Aus dem Institut für Ethik, Geschichte und Theorie der Medizin
Institut der Ludwig-Maximilians-Universität München

KI-Systeme zur Unterstützung ärztlicher Entscheidungen:

eine ethische Bewertung

Dissertation
zum Erwerb des Doktorgrades der Medizin
an der Medizinischen Fakultät der
Ludwig-Maximilians-Universität München

vorgelegt von
Florian Michael Englich

aus
Bonn

2025

Mit Genehmigung der Medizinischen Fakultät der

Ludwig-Maximilians-Universität München

Erstes Gutachten:	Prof. Dr. Georg Marckmann, MPH
Zweites Gutachten:	Prof. Dr. Martin Fischer
Drittes Gutachten:	Prof. Dr. Michael Ingrisch
ggf. weitere Gutachter:	_____
	_____
Dekan:	Prof. Dr. med. Thomas Gudermann
Tag der mündlichen Prüfung:	27.02.2025

Inhaltsverzeichnis

Inhaltsverzeichnis *3*

Zusammenfassung *7*

Abstract (English) *9*

Abbildungsverzeichnis *10*

Tabellenverzeichnis *11*

Abkürzungsverzeichnis *12*

1. Einleitung *13*

1.1 Medizinethischer Forschungsstand **14**

1.2 Forschungsfragen **18**

1.3 Aufbau der Arbeit **18**

2. Methodik *19*

2.1 Instrumente zur ethischen Evaluation digitaler medizinischer Technologien **19**

2.2 Ethische Bewertungsmatrix für eHealth-Anwendungen **21**

2.2.1 Methodisches Vorgehen 22

2.2.2 Normative Begründung der Bewertung 23

2.2.3 Darstellung der Bewertungsmatrix 24

2.3 Spezifizierung der Bewertungskriterien **27**

2.4 Ethische Beurteilungskriterien zur Bewertung von AI-CDSS **28**

3. KI-Systeme zur Unterstützung ärztlicher Entscheidungen *30*

3.1 Technologische Grundlagen **30**

3.1.1 Aufbau eines AI-CDSS 30

3.1.2 Machine Learning 32

3.1.2.1 Formen von Machine Learning 32

3.1.2.2 Entwicklungsprozess eines ML-Systems 33

3.2 Vorstellung beispielhafter Anwendungen **34**

3.2.1 Diagnostik 34

3.2.2 Prognostik 36

3.2.3 Therapieempfehlung 39

3.3 Zieldefinition von AI-CDSS **42**

3.4 Charakteristische Eigenschaften von AI-CDSS ... 43
3.4.1 Opazität und Erklärbarkeit ... 43
3.4.2 Abhängigkeit von Trainingsdaten und eingeschränkte Generalisierbarkeit ... 45
3.4.3 Zwischen Entscheidungsunterstützung und -übernahme ... 47
3.4.4 Transformation der Medizin und des ärztlichen Berufs ... 51
4. Ethische Bewertung von AI-CDSS ... 54
4.1 Funktionsfähigkeit ... 54
4.1.1 Machbarkeit ... 54
4.1.2 Wirksamkeit ... 58
4.1.3 Brauchbarkeit ... 62
4.2 Mögliche Alternativen ... 67
4.3 Nutzenpotenzial für die Patientinnen und Patienten ... 69
4.4 Schadenspotenzial für die Patientinnen und Patienten ... 72
4.4.1 Schadenspotenzial durch die Ausgabe falscher Systemergebnisse ... 73
4.4.2 Schadenspotenzial durch Anwendungsfehler ... 77
4.4.2.1 Anwendung eines AI-CDSS trotz fehlender Anwendbarkeit ... 77
4.4.2.2 Übernahme falscher Systemergebnisse ... 79
4.4.2.3 Überdiagnostik ... 82
4.5 Wahrung und Förderung der Patientenautonomie ... 83
4.5.1 Informierte Einwilligung der Patientinnen und Patienten ... 84
4.5.2 Förderung der Gesundheitsmündigkeit der Patientinnen und Patienten ... 86
4.5.3 AI-CDSS und die Beachtung von Patientenpräferenzen ... 87
4.6 Wahrung der ärztlichen Entscheidungsautonomie ... 91
4.7 Auswirkung auf die ärztliche Entscheidungskompetenz ... 94
4.8 Zuschreibbarkeit von Verantwortung beim Einsatz von AI-CDSS ... 96
4.9 Wahrung der Integrität der Arzt-Patient-Beziehung ... 101
4.9.1 Vertrauen in KI-unterstützte Ärztinnen und Ärzte ... 102
4.9.2 Auswirkungen auf die Empathie und den nicht-reduktionistischen, bio-psycho-sozialen Fokus der Ärztinnen und Ärzte ... 105
4.10 Datenschutz und Datenverfügbarkeit ... 107
4.11 Effizienz ... 112
4.12 Gerechtigkeit ... 115
5. Übergreifende ethische Bewertung von AI-CDSS ... 121

6. Empfehlungen 127

6.1 Empfehlungen für die Entwicklung von AI-CDSS 127

6.1.1 Zusammensetzung des Entwicklungsteams 127

6.1.2 Identifizierung eines Bedarfs an Entscheidungsunterstützung 128

6.1.3 Vergleich des designierten AI-CDSS mit Alternativen 128

6.1.4 Machbare, klare und evaluierbare Zieldefinition 129

6.1.5 Auswahl eines geeigneten Trainingsdatensatzes 130

6.1.6 Umweltfreundliche Entwicklung 130

6.1.7 Lokale Validierung und Rekalibrierung 131

6.1.8 Gewährleistung von Flexibilität für Patientenpräferenzen 131

6.1.9 Integrierbarkeit in den Workflow 132

6.1.10 Benutzerfreundlichkeit 132

6.1.11 Verhinderung von *alert fatigue* 133

6.1.12 Ausgabe mehrerer Ergebnisse mit jeweiligem Konfidenzgrad 134

6.1.13 Ausgabe der Outputs an die Patientinnen und Patienten 135

6.1.14 Datenschutzmaßnahmen 136

6.1.15 Sicherung von Fairness 136

6.1.16 Gewährleistung von Erklärbarkeit 137

6.1.17 Vermeidung falscher Systemergebnisse 138

6.1.18 Rigorose Evaluation und Testung 139

6.1.18.1 Wirksamkeit 140

6.1.18.2 Brauchbarkeit 141

6.1.18.3 Nutzenpotenzial für die Patientinnen und Patienten 141

6.1.18.4 Fehler und Schadenspotenzial 142

6.1.18.5 Weitere Aspekte der Evaluation und Testung 142

6.1.19 Klare Anforderungen an die Systemnutzung 143

6.1.20 Wartung und Aktualisierung der Systeme 143

6.2 Empfehlungen für die Schaffung geeigneter Rahmenbedingungen für die Entwicklung und Anwendung von AI-CDSS 144

6.2.1 Transparente rechtliche Anforderungen für die Entwicklung und Nutzung von AI-CDSS 144

6.2.2 Einteilung von AI-CDSS in Risikoklassen 144

6.2.3 Ermöglichung von klarer Verantwortungszuschreibung 144

6.2.4 Standardisierung von AI-CDSS und der IT-Infrastruktur 145

6.2.5 Generierung und Bereitstellung von qualitativ hochwertigen Datensätzen 146

6.2.6 Förderung der Entwicklung von AI-CDSS 146
6.2.7 Förderung der Nutzung von AI-CDSS 147
6.2.8 Anpassung der ärztlichen Ausbildung 147
6.2.9 Ethische Sensibilisierung der Entwicklerinnen und Entwickler von AI-CDSS 148
6.2.10 Gesellschaftliche Aufklärung über KI 149
6.3 Empfehlungen für die Nutzung von AI-CDSS 150
6.3.1 Überprüfung des Bedarfs an Entscheidungsunterstützung 150
6.3.2 Vergleich mit Alternativen 150
6.3.3 Überprüfung der Eignung des Systems für die spezifischen Anwendungsbedingungen . 150
6.3.4 Probatorische Nutzung vor der langfristigen Inbetriebnahme 151
6.3.5 Etablierung einer passenden IT-Umgebung und Einbettung des AI-CDSS 151
6.3.6 Umweltfreundliches Nutzungsverhalten 152
6.3.7 Schulung und Weiterbildung der Nutzerinnen und Nutzer 152
6.3.8 Aufklärung und Einwilligung der Patientinnen und Patienten 154
6.3.9 Verhinderung von Zeitdruck und von Ablenkung bei der Verwendung von AI-CDSS 154
6.3.10 Verhinderung von Überdiagnostik 155
6.3.11 Überprüfung der Eignung des AI-CDSS für den spezifischen Anwendungsfall 155
6.3.12 Kontrolle und Kontextualisierung der Outputs von AI-CDSS 156
6.3.13 Umgang mit personenbezogenen Daten 157
6.3.14 Patientenorientierte Nutzung der frei gewordenen Ressourcen 158
6.3.15 Kontinuierliche Überprüfung des Systems und der sachgemäßen Nutzung 159
7. Literaturverzeichnis 160
8. Affidavit 191
9. Danksagung 192

Zusammenfassung

Anwendungen aus dem Bereich der Künstlichen Intelligenz, die insbesondere auf der Technologie des Maschinellen Lernens beruhen, breiten sich in der Medizin zunehmend im Gebiet der Entscheidungsunterstützung aus. Einige dieser Systeme unterstützen Ärzte[1], indem sie Diagnosen, Prognosen sowie Therapieempfehlungen abgeben. Diese Anwendungen können auch als *artificial intelligence-based clinical decision support systems* (AI-CDSS) bezeichnet werden. Indem die Systeme die klinische Entscheidungsfindung beeinflussen, dringen sie in den Kernbereich der Medizin und des ärztlichen Tuns und Denkens vor. In diesem Zusammenhang entstehen einige ethische Fragen: Wie soll mit dem nicht eliminierbaren Fehlerpotenzial von AI-CDSS umgegangen werden? Wer trägt die Verantwortung für falsche Entscheidungen, die auf Systemempfehlungen beruhen? Und wie verändert sich die Arzt-Patient-Beziehung, wenn solche Anwendungen benutzt werden?

Diese und weitere ethische Fragestellungen, die bei der Verwendung der Systeme auftreten, wurden bereits einzeln und in Verbindung mit anderen ethischen Implikationen der Anwendungen untersucht. Jedoch ist die Zahl der Arbeiten, die AI-CDSS als Gruppe ethisch bewerten, bisher gering. Diese Bewertungen sind außerdem meist knapp gehalten und gehen wenig systematisch vor.

Die vorliegende Arbeit greift zur systematischen ethischen Bewertung von AI-CDSS auf eine ethische Bewertungsmatrix für eHealth-Anwendungen zurück. Zunächst werden die Systeme, ihre Funktionsweise und ihre Charakteristika beschrieben. Die darauffolgende ethische Bewertung von AI-CDSS erfolgt anhand von zwölf ethischen Beurteilungskriterien: Funktionsfähigkeit, Nutzenpotenzial für die Patientinnen und Patienten, mögliche Alternativen, Schadenspotenzial für die Patientinnen und Patienten, Wahrung und Förderung der Patientenautonomie, Zuschreibbarkeit von Verantwortung beim Einsatz von AI-CDSS, Wahrung der Integrität der Arzt-Patient-Beziehung, Datenschutz und Datenverfügbarkeit, Effizienz und schließlich Gerechtigkeit.

In der übergreifenden Bewertung werden die Einzelbewertungen zusammengeführt und gegeneinander abgewogen. Es wird dabei deutlich, dass die Anwendungen nicht

[1] In der vorliegenden Arbeit wird aus Gründen der Lesbarkeit die männliche Form verwendet; die Angaben beziehen sich auf alle Geschlechter und sollen niemanden ausschließen.

kategorisch positiv oder negativ zu beurteilen sind, sondern dass deren ethische Evaluation von der jeweiligen Entwicklungs- und Nutzungsweise abhängt. Abschließend folgen ethisch begründete Empfehlungen zur Entwicklung und Anwendung von AI-CDSS sowie zur Schaffung dafür geeigneter Rahmenbedingungen.

Abstract (English)

Applications from the field of artificial intelligence, mostly based on machine learning, are increasingly spreading in clinical decision support. Some of these systems support doctors by providing prognoses, diagnoses and therapeutic recommendations. They can also be referred to as artificial intelligence-based clinical decision support systems (AI-CDSS). As these applications support and influence clinical decision-making, they penetrate the core of medicine. This raises a number of ethical questions: How should the systems' potential for errors, which cannot be eliminated, be dealt with? Who bears responsibility for incorrect decisions based on the recommendations of AI-CDSS? And how is the doctor-patient relationship affected by the use of these systems?

These and other ethical issues that arise when doctors utilize AI-CDSS have already been addressed individually and in conjunction with other ethical implications of the systems. However, there is still little work that evaluates AI-CDSS ethically as a whole. Moreover, these works are mostly of short length and not very systematic.

This dissertation employs a structured ethical evaluation tool for eHealth applications to ethically evaluate AI-CDSS. First, the systems, their mode of operation and their characteristics are described. The subsequential ethical evaluation of AI-CDSS is based on twelve ethical evaluation criteria: functionality, potential benefits for patients, possible alternatives, potential harm for patients, preservation and promotion of patient autonomy, attributability of responsibility in the use of the systems, preservation of the integrity of the doctor-patient relationship, data protection and data availability, efficiency and finally justice.

In the overarching evaluation, the individual assessments are synthesized and balanced against each other. It becomes evident that the ethical evaluation of AI-CDSS is neither categorically positive nor categorically negative, but contingent on how the systems are used. Lastly, ethically justified recommendations are provided for the development and application of AI-CDSS as well as for the creation of suitable framework conditions.

Abbildungsverzeichnis

Abbildung 1: *Aufbau eines AI-CDSS* 31

Tabellenverzeichnis

Tabelle 1: *Ethische Bewertungsmatrix zur Bewertung von eHealth-Anwendungen* 25

Tabelle 2: *Ethische Bewertungsmatrix zur Bewertung von AI-CDSS* 29

Tabelle 3: *Nützliche und schädliche Faktoren in Bezug auf alert fatigue* 134

Abkürzungsverzeichnis

AI-CDSS ..... *artificial intelligence-based clinical decision support system*
aPTT ..... aktivierte partielle Thromboplastinzeit
CDSS ..... *clinical decision support system*
DL ..... Deep Learning
EBEA ..... Ethische Bewertungsmatrix für eHealth-Anwendungen
KI ..... Künstliche Intelligenz
LVEF ..... linksventrikuläre Ejektionsfraktion
ML ..... Machine Learning
UH ..... unfraktioniertes Heparin
WFO ..... *IBM Watson for Oncology*

1. Einleitung

Kaum ein Patient wird heutzutage behandelt, ohne dass bei der klinischen Entscheidungsfindung eine digitale Technologie verwendet wird. Über Computer, Tablets, Smartphones und viele weitere Systeme kann man etwa Fachinformationen zu Erkrankungen, Differentialdiagnosen oder therapeutischen Maßnahmen abrufen. Anwendungen wie die elektronische Patientenakte aggregieren verschiedene behandlungsrelevante Informationen und bereiten diese auf. Digitale Technologien haben also einen festen Platz in der klinischen Entscheidungsfindung. Diese herkömmlichen Formen der computerbasierten Entscheidungsunterstützung unterstützen die Diagnostik, Prognostik und die Therapieentscheidung vor allem durch eine Optimierung der *Entscheidungsgrundlage*. Darüber hinaus gibt es auch ‚intelligente' Systeme, die selbst Diagnosen, Prognosen oder therapeutische Maßnahmen vorschlagen. Insofern Diagnostik, Prognostik und Therapieentscheidung zu den Kernbestandteilen des ärztlichen Tuns und Denkens gehören, werfen diese Anwendungen mit zunehmender Leistungsfähigkeit die Frage nach der Ersetzbarkeit des Arztes auf. So sind diese Systeme mit potenziell tiefgreifenden Transformationen verbunden und aus ethischer Perspektive implikationsreich (Marckmann, 2003).

Der geschichtliche Rückblick zeigt, dass der Versuch nicht neu ist, Computersysteme zu entwickeln, welche die klinische Problemlösefähigkeit des Arztes imitieren oder gar übertreffen sollen: Als eine erste historisch relevante, größere Gruppe solcher Anwendungen kann man die ab 1976 entwickelten medizinischen Expertensysteme anführen (Peiffer-Smadja et al., 2020). Deren technologische Grundlage wird auch als wissensbasierte Künstliche Intelligenz (KI) bezeichnet (van Baalen et al., 2021): Relevantes Wissen eines Teilgebiets wird so digitalisiert, dass daraus Schlussfolgerungen für Einzelfälle gezogen werden können (Wright & Sittig, 2008).

Obwohl – vor allem im Bereich der Diagnostik – viele medizinische Expertensysteme entwickelt wurden und diesen anfangs viel Enthusiasmus entgegengebracht wurde, kam es nie zu einer weiten Verbreitung in die klinische Praxis (Kulikowski, 2019; Sonar & Weber, 2020).

In letzter Zeit werden wieder vermehrt Systeme entwickelt, die dem Begriff der KI zugeordnet werden und Diagnosen, Prognosen oder auch Therapieempfehlungen geben können. Grund für diesen Boom sind vor allem Fortschritte auf dem Gebiet des Maschinellen Lernens (ML), die insbesondere auf die wachsende Verfügbarkeit großer Datenmengen, die Zunahme von Rechenkapazitäten und verbesserte Algorithmen zurückzuführen sind (Kaul et al., 2020; Rosengrün, 2021).

