022). Exploring ADM in clinical decision-making. Healthcare experts encountering digital automation. In S. Pink, M, Berg, D. Lupton, M. Ruckenstein (Hrsg.), *Everyday automation. Experiencing and anticipating emerging technologies* (S. 140–153). Routledge.

Berlin, I. (2002). Two concepts of liberty. In *Four essays on liberty*. Oxford University Press.

Binns, R. (2018). Fairness in machine learning. Lessons from political philosophy. *Proceedings of Machine Learning Research*, *81*, 1–11.

Bottrighi, A., Grosso, F., Ghiglione, M., Maconi, A., Nera, S., Piovesan, L., Raina, E., Roveta, A., Terenziani, P. (2025). A symbolic AI approach to medical training. *Journal of Medical Systems, 49*(1), 2. https://doi.org/10.1007/s10916-024-02139-y

Boyle, M. (2017). Wesentlich vernünftige Tiere. In A. Kern, C. Kietzmann (Hrsg.), *Selbstbewusstes Leben. Texte zu einer transformativen Theorie der menschlichen Subjektivität*. Suhrkamp.

Boyle, M. (2024). *Transparency and reflection. A study of self-knowledge and the nature of mind*. Oxford University Press.

Bradley, B. (2015). *Well-Being*. Polity Press.

Brandt, W., Fritz, A., Kießig, A., Lerch, P. (2025). KI in der bildgebenden Diagnostik verantwortet vertrauen. Erfahrungen aus Radiologie und Pathologie ethisch diskutiert. *Ethik in der Medizin*, *37*, 533–552. https://doi.org/10.1007/s00481-025-00878-1

Bruijne, M. (2016). Machine learning approaches in medical image analysis. From detection to diagnosis. *Medical Image Analysis, 33*, 94–97.

Burrell, J. (2016). How the machine ›thinks‹. Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1). https://doi.org/10.1177/2053951715622512

Caro, M., Vaccarezza, M. S. (Hrsg.) (2021). *Practical wisdom. Philosophical and psychological perspectives*. Routledge.

Chan, B. (2023). Black-box assisted medical decisions: AI power vs. ethical physician care. *Medicine, Health Care and Philosophy, 26*, 285–292. https://doi.org/10.1007/s11019-023-10153-z

Chaplin, R. (2023). Personal reactive attitudes and partial responses to others. A partiality-based approach to Strawson's reactive attitudes. *Journal of Ethics and Social Philosophy, 25*(2), 323–345. https://doi.org/10.26556/jesp.v25i2.1726

Christman, J. (Hrsg.) (2014). *The inner citadel. Essays on individual autonomy*. Echo Point Books & Media.

Churchland, P. (2013). *Matter and consciousness*. MIT Press.

Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Gigante, A., Valencia, A., Rementeria, M. J., Chadha, A. S., Mavridis, N. (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *npj Digital Medicine, 3*(81). https://doi.org/10.1038/s41746-020-0288-5.

Coeckelbergh, M. (2012). Can we trust robots? *Ethics and Information Technology 14,* 53–60. https://doi.org/10.1007/s10676-011-9279-1

Coeckelbergh, M. (2020a). *AI ethics*. MIT Press.

Coeckelbergh, M. (2020b). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics 26*, 2051–2068. https://doi.org/10.1007/s11948-019-00146-8

Clark, A. (2001). *Mindware. An introduction to the philosophy of cognitive science*. Oxford University Press.

Cross, J. L., Choma, M. A., Onofrey, J. A. (2024). Bias in medical AI. Implications for clinical decision-making. *PLOS Digital Health 3*(11). https://doi.org/10.1371/journal.pdig.0000651

Crowston, K., Bolici, F. (2025). Deskilling and upskilling with generative AI systems. *Information Research an International Electronic Journal*, 30(iConf), 1009–1023. https://doi.org/10.47989/ir30iConf47143

Danaher, J. (2020). Welcoming robots into the moral circle: A defence of ethical behaviourism. *Science and Engineering Ethics*, *26*, 2023–2049. https://doi.org/10.1007/s11948-019-00119-x

Daniels, D. (1979): Wide reflective equilibrium and theory acceptance in ethics. *The Journal of Philosophy*, *76*(5), 256–282. https://doi.org/10.1017/CBO9780511624988.003

DePaul, M. R. (1993). *Balance and refinement beyond coherence methods of moral inquiry*. Routledge.

