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