Dadurch ist ein neues Leistungsniveau erreichbar geworden: Manche KI-Systeme können bei der Diagnostik oder bei der Prognostik in einzelnen, eng abgegrenzten Anwendungsbereichen der menschlichen Leistungsfähigkeit nahekommen oder diese sogar übertreffen (Asch et al., 2021; Bologheanu et al., 2023; Haenssle et al., 2020; Huang et al., 2020; Sarayar et al., 2023). Auch im therapeutischen Bereich gibt es Anwendungen. Diese können etwa Behandlungsoptionen empfehlen, die an individuelle Charakteristika des Patienten angepasst sind, oder Chirurgen nächste Operationsschritte nahelegen (Navarrete-Welton & Hashimoto, 2020; Pan et al., 2019; Sheng et al., 2019).

Die hier beschriebenen Systeme werden üblicherweise zum Bereich der KI in der Medizin gezählt. Zur genaueren Differenzierung innerhalb dieses großen, heterogenen und ungenau definierten Feldes werden die Anwendungen auch dem Gebiet der KI-basierten klinischen Entscheidungsunterstützung zugeordnet und etwa als *artificial intelligence-based clinical decision support systems* (AI-CDSS) bezeichnet (Shaikh et al., 2021; Yang et al., 2020).[2,3] Unter dieser Bezeichnung – AI-CDSS – werden die Systeme auch im Folgenden behandelt.

1.1 Medizinethischer Forschungsstand

Angesichts der zunehmenden Bedeutung und des transformativen Charakters der AI-CDSS ist es wenig verwunderlich, dass sich die medizinethische Forschung mit diesen Systemen bereits in vielerlei Hinsicht beschäftigt hat. Eine grundlegende Frage betrifft

[2] Zur Bezeichnung der Systeme wird auch der Begriff ‚*clinical decision support system*' (CDSS) verwendet (ZEKO, 2021). Das Feld der CDSS ist jedoch weit und heterogen: Unter CDSS fallen auch Anwendungen, die nicht dem Bereich der KI zuzuordnen sind (Kwan et al., 2020; Wasylewicz & Scheepers-Hoeks, 2019; Whelton et al., 2018). Zur Gewährleistung begrifflicher Schärfe wird in der vorliegenden Arbeit daher der engere Begriff ‚AI-CDSS' verwendet.

[3] Die Systeme können wie folgt definiert werden: „AI-driven decision support systems (AI-DSS) take various patient data and information about clinical presentation as input, and provide diagnoses, predictions or treatment recommendations as output." (Braun et al., 2020, S.1)

dabei deren Funktion. Auch weil bei den Systemen die Grenze zwischen Entscheidungsunterstützung und -übernahme fließend ist, wird gefragt, ob diese nicht möglicherweise die klinische Entscheidungsfindung übernehmen und vielleicht sogar Ärzte ersetzen könnten (Zentrale Ethikkommission [ZEKO], 2021). Dem wird jedoch entgegengehalten, dass für das ärztliche Denken und Handeln genuin humane Fähigkeiten wie Empathie und ein ganzheitliches Verständnis des Patienten essentiell seien (Goldhahn et al., 2018; Karches, 2018). So ist man sich weithin darüber einig, dass KI-Systeme Ärzte – zumindest derzeit – nicht gänzlich ersetzen, sondern diese nur unterstützen und ergänzen können (Funer, 2021; Verghese et al., 2018).

Neben dieser grundsätzlichen Analyse der Möglichkeiten und Grenzen von AI-CDSS untersucht die medizinethische Forschung auch Chancen und Risiken von deren Verwendung. Mit den Systemen werden vielerlei Hoffnungen verbunden. So wird mit Blick auf die oft hohe Leistungsfähigkeit der KI-Anwendungen darauf hingewiesen, dass deren Nutzung das Potenzial besitze, die Qualität der Patientenversorgung zu verbessern (Deutscher Ethikrat, 2023; ZEKO, 2021).

Außerdem bestünde die Chance, dass Ärzte durch die Verwendung der KI-Systeme entlastet werden könnten, wodurch diese die so entstandenen zeitlichen Freiräume für eine Verbesserung der Arzt-Patient-Beziehung und eine empathischere Behandlung nutzen könnten (Topol, 2019). Hinsichtlich der Entlastung der Ärzte und der Verbesserung der Entscheidungsfindung weisen manche auch auf das ökonomische Potenzial der Systeme hin (Al Meslamani, 2023; Liu et al., 2018).

Weitere Chancen werden im gerechtigkeitsethischen Zusammenhang gesehen: Die Nutzung von KI-Systemen könne etwa die Behandlung seltener Erkrankungen verbessern und die Versorgungsqualität in benachteiligen Weltregionen steigern (Panesar et al., 2019; ZEKO, 2021).

Gegenüber diesen Chancen weist die medizinethische Forschung auch auf Risiken der AI-CDSS hin. Auch wenn die Systeme Ärzte nicht ersetzen, können sie dennoch deren klinische Entscheidungsfindung maßgeblich beeinflussen. So besteht die Befürchtung, dass fehlerhafte Systemergebnisse von Nutzern nicht erkannt werden, zu falschen

klinischen Entscheidungen führen und so die Gesundheit der Patienten beeinträchtigen könnten (Sutton et al., 2020).

Kommt es zu falschen Empfehlungen und Entscheidungen, werden Fragen nach der Verantwortung aufgeworfen. Ein Problem stellt in diesem Zusammenhang die Opazität vieler ML-Anwendungen dar: Man kann nicht immer nachvollziehen, wie diese zu einem bestimmten Ergebnis kommen (‚Black-Box-Systeme') (Antoniadi et al., 2021). Unter anderem, um bei der Nutzung von KI-Systemen Verantwortung klar zuschreiben zu können, verlangen daher manche, diese sollten ihre Ergebnisse erklären können (Floridi et al., 2018; Ursin et al., 2021). Shortliffe und Sepúlveda (2018) fordern sogar den Verzicht auf Black-Box-Systeme. Zwar gibt es einige technologische Möglichkeiten, Outputs nachvollziehbar zu machen (Schaaf et al., 2021). Die Reichweite dieser Erklärungen ist jedoch beschränkt. Weil das Beharren auf der Forderung nach Transparenz oder strikter Erklärbarkeit die Verwendung von Black-Box-Systemen verunmöglichen würde, argumentieren viele für einen differenzierten Umgang mit opaken Anwendungen, um deren Chancen nutzen zu können (Liedtke & Langanke, 2021; London, 2019).

Darüber hinaus gibt es eine Debatte darüber, inwiefern die Verwendung von AI-CDSS die Patientenautonomie unterminieren kann. So wird gefragt, ob dem Patienten das Recht eingeräumt werden sollte, der Nutzung von KI-Systemen zu widersprechen (Ploug & Holm, 2020b). Wenn KI-Systeme bei der Behandlungsentscheidung verwendet werden, wird außerdem eine besondere Gefahr bezüglich der Wahrung und Förderung der Patientenautonomie gesehen. Diese hängt damit zusammen, dass bei der therapeutischen Entscheidungsfindung die evaluativen Präferenzen der Patienten zu beachten sind. Da es für AI-CDSS generell eine Schwierigkeit darstellt, individuelle Werte, Überzeugungen und Wünsche in die Berechnungen einzubeziehen, weist etwa McDougall (2019) auf die Gefahr hin, dass Patientenpräferenzen bei einer KI-unterstützten Therapieentscheidung möglicherweise nicht hinreichend Beachtung finden. Dadurch könnte ein neuer Paternalismus in die Medizin einziehen, der schon überwunden geglaubt war. Denn heutzutage besteht grundsätzlich Einigkeit darüber, dass der therapeutische Entscheidungsprozess nicht vom Arzt alleine, sondern im Verbund mit dem Patienten durchzuführen ist (Charles et al., 1997; Moumjid et al., 2007).

Um eine partizipative Entscheidungsfindung zu ermöglichen, ist die Kommunikation zwischen Arzt und Patient zentral. Im Zusammenhang damit steht auch die Gefahr, dass

Ärzte vom Patienten abgelenkt werden könnten, wenn sie KI-Systeme verwenden (Arnold, 2021; Loder & Nicholas, 2018). Das könne negative Konsequenzen für die Qualität der Arzt-Patient-Beziehung nach sich ziehen (Triberti et al., 2020). Außerdem bestehe das Risiko, dass sich das Vertrauen des Patienten auf den Arzt wegen möglicher Vorbehalte gegenüber KI-Systemen verschlechtert (Funer, 2021).

Hinzu kommt das Diskriminierungsrisiko durch AI-CDSS. Diese können zu falschen Ergebnissen kommen, wenn sich die Daten, die bei der Nutzung in das System eingespeist werden, zu stark von den Daten unterscheiden, anhand derer es entwickelt wurde (Trainingsdaten). Falls bei der Entwicklung einer Anwendung Datensätze verwendet werden, die nicht ausgeglichen sind, sondern Verzerrungen beinhalten, kann diese bei der Behandlung von bestimmten Patientengruppen übermäßig häufig zu falschen Ergebnissen kommen (World Health Organization [WHO], 2021).

Im Hinblick auf die zahlreichen ethischen Implikationen von AI-CDSS lässt sich also feststellen, dass deren Auswirkungen ambivalent sind: Mit der Verwendung dieser Systeme sind sowohl Chancen als auch Risiken verbunden. Da die Nutzung einer Technologie in der Medizin prima facie ethisch geboten ist, wenn sie nachweislich die Patientenversorgung verbessern kann, sollte die Ethik auf Potenziale von AI-CDSS hinweisen, ethische Implikationen der Systeme analysieren und Hinweise zu deren Abwägung geben (Marckmann, 2003). Das Ziel muss es sein, die Anwendungen so zu nutzen, dass nach Möglichkeit deren Potenziale maximiert und gleichzeitig die damit verbundenen Risiken minimiert werden (Marckmann, 2016).

Hierzu ist es nötig, nicht nur einzelne ethische Implikationen der Systeme ins Auge zu fassen, sondern die Anwendungen möglichst umfassend ethisch zu bewerten. AI-CDSS stellen einen beträchtlichen Teil des übergeordneten Bereichs der KI in der Medizin dar und diese macht wiederum einen der größten Anwendungsfelder dieser neuen Technologie aus. So beziehen sich einige der Arbeiten aus dem Gebiet der KI-Ethik – der allgemeinen und der medizinischen im Speziellen – auch und teilweise sogar besonders auf die Systeme, die hier als AI-CDSS bezeichnet werden (Deutscher Ethikrat, 2023; Gómez-González et al., 2020; Topol, 2019; WHO, 2021). Da der Bereich der KI in der Medizin neben KI-Systemen zur Unterstützung ärztlicher Entscheidungen noch andere Anwendungen und Technologien beinhaltet, kann diese Forschungsliteratur aber nur bedingt als ethische Bewertung von AI-CDSS gelten.

Darüber hinaus gibt es auch einige Arbeiten, die – unter verschiedenen Bezeichnungen – insbesondere KI-Systeme zur Unterstützung ärztlicher Entscheidungen in der Patientenversorgung ethisch behandeln und bewerten (Braun et al., 2020; Grote & Berens, 2020; Nsier, 2023; Sikma et al., 2020; Smith et al., 2024; ZEKO, 2021). Teilweise ist diese Literatur aber auf einen bestimmten medizinischen Bereich, beispielsweise auf die Kinderheilkunde, beschränkt (Nsier, 2023). Des Weiteren ist der Umfang vieler ethischer Bewertungen begrenzt; somit werden nicht alle ethisch relevanten Aspekte von AI-CDSS in der Tiefe behandelt (Funer, 2021; Smith et al., 2024). Außerdem gehen viele Arbeiten – wohl auch aufgrund ihrer Kürze – nur begrenzt systematisch und transparent vor (Grote & Berens, 2020; ZEKO 2021). Gerade zur Bewertung einer ethisch implikationsreichen und nicht unumstrittenen Technologie kommt der Nachvollziehbarkeit der Ausführungen aber große Bedeutung zu. Demnach lässt sich ein Bedarf an einer möglichst umfassenden, systematischen und nachvollziehbaren Bewertung von AI-CDSS feststellen.

1.2 Forschungsfragen

In der vorliegenden Arbeit sollen verschiedene Forschungsfragen beantwortet werden: Welche ethischen Implikationen haben AI-CDSS? Wie ist deren Einsatz übergreifend ethisch zu bewerten? Welche Empfehlungen ergeben sich daraus? Zur Beantwortung dieser Fragen wird auf die Ethische Bewertungsmatrix für eHealth-Anwendungen (EBEA) zurückgegriffen (Marckmann, 2016). Das Ziel der Arbeit besteht darin, einen Beitrag zur ethisch verantwortbaren Entwicklung und Anwendung der Systeme zu leisten.

1.3 Aufbau der Arbeit

Zunächst wird die in dieser Arbeit verwendete Methodik dargestellt (2. Kapitel). Diese sieht vor, den Gegenstand der ethischen Bewertung – AI-CDSS – genauer zu beschreiben (3. Kapitel). Anschließend werden die Systeme anhand von zwölf ethischen Beurteilungskriterien bewertet (4. Kapitel). Im 5. Kapitel werden die Einzelbewertungen zusammengeführt und gegeneinander abgewogen. Den Abschluss der Arbeit bilden ethisch begründete Empfehlungen für die Entwicklung und Nutzung von AI-CDSS sowie für die Schaffung geeigneter Rahmenbedingungen (Kapitel 6).

2. Methodik

Wie im vorangegangenen Kapitel erklärt, soll die vorliegende Arbeit AI-CDSS systematisch und nachvollziehbar bewerten. Dafür ist zunächst eine passende Methodik auszuwählen (vgl. Kapitel 2.1), die es anschließend zu erläutern gilt (vgl. Kapitel 2.2). Danach wird die Methodik für die Eigenheiten der vorliegenden Arbeit spezifiziert (vgl. Kapitel 2.3) und in der angepassten Form dargestellt (vgl. Kapitel 2.4).

2.1 Instrumente zur ethischen Evaluation digitaler medizinischer Technologien

Um für die ethische Bewertung von AI-CDSS die am besten geeignete Methodik auszuwählen, sind zunächst die verschiedenen in Frage kommenden Bewertungsinstrumente darzulegen. Da es sich bei den Systemen um eine Gruppe von digitalen Anwendung handelt, die innerhalb der Medizin Verwendung finden, werden im Folgenden nur solche ethischen Bewertungsinstrumente aufgeführt, die für die Bewertung digitaler medizinischer Technologien entwickelt wurden. Vier Instrumente zur ethischen Evaluation digitaler Technologien im Gesundheitswesen identifizierten Behrens et al. (2021) in einer systematischen Literaturrecherche. Aufgrund von deren systematischer Vorgehensweise und deren Aktualität bietet es sich an, in der vorliegenden Arbeit eines dieser Evaluationsinstrumente zu verwenden. Damit aus dieser Gruppe diejenige Methodik *begründet* ausgewählt werden kann, die für die Bewertung von AI-CDSS am besten geeignet ist, werden diese Bewertungsinstrumente nachfolgend in Anlehnung an Behrens et al. (2021) jeweils kurz erläutert:

1. Das **Model for Assessment of Telemedicine (MAST)** wurde zur Bewertung von telemedizinischen Anwendungen entwickelt und verfolgt einen dreistufigen Aufbau, bestehend aus vorausgehenden Überlegungen, multidisziplinärem Assessment und Überprüfung der Übertragbarkeit (Kidholm et al., 2012).

2. Das Instrument **MEESTAR** ist zur Bewertung von altersgerechten Assistenzsystemen entstanden. Man kann es jedoch auch verwenden, um andere Technologien zu evaluieren. MEESTAR sieht die Veranstaltung eines Workshops vor, in dem potenzielle Probleme identifiziert und evaluiert werden, die bei der Verwendung der zu bewertenden Technologie auftreten könnten. Als

Charakteristikum dieses Instrumentes kann der Fokus auf die Nutzerperspektive beschrieben werden (Manzeschke, 2015).

3. Die **mHealth Assessment Guidelines (first draft)** wurden im Auftrag der Europäischen Kommission entwickelt. Dieses Instrument fokussiert besonders die Beurteilung von Risiken. Darüber hinaus wird die Tauglichkeit der betreffenden Technologie explizit bewertet (Ruck et al., 2016).

4. Die **Ethische Bewertungsmatrix für eHealth-Anwendungen** stellt eine Spezifizierung einer Methodik dar, die im Rahmen der Bewertung medizinischer Expertensysteme entwickelt wurde (Marckmann, 2003). Die Matrix besteht aus dreizehn Bewertungskriterien, die vorwiegend aus dem Bereich der Technik- und der Medizinethik stammen (Marckmann, 2016).

Die oben genannten Instrumente weisen zur ethischen Bewertung von AI-CDSS unterschiedliche Vor- und Nachteile auf. Diese sollen im Anschluss an Behrens et al. (2021) nachfolgend in gebotener Kürze erörtert werden.

Bezüglich MAST ist besonders die Beachtung der Nutzerperspektive als Stärke zu erwähnen. Auch die Vielseitigkeit – dieses Evaluationsinstrument bezieht sich unter anderem auch auf ökonomische und organisatorische Aspekte – ist ein Vorteil von MAST. Zu einer *ethischen* Evaluation eignet sich das Instrument jedoch weniger, da es „nicht originär auf ethische Aspekte“ (Behrens et al., 2021, S. 562) abzielt. Als Methodik der vorliegenden *ethischen* Bewertung von AI-CDSS wird MAST darum nicht verwendet.