Diakopolous, N. (2020). Transparency. In M. D. Dubber, F. Pasquale, S. Das (Hrsg.), *The Oxford Handbook of Ethics of AI* (S. 197–213). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.11

Durán, J. M., Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis for trust in medical AI. *Journal of Medical Ethics, 47,* 329–335. https://doi.org/10.1136/medethics-2020-106820

Dworkin, G. (1998). *The theory and practice of autonomy.* Cambridge University Press.

Fabris, A. (Hrsg.) (2020): *Trust. A philosophical approach*. Springer.

Faden, R., Beauchamp, T. (1986). *A history and theory of informed consent*. Oxford University Press.

Favier, M., Calders, T., Pinxteren, S., Meyer, J. (2023). How to be fair? A study of label and selection bias. *Machine Learning*, *112*, 5081–5104. https://doi.org/10.1007/s10994-023-06401-1

Felzmann, H., Fosch-Villaronga, E., Lutz, C., Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics, 26*, 3333–3361. https://doi.org/10.1007/s11948-020-00276-4

Feuerriegel, S., Dolata, M., Schwabe, G. (2020). Fair AI. Challenges and opportunities. *Business & Information Systems Engineering, 62*(4), 379–384. https://doi.org/10.1007/s12599-020-00650-3

Frankish, K., Ramsey, W. M. (Hrsg.) (2012). *The Cambridge handbook of cognitive science*. Cambridge University Press.

Freed, S. (2020). *AI and human thought and emotion*. CRC Press.

Filippi, C. G., Stein, J. M., Wang, Z., Bakas, S., Liu, Y., Chang, P. D., Lui, Y., Hess, C., Barboriak, D. P., Flanders, A. E., Wintermark, M., Zaharchuk, G., Wu, O. (2023). Ethical considerations and fairness in the use of artificial intelligence for neuroradiology. *American Journal of Neuroradiology 44*(11), 242–1248. https://doi.org/10.3174/ajnr.A7963

Flasiński, Mariusz (2011). *Introduction to artificial intelligence*. Springer Nature.

Floridi, L. (2015). *The ethics of information*. Oxford University Press.

Frankfurt, H. (1969). Alternate possibilities and moral responsibility. In *The Journal of Philosophy, 66*(23), 829–839.

Frankfurt, H. (2009). Freedom of the will and the concept of a person. In H. Frankfurt, *The importance of what we care about. Philosophical essays*. Cambridge University Press.

Fodor, J. (1975). *The language of thought*. Harvard University Press.

Funer, F., Wiesing, U. (2024). Physician's autonomy in the face of AI support: walking the ethical tightrope. *Frontiers in Medicine 11*. https://doi.org/10.3389/fmed.2024.1324963

Gallagher, S. (2020). *Action and interaction*. Oxford University Press.

Gardner, A., Smith, A. L., Steventon, A., Coughlan, E., Oldfield, M. (2022). Ethical funding for trustworthy AI: proposals to address the responsibilities of funders to ensure that projects adhere to trustworthy AI practice. *AI Ethics, 2,* 277–291. https://doi.org/10.1007/s43681-021-00069-w

Goel, A. K. (2021). Looking back, looking ahead: Symbolic versus connectionist AI. *AI Magazine, 42*(4), 83–85. https://doi.org/10.1609/aaai.12026

Goldberg, S. (2020). Trust and reliance. In J. Simon (Hrsg.), *The Routledge handbook of trust and philosophy*. Routledge.

Griffin, J. (1986). *Well-being. Its meaning, measurement, and moral importance*. Oxford University Press.

Grote, T., Keeling, G. (2022). Enabling fairness in healthcare through machine learning. *Ethics and Information Technology, 24*, 39. https://doi.org/10.1007/s10676-022-09658-7

Halldenius, L. (2005). Dissecting »discrimination«. *Cambridge Quarterly of Healthcare Ethics, 14*(4), 455–463.

Heinrichs, B. (2021). Discrimination in the age of artificial intelligence. *AI&Society, 37*, 143–154. https://doi.org/10.1007/s00146-021-01192-2

Heinrichs, B., Heinrichs, J. H., Rüther, M. (2022). *Künstliche Intelligenz*. de Gruyter.