Die mHealth Assessment Guidelines zeichnen sich dadurch aus, dass im Rahmen dieser Methodik verschiedene Gruppen an der Bewertung beteiligt werden, die mit der Technologie verbunden sind. Wie MAST, so eignen sich auch die mHealth Assessment Guidelines nur bedingt für eine *ethische* Bewertung, weil der Fokus hier eher „technisch-funktionell[…]“ (Behrens et al., 2021, S. 569) ist. Außerdem kann die Gruppe der AI-CDSS höchstens teilweise dem Bereich von mHealth zugeordnet werden.

Ein Vorteil von MEESTAR besteht in der Offenheit dieses Instruments für verschiedene evaluative Präferenzen und Weltanschauungen. Insofern es ausdrücklich über altersgerechte Assistenzsysteme hinaus auch zur Evaluation weiterer Technologien angewandt werden kann, ist es zwar grundsätzlich für die ethische Bewertung von AI-CDSS geeignet. Problematisch ist allerdings, dass nicht konkret beschrieben wird, wie

die Evaluationsergebnisse der Arbeitsgruppen zusammengeführt werden sollen (Behrens et al., 2021).

Die EBEA ist eine genuin ethische Bewertungsmatrix und mit ihren dreizehn Kriterien vielseitig – das ist gerade in Anbetracht der Heterogenität von AI-CDSS vorteilhaft:

> „Insgesamt stellt die Ethische Bewertungsmatrix für eHealth-Anwendungen ein umfassendes Instrument zur ethischen Evaluation dar. Die Bewertungskriterien ermöglichen eine detaillierte Betrachtung verschiedener technischer Voraussetzungen und eine Abwägung der jeweiligen Relevanz von Beurteilungskriterien. Auf dieser Basis können fest eingeplante Empfehlungen erarbeitet werden.“ (Behrens et al., 2021, S. 567)

Weiterhin ist der klare, systematische Aufbau der EBEA gewinnbringend für die vorliegende Arbeit. Die vielfältigen ethischen Implikationen von AI-CDSS können somit systematisiert werden. Außerdem eignet sich die EBEA besonders für die ethische Bewertung von KI-Systemen zur Unterstützung ärztlicher Entscheidungen, weil die Wurzeln dieses Instruments in einer ethischen Evaluation von medizinischen Expertensystemen liegen, die historisch als Vorläufer der AI-CDSS betrachtet werden können (Marckmann, 2003). Für die Verwendung der EBEA in der vorliegenden Arbeit spricht außerdem, dass diese Methodik sich bereits für die ethische Bewertung verschiedener medizinischer Technologien als leistungsfähig erwiesen hat (Groß & Schmidt, 2018; Marckmann, 2020). Zwar sieht die EBEA es nicht vor, Nutzer oder Patienten etwa durch Interviews direkt in die Bewertung zu integrieren. Weil AI-CDSS in heterogenen, sich stark voneinander unterscheidenden Kontexten verwendet werden, erscheint es aber ohnehin fragwürdig, ob eine Befragung von Nutzern und Patienten für die Bewertung der Systeme sinnvoll und machbar wäre. Wenn etwa AI-CDSS in einem spezifischen Anwendungsfeld ethisch bewertet werden sollen, könnte sich hierfür ein Instrument anbieten, das die Perspektive von diesen direkt einbezieht.

Somit kann festgestellt werden, dass sich für eine möglichst umfassende, systematische und nachvollziehbare ethische Bewertung von AI-CDSS die EBEA am besten eignet.

2.2 Ethische Bewertungsmatrix für eHealth-Anwendungen

Eine ethische Untersuchung, die sich methodisch auf die EBEA stützt, soll nicht nur ethische Implikationen erläutern und moralische Fragen darlegen, sondern sie soll im

Sinne der normativen oder präskriptiven Ethik auch Hinweise zur Abwägung einzelner ethischer Aspekte geben und Empfehlungen für die ethisch vertretbare Verwendung von eHealth-Anwendungen aussprechen (Marckmann, 2003, 2016). Diese ethische Bewertung und die ethisch begründeten Empfehlungen bedürfen einer normativen Begründung. In Kapitel 2.2.2 wird diese erläutert. Davor stellt das Kapitel 2.2.1 das methodische Vorgehen der EBEA dar.

2.2.1 Methodisches Vorgehen

Zur Bewertung von eHealth-Anwendungen sieht die EBEA folgendes Vorgehen vor:

1. **Beschreibung der Technologie** (vgl. Kapitel 3): In einem ersten Arbeitsschritt ist die betreffende Technologie zu erläutern. Im Zuge dessen soll diese etwa hinsichtlich ihres Ziels, ihrer Funktionsweise und ihres Anwendungsbereichs charakterisiert werden.

2. **Spezifizierung der Beurteilungskriterien** (vgl. Kapitel 2.3): Die Beurteilungskriterien der EBEA sind nicht starr, sondern sollen an Spezifika der zu untersuchenden Technologie angepasst werden.

3. **Anwendung der einzelnen Beurteilungskriterien** (vgl. Kapitel 4): Im Rahmen einer Einzelbewertung werden die Beurteilungskriterien auf die Technologie angewandt.

4. **Synthese** (vgl. Kapitel 5): In einer übergreifenden Bewertung werden die Einzelbewertungen zusammengebracht und gegeneinander abgewogen.

5. **Empfehlungen** (vgl. Kapitel 6): Schließlich werden aus der Bewertung ethisch begründete Empfehlungen für die Entwicklung und Anwendung der Technologie abgeleitet.

6. **Monitoring**: Im zeitlichen Verlauf ist zu überprüfen, ob die getroffene ethische Bewertung (noch) korrekt ist. Gegebenenfalls ist diese entsprechend zu modifizieren. Innerhalb der vorliegenden Arbeit kann das Monitoring freilich noch nicht durchgeführt werden. Die Forschungsergebnisse sind im weiteren zeitlichen Verlauf immer wieder auf ihre Aktualität hin zu überprüfen. Im vorliegenden Fall dürfte das besonders deshalb notwendig werden, weil sich die

dynamischen Entwicklungen im Bereich der KI in der Medizin voraussichtlich in nächster Zeit fortsetzen dürften.

2.2.2 Normative Begründung der Bewertung

Was das normative Fundament angeht, so ist die EBEA dem problemorientierten Kohärentismus zuzuordnen. Deren Kriterien stammen demnach nicht nur aus einer moralphilosophischen Konzeption. Vielmehr ist sie offen für Überlegungen verschiedener moralphilosophischer Entwürfe (etwa Tugendethik, deontologische Ethik und Umweltethik). Ziel ist es dabei, die Überlegungen in einen *kohärenten* Zusammenhang zu bringen (Marckmann, 2016).

Der normative Gehalt der EBEA wird vor allem aus zwei Bereichsethiken gewonnen, in deren Schnittfeld die Bewertung von eHealth-Anwendungen liegt (Behrens et al., 2021). Da eHealth-Anwendungen dem Gesundheitsbereich beziehungsweise der Medizin zuzuordnen sind, ist es nachvollziehbar, dass diese ethische Bewertungsmatrix auf die normative Basis der Medizinethik zurückgreift. Eine zweite Bereichsethik, aus der normative Inhalte in die EBEA eingehen, ist die Technikethik beziehungsweise die Technikbewertung. Bevor in Kapitel 2.2.3 die Bewertungsmatrix dargestellt wird, in der die normativen Kriterien zusammengefasst sind, sollen die beiden einschlägigen Bereichsethiken kurz vorgestellt werden.

Hinsichtlich der Medizinethik bezieht sich die EBEA explizit auf das heutzutage einflussreichste medizinethische Modell: die Prinzipienethik nach Beauchamp und Childress (2019). Im Gegensatz zu den höchsten Moralprinzipien anderer ethischer Konzeptionen sind die hier beschriebenen vier bioethischen Prinzipien auf der mittleren ethischen Abstraktionsebene angesiedelt. Sie sind prima facie gleichberechtigt und nicht rein formal, sondern bis zu einem gewissen Grad mit Inhalt gefüllt.

Im Folgenden sollen die Prinzipien kurz erläutert werden: Das Prinzip des Wohltuns gebietet zunächst, dass medizinische Maßnahmen das Wohl des Patienten fördern sollen. Gewissermaßen als negative Wendung dessen kann man das Prinzip des Nichtschadens beschreiben. Auch wenn sich in der Medizin die Entstehung von Schaden nicht immer vermeiden lässt, ist dieser möglichst zu minimieren. Darüber hinaus soll die Autonomie des Patienten respektiert und gefördert werden (Prinzip des Respekts der Patientenautonomie). Schließlich sind auch Auswirkungen gegenüber Dritten zu beachten (Prinzip der Gerechtigkeit).

Neben der Medizinethik ist eine weitere normative Quelle der EBEA die Technikethik. Charakteristisch für die ethische Technikbewertung beziehungsweise die Technikfolgenabschätzung sind drei Funktionen: die Erforschung der Folgen einer Technologie, deren Bewertung und schließlich die Beratung (Marckmann, 2003).

Die technikethische Basis der EBEA steht unter besonderem Einfluss der ethischen Matrix von Ott (1997), die mit „1) instrumentellen Urteilen, 2) Werturteilen, 3) Rechtsnormen, 4) Moralnormen, 5) technikethischen Praxisnormen und 6) ethischen Prinzipien“ (Marckmann, 2016, S. 91) – wie die EBEA – kategoriell unterschiedliche Beurteilungskriterien in sich vereint.

2.2.3 Darstellung der Bewertungsmatrix

Nachdem die normative Begründung der ethischen Bewertungsmatrix für eHealth-Anwendungen erläutert wurde, kann diese dargestellt werden. Davor ist noch kurz auf deren zwei Funktionen einzugehen: Zum einen dient sie als normative Begründung für die ethische Technikbewertung und für die ethisch begründeten Empfehlungen zur Entwicklung und Anwendung von eHealth-Anwendungen. Zum anderen hilft die Matrix dabei, ethische Implikationen der betreffenden Anwendungen einzuordnen und zu systematisieren. Somit kommt ihr auch die Funktion einer „‚Suchmatrix‘“ (Marckmann, 2016, S. 93) zu.

Aus den beschriebenen normativen Grundlagen werden die in Tabelle 1 dargestellten Beurteilungskriterien abgeleitet.

Tabelle 1:

Ethische Bewertungsmatrix zur Bewertung von eHealth-Anwendungen

Bewertungskriterium	Ethische Begründung
Funktionsfähigkeit - Zielsetzung der Technologie - Grad der Zielerreichung („Wirksamkeit“) - Technische Effizienz	Zweck-Mittel-Rationalität; Prinzip des Nichtschadens; Prinzip des Wohltuns
Mögliche Alternativen	Zweck-Mittel-Rationalität
Nutzenpotenzial für die Patientinnen und Patienten - Verbesserung von Mortalität, Morbidität und Lebensqualität - Validität (Evidenzgrad)	Prinzip des Wohltuns
Schadenspotenzial für die Patientinnen und Patienten - Sicherheit, geringe Fehleranfälligkeit - Belastungen & gesundheitliche Risiken - Validität (Evidenzgrad)	Prinzip des Nichtschadens
Wahrung der Integrität der Arzt-Patient-Beziehung	Respekt der Autonomie; Prinzip des Wohltuns
Wahrung bzw. Förderung der Patientenautonomie - Möglichkeit der informierten Einwilligung - Auswirkung auf Entscheidungsfreiheit - Förderung der Gesundheitsmündigkeit	Respekt der Autonomie
Schutz vertraulicher Patientendaten vor unautorisiertem Zugriff (Datenschutz)	Informationelle Selbst-bestimmung; Respekt der Autonomie
Sicherheit vor systembedingtem Verlust der Integrität von Patientendaten (Datensicherheit)	Prinzip des Nichtschadens
Effizienz - (inkrementelles) Kosten-Nutzen-Verhältnis - Validität der Effizienzmessung	Verteilungsgerechtigkeit bei knappen Ressourcen; Zweck-Mittel-Rationalität
Wahrung der ärztlichen Entscheidungsautonomie	Prinzip des Wohltuns
Auswirkung auf die ärztliche Entscheidungskompetenz	Prinzip des Nichtschadens; Prinzip des Wohltuns
Zuschreibbarkeit von Verantwortung beim Einsatz der Technologie	Prinzip des Nichtschadens
Gerechtigkeit - Nicht-diskriminierender Zugang zur Technologie - Verteilung der gesundheitlichen Nutzen- und Schadenspotenziale	Prinzip der Gerechtigkeit

Anmerkung. Übernommen aus „Ethische Aspekte von eHealth“ von G. Marckmann, in F. Fischer und A. Krämer (Hrsg.), *eHealth in Deutschland* (S. 92), 2016, Springer (https://doi.org/10.1007/978-3-662-49504-9_4).

Am Anfang der Bewertungsmatrix stehen mit ‚Funktionsfähigkeit' und ‚Nutzenpotenzial für die Patientinnen und Patienten' zwei Kriterien, die primär anhand von empirischen Evaluationsstudien zu beurteilen sind. Die Funktionsfähigkeit wird hier mit ‚Wirksamkeit' verbunden. Es geht also darum, inwiefern die Nutzung einer bestimmten Technologie die damit verbundenen Ziele erreichen kann.

Mit der Funktionsfähigkeit ist auch das folgende Kriterium ‚**mögliche Alternativen**' verbunden: Meist sind bestimmte Ziele nicht nur durch die Verwendung einer bestimmten eHealth-Anwendung zu erreichen. Der Zweck-Mittel-Rationalität entsprechend sollten die verschiedenen Möglichkeiten miteinander verglichen werden.

‚**Nutzenpotenzial für die Patientinnen und Patienten**' und ‚**Schadenspotenzial für die Patientinnen und Patienten**' fungieren jeweils als eigenständiges ethisches Bewertungskriterium. Die beiden Kriterien sind direkt nach den möglichen Alternativen zu Beginn der ethischen Bewertung zu untersuchen: Wenn eine Technologie keinen Nutzen oder ein unverhältnismäßiges Nutzen-Schaden-Verhältnis aufweist, ist es unvertretbar, diese zu verwenden. Die Untersuchung der restlichen Kriterien der EBEA ist in diesem Fall hinfällig.

Darüber hinaus fordert das Prinzip des Respekts der Patientenautonomie, zu untersuchen, ob die **Wahrung bzw. Förderung der Patientenautonomie** gewährleistet werden kann.

Unter anderem vom Prinzip des Respekts der Patientenautonomie lässt sich auch das Kriterium des **Datenschutzes** ableiten. Davon zu trennen ist ‚**Datensicherheit**'. Dieses Bewertungskriterium bezieht sich auf die Vermeidung von systembedingtem Datenverlust.

In einem gemeinschaftlich finanzierten Gesundheitssystem gilt es aus Gründen der Gerechtigkeit auch die **Effizienz** der nicht selten kostspieligen eHealth-Anwendungen zu untersuchen. Die EBEA sieht unter dem Aspekt der Effizienz unter anderem eine Auseinandersetzung mit dem Kosten-Nutzen-Verhältnis der betreffenden Technologie vor.

Die folgenden beiden Kriterien sind miteinander verbunden. Besonders hinsichtlich zunehmend leistungsfähiger Technologie stellt sich die Frage nach der **Entscheidungsautonomie des Arztes**. Trifft der Arzt immer weniger Entscheidungen selbst, so mag seine **Entscheidungskompetenz** darunter leiden.

Im Zusammenhang mit der ärztlichen Autonomie steht auch die Fähigkeit des Arztes, Verantwortung zu übernehmen. Es entspricht daher einer inneren Logik, dass in der EBEA anschließend die Auseinandersetzung mit der ‚**Zuschreibbarkeit von Verantwortung**' folgt, bevor das Kriterium der **Gerechtigkeit** den Abschluss der Matrix markiert.

Es lässt sich feststellen, dass die Kriterien der EBEA keinen auf den ersten Blick erkennbaren systematischen Zusammenhang aufweisen. Das liegt zum einen in der Spezifizierung, zum anderen im normativen Kohärentismus begründet: Die Bewertungsmatrix umfasst sowohl „moralisch-technische Urteile, als auch Werturteile und moralische Normen" (Marckmann, 2016, S. 91).

2.3 Spezifizierung der Bewertungskriterien

Die Methodik, die dieser Arbeit zugrunde liegt, fordert explizit, die Bewertungsmatrix dem problemorientierten Kohärentismus entsprechend für die ethische Evaluation der jeweiligen eHealth-Anwendung zu spezifizieren. Dafür ist es naheliegend, auf die KI-Ethik Bezug zu nehmen, weil diese AI-CDSS vornehmlich untersucht.

Eine der größten ethischen Herausforderungen von KI-Anwendungen ist damit verbunden, dass man bei einigen davon nicht erklären kann, wie diese zu bestimmten Ergebnissen kommen (Opazität, Black-Box-Problem). Es wird daher oft gefordert, dass KI-Systeme erklärbar sein sollen. Innerhalb der KI-Ethik nimmt Erklärbarkeit sogar teilweise die Rolle eines Prinzips ein (Floridi et al., 2018; Ursin et al., 2021). In Anbetracht dessen scheint es sich für die ethische Bewertung von AI-CDSS auf den ersten Blick anzubieten, ‚Erklärbarkeit' als eigenes ethisches Bewertungskriterium in die Matrix zu integrieren.