Heinrichs, B., Wagner, R. (2024). Four notions of autonomy. Pitfalls of conceptual pluralism. *Human-Machine Communication, 9*, 37–50.

Herzog, C. (2022). On the risk of confusing interpretability with explicability. *AI Ethics, 2*, 219–225. https://doi.org/10.1007/s43681-021-00121-9

Holman, J. G., Cookson, M. J. (1987). Expert systems for medical applications. *Journal of Medical Engineering & Technology, 11*(4), 151–159.

Holton, R. (1994). Deciding to trust, coming to believe. *Australasian Journal of Philosophy, 72*, 63–76. https://doi.org/10.1080/00048409412345881

Hooker, S. (2021). Moving beyond »algorithmic bias is a data problem«. *Patterns, 2*(4). https://doi.org/10.1016/j.patter.2021.100241.

Horn, C., Löhrer, G. (Hrsg.) (2010). *Gründe und Zwecke. Texte zur aktuellen Handlungstheorie*. Suhrkamp.

Huang, J., Fox, J., Gordon, C., Jackson-Smale, A. (1993). Symbolic decision support in medical care. *Artificial Intelligence in Medicine, 5*, 415–430.

Johnson, M. (2008). *The meaning of the body. Aesthetics of Human Understanding*. University of Chicago Press.

Jones, K. (1996). Trust as an affective attitude. *Ethics, 107*(1), 4–25. https://doi.org/10.1086/233694

Jordan, M. I., Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260.

Kamiran, F., Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*, 1–33. https://doi.org/10.1007/s10115-011-0463-8.

Kawamleh, S. (2023). Against explainability requirements for ethical artificial intelligence in health care. *AI Ethics, 3*, 901–916. https://doi.org/10.1007/s43681-022-00212-1

Kern, A. (2006). *Quellen des Wissens. Zum Begriff vernünftiger Erkenntnisfähigkeiten*. Suhrkamp.

Koçak, B., Ponsiglione, A., Stanzione, A., Bluethgen, C., Santinha, J., Ugga, L., Huisman, M., Klontzas, M. E., Cannella, R., Cuocolo, R. (2025). Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. *Diagnostic and Interventional Radiology, 31*(2), 75–88. https://doi.org/10.4274/dir.2024.242854

Kokol, P., Stiglic, B., Rozman, I. (2002). Decision tree: an overview and their use in medicine. *Journal of Medical Systems, 26*, 445–463.

Lakoff, G, Johnson, M. (1999). *Philosophy in the flesh. Embodied mind and its challenges to western thought*. Basic Books.

Laitinen, A., Sahlgren, O. (2021). AI systems and respect for human autonomy. *Frontiers in Artificial Intelligence, 4*. https://doi.org/10.3389/frai.2021.705164.

Lippert-Rasmussen, K. (2013). *Born Free and Equal? A Philosophical Inquiry into the Nature of Discrimination*. Oxford University Press.

Liua, R., Rong, Y., Peng, Z. (2020). A review of medical artificial intelligence. *Global Health Journal*, *4*(2). https://doi.org/10.1016/j.glohj.2020.04.002

Lombi, L., Rossero, E. (2024). How artificial intelligence is reshaping the autonomy and boundary work of radiologists. A qualitative study. *Sociology of Health & Illness, 46*(2), 200–218. https://doi.org/10.1111/1467-9566.13702.

London, Alex J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report*, *49*(1), 15–21. https://doi.org/10.1002/hast.973.

MacCallum, G. (1967). Negative and positive freedom. *The Philosophical Review, 76*(3), 312–334.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*, 175–183. https://doi.org/10.1007/s10676-004-3422-1

Mayr, E. (2018). *Understanding human agency*. Oxford University Press.

Menke, C. (2018). *Autonomie und Befreiung. Studien zu Hegel*. Suhrkamp.

Metzinger, T. (2019, 8. April). EU guidelines: Ethics washing made in Europe. *Der Tagesspiegel Online*. https://www.tagesspiegel.de/politik/ethics-washing-made-in-europe-5937028.html

Munn, N., Weijers, D. (2023). Corporate responsibility for the termination of digital friends. *AI & Society, 38*, 1501–1502. https://doi.org/10.1007/s00146-021-01276-z

Natali, C., Marconi, L., Duran, L. D. D., Miglioretti, M., Cabitza, F. (2025). *AI-induced deskilling in medicine: a mixed method literature review for setting a new research agenda*. http://dx.doi.org/10.2139/ssrn.5166364

Newen, A., Bruin, L., Gallagher, S. (Hrsg.) (2018). *The Oxford handbook of 4E cognition*. Oxford University Press.