Zurecht wird jedoch davor gewarnt, durch einen zu starken Fokus auf die Erklärbarkeit von KI-Systemen illegitime Doppelstandards in die Medizinethik einzuführen: Die Situation, dass Ärzte nicht erklären können, wie sie zu bestimmten Ergebnissen kommen, wird nicht erst durch die Verwendung von opaken KI-Systemen hervorgerufen. Nicht selten ist die ärztliche Entscheidungsfindung auch ohne den Einsatz von KI-Anwendungen nicht (gänzlich) erklärbar (London, 2019). Daher verzichtet die vorliegende Arbeit darauf, Erklärbarkeit als eigenständiges ethisches Bewertungskriterium zu verwenden. Erklärbarkeit ist dennoch ein relevanter Aspekt, der

bei einer ethischen Auseinandersetzung mit AI-CDSS nicht ausgeklammert werden kann und deshalb im deskriptiven Teil dieser Arbeit behandelt wird (vgl. Kapitel 3.4.1). Die normativ relevanten Aspekte, die mit Erklärbarkeit verbunden sind, werden im bewertenden Teil der Arbeit jeweils einzeln untersucht und evaluiert (vgl. Kapitel 4.4.2.2, 4.5.2, 4.5.3, 4.8 und 4.9.1)

Weiterhin ist es sinnvoll, die Matrix in Bezug auf das Kriterium der Datensicherheit zu modifizieren. Dieses fokussiert den Schutz vor systembedingtem Datenverlust (Marckmann, 2016). Gewiss ist bei der Anwendung von AI-CDSS wie bei der Verwendung von wohl jeder IT-Technologie darauf zu achten, die Daten vor systembedingtem Verlust zu sichern. Wie sich der Umgang mit der Datensicherheit im Zusammenhang mit KI-Systemen zur Unterstützung ärztlicher Entscheidungen im Vergleich zur Entwicklung und zum Gebrauch von anderen eHealth-Anwendungen unterscheiden soll, ist aber nicht erkennbar. Bei der Bewertung von AI-CDSS einzeln auf die Datensicherheit einzugehen, erscheint also nicht erforderlich.

Teilweise modifiziert wurde außerdem die Reihenfolge der Bewertungskriterien. Aufgrund der besonderen Bedeutung des Autonomieprinzips steht das Kriterium ‚Wahrung und Förderung der Patientenautonomie' weiter am Anfang der hier verwendeten Matrix als ursprünglich in der EBEA (Schmietow & Marckmann, 2019) (vgl. Kapitel 4.5). Insofern es sich bei der Autonomie des Arztes ebenfalls um eine Form der Freiheit handelt, bietet es sich thematisch an, das Kriterium ‚Wahrung der ärztlichen Entscheidungsautonomie' direkt daran anschließend zu behandeln (vgl. Kapitel 4.6) – gefolgt von ‚Auswirkung auf die ärztliche Entscheidungskompetenz' (vgl. Kapitel 4.7). Inhaltlich damit verbunden ist die Frage nach der Verantwortung für Entscheidungen, die mithilfe von AI-CDSS getroffen wurden. Daher schließt sich in Kapitel 4.8 das Kriterium ‚Zuschreibbarkeit von Verantwortung bei der Nutzung von AI-CDSS' an. Die noch übrigen drei Kriterien der EBEA werden wiederum in der ursprünglichen Reihenfolge der EBEA untersucht (vgl. Kapitel 4.9 bis 4.12).

2.4 Ethische Beurteilungskriterien zur Bewertung von AI-CDSS

Aus den oben beschriebenen Modifikationen der EBEA ergibt sich die in Tabelle 2 dargestellte Matrix.

Tabelle 2:

Ethische Bewertungsmatrix zur Bewertung von AI-CDSS

Ethisches Beurteilungskriterium	Ethische Begründung
Funktionsfähigkeit (vgl. Kapitel 4.1)	Zweck-Mittel-Rationalität; Prinzip des Nichtschadens; Prinzip des Wohltuns
Mögliche Alternativen (vgl. Kapitel 4.2)	Zweck-Mittel-Rationalität
Nutzenpotenzial für die Patientinnen und Patienten (vgl. Kapitel 4.3)	Prinzip des Wohltuns
Schadenspotenzial für die Patientinnen und Patienten (vgl. Kapitel 4.4)	Prinzip des Nichtschadens
Wahrung und Förderung der Patientenautonomie (vgl. Kapitel 4.5)	Respekt der Autonomie
Wahrung der ärztlichen Entscheidungsautonomie (vgl. Kapitel 4.6)	Prinzip des Wohltuns
Auswirkung auf die ärztliche Entscheidungskompetenz (vgl. Kapitel 4.7)	Prinzip des Nichtschadens; Prinzip des Wohltuns
Zuschreibbarkeit von Verantwortung (vgl. Kapitel 4.8)	Prinzip des Nichtschadens
Wahrung der Integrität der Arzt-Patient-Beziehung (vgl. Kapitel 4.9)	Respekt der Autonomie; Prinzip des Wohltuns
Datenschutz und Datenverfügbarkeit (vgl. Kapitel 4.10)	Informationelle Selbstbestimmung; Respekt der Autonomie; Prinzip des Nichtschadens; Prinzip des Wohltuns
Effizienz (vgl. Kapitel 4.11)	Verteilungsgerechtigkeit bei knappen Ressourcen; Zweck-Mittel-Rationalität
Gerechtigkeit (vgl. Kapitel 4.12)	Prinzip der Gerechtigkeit

Anmerkung. In Anlehnung an „Ethische Aspekte von eHealth“ von G. Marckmann, in F. Fischer und A. Krämer (Hrsg.), *eHealth in Deutschland* (S. 92), 2016, Springer (https://doi.org/10.1007/978-3-662-49504-9_4).

3. KI-Systeme zur Unterstützung ärztlicher Entscheidungen

Bevor im 4. Kapitel die ethische Bewertung von AI-CDSS erfolgen kann, gilt es, diese Systeme im vorliegenden Kapitel näher zu beschreiben. Hierzu sollen zunächst deren technologische Grundlagen dargelegt werden (vgl. Kapitel 3.1). Um ein Bild von den konkreten Anwendungsmöglichkeiten zu zeichnen, werden darauf einzelne beispielhafte AI-CDSS erläutert (vgl. Kapitel 3.2). Aus dem bis dahin Dargestellten werden die Zieldefinition der Systeme (vgl. Kapitel 3.3) und deren charakteristische Eigenschaften (vgl. Kapitel 3.4) abgeleitet.

3.1 Technologische Grundlagen

Da ethische Implikationen bestimmter technologischer Anwendungen maßgeblich von der Beschaffenheit der zugrundeliegenden Technologien abhängen, sollen in diesem Kapitel die technologischen Grundlagen von AI-CDSS erläutert werden. Es ist klar, dass solche komplexen Systeme in der vorliegenden Arbeit in technologischer Hinsicht nicht erschöpfend behandelt werden können.[4] Die folgenden Ausführungen beschränken sich darum auf die technologischen Aspekte, die für das Verständnis der später behandelten ethischen Implikationen notwendig sind.

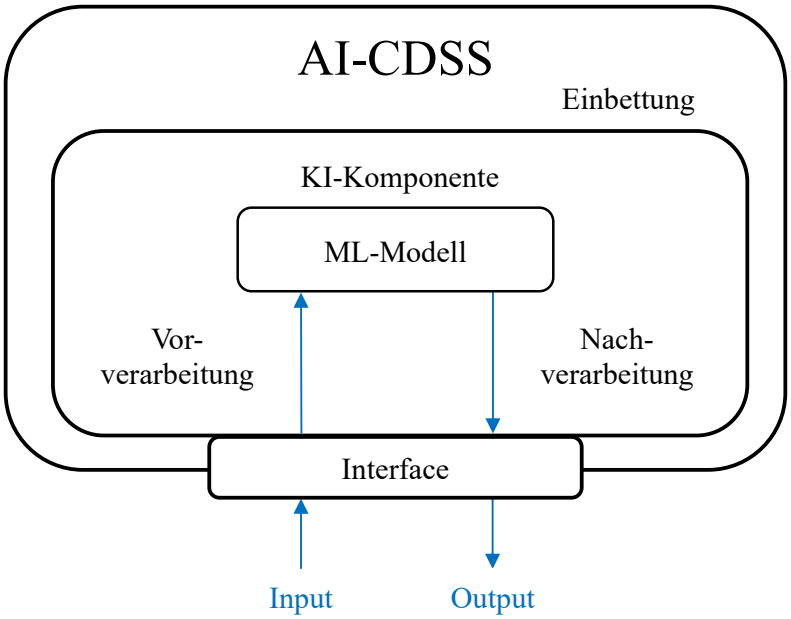
3.1.1 Aufbau eines AI-CDSS

Auch wenn die Gruppe der AI-CDSS von Heterogenität geprägt ist, so sind die unterschiedlichen Systeme, insoweit es sich bei diesen um KI-Anwendungen handelt, doch in ihrem grundlegenden Aufbau vereint. Dieser soll im vorliegenden Unterkapitel anhand der Abbildung 1 erläutert werden.

[4] Eine Einführung in das Gebiet der Künstlichen Intelligenz aus technologischer Sicht bieten etwa Russell und Norvig (2012).

Abbildung 1:

Aufbau eines AI-CDSS



Anmerkung. In Anlehnung an *Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz: KI-Prüfkatalog* (S. 18), von M. Poretschkin et al., 2021, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS (https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/KI-pruefkatalog/202107_KI-Pruefkatalog.pdf). Copyright durch das Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, 2021.

Im Zentrum eines AI-CDSS steht die KI-Komponente. Diese besteht, wenn es sich um ein ML-System handelt, aus einem ML-Modell, flankiert von Einheiten zur Vor- und Nachverarbeitung von Daten. Die Entwicklung eines solchen Modells wird in Kapitel 3.1.2.2 erläutert.

Innerhalb des AI-CDSS ist die KI-Komponente in andere Softwarebestandteile eingebettet. Hierzu gehört unter anderem das Interface. Darüber kann der Nutzer Daten eingeben (Input), die die KI-Komponente anschließend in Berechnungen einbezieht. Schließlich gibt das System das Ergebnis (Output) an den Nutzer über das Interface aus (Poretschkin et al., 2021).

3.1.2 Machine Learning

Nachdem dargelegt wurde, dass die KI-Komponente beziehungsweise das ML-Modell den Kern eines AI-CDSS bildet, soll im vorliegenden Kapitel ML in den Blick genommen werden. Hierzu sollen zunächst Formen davon erläutert werden, die im Zusammenhang mit AI-CDSS relevant sind (vgl. Kapitel 3.1.2.1). Das ist auch deshalb erforderlich, weil diese verschiedenen Arten von ML mit unterschiedlichen ethischen Implikationen – etwa bezüglich des Schadenspotenzials (vgl. Kapitel 4.4) – einhergehen.

Dass anschließend noch in gebotener Kürze erläutert wird, wie ML-basierte AI-CDSS entwickelt werden (vgl. Kapitel 3.1.2.2), ist zum einen geboten, weil sich manche ethischen Implikationen der Systeme anhand des Entwicklungsprozesses besser verstehen lassen. Zum anderen ist es erforderlich, auf den Entwicklungsprozess von ML-Systemen einzugehen, da am Ende der vorliegenden Arbeit auch ethisch begründete Empfehlungen zur Entwicklung von AI-CDSS zu geben sind (vgl. Kapitel 6.1).

3.1.2.1 Formen von Machine Learning

Klassischerweise unterscheidet man drei Formen des Maschinellen Lernens: das überwachte, das unüberwachte und das bestärkende Lernen.

Für das **überwachte Lernen** wird ein Trainingsdatensatz benötigt, in dem jeder Eingangswert mit einem zugehörigen Ausgabewert versehen ist. Soll ein ML-Modell beispielsweise Bilder klassifizieren, muss zu jeder Aufnahme, mit dem das Modell trainiert wird, bereits eine Klassifikation – etwa die korrekte Diagnose – vorliegen. Diese Zuordnung wird auch als Grundwahrheit (*ground truth*) bezeichnet. Anhand dieser Daten ‚lernt' das ML-Modell: Es verändert sich, sodass dessen Outputs mit den vorgesehenen Ausgabewerten übereinstimmen (Datenschutzkonferenz [DSK], 2019; Plattform Lernende Systeme, 2019).

Im Gegensatz zum überwachten Lernen gibt es beim **unüberwachten Lernen** im Trainingsdatensatz keine vordefinierten Outputs, die das ML-Modell bei bestimmten Eingabewerten ausgeben soll. Stattdessen werden in den Eingabewerten, etwa anhand von ähnlichen Mustern, Gruppen identifiziert – ohne, dass diese gleich unter eine bestimmte Bezeichnung gefasst würden (DSK, 2019).

Eine dritte Form von ML, die teilweise auch als Unterform des überwachten Lernens bezeichnet wird, ist das **bestärkende Lernen** (*reinforcement learning*). Hierbei simuliert

ein Algorithmus verschiedene Vorgehensweisen, um ein bestimmtes Ziel in einer virtuellen Umgebung zu erreichen. Je nachdem, wie erfolgreich das Vorgehen ist, erhält der Algorithmus als Feedback Belohnungen oder Bestrafungen. Indem er nach der Maximierung der Belohnungen ‚strebt', verbessert sich der Algorithmus kontinuierlich (Neumuth, 2020; Russell & Norvig, 2012).

Eine Form von ML, die hier noch Erwähnung finden soll, ist das sogenannte **Deep Learning** (DL). Als Charakteristikum von DL kann man die strukturelle Ähnlichkeit zum menschlichen Lernen beschreiben: Es werden künstliche neuronale Netze verwendet, die besonders viele verdeckte Schichten haben. Deshalb neigen DL-Systeme zu Opazität (Plattform Lernende Systeme, 2019). Viele Fortschritte im Bereich von ML sind mit dieser Form davon verbunden. Auch für die Medizin wird DL als zukunftsträchtig aufgefasst (Zhou et al., 2021).

Eine besondere Herausforderung bezüglich ML ist mit dem Umstand verknüpft, dass die darauf basierenden Systeme anhand von Daten trainiert werden: Gibt es einen zu großen Unterschied zwischen den Trainingsdaten und den Daten, die bei der Nutzung in das ML-System eingespeist werden, kann es zu falschen Ergebnissen kommen (vgl. Kapitel 3.4.2 und 4.4.2.1). In diesem Fall spricht man von einer Verzerrung (Bias) des Trainingsdatensatzes.

3.1.2.2 Entwicklungsprozess eines ML-Systems

Der Entwicklungsprozess eines ML-Systems beginnt mit der Sammlung und Vorbereitung der Trainingsdaten. Darauf folgt das Training des ML-Modells. Den Abschluss bildet die Evaluierung und Validierung des Systems.

Es kann als bisheriger Standard bezeichnet werden, dass der Entwicklungsprozess vor der Anwendungsphase des ML-Systems gestoppt wird (*locked algorithm*). ML-Modelle können aber auch so verfasst sein, dass sie sich in der Praxis kontinuierlich weiterentwickeln (*continual learning*). Ein großer Vorteil von *continual learning* besteht in der hohen Anpassungsfähigkeit der darauf basierenden Systeme. Besser kontrollieren lässt sich die Qualität eines ML-Modells jedoch, wenn es in abgeschlossener Form vorliegt (Futoma et al., 2020; C. S. Lee & A. Y. Lee, 2020)

3.2 Vorstellung beispielhafter Anwendungen

Die ethisch relevanten Aspekte von AI-CDSS ergeben sich nicht einzig aus deren Funktionsweise. Bedeutend ist auch, in welchen Situationen, wie und wozu die KI-Systeme verwendet werden. Damit davon ein Eindruck entstehen kann, sollen im Folgenden einzelne beispielhafte Anwendungen vorgestellt werden. Um der Heterogenität der Systeme bei der späteren ethischen Bewertung durch analytische Trennschärfe gerecht werden zu können und um ein möglichst übersichtliches Bild von AI-CDSS zu zeichnen, werden die einzelnen Anwendungen nach den drei Anwendungsbereichen Diagnostik, Prognostik und Therapieempfehlung getrennt vorgestellt.[5]

3.2.1 Diagnostik

Wird eine Erkrankung nicht hinreichend frühzeitig oder gar nicht diagnostiziert, kann der betreffende Patient möglicherweise erst zu spät oder überhaupt nicht behandelt werden. In vielen Fällen ist deshalb die möglichst frühe Erkennung einer Krankheit wichtig für ein optimales Ergebnis der Behandlung.

Das gilt etwa für das maligne Melanom (Winkler et al., 2020). Herkömmlicherweise detektieren Dermatologen diesen Tumor durch die dermatoskopische Untersuchung auffälliger Hautläsionen (Deutsche Dermatologische Gesellschaft et al., 2020). Darüber hinaus wurden zur Detektion maligner Melanome KI-basierte Entscheidungsunterstützungssysteme entwickelt (Du-Harpur et al., 2020). Mit *Moleanalyzer pro* liegt ein System vor, das für die Benutzung in der Europäischen Union zugelassen ist (Winkler et al., 2020). In einer Evaluationsstudie zeigte sich, dass dieses AI-CDSS die diagnostische Treffsicherheit von Ärzten übertreffen kann, die bei der Detektion maligner Melanome wenig erfahren sind (Haenssle et al., 2020).