Nozick, R. (2013). *Anarchy, State, Utopia*. Basic Books.

Nyholm, S. (2018). Attributing agency to automated systems – Reflections on human-robot collaborations and responsibility. *Science and Engineering Ethics, 24*, 1201–1219. https://doi.org/10.1007/s11948-017-9943-x

Nyholm, S. (2020). *Humans and robots. Ethics, agency, and anthropomorphism*. Rowman & Littlefield.

O'Neill, O. (2002a). *Autonomy and trust in bioethics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511606250

O'Neill, O. (2002b). *A question of trust: The BBC Reith Lectures 2002*. Cambridge University Press.

O'Neill, O. (2003). Autonomy: The emperor's new clothes. In *Aristotelian Society Supplementary, 77(1)*, 1–21.

Oshana, M. (Hrsg.) (2015). *Personal autonomy and social oppression. Philosophical perspectives*. Routledge.

Panch, T., Mattie, H., Atun, R. (2019). Aritificial intelligence and algorithmic bias: implications for health systems. *Journal of Global Health*, *9*(2), 020318. https://doi.org/10.7189/jogh.09.020318

Peteet, J.R., Witvliet, C.V., Glas, G., Frush, B. W. (2023). Accountability as a virtue in medicine: from theory to practice. *Philosophy, Ethics, and Humanities in Medicine, 18*, 1. https://doi.org/10.1186/s13010-023-00129-5

Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds & Machines, 30*, 385–409. https://doi.org/10.1007/s11023-020-09537-4

Potter, N. N. (2003). *How can I be trusted? A virtue theory of trustworthiness*. Rowman & Littlefield.

Quiñones, J. L. G. (2025). Using artificial intelligence to enhance patient autonomy in healthcare decision-making. *AI & Society, 40*, 1917–1926. https://doi.org/10.1007/s00146-024-01956-6

Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N., Fanos, V. (2020). Comparison of conventional satistical methods with machine learning in medicine: Diagnosis, drug development, and treatment. *Medicina, 56*, 455. https://doi.org/10.3390/medicina56090455

Rathkopf, C., Heinrichs, B. (2024). Learning to live with strange error: Beyond trustworthiness in artificial intelligence ethics. *Cambridge Quarterly of Healthcare Ethics*, *33*(3), 333–345.

Rebera, A. P. (2024). Reactive attitudes and AI-agents – Making sense of responsibility and control gaps. *Philosophy & Technology, 37*, 126. https://doi.org/10.1007/s13347-024-00808-x

Reeves, K. (2024). AI's diversity problem in radiology: Addressing algorithm bias. *Applied Radiology, 1*, 44–45.

Roberson, T, Bornstein, S., Liivoja, R., Ng, S., Scholz, J., Devitt, K. (2022). A method for ethical AI in defence: A case study on developing trustworthy autonomous systems. *Journal of Responsible Technology, 11*. https://doi.org/10.1016/j.jrt.2022.100036

Ross, D. (2002). *The right and the good*. Oxford University Press.

Rowlands, M. (2010). *The new science of the mind. From extended mind to embodied phenomenology*. MIT Press.

Rubeis, G. (2024). *Ethics of medical AI*. Springer Nature.

Rubel, A., Castro, C., Pham, A. (2021). *Algorithms and autonomy. The ethics of automated decision systems.* Cambridge University Press. https://doi.org/10.1017/9781108895057

Ryan, M. (2020). In AI we trust: Ethics, artifcial intelligence, and reliability. *Science and Engineering Ethics, 26*, 2749–2767. https://doi.org/10.1007/s11948-020-00228-y

Scheman, N. (2020). Trust and trustworthiness. In J. Simon (Hrsg.), *The Routledge Handbook of Trust and Philosophy*. Routledge.

Schubbach, A. (2021). Judging machines: philosophical aspects of deep learning. *Synthese, 198*, 1807–1827. https://doi.org/10.1007/s11229-019-02167-z

Seligman, M. (2011). *Flourish. A visionary new understanding of happiness and well-being*. Free Press.

Shapiro, L., Spaulding, S. (Hrsg.) (2024). *The Routledge handbook of embodied cognition*. Routledge.

Simon, J. (Hrsg.) (2020). *The Routledge handbook of trust and philosophy*. Routledge.

Simpson, T. W. (2023). *Trust. A philosophical study*. Oxford University Press.

Sridhar, S., Khamaj, A., Asthana, M. K. (2023). Cognitive neuroscience perspective on memory: overview and summary. *Frontiers in Human Neuroscience, 17*. https://doi.org/10.3389/fnhum.2023.1217093

Starke, G., Brule, R., Elger, B. S., Haselager, P. (2022). Intentional machines: A defence of trust in medical artificial intelligence. *Bioethics, 36*(2), 154–161. https://doi.org/10.1111/bioe.12891

Steckmann, U., Heinrichs, B. (2023) Künstliche Intelligenz und menschliches Maß. In J. Loh, T. Grote (Hrsg.), *Medizin – Technik – Ethik. Spannungsfelder zwischen Theorie und Praxis (S. 17–36).* Metzler.

Steinfath, H. (Hrsg.) (2016). *Autonomie und Vertrauen. Schlüsselbegriffe der modernen Medizin*. Springer VS.

Strawson, P. F. (2008). Freedom and resentment. In P. F. Strawson, *Freedom and Resentment, and other Essays*. Routledge.

Talbert, M. (2016). *Moral responsibility*. Polity Press.

Thompson, E. (2007). *Mind in life. Biology, phenomenology, and the sciences of mind*. Harvard University Press.

Tigard, D. W. (2021). There is no techno-responsibility gap. *Philosophy & Technology, 34*, 589–607. https://doi.org/10.1007/s13347-020-00414-7

Topol, E. (2019). *Deep medicine. How artificial intelligence can make healthcare human again*. Basic Books.

Varela, F. J., Thompson, E., Rosch, E. (2016). *The emobied mind. Cognitive science and human experience*. MIT Press.

Vasiliou, I. (2025). *Varieties of happiness: Eudaimonism and greek ethical theory*. Oxford University Press. https://doi.org/10.1093/9780197645093.001.0001

Waite, S., Scott, J. (2021). Narrowing the gap: Imaging disparities in radiology. *Radiology, 299,* 27–35. https://doi.org/10.1148/radiol.2021203742

Wallace, R. J. (2022). Responsibility and reactive attitudes. In D. K. Nelkin, D. Pereboom (Hrsg.), *The Oxford Handbook of Moral Responsibility*. Oxford University Press.

Walmsley, J. (2021). Artificial intelligence and the value of transparency. *AI & Society, 36*, 585–595. https://doi.org/10.1007/s00146-020-01066-z

Winter, P. D., Carusi, A. (2023). (De)troubling transparency: artificial intelligence (AI) for clinical applications. *Journal of Medical Humanities*, *49*(1), 17–26. https://doi.org/10.1136/medhum-2021-012318

Woleński, J. (2004). The history of epistemology. In I. Niiniluoto, M. Sintonen, J. Woleński (Hrsg.), *Handbook of Epistemology*. Kluwer Academic Publishers.

Wolfensberger, M., Wrigley, A. (2019). *Trust in medicine. Its nature, justification, significance, and decline*. Cambridge University Press.

Wood, S. (Hrsg.) (2024). *The degradation of work? Skill, de-skilling and the labour process*. Routledge.

Xu, J., Xiaob, Y., Wangc, W. H., Ningc, Y., Shenkmana, E. A., Biana, J., Wang, F. (2022). Algorithmic fairness in computational medicine. *eBioMedicine, 84,* 104250. https://doi.org/10.1016/j.ebiom.2022.104250

Zahavi, D. (2005). *Subjectivity and selfhood. Investigating the first-person perspective*. MIT Press.

Zanotti G., Petrolo, M., Chiffi, D., Schiaffonati, V. (2024). Keep trusting! A plea for the notion of trustworthy AI. *AI & Society, 39*, 2691–2702. https://doi.org/10.1007/s00146-023-01789-9

Zhang, Y., Li, B., Ling, Z., Zhou, G. (2024). Mitigating label bias in machine learning: Fairness through confident learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(15), 16917–16925. https://doi.org/10.1609/aaai.v38i15.29634