Besonders im Bereich der Tumorerkennung sind einige weitere Systeme vorhanden. Diese können Ärzte etwa dabei unterstützen, Mammakarzinome oder Lungenkarzinome zu detektieren (Baltzer, 2021; Sechopoulos et al., 2021; Svoboda, 2020).

[5] Die Einteilung von AI-CDSS nach diesen drei Bereichen kann als üblich gelten (Braun et al., 2020; Esmaeilzadeh, 2020; ZEKO, 2021).

Im Gebiet der Bilddiagnostik gibt es generell viele AI-CDSS. Das hängt nicht zuletzt damit zusammen, dass in diesem Feld hinreichend große Datenmengen für die Entwicklung von KI-Systemen zur Verfügung stehen. Es sind aber auch jenseits der Bilderkennung KI-Anwendungen zur Erkennung von Krankheiten vorhanden.

Als Beispiel kann hierfür ein System genannt werden, das Ärzte dabei unterstützen kann, Autismus-Spektrum-Störungen bei Kleinkindern zu diagnostizieren (Abbas et al., 2020). Es rekurriert einerseits auf Ergebnisse eines Fragebogens, den die Eltern des betreffenden Kindes ausfüllen, und andererseits auf Einschätzungen der behandelnden Ärzte. Außerdem bezieht sich das AI-CDSS auf eine Liste von Verhaltensauffälligkeiten, die aus Videoaufnahmen des Kindes hervorgehen. Es besteht die Hoffnung, durch die Verwendung des Systems Autismus bei Kindern früher zu erkennen. Bislang gelingt diese Diagnose erst recht spät: Der Median des Alters, in dem Autismus bei Kindern diagnostiziert wird, liegt bei 52 Monaten (Baio et al., 2018). Gleichzeitig ist die frühzeitige Detektion von Autismus besonders wichtig, um eine möglichst gute Entwicklung von Kindern mit Autismus gewährleisten zu können (Dawson et al., 2010; Dawson & Bernier, 2013). Daher birgt das System für diese Patienten ein Potenzial zur Verbesserung der Versorgungsqualität.

Die Detektion einer Erkrankung stellt keinen Selbstzweck dar. Vielmehr soll eine Diagnosestellung dazu dienen, einem Patienten die optimale Behandlung zukommen zu lassen. Dazu reicht es nicht aus, eine bestimmte Erkrankung nur festzustellen. Um eine passende Therapie auswählen zu können, ist die detektierte Pathologie näher zu beurteilen. Zur Charakterisierung von Krankheiten oder des Gesundheitszustands eines Patienten kann man verschiedene Parameter verwenden, deren Bestimmung teilweise recht aufwändig ist. Als Beurteilungskriterium der Leistungsfähigkeit des Herzens kann man beispielsweise die linksventrikuläre Ejektionsfraktion (LVEF) verwenden. Anhand der LVEF unterteilt man etwa die Herzinsuffizienz in drei Formen, die jeweils unterschiedlich zu behandeln sind (Berliner et al., 2020). Die LVEF kann man unter anderem am Patientenbett mithilfe der Echokardiographie abschätzen. Ein diesbezüglicher Nachteil besteht jedoch in der hohen Interobservervariabilität (Kusunose et al., 2018; Thavendiranathan et al., 2013). Eine Alternative bietet ein ML-System, das anhand von Echokardiogrammen die LVEF bestimmt und *at the point of care* verwendet werden kann. Was die diagnostische Genauigkeit betrifft, so geben die Entwickler an,

dass dieses AI-CDSS die LVEF beinahe so akkurat wie Ärzte bestimmt (Asch et al., 2021).

Darüber hinaus gibt es weitere Systeme, die die Charakterisierung von Krankheiten unterstützen. KI-Systeme kann man etwa auch zur Segmentierung von Lungentumoren oder zur Berechnung des Volumens von Prostatatumoren verwenden (Tian et al., 2017; Vivanti et al., 2015).

3.2.2 Prognostik

Neben der Diagnostik ist die Prognostik ein Anwendungsfeld vieler AI-CDSS. Was die KI-Systeme vorhersagen, ist vielfältig: Manche Anwendungen schätzen zum Beispiel die Wahrscheinlichkeit dafür ab, dass ein Patient sich eine bestimmte Krankheit zuzieht oder dass er verstirbt. Andere Systeme treffen Vorhersagen darüber, wie bestimmte körperliche Reaktionen eines Patienten wahrscheinlich ausfallen werden und wie dieser auf gewisse therapeutische Maßnahmen ansprechen wird.

Eine medizinische Fachdisziplin, in der die Abschätzung der Sterbe- beziehungsweise Überlebenswahrscheinlichkeit eines Patienten eine besondere Rolle spielt, ist die Onkologie: Welche therapeutische Option man ergreifen sollte und welche damit einhergehenden Nebenwirkungen und Belastungen man akzeptieren sollte, hängt schließlich auch davon ab, wie gefährlich die onkologische Erkrankung für den individuellen Patienten einzuschätzen ist. Die Überlebenswahrscheinlichkeit von Patienten mit kolorektalem Karzinom kann etwa anhand der Klassifizierung der *Union international contre le cancer* abgeschätzt werden (Deutsche Gesellschaft für Gastroenterologie, Verdauungs- und Stoffwechselkrankheiten et al., 2021).

Präziser als mithilfe dieser Einteilung kann man die Sterblichkeit von Patienten mit kolorektalem Karzinom unter Zuhilfenahme eines KI-Systems vorhersagen, das histologische Schnittbilder analysiert (Kather et al., 2019). Durch eine verbesserte Risikostratifizierung könnte man einerseits verhindern, dass Patienten eine zu aggressive Therapie erhalten und somit unnötigen Belastungen ausgesetzt werden. Außerdem besteht durch eine Optimierung der Risikostratifizierung die Chance, dass Patienten mit einer besonders hohen Sterbewahrscheinlichkeit eine hinreichend intensive Behandlung erfahren.

Darüber hinaus berechnen manche AI-CDSS die Wahrscheinlichkeit dafür, dass bei einer bestimmten Person in Zukunft eine spezifische Krankheit auftreten wird. Die Vorhersage der Erkrankung ist besonders dann nützlich, wenn sich diese nicht im Voraus merklich abzeichnet. Solch ein unvermitteltes Auftreten ist ein Charakteristikum von kardiovaskulären Ereignissen. Um deren Wahrscheinlichkeit abzuschätzen, kann man Risikoscores wie denjenigen verwenden, den das *American College of Cardiology* mit der *American Heart Association* entwickelt hat (Goff et al., 2014). Genauer als mithilfe dieses Risikoscores kann man kardiovaskuläre Ereignisse allerdings unter Zuhilfenahme eines ML-Systems prognostizieren (Weng et al., 2017).

Zur effektiven Vorbeugung von kardiovaskulären Ereignissen sind entsprechende Maßnahmen langfristig durchzuführen. Eine diesbezügliche Vorhersage sollte daher idealerweise möglichst lange im Voraus geschehen. Krankheiten zu prognostizieren kann aber auch nützlich sein, wenn diese kurz nach dem Zeitpunkt eintreten, zu dem die Prognose getroffen wurde. Beispiele für Erkrankungen beziehungsweise Zustände, bei denen rasches Eingreifen nötig ist, sind Nachblutungen und Nierenversagen nach Operationen am Herz. Diese können unvorhergesehen auftreten. Doch wenn entsprechende Symptome erkennbar sind, kann der Zeitpunkt bereits verstrichen sein, zu dem man noch einen rettenden Eingriff hätte durchführen kann. Hier kann ein AI-CDSS Abhilfe schaffen, das auf der Basis von Deep Learning entwickelt wurde und Nierenversagen und Nachblutungen nach Operationen am Herzen möglichst frühzeitig prognostizieren soll (Meyer et al., 2018). Auch die Entwicklung von intraoperativen Hypotonien ist Gegenstand eines prognostischen KI-Systems (Kendale et al., 2018).

Ein großes Feld der KI-basierten Vorhersagen betrifft körperliche Reaktionen. Um zu prognostizieren, wie sich der Körper eines Patienten in einer bestimmten Situation verhalten wird, können sich Ärzte zwar auf ihr physiologisches Wissen stützen. Darauf fußende Vorhersagen sind aber nicht zuletzt aufgrund der Individualität der Patienten in ihrer Genauigkeit begrenzt. Ein Beispiel für eine körperliche Reaktion, die für die klinische Entscheidungsfindung relevant ist, kann man im Bereich der Diabetologie finden: Wie sich der Blutzuckerspiegel nach einer Mahlzeit entwickelt, hat unter anderem Auswirkungen auf die Sterblichkeit von Patienten mit Diabetes mellitus Typ 2 – diese wird durch postprandiale Hyperglykämien erhöht (Cavalot et al., 2011). Obwohl der Reduktion solcher hoher Blutzuckerspiegel nach Mahlzeiten folglich Relevanz zukommt,

ist es schwierig, diese vorherzusagen (Zeevi et al. 2015). Wie sich die Blutzuckerkonzentration postprandial verändert, kann zwar mit Bezug auf das konsumierte Lebensmittel abgeschätzt werden (Bao et al., 2011; Jenkins et al., 1981). Die Qualität solcher Prognosen ist aber beschränkt: Die Blutzuckerreaktion hängt nämlich nicht nur vom konsumierten Lebensmittel, sondern auch vom jeweiligen Patienten ab (Zeevi et al., 2015).

Um dieser individuellen Dimension der glykämischen Reaktion gerecht zu werden, wurde eine ML-Anwendung entwickelt, in die unter anderem individuelle Blutwerte und Informationen bezüglich der Essgewohnheiten und des Mikrobioms des Patienten eingehen (Zeevi et al., 2015). Diese große Menge an Daten ermöglicht eine personalisierte Bestimmung der postprandialen Blutzuckerreaktion. Damit ist dieses System ein Beispiel für die Verbindung von AI-CDSS zur sogenannten Präzisionsmedizin beziehungsweise zur individualisierten Medizin (vgl. Kapitel 3.4.4).

Eine letzte Form von KI-basierten Vorhersagen, die hier vorgestellt werden soll, betrifft das therapeutische Ansprechen: Es gibt Systeme, die prognostizieren, wie ein Patient wahrscheinlich auf eine Therapie reagieren wird. Das kann nützlich sein, weil es für die meisten Patienten mehrere geeignete Behandlungsmöglichkeiten gibt. Auf welche dieser therapeutischen Maßnahmen der Patient am besten ansprechen wird, ist allerdings oft schwierig abzuschätzen.

Das gilt auch für die antiretrovirale Therapie von HIV. Eine besondere Herausforderung dieser Behandlung besteht in der zunehmenden Entwicklung von Resistenzen gegenüber solchen Wirkstoffen (WHO, 2017). Wenn eine derartige Therapie scheitert, sollte zur Suppression des HI-Virus eine neue, besser wirksame Kombination antiretroviraler Medikamente ausgewählt werden. Die beste Therapie für einen bestimmten Patienten kann man unter anderem anhand einer Genotypisierung des HI-Virus auswählen (Deutsche Aids-Gesellschaft et al., 2020). Diese ist aber kostenintensiv und kann daher insbesondere in strukturschwachen Regionen der Welt nicht immer herangezogen werden, um die Kombination der antiretroviralen Medikamente zu selektieren (Revell et al., 2018). Das ist problematisch, weil HIV in ökonomisch schwachen Weltregionen eine hohe Prävalenz aufweist (Mbirimtengerenji, 2007).

Gerade um die Auswahl der antiretroviralen Medikamente in diesen ressourcenarmen Gesundheitssystemen zu verbessern, haben Revell et al. (2018) eine KI-Anwendung entwickelt, die das therapeutische Ansprechen auf verschiedene antiretrovirale HIV-Medikamente prognostiziert, ohne dass dafür ein genotypischer Resistenztest durchgeführt werden müsste.

3.2.3 Therapieempfehlung

Während die bisher vorgestellten Systeme im Bereich der Diagnostik und der Prognostik die Behandlung präformieren mögen, geben die nachfolgenden Systeme konkrete Therapieempfehlungen und haben somit tendenziell einen besonders starken Einfluss auf die Behandlung des Patienten.

Im Bereich der Therapieempfehlung können AI-CDSS etwa die ideale Dosierung von Medikamenten unterstützen. Diese hängt generell von verschiedenen, sich zeitlich verändernden Parametern ab, weshalb die Verabreichungsmenge des Medikaments entsprechend der jeweiligen situativen Begebenheiten zu überprüfen und gegebenenfalls zu ändern ist. Die kontinuierliche Anpassung der Dosierung kann sich für das ärztliche Personal zeitaufwändig und schwierig gestalten: Unterhalb einer gewissen Wirkstoffkonzentration erreichen Medikamente nicht die gewünschte Wirkung, bei einer zu hohen Dosierung drohen jedoch Nebenwirkungen und Schäden für den Patienten.

Ein Beispiel für ein schwierig zu dosierendes Medikament ist unfraktioniertes Heparin (UH). Es ist durch ein enges therapeutisches Fenster gekennzeichnet: Je höher UH im Blut des Patienten konzentriert ist, desto stärker steigt das Risiko für Blutungen; je niedriger dessen Blutkonzentration ist, desto geringer ist der protektive Effekt gegenüber der Entstehung von Blutgerinnseln und desto größer ist demnach die Gefahr einer Blutgerinnselbildung (Landefeld et al., 1987). Um die Risiken zu minimieren, die mit einer zu hohen oder zu niedrigen Dosierung von UH einhergehen, kann ein Monitoring der aktivierten partiellen Thromboplastinzeit (aPTT) betrieben werden (Eikelboom & Hirsh, 2006). Je nachdem, wie sich dieser Wert entwickelt, kann man die Dosierung von UH anpassen, um die Gefahr der Blutung und der Blutgerinnselentstehung zu minimieren. Dieses Monitoring ist aber mit einem nicht unbeträchtlichen Aufwand verbunden. Anstatt die aPTT zu überwachen, können Ärzte zur Dosierung von UH auch ein AI-CDSS verwenden, das die Zeitspanne maximieren soll, in der die aPTT im therapeutischen Bereich liegt (Nemati et al., 2016).

Ein weiteres KI-System, das die Dosierung von Medikamenten unterstützt, kann bei der Behandlung der Anämie von hämodialysepflichtigen Patienten verwendet werden. Während die Therapie der Anämie herkömmlicherweise an recht starre Protokolle gebunden ist, bietet ein *reinforcement-learning*-basiertes AI-CDSS die Möglichkeit, intra- und interindividuellen Differenzen der Patienten gerecht zu werden (Escandell-Montero et al., 2014). Den Entwicklern zufolge führt die Anwendung des Systems dazu, dass im Vergleich zur Therapie gemäß dem Standardprotokoll die Hämoglobinwerte von einem 27,6 % größeren Anteil der Patienten im Zielbereich liegen. Ein weiterer Vorteil des Systems ist gemäß den Angaben der Entwickler, dass dessen Nutzung 5,13 % der verwendeten Menge der Erythropoese-stimulierenden Agenzien einsparen kann. Somit bietet es das Potenzial, die Behandlungskosten zu senken.

Darüber hinaus können AI-CDSS die Planung der therapeutischen Bestrahlung unterstützen. Eine Form der Radiotherapie bei Brustkrebspatientinnen ist die Ganzbrustbestrahlung (Deutsche Gesellschaft für Gynäkologie und Geburtshilfe [DGGG] & Deutsche Krebsgesellschaft [DKG], 2021). Nicht zuletzt aufgrund interindividueller anatomischer Unterschiede kann der Planungsprozess viel Zeit in Anspruch nehmen (Yoo et al., 2021). Unter anderem, um den zeitlichen Aufwand der Bestrahlungsplanung zu minimieren, haben Sheng et al. (2019) ein ML-System zur automatischen Entwicklung von Behandlungsplänen entwickelt. Die Autoren geben an, dass die KI-basierten Therapiepläne im Rahmen einer Evaluation in 19 von 20 Fällen dieselbe Energie wie die klinischen Bestrahlungspläne aufwiesen. Ein Vorteil des Systems ist die Optimierung der zeitlichen Effizienz: Durch die Nutzung des AI-CDSS verringerte sich den Entwicklern zufolge die Zeit zur Erstellung von Bestrahlungsplänen von bis zu vier Stunden auf unter fünf Minuten (Sheng et al., 2019).

Auch im operativen Bereich gibt es KI-Systeme, die Ärzten Empfehlungen zu therapeutischen Entscheidungen unterbreiten. Eine derartige Anwendung kann Chirurgen etwa nächste Schritte bei einer Operation empfehlen (Navarrete-Welton & Hashimoto, 2020).

Die bisher vorgestellten Systeme, die therapeutische Maßnahmen empfehlen, beziehen sich unter anderem auf die Auswahl der Dosierung eines bereits ausgewählten

Medikaments oder auf die Anpassung einer schon gewählten Therapieform. AI-CDSS können jedoch nicht nur bereits gewählte therapeutische Maßnahmen optimieren, sie können auch die Auswahl einer therapeutischen Option unterstützen.

Ein über die Grenzen der Medizin hinaus bekanntes Beispiel dieser Systemgruppe ist *IBM Watson for Oncology* (WFO). Das System ordnet Optionen zur Behandlung verschiedener Tumorerkrankungen nach ihrem Evidenzgrad in drei Kategorien an: ‚*recommended*', ‚*for consideration*' und ‚*not recommended*' (Somashekhar et al., 2018). WFO gibt es in verschiedenen Ausführungen zur Therapie unterschiedlicher onkologischer Erkrankungen. Diese Systeme stützen sich auf vielerlei Daten und Informationen. Als Basis der Anwendung zur Behandlung von Brustkrebs dienen beispielsweise folgende Quellen: Texte aus mehr als dreihundert medizinischen Fachzeitschriften, klinische Leitlinien der US-amerikanischen Klinik *Memorial Sloan Kettering* und weitere Literatur, die von Ärzten dieses Krankenhauses ausgewählt wurde (Pan et al., 2019). Bei der Entwicklung des Systems wurden außerdem Daten von mehr als 550 Brustkrebspatientinnen verwendet. Zur Anpassung der Empfehlungen an die jeweiligen Patientinnen werden Informationen über die Patientin – beispielsweise aus der elektronischen Patientenakte –, Laborergebnisse und Charakteristika des zu behandelnden Tumors in das System übertragen (Pan et al., 2019). WFO gibt zu jeder Behandlungsoption auch die sie unterstützende Evidenz an (Somashekhar et al., 2018).

Die Empfehlungen des Systems weisen in vielen Evaluationsstudien eine hohe Übereinstimmungsrate mit der Therapiewahl von Experten auf (Jie et al., 2021). Abweichungen der Systemoutputs von den Expertenempfehlungen können unter anderem in nationalen Unterschieden der Gesundheitssysteme begründet sein: Manche therapeutische Möglichkeiten, die in den USA als dem Entstehungsland von WFO verfügbar sind, können Ärzte in anderen Ländern nicht verwenden – etwa aufgrund differierender klinischer Leitlinien oder wegen ökonomischer Restriktionen (Somashekhar et al., 2018). Ein weiterer Grund für Abweichungen zwischen den Systemempfehlungen und der Behandlungsoption, die ein Arzt wählt, kann außerdem in individuellen Begebenheiten des vorliegenden Falls oder bestimmten Vorlieben und Werten des Patienten liegen, die das KI-System nicht beachtet (Somashekhar et al., 2018).

Da WFO individuell angepasste therapeutische Maßnahmen vorschlägt und sich dabei auf eine große Menge von Daten bezieht, die ein Mensch kaum überblicken kann, ist es verständlich, dass das System große Erwartungen geweckt hat. Diese wurden aber zu

einem nicht geringen Teil enttäuscht (O'Leary, 2022): Am *MD Anderson Cancer Center* wurde beispielsweise die Verwendung des Systems abgebrochen, nachdem es Empfehlungen abgegeben hatte, die als falsch aufgefasst wurden und eine Gefährdung für die Patienten darstellten (Strickland, 2019). Dass WFO insgesamt weniger als Erfolg gelten kann, lässt sich auch daran ablesen, dass IBM Anfang 2022 die zugehörige Sparte *Watson Health* verkauft hat (Lohr, 2022).

3.3 Zieldefinition von AI-CDSS

Um eine Technologie zu verstehen, ist es zentral zu wissen, wozu man diese einsetzt. So lässt es sich nachvollziehen, dass die EBEA fordert, im Rahmen der Beschreibung der zu bewertenden Technologie auch ihr Einsatzziel darzulegen. Angesichts der Vielfalt von AI-CDSS bezüglich der Anwendungsbereiche und Funktionen stellt sich aber die Frage, ob überhaupt *ein* Ziel von deren Verwendung bestimmt werden kann. Schließlich hängt dieses von der Systemnutzung, vom konkreten Anwendungsfall und vom jeweiligen Nutzer ab (Morley et al., 2021). Um das Verständnis der Anwendungen weiter zu schärfen, soll dennoch nachfolgend versucht werden – eingedenk der erläuterten Limitationen – das Ziel beziehungsweise die Ziele des Einsatzes von AI-CDSS zu definieren.

In allgemeiner Form kann als übergreifendes Ziel der Nutzung dieser Systeme die Verbesserung der Qualität des ärztlichen Entscheidungsprozesses und somit die Optimierung der Qualität der Patientenversorgung beschrieben werden. Im Bereich der Diagnostik und der Prognostik soll primär die Genauigkeit erhöht werden. Auch der Zeitpunkt, zu dem Krankheiten entdeckt werden, soll durch den Einsatz der Anwendungen günstig beeinflusst werden. AI-CDSS, die therapeutische Maßnahmen empfehlen, zielen insbesondere darauf ab, dass die Therapieentscheidung stärker an aktueller Evidenz orientiert wird sowie im Sinne der Präzisionsmedizin an individuelle Charakteristika der Patienten angepasst wird. Durch eine Entlastung der Ärzte sollen zeitliche Ressourcen etwa für die Interaktion mit dem Patienten freigesetzt werden. Der gesamte klinische ärztliche Entscheidungsprozess soll durch die Verwendung der Systeme zeitlich und finanziell effizienter gestaltet werden. Auch eine Optimierung des Ressourcenverbrauchs wird durch die Nutzung von AI-CDSS angestrebt.

3.4 Charakteristische Eigenschaften von AI-CDSS

Nachdem die Funktionsweise der Systeme dargelegt wurde (vgl. Kapitel 3.1), beispielhafte Anwendungen vorgestellt wurden (vgl. Kapitel 3.2) und anhand dessen das Ziel des Einsatzes von AI-CDSS definiert wurde (vgl. Kapitel 3.3), sollen im vorliegenden Kapitel deren typischen Eigenschaften erläutert werden. Diese sind in den bisherigen Untersuchungen schon an verschiedenen Stellen angeklungen. Im Folgenden sollen diese Charakteristika von AI-CDSS noch explizit herausgearbeitet werden, bevor die Systeme im 4. Kapitel ethisch bewertet werden.

3.4.1 Opazität und Erklärbarkeit

Eine Eigenschaft, die als typisch für viele ML-Systeme gelten kann, ist Opazität. Um dieses Phänomen zu erklären, liegt es nahe, abgrenzend auf die herkömmliche Intransparenz vieler technologischer Systeme einzugehen: Angesichts der Komplexität moderner Technologien ist anzunehmen, dass Nutzer oft nicht verstehen, wie diese funktionieren. Wenn auch nicht jeder die Funktionsweise herkömmlicher Anwendungen nachvollziehen kann, so ist aber dennoch grundsätzlich anzunehmen, dass es zumindest jemanden – mit der entsprechenden Qualifikation – gibt, der versteht, wie ein solches System funktioniert und wie es zu bestimmten Ergebnissen kommt. Hierin besteht der zentrale Unterschied zu manchen ML-Anwendungen (Bjerring & Busch, 2021): Auch wenn der Programmcode eines ML-Systems bekannt ist, kann man nicht immer genau erklären, warum es ein bestimmtes Output ausgibt (Zweig, 2018). Eine derart opake ML-Anwendung wird auch als Black-Box-System bezeichnet (Antoniadi et al., 2021). Wie opak eine Anwendung ist, hängt maßgeblich von der zugrundeliegenden Form von ML ab (O'Sullivan, 2020). Deep Learning etwa gilt als besonders opak (Holzinger, 2018).

Zum Verständnis von Opazität kann es auch beitragen, sich vor Augen zu führen, wie eine Anwendung beschaffen sein muss, die als Gegenteil eines Black-Box-Systems – als *white box* – beschrieben werden kann. Nach Lipton (2018) fällt ein System in die Kategorie der *white box*, wenn es drei Anforderungen erfüllt:

1. **Simulierbarkeit**: Ein Mensch kann anhand der Inputdaten und der Parameter des Modells in einer angemessenen Zeit alle Berechnungen ‚simulieren' und somit die Berechnungen des Systems überprüfen.

2. **Zerlegbarkeit**: Nicht nur das ganze System, sondern auch dessen einzelne Teile weisen eine eingängige Verständlichkeit auf. Das kann sich beispielsweise darin zeigen, dass ein Knoten in einem Modell einer Entität in der Realität entspricht (zum Beispiel ‚alle Patienten mit einem systolischen Blutdruckwert von mehr als 150 mmHg').

3. **Algorithmische Transparenz**: Die Nachvollziehbarkeit des Lernalgorithmus muss gegeben sein.

Da Opazität aus verschiedenen Gründen problematisch sein kann, wurden und werden diverse Methoden entwickelt, um Systemergebnisse erklärbar zu gestalten. Was mit Erklärbarkeit gemeint ist, kann anhand der drei Anforderungen von Biecek und Burzykowski (2021) verstanden werden, denen erklärbare KI-Anwendungen genügen müssen:

1. **Vorhersagevalidierung**: Die Evidenz, auf der ein bestimmtes Output beruht, sollte nachvollzogen werden können.

2. **Vorhersagerechtfertigung**: Es muss klar sein, welche Variablen das Zustandekommen des Outputs in welchem Ausmaß beeinflussen.

3. **Vorhersagespekulation**: Man sollte nachvollziehen können, wie sich die Vorhersage eines Modells verändern würde, wenn sich die Werte verändern würden.

Die Entwicklung von Methoden zur Erklärung der Ergebnisse von KI-Systemen ist Gegenstand eines wachsenden Forschungsbereichs und verläuft recht dynamisch (Schaaf et al., 2021). Um ein Verständnis der praktischen Möglichkeiten und Grenzen der Erklärbarkeit zu entwickeln, soll nun ein kurzer Einblick in Methoden zu deren Steigerung folgen.

Erstens kann man ein Black-Box-Modell durch ein anderes Modell (Surrogatmodell) erklären, das seinerseits erklärbar ist (Schaaf et al., 2021). Da für die Nutzer von AI-CDSS die Frage zentral ist, aus welchen Gründen ein bestimmtes Ergebnis zustande kommt, können zur Erklärung der Systemergebnisse auch *attribution-based explanations* verwendet werden. Diese geben an, welche Faktoren das Output in welchem Ausmaß beeinflusst haben (Markus et al., 2021). Weil die Beschaffenheit von ML-Systemen maßgeblich vom zugrundeliegenden Trainingsdatensatz abhängt (vgl. Kapitel 3.4.2),

kann man aus dessen Eigenschaften grundsätzlich Schlüsse auf die Beschaffenheit des Systems ziehen. Das kann dadurch erleichtert werden, dass ‚Prototypen' des Trainingsdatensatzes aufgeführt werden, die für diesen besonders repräsentativ sind (Kim et al., 2016).

3.4.2 Abhängigkeit von Trainingsdaten und eingeschränkte Generalisierbarkeit

Wenn AI-CDSS auf ML basieren, sind sie durch eine typische Eigenschaft gekennzeichnet: Die Qualität von ML-Anwendungen hängt wesentlich von der Beschaffenheit der zugrundeliegenden Trainingsdaten ab. Unterscheiden sich die Inputdaten zu stark von den Trainingsdaten, kann das System falsche Ergebnisse ausgeben (Challen et al., 2019; ZEKO, 2021). Man spricht in diesem Zusammenhang auch von Verzerrungen oder von Bias der Trainingsdaten.

In Bezug auf ML-Systeme lassen sich verschiedene Arten von Bias identifizieren, die unterschiedlich definiert und eingeteilt werden können (Prinz, 2021). Die folgende, auf Futoma et al. (2020) beruhende Darstellung von Bias kann daher nicht den Anspruch auf Vollständigkeit erheben, sie soll aber die im Zusammenhang mit AI-CDSS wichtigsten Bias aufführen.

1. **Genotypische und phänotypische Variationen**: Wenn die Performanz einer ML-Anwendung an bestimmte genotypische oder phänotypische Ausprägungen einer Krankheit gebunden ist, die in der Population bestehen, in der die Trainingsdaten gewonnen wurden, kann das System bei der Anwendung in einer anderen Populationen zu fehlerhaften Ergebnissen kommen. Wird ein AI-CDSS zur Detektion von malignen Melanomen zum Beispiel nur mit Daten von Patienten mit heller Hautfarbe trainiert, so besteht die Gefahr, dass es bei der Behandlung von Personen mit dunkler Hautfarbe eine niedrigere Leistungsfähigkeit aufweist und zu fehlerhaften Ergebnissen kommt.

2. **Demographische Unterschiede zwischen den Patienten des Trainings- und des Anwendungsdatensatzes**: Entwickelt man beispielsweise ein AI-CDSS zur Prädiktion der Mortalität einer bestimmten Erkrankung in einem Land mit einer vergleichsweise alten Bevölkerung, kann das System überdurchschnittlich häufig zu falschen Ergebnissen kommen, wenn Daten von jüngeren Patienten eingegeben werden.

3. **Unterschiede bezüglich der Umgebung, in der Trainings- und Anwendungsdaten gewonnen werden**: Als Beispiel hierfür kann ein System genannt werden, das diabetische Retinopathie anhand von Bildaufnahmen detektieren soll. Es wurde in den USA entwickelt und wies dort eine hohe Performanz auf (Beede et al., 2020). In Thailand erreichte das System diese Leistungsfähigkeit hingegen nicht. Als Ursache dafür identifizierte man die oft nur spärliche Beleuchtung in thailändischen Krankenhäusern: Die Aufnahmen, die in Thailand in das AI-CDSS eingespeist wurden, waren schwächer beleuchtet als die in den US-amerikanischen Kliniken aufgenommenen Bilder, die für das Training der Systeme verwendet wurden (Beede et al., 2020). Da der Unterschied zwischen den amerikanischen Trainingsbildern und den thailändischen Aufnahmen also zu groß war, konnte das System die in den USA erreichte diagnostische Treffsicherheit in Thailand nicht erreichen.

4. **Unterschiede hinsichtlich der Hardware und der Software, die zur Datenverarbeitung verwendet werden**: Beispielsweise kann die Leistungsfähigkeit eines KI-Systems zur Interpretation von CT-Bildern von der Herstellerfirma und dem Modell des Computertomographen abhängen, mit dem die zu interpretierenden Bilder aufgenommen werden. Falls in ein solches KI-System CT-Bilder eingespeist werden, die etwa von einem Computertomographen einer anderen Herstellerfirma stammen, kann das KI-System falsche Ergebnisse ausgeben.

5. **Unterschiede bezüglich anderer Determinanten von Krankheit und Gesundheit**: In besonderen Situationen (zum Beispiel transiente Mittelknappheit, Pandemie) kann sich die Beziehung zwischen den Risikofaktoren für eine Erkrankung und den klinischen Ereignissen verändern – etwa durch eine Zunahme des Risikos von Krankenhausinfektionen aufgrund verschlechterter hygienischer Zustände. Wenn in dieser Situation ein ML-System angewandt wird, das mit Trainingsdaten entwickelt wurde, die zeitlich vor der Entstehung dieser besonderen Situation erhoben wurden, kann die Leistungsfähigkeit dieses Systems in der vorliegenden Ausnahmesituation eingeschränkt sein.

Anhand dieser Beispiele dürfte deutlich geworden sein, dass die Beschaffenheit der Trainingsdaten bezüglich der Frage relevant ist, ob ein KI-System in bestimmten Situationen oder Populationen zu richtigen Ergebnissen kommt. In diesem

Zusammenhang ist das Konzept der Generalisierbarkeit zentral. Diese ist nicht binär zu verstehen: Eine ML-Anwendung ist nicht entweder generalisierbar oder nicht generalisierbar.

Es gibt stattdessen einige Grade und Formen von Generalisierbarkeit (Futoma et al., 2020): So kann ein ML-System als generalisierbar aufgefasst werden, wenn es ohne maßgeblichen Verlust der Leistungsfähigkeit außerhalb des engen Kontextes angewandt werden kann, in dem es entwickelt wurde – etwa in verschiedenen Kliniken oder Ländern. Generalisierbarkeit kann man auch einer ML-Anwendung zuschreiben, deren Leistungsfähigkeit nicht signifikant abnimmt, wenn sie mit Daten aus Personengruppen gespeist wird, die sich von der Population maßgeblich unterscheiden, aus der die Trainingsdaten stammen. Außerdem kann Generalisierbarkeit bedeuten, dass man ein System an dem Ort, an dem es entwickelt wurde, zu einem späteren Zeitpunkt ohne relevante Einbußen bezüglich der Leistungsfähigkeit verwenden kann. Generalisierbarkeit kann also verschieden ausgeprägt sein; deren Kontinuum ist aber begrenzt: Es ist eher der Normalfall als eine Ausnahme, dass universale Generalisierbarkeit nicht erreichbar ist. Ein Charakteristikum von AI-CDSS ist deshalb deren eingeschränkte Generalisierbarkeit. Diesen Umstand beschreiben Futoma et al. (2020) treffend mit folgenden Worten:

> „Machine learning systems are not like thermometers, reliably measuring the temperature via universal rules of physics; nor are they like trained clinicians, gracefully adapting to new circumstances. Rather, these systems should be viewed as a set of rules that were trained to operate under certain contexts and rely on certain assumptions, and might work seamlessly at one centre but fail altogether somewhere else.“ (S. 491)

Die eingeschränkte Generalisierbarkeit stellt eine charakteristische Eigenschaft von AI-CDSS dar, die die vorliegende Arbeit bezüglich der Machbarkeit (vgl. Kapitel 4.1.1), der Wirksamkeit (vgl. Kapitel 4.1.2) und des Schadenspotenzials der Systeme (vgl. Kapitel 4.4) sowie hinsichtlich der Gerechtigkeit (vgl. Kapitel 4.12) aufgreift.

3.4.3 Zwischen Entscheidungsunterstützung und -übernahme

AI-CDSS stellen Diagnosen sowie Prognosen und geben Therapieempfehlungen ab. Sie üben somit Aufgaben aus, die herkömmlicherweise im Zuständigkeitsbereich des Arztes liegen. Nicht zuletzt, weil die Systeme menschliche Fähigkeiten teilweise übertreffen,

stellt sich die Frage, ob die Anwendungen die ärztliche Entscheidungsfindung übernehmen und automatisieren können und werden. Auch aus einer Definition von KI, die vom Bundesamt für Sicherheit in der Informationstechnik (2022) stammt, geht hervor, dass die Automatisierung von Abläufen als ein Charakteristikum dieser Technologie beschrieben werden kann. Wie hoch der Automatisierungsgrad von AI-CDSS einzuschätzen ist, wird oft mit der Frage verknüpft, ob und wie autonom diese Systeme sind. Der Autonomie beziehungsweise der Automatisierung dieser Anwendungen sollen sich die nachstehenden Untersuchungen in Anlehnung an die Ausführungen von Weber und Zoglauer (2019) als erstes von der Seite der Autonomie nähern.

Dazu ist zunächst zu klären, wie diese im Zusammenhang mit KI-Systemen zu verstehen ist. Zum einen kann der Begriff der Autonomie in einem moralischen Sinne verwendet werden. Hiermit meint man meist die mit Verantwortungsfähigkeit verbundene Befähigung, selbstbestimmt Entscheidungen zu treffen (Weber & Zoglauer, 2019; ZEKO, 2021). Ob Intentionalität und Bewusstsein als Bedingungen moralischer Autonomie ausreichen – wie von Weber und Zoglauer (2019) dargestellt –, kann im Rahmen der vorliegenden Arbeit nicht untersucht werden. Hier soll darum der Verweis auf die herrschende Meinung genügen, dass jedenfalls den derzeit verfügbaren AI-CDSS Autonomie nicht in einem moralischen Sinne zukommt und dass diese Systeme keine Verantwortung tragen können (ZEKO, 2021).

Besonders in der Robotik und in der KI-Forschung kann jenseits des moralischen Begriffs der Autonomie noch ein anderes, technisches Verständnis identifiziert werden: Autonomie bedeutet hier (mehr oder weniger weitgehende) Unabhängigkeit von menschlicher Kontrolle (Weber & Zoglauer, 2019). In diesem Sinne hat ‚Autonomie' eine enge Verbindung zu ‚Automatisierung': Herkömmlicherweise wird der Automatisierungsgrad eines Systems nämlich umso höher eingestuft, je weniger Eingriffe des Menschen bei der Systemnutzung erforderlich sind, also je unabhängiger das System vom Menschen agiert (Parasuraman et al., 2000). Auch wenn ‚Automatisierung' und ‚Autonomie' teilweise beinahe synonym verwendet werden, kann man beide Begriffe differenzieren, indem man das jeweils zugrunde liegende Systemverständnis betrachtet: *Automatisierte* Prozesse kann man vorhersehen, *autonome* Vorgänge hingegen nicht (Adler, 2019).

An dieser Stelle wird auch sichtbar, wie technische und moralische Autonomie im Auge des Betrachters einer technisch autonomen Maschine konfluieren mögen: Je weniger

kontrollierbar und vorhersehbar ein System ist, desto eher dürfte man dazu neigen, ihm nicht nur technische, sondern auch moralische Autonomie und Verantwortungsfähigkeit zuzuschreiben (Sparrow, 2007; Weber & Zoglauer, 2019). Da die Begrifflichkeiten nun klarer differenziert wurden, kann nachfolgend die Frage behandelt werden, inwiefern AI-CDSS – in einem nicht-moralischen Sinne – als autonom beziehungsweise automatisiert bezeichnet werden können.

Zur Einstufung der Autonomie und der Automatisierung von KI-Systemen in der Medizin gibt es verschiedene Klassifizierungen. Braun et al. (2020) kennen etwa drei Kategorien von KI-basierten CDSS, die sich darin unterscheiden, ob und in welchem Ausmaß beim Betrieb des betreffenden Systems menschliches Eingreifen notwendig ist. Angelehnt an eine Einteilung verschiedener Stufen des autonomen Fahrens werden dagegen häufig fünf verschiedene Grade der Autonomie beziehungsweise der Automatisierung definiert, die von vollständiger bis zu gar nicht vorhandener Abhängigkeit vom Menschen reichen (Bitterman et al., 2020; Kazzazi, 2021).

Welche Anforderungen ein medizinisches KI-System erfüllen muss, um in eine bestimmte Kategorie zu fallen, bleibt zwischen den verschiedenen Klassifizierungen umstritten. Für Kazzazi (2021) muss ein vollständig autonomes medizinisches KI-System emotionale Fähigkeiten besitzen und so beschaffen sein, dass dessen Outputs nicht vom Arzt kontrolliert werden müssen. Von Ersterem ist bei Bitterman et al. (2020) hingegen keine Rede. Der Autonomiegrad hängt hier etwa davon ab, ob der Nutzer oder der Entwickler für Fehler des Systems Verantwortung trägt. Die einzelnen Grade der Autonomie werden also in den verschiedenen Klassifizierungen nicht übereinstimmend definiert. Darüber hinaus bleiben die Einteilungen auch in einigen Punkten recht vage, wodurch deren Anwendbarkeit eingeschränkt wird.

Obwohl hier also AI-CDSS nicht nach verschiedenen, fein aufgegliederten Graden der Autonomie oder der Automatisierung eingeteilt werden können, erscheint es dennoch angebracht der Frage nachzugehen, inwiefern man entscheidungsunterstützende KI-Systeme als autonom beziehungsweise als automatisiert beschreiben kann.

Liedtke und Langanke (2021) vertreten den Standpunkt, dass diese „‚lediglich' der Unterstützung der jeweiligen Entscheidungsträger bei diagnostischen oder

therapeutischen Maßnahmen [dienen]. Entscheidungen werden gerade *nicht* in eine Maschine oder einen Algorithmus hineinverlagert“ (Liedtke & Langanke, 2021, S. 280). Sieht man nur die zwei Pole – auf der einen Seite Autonomie beziehungsweise Automatisierung und somit eine Ersetzung des Arztes, auf der anderen Seite Unterstützung – so ist Liedtke und Langanke Recht zu geben: Die klinische Entscheidungsfindung wird von AI-CDSS nicht übernommen, sondern unterstützt.

Auch wenn eine dichotome Unterscheidung zwischen Entscheidungsassistenz auf der einen Seite und -übernahme auf der anderen Seite in der Theorie trennscharf wirken mag, zeigt sich mit Blick auf die Praxis jedoch, dass es einen Graubereich verschiedener Schattierungen der Automatisierung und der Autonomie gibt (ZEKO, 2021). Hinsichtlich AI-CDSS lassen sich ‚Entscheidungsunterstützung‘ und ‚Entscheidungsübernahme‘ darum als Pole auf einem Kontinuum der Automatisierung und der Autonomie begreifen. Für die Anwendungen ist es typisch, dass diese sich nicht restlos einem der beiden Pole zuordnen lassen. Folgender Grundaussage der oben genannten Klassifizierungen ist daher zuzustimmen: Es gibt nicht nur zwei Kategorien der Autonomie beziehungsweise der Automatisierung, sondern mehrere Zwischenstufen zwischen der Entscheidungsübernahme und der Entscheidungsunterstützung.

Da keine allgemein anerkannte und hinreichend praktikable Systematisierung der Autonomie oder der Automatisierung von AI-CDSS zu finden ist, soll hier nicht der (fragwürdige) Versuch unternommen werden, die Anwendungen nach ihrem Autonomie- oder Automatisierungsgrad einzuteilen. Stattdessen soll als Erkenntnis der obigen Ausführungen festgehalten werden, dass bezüglich der Autonomie und der Automatisierung gerade der (diffuse) Zwischenzustand – zwischen Entscheidungsunterstützung und -übernahme – als ein Charakteristikum von AI-CDSS beschrieben werden kann.

Daraus ergeben sich ethische Implikationen: Insofern AI-CDSS automatisiert oder autonom agieren, kann die Verwendung der Systeme die Entscheidungsautonomie des Arztes einschränken (vgl. Kapitel 4.6). Je autonomer ein System auf den Arzt wirkt, desto eher mag er außerdem dazu neigen, dem System Verantwortungsfähigkeit zuzuschreiben, wodurch bei der Nutzung von AI-CDSS Verantwortungsdiffusion entstehen kann (vgl. Kapitel 4.8).

3.4.4 Transformation der Medizin und des ärztlichen Berufs

Historisch zeigt sich, dass die Anwendung neuer Technologien in der Medizin mit einer Veränderung des ärztlichen Berufsbilds einhergeht. Das kann etwa an der digitalen Transformation der Medizin verdeutlicht werden: Die Diffusion von Computern und anderen digitalen Technologien in die ärztliche Praxis hat die Medizin und den Arztberuf in vielerlei Hinsicht verändert (Alt & Zimmermann, 2021; Mesko & Győrffy, 2019).

Vor diesem Hintergrund ist es verständlich, dass auch mit der Verbreitung von KI-Systemen Veränderungen erwartet werden. Diese sind derart zahlreich und weitreichend, dass man auch von der *vierten Revolution* spricht (Schwab, 2016). Da es sich bei AI-CDSS um eine Form von KI-Anwendungen handelt, überrascht es nicht, dass mit der Verbreitung dieser Systeme eine Vielzahl von Transformationen verbunden werden. Diese sind so tiefgreifend, dass man das transformative Potenzial von AI-CDSS als ein Charakteristikum der Systeme betrachten kann. Es ist klar, dass nicht alle der verschiedenen Veränderungen, die mit den Anwendungen verknüpft werden, hier dargelegt werden können. Im Folgenden sollen darum einige besonders relevante dieser erwarteten Transformationen erläutert werden.

Bei der Vorstellung beispielhafter AI-CDSS in Kapitel 3.2 wurde an verschiedenen Stellen deutlich, dass die Nutzung der Systeme dazu beitragen kann, Diagnostik, Prognostik und Therapieempfehlungen stärker an individuellen Eigenschaften der Patienten auszurichten. Somit sind die Anwendungen in der Nähe der Präzisionsmedizin zu verorten. Darunter ist eine Form der Medizin zu verstehen, die therapeutische Maßnahmen hochpräzise an individuelle Merkmale eines Patienten wie genetische, molekulare oder proteomische Marker anpasst, die oft zum Phänomen Big Data gezählt werden (Reinhardt et al., 2020).

Dieses ist nach einer verbreiteten Definition durch die Zunahme der Datenmenge (*volume*), die immer schnellere Verarbeitung (*velocity*) und die steigende Vielfalt von Daten (*variety*) gekennzeichnet (Gartner, o. D.). Zu den Feldern, in denen Big Data besonders relevant ist, wird die Medizin gezählt: Riesige Datenmengen werden etwa in der Forschung und bei der Bildgebung produziert und verarbeitet. Auch die Digitalisierung der medizinischen Praxis trägt zu einer zunehmenden Genese von medizinischen Daten bei. In elektronischen Patientenakten werden etwa Laborergebnisse,

Diagnosen und anamnestische Informationen gespeichert. Schließlich fallen auch außerhalb von Kliniken und Arztpraxen gesundheitsbezogene Daten zum Beispiel durch die Nutzung von Gesundheits-Apps und Wearables an (Deutscher Ethikrat, 2018a).

Um aus diesen großen Datenmengen relevante Informationen zu generieren, sind lange und komplizierte Rechenaufgaben durchzuführen. Diese mithilfe von herkömmlichen Computern zu bearbeiten, bedeutet einen großen Aufwand. Hier wird das Potenzial von ML-Systemen deutlich: Diese Anwendungen können gewaltige Datenmengen verarbeiten und sind dabei weitgehend nicht auf menschliches Eingreifen angewiesen (Qiu et al., 2016). Erschließen AI-CDSS das Potenzial von Big Data in der Medizin, können sie die Präzisionsmedizin fördern und die klinische Entscheidungsfindung individualisierter gestalten. Hierin liegt ein Nutzenpotenzial von KI-Systemen (vgl. Kapitel 4.3).

Weil AI-CDSS mit der klinischen Entscheidungsfindung eine genuin ärztliche Aufgabe unterstützen und der Übergang zwischen Entscheidungsunterstützung und -übernahme fließend ist (vgl. Kapitel 3.4.3), gehen manche davon aus, dass Ärzte von KI-Systemen ersetzt werden könnten. Das kann man plausibel machen, indem man auf den Umstand verweist, dass bereits jetzt KI-Anwendungen Ärzte bei der Durchführung mancher, eng abgegrenzter Aufgaben übertreffen (vgl. Kapitel 4.1.2). So prognostizieren Liu et al. (2018), dass in Zukunft vor allem Routineaufgaben, die wenig riskant sind, von KI-Systemen übernommen werden. Die potenzielle Ersetzung des Arztes und der mögliche Abbau ärztlicher Arbeitsplätze werfen gerechtigkeitsethische Fragen auf (vgl. Kapitel 4.12).

Eine weitere durch AI-CDSS hervorgerufene Veränderung der Medizin besteht darin, dass – wenn die Systeme in der Privatwirtschaft entwickelt werden – profitorientierte Akteure mit den Anwendungen in den Kernbereich der Medizin vordringen (Ploug & Holm, 2020b). Da die Systementwickler generell nicht an das ärztliche Berufsethos gebunden sind, die teilweise Autonomie von AI-CDSS aber die Entscheidungsfreiheit der Ärzte möglicherweise in gewissem Ausmaß untergraben könnte (vgl. Kapitel 3.4.3 und 4.6), werden durch die Integration privatwirtschaftlicher Akteure in die Medizin ethische Fragen bezüglich der ärztlichen Entscheidungsautonomie aufgeworfen (vgl. Kapitel 4.6).

Eine letzte mögliche Transformation der Medizin durch AI-CDSS, die hier angeführt werden soll, ist die Förderung der Telemedizin (Hopkins et al., 2020; Li et al., 2021). KI-Systeme, die beispielsweise medizinische Bildaufnahmen analysieren, kann ein Arzt auch in weiter Entfernung von dem betreffenden Patienten nutzen. Hierdurch kann etwa in strukturschwachen Regionen der Welt die Patientenversorgung verbessert werden (vgl. Kapitel 4.12).

4. Ethische Bewertung von AI-CDSS

4.1 Funktionsfähigkeit

In der EBEA steht das Kriterium der Funktionsfähigkeit an erster Stelle. Das unterstreicht deren Bedeutung für die ethische Bewertung von eHealth-Anwendungen: Die Funktionsfähigkeit ist eine notwendige, wenn auch nicht hinreichende Bedingung für die Nützlichkeit einer Technologie (vgl. Kapitel 4.3). Diese ist wiederum zentral für die ethische Legitimität der Technologienutzung. Falls sich herausstellt, dass eine bestimmte Technologie nicht nützlich ist beziehungsweise mit keinem zusätzlichen Nutzen verbunden ist, ist ihr Einsatz ethisch nicht vertretbar. In diesem Fall wird die Überprüfung der anderen Kriterien der EBEA hinfällig. Bevor also die Nützlichkeit von AI-CDSS in Kapitel 4.3 bewertet wird, soll die Funktionsfähigkeit der Systeme in diesem Kapitel in den Fokus gerückt werden.

Entsprechend der Arbeit, die als Grundlage für die Entwicklung der EBEA gedient hat, soll die Funktionsfähigkeit von AI-CDSS anhand der drei Kriterien Machbarkeit, Brauchbarkeit und Wirksamkeit evaluiert werden (Marckmann, 2003). Zur Bestimmung der Funktionsfähigkeit der Systeme ist zunächst auszumachen, wo die Möglichkeiten und Grenzen der Anwendungen liegen (Machbarkeit, vgl. Kapitel 4.1.1). Daraufhin ist zu erörtern, inwiefern diese Systeme die Zielsetzung ihres Einsatzes erreichen (Wirksamkeit, vgl. Kapitel 4.1.2). Wenn Art und Umfang der Fähigkeiten von AI-CDSS bestimmt sind, gilt es zu überprüfen, ob diese einen Bedarf der Ärzte erfüllen (Brauchbarkeit, vgl. Kapitel 4.1.3).

4.1.1 Machbarkeit

Bei der Vorstellung einiger beispielhafter Systeme wurde bereits teilweise sichtbar, wozu diese in der Lage sind (vgl. Kapitel 3.2). Aus ethischer Perspektive interessieren neben den Möglichkeiten der Anwendungen insbesondere auch die Grenzen der KI-basierten Diagnostik, Prognostik und Therapieempfehlung. Denn eine ethisch vertretbare Nutzung von AI-CDSS muss deren Begrenzungen beachten. Da die Systeme mit der ärztlichen Entscheidungsfindung einen Prozess unterstützen, der ansonsten von Menschen ausgeführt wird, liegt es nahe, die Grenzen der KI-Systeme anhand von Unterschieden zwischen diesen und Ärzten zu untersuchen.

Als bedeutender Unterschied fällt zunächst auf, dass AI-CDSS auf die Unterstützung eines zumeist eng abgegrenzten Teils des Entscheidungsprozesses beschränkt sind. Sie können etwa nur Diagnosen *oder* Prognosen *oder* therapeutische Maßnahmen in einem bestimmten medizinischen Bereich empfehlen. Ein System, das den gesamten klinischen ärztlichen Entscheidungsprozess – auch nur in einem bestimmten Fachgebiet – unterstützen oder gar übernehmen könnte, gibt es bislang nicht. Die Menge der Erkrankungen, auf die sich eine einzelne Anwendung bezieht, ist im Vergleich zu der Zahl an Erkrankungen, die ein Arzt normalerweise behandeln kann, begrenzt: Obgleich Ärzte heutzutage oft hochspezialisiert sind, besitzen sie generell über ihren Fachbereich hinaus Wissen und Kompetenzen. AI-CDSS unterstützen hingegen typischerweise nur die Entscheidungsfindung im Zusammenhang mit einer bestimmten Pathologie oder maximal einer bestimmten Gruppe von Krankheiten. Ein System kann beispielsweise anhand von Bildern des Augenhintergrundes eine diabetische Retinopathie erkennen, zur Detektion einer anderen Erkrankung ist es aber nicht imstande (Beede et al., 2020).

Problematisch ist, dass AI-CDSS charakteristischerweise nicht ‚merken', wenn sie in Kontexten angewandt werden, in denen ihre Funktionsfähigkeit eingeschränkt ist. Die Systeme können sich an solche Änderungen auch nicht selbständig anpassen (Futoma et al., 2020; Y. Wang et al., 2020). Daher kommt die Aufgabe der Identifikation und der Berichtigung fehlerhafter Ergebnisse den Nutzern zu (vgl. Kapitel 4.4.2.2).

Doch nicht nur die Funktion und die Anwendungsdomäne eines Systems sind beschränkt, auch die Anzahl und die Heterogenität der Quellen, aus denen diese Informationen und Daten beziehen und verarbeiten können, ist begrenzt. Ärzte können sich bei der klinischen Entscheidungsfindung auf eine Vielzahl unterschiedlicher Informationsquellen beziehen. Unter anderem recht objektive Informationen, etwa aus der Laborchemie, aber auch subjektive Erkenntnisse, die bei der körperlichen Untersuchung oder bei der Anamnese gewonnen wurden, können in das klinische Denken eingehen. Somit ist der Arzt in der Lage, den Patienten ganzheitlich zu betrachten (van Baalen et al., 2021).

AI-CDSS beziehen sich hingegen auf eine Auswahl von Patientendaten (ZEKO, 2021). Das wird beispielsweise deutlich im Hinblick auf Systeme, die auf der Grundlage eines Bildausschnitts eine Diagnose stellen.

Diese Einschränkungen bedeuten aber nicht, dass sich AI-CDSS nur auf eine Quelle oder eine Form von Daten beziehen können. Auch in die KI-gestützten Berechnungen können Informationen und Daten unterschiedlicher Art und verschiedener Herkunft einfließen. Das System WFO empfiehlt Therapien etwa anhand von mehreren Faktoren wie Alter, Geschlecht und individuellen biomolekularen Charakteristika des Patienten (Pan et al., 2019; Somashekhar et al., 2018).

Die Spannweite und Heterogenität der von AI-CDSS verarbeitbaren Daten ist aber beschränkt: Prinzipiell ist der Umfang von Daten und Informationen, die von den Systemen verwendet werden können, auf solche begrenzt, die sich überhaupt digital darstellen lassen. Zu den Bereichen, die sich nur schwierig oder nicht umfänglich in digitaler Form abbilden lassen, gehört etwa das Soziale und das Emotionale, das Psychische sowie das Spirituelle (Funer, 2021).[6] Da AI-CDSS diese Dimensionen und Aspekte der klinischen Entscheidungsfindung (jedenfalls bislang) nicht hinreichend beachten, ist es die Aufgabe des Arztes zu gewährleisten, dass das ärztliche Tun nicht reduktionistisch verkürzt wird und dass die ‚weichen Faktoren' bei der klinischen Entscheidungsfindung hinreichend Beachtung finden (vgl. Kapitel 4.9.2).

In diesem Zusammenhang ist noch besonders auf Patientenpräferenzen einzugehen. Während die Diagnostik und die Prognostik grundsätzlich im Aufgabenbereich des Arztes liegen, soll die Therapieentscheidung nicht nur vom Arzt, sondern gemeinsam mit dem Patienten getroffen werden (Shared Decision Making) (Elwyn et al., 2012). Schließlich ist die Frage, welche der therapeutischen Möglichkeiten als ‚die beste' auszuwählen ist, nicht nur von Evidenz abhängig. Die Wahl einer Behandlungsmaßnahme erfordert nämlich Abwägungen zwischen verschiedenen Nutzen- und Schadenspotenzialen, wobei evaluative Urteile zu treffen sind (Hazlewood et al., 2018). Das Prinzip des Respekts der Patientenautonomie verbietet es, dabei Werte und Wünsche der Patienten zu übergehen. Stattdessen ist im Sinne des Shared Decision Makings Flexibilität für Patientenpräferenzen geboten (Charles et al., 1997).

Bisherige Systeme sind grundsätzlich nicht dazu imstande, auf Patientenpräferenzen einzugehen (McDougall, 2019). Es wurde darüber hinaus überzeugend dafür

[6] Es gibt zwar verschiedene Versuche im Rahmen von *affective computing*, Emotionen zu erkennen und hervorzurufen (Kächele et al., 2014; Mai et al., 2021). Bei der ärztlichen Entscheidungsunterstützung spielen diese Technologien allerdings (bisher) keine wesentliche Rolle.

argumentiert, dass KI-Systeme auch in Zukunft die evaluativen Präferenzen von Patienten nicht in einer Weise behandeln können, die für die Anforderungen des Shared Decision Makings hinreichend wäre und gleichzeitig ethisch vertretbar wäre: Ploug und Holm (2020b) zeigen auf, dass eine Herleitung von Patientenpräferenzen mithilfe von Big Data zwar bedingt möglich, ethisch aber kaum vertretbar wäre, weil der Patient dadurch auf seine (digital darstellbare) Vergangenheit festgelegt wäre. Eine weitere Möglichkeit zur Integration von Patientenpräferenzen in die Systemarchitektur bestünde darin, sich auf evaluative Präferenzen zu beziehen, die in dem Milieu oder der Gruppe des Patienten vorwiegen. Da sich die Werte und Wünsche eines Individuums von denjenigen seiner Umgebung nicht selten unterscheiden, könnte diese Vorgehensweise aber kaum dem Gebot des Respekts der Patientenautonomie genügen (Rajput et al., 2020).

Auch wenn es sich also als ethisch fragwürdig erweist, die Patientenpräferenzen von der Vergangenheit des Patienten oder seiner Umgebung abzuleiten, stellt sich die Frage, ob ein KI-System nicht die Wünsche und Werte des Patienten bei der klinischen Entscheidungsfindung erfragen und zusammen mit ihm explorieren könnte. Obwohl es jedenfalls denkbar ist, ein solches System zu entwickeln, ist festzuhalten, dass KI-basiertes Shared Decision Making derzeit mit Herausforderungen verbunden ist und es ins Aufgabenfeld des Arztes fällt, den Patienten und seine evaluativen Präferenzen in die Entscheidungsfindung einzubeziehen.

Bezüglich der übergeordneten Frage nach der Machbarkeit von AI-CDSS kann man daher Folgendes festhalten: Insofern Therapie*entscheidungen* an Patientenpräferenzen orientiert sein sollen, können die derzeitigen KI-Systeme diese nicht (auf ethisch vertretbare Weise) treffen. Die Outputs von AI-CDSS im therapeutischen Bereich können jedoch als Therapie*empfehlungen* aufgefasst werden, die in den Prozess der gemeinsamen Entscheidungsfindung von Arzt und Patient einfließen können – wobei die Entscheidung letztlich beim Patienten liegt.

Da Werte und Wünsche der Patienten bei der Diagnose- und der Prognosestellung an sich keine Rolle spielen, ist auch die Machbarkeit der KI-gestützten Diagnostik und Prognostik in dieser Hinsicht nicht eingeschränkt. Die oben genannten grundsätzlichen Einschränkungen von KI-Systemen, die mit deren reduktionistischer Natur verbunden sind, gelten aber auch für die Anwendungen, die Diagnosen und Prognosen stellen.

Aufgrund der Begrenzungen der AI-CDSS erscheint es auch im diagnostischen und prognostischen Bereich zumindest zum jetzigen Zeitpunkt generell ethisch geboten, deren Outputs als Empfehlungen zu begreifen. Wenn sich nachweisen lässt, dass ein KI-System bei einer spezifischen Fragestellung zu besseren Ergebnissen als ein bestimmter Arzt kommt, sollte in einem solchen Fall aus ethischer Perspektive grundsätzlich – nach einer Plausibilitätsprüfung – das Systemergebnis verwendet werden (vgl. Kapitel 6.3.12).

AI-CDSS sollen folglich keine Entscheidungen treffen, sondern die klinische ärztliche Entscheidungsfindung unterstützen. Damit Empfehlungen von AI-CDSS in den Augen der Nutzer nicht zu Entscheidungen werden (vgl. Kapitel 3.4.3), kommt der Überprüfung und Anpassung der Systemergebnisse aus ethischer Sicht besondere Bedeutung zu (vgl. Kapitel 6.3.12).

4.1.2 Wirksamkeit

Im vorangegangenen Kapitel wurde deutlich, wozu AI-CDSS imstande sind und wo ihre Grenzen liegen. Dadurch wurde auch eingegrenzt, welche Ziele die Systeme erreichen können. Das vorliegende Kapitel soll der Frage nachgehen, wie wirksam die Anwendungen sind, inwiefern sie also das jeweilige Einsatzziel erreichen.

Die Wirksamkeit von AI-CDSS kann nicht im Abstrakten bestimmt werden – hierfür sind Evaluationsstudien nötig. Da man diesen Parameter in den Bereichen der Diagnostik und der Prognostik ähnlich bestimmt, sollen die Systeme aus diesen Feldern gemeinsam behandelt werden, bevor die Wirksamkeit der Anwendungen untersucht wird, die Therapieempfehlungen abgeben.

Zur Evaluierung der Wirksamkeit bestimmt man in den Feldern der Prognostik und der Diagnostik die *Genauigkeit* der Systeme. Hierzu verwendet man insbesondere die Parameter Spezifität, Sensitivität und *accuracy*. Während die ersten beiden Größen im medizinischen Bereich verbreitet sind und Ärzte um deren Bedeutung wissen dürften, ist zu bezweifeln, ob allen potenziellen Nutzern von AI-CDSS klar ist, wie *accuracy* berechnet wird und welche Schlüsse aus diesem Parameter gezogen werden können.[7] Falls die Anwender Evaluationsstudien zur Wirksamkeit einer Anwendung nicht

[7] Diese Größe wird aus dem Verhältnis der Anzahl der von einem System richtig klassifizierten Exemplare und der Anzahl aller klassifizierter Exemplare berechnet: *accuracy*=(TP+TN)/(TP+TN+FP+FN) (Baratloo et al., 2015).

verstehen, wäre deren Nutzung ethisch nicht vertretbar. Es sind darum Maßnahmen zu treffen, damit Ärzte nachvollziehen können, wie wirksam ein System ist (vgl. Kapitel 6.1.18.1).

Da nun die Evaluationsparameter von AI-CDSS im Bereich der Diagnostik und der Prognostik erläutert und problematisiert wurden, kann mit dem Blick auf Evaluationsstudien der Frage nachgegangen werden, wie die Wirksamkeit der Systeme – gemessen anhand der zuvor behandelten Parameter – einzuschätzen ist. Wie bereits bei der Vorstellung beispielhafter AI-CDSS deutlich wurde (vgl. Kapitel 3.2), erreichen KI-Systeme im Gebiet der Diagnostik und der Prognostik oft hohe Genauigkeitswerte. Das zeigt auch eine Übersichtsarbeit zur Genauigkeit von 14 diagnostischen Deep-Learning-Anwendungen: Die gepoolte Sensitivität wird hier mit 87,0 % und die gepoolte Spezifität gar mit 92,5 % angegeben (Liu et al., 2019). Auch im prognostischen Bereich sind AI-CDSS leistungsfähig. Zum Beispiel beträgt der durchschnittliche *accuracy*-Wert von ML-Systemen zur Prädiktion von Outcomes in der plastischen Chirurgie 86,11 % (Mantelakis et al., 2021).

Interessanterweise kann die Spannweite der Genauigkeitswerte von Systemen aus demselben Anwendungsbereich groß sein: In einer Übersichtsarbeit wurden KI-Anwendungen zum Screening von Mammographie-Aufnahmen untersucht. Die *accuracy*-Werte reichen hier von 69,2 % bis 97,8 % (Houssami et al., 2019). Hieran wird deutlich, dass die pauschale Bestimmung der Genauigkeit von AI-CDSS in bestimmten Anwendungsbereichen nur eingeschränkt sinnvoll ist. Es gilt stattdessen die Wirksamkeit jedes Systems einzeln zu evaluieren (vgl. Kapitel 6.1.18.1). Wie Genauigkeitswerte einzuschätzen sind, wird im Zusammenhang mit der Brauchbarkeit der Systeme diskutiert (vgl. Kapitel 4.1.3).

Nachdem der vorherige Abschnitt auf Evaluationsstudien von KI-Systemen Bezug genommen hat, soll an dieser Stelle nicht unerwähnt bleiben, dass die Qualität solcher Studien oft als methodisch mangelhaft beschrieben wird. Zu kritisieren ist etwa, dass KI-Anwendungen teilweise nur an Trainingsdatensätzen (interne Validierung), nicht aber an externen Datensätzen validiert werden (externe Validierung). Liu et al. (2019) weisen darauf hin, dass von 82 untersuchten Evaluationsstudien von KI-Systemen nur 24 % extern evaluiert wurden.

Das ist insbesondere deshalb zu kritisieren, weil die externe Validierung wichtig ist, um auf die Wirksamkeit von Systemen in der Praxis schließen zu können. Wenn Nutzer nur auf Ergebnisse einer internen Validierung zurückgreifen können, können sie von einer unangemessen hohen Leistungsfähigkeit ausgehen, die die Systeme bei der Anwendung in der Praxis nicht erreichen. Die Überschätzung von deren Wirksamkeit kann die gefährliche Folge nach sich ziehen, dass sich Nutzer übermäßig auf die Systemergebnisse verlassen, diese möglicherweise nicht hinreichend kontrollieren und den Patienten aufgrund falscher Entscheidungen schaden (vgl. Kapitel 4.4.2.2). Zur Steigerung der Patientensicherheit sollten AI-CDSS darum extern validiert werden (vgl. Kapitel 6.1.18.1).

Um die Wirksamkeit der KI-Systeme zu beurteilen, die den Arzt nicht bei der Diagnostik oder bei der Prognostik, sondern bei der Therapieentscheidung unterstützen, sind zwei Gruppen zu differenzieren.

Es gibt Anwendungen, deren Wirksamkeit man anhand ihrer Auswirkungen auf bestimmte Zielwerte evaluiert. Als Stellvertreter dieser ersten Gruppe soll hier ein System zur Unterstützung der Dosierung von unfraktioniertem Heparin (UH) angeführt werden. Die Wirksamkeit dieses Systems wurde evaluiert, indem die Zeitspanne gemessen wurde, in der die Zielgröße (aPTT), im therapeutischen Bereich lag (Nemati et al., 2016).

Diese Art der Wirksamkeitsbestimmung lässt sich auch bei einem anderen AI-CDSS finden, das in Kapitel 3.2 vorgestellt wurde (Escandell-Montero et al., 2014). Dieses System unterstützt Ärzte bei der Therapie von hämodialysepflichtigen anämischen Patienten. Zur Evaluation der Wirksamkeit ist die Auswirkung der Systemnutzung auf die Hämoglobinwerte beziehungsweise auf deren Zielbereich bestimmt worden.

In beiden Fällen – demjenigen des Systems zur Dosierung von UH als auch im Fall der Anwendung zur Behandlung von hämodialysepflichtigen anämischen Patienten – werden also Auswirkungen der Systemnutzung auf bestimmte Zielwerte beziehungsweise Surrogatparameter zur Bestimmung der Wirksamkeit untersucht. Was die Wirksamkeit dieser Systeme angeht, so geben Escandell-Montero et al. (2014) an, dass durch die Benutzung ihres Algorithmus gegenüber der Behandlung nach dem Standardprotokoll der Anteil von hämodialysepflichtigen Patienten mit normwertigen Hämoglobinwerten um 27,6 % gesteigert wurde.

Die Wirksamkeit von AI-CDSS im therapeutischen Bereich wird nicht immer auf die oben geschilderte Weise bestimmt. Eine zweite Gruppe von KI-Systemen wird anhand der Konkordanz evaluiert. Dieser Parameter gibt an, inwieweit die Outputs eines Systems den Therapieempfehlungen von Experten entsprechen (Tupasela & Di Nucci, 2020).

Hinsichtlich der Konkordanz stellt sich an dieser Stelle die Frage, ob anhand dieser Größe die Wirksamkeit eines AI-CDSS auf sinnvolle Weise evaluiert werden kann. Um dem nachzugehen, ist auf die Bedeutung der Wirksamkeit hinzuweisen: Die Wirksamkeit gibt an, inwiefern ein System das intendierte Einsatzziel erreicht (Marckmann, 2003, 2016). Wäre Konkordanz dazu geeignet, die Wirksamkeit eines AI-CDSS zu evaluieren, so müsste der Zielerreichungsgrad des Systems von der Konkordanz abgebildet werden.

Es gilt also zu fragen, die Erreichung welchen Ziels die Konkordanz abbildet. Weil diese Größe die Übereinstimmung zwischen Systemoutputs und den Ergebnissen bestimmter Experten angibt, lässt die Konkordanz Rückschlüsse darauf zu, inwieweit ein System eine Gruppe von Experten imitieren beziehungsweise ersetzen kann. Somit kann man die Nachahmung oder Ersetzung fachlich kompetenter Ärzte als Ziel bezeichnen, dessen Erreichung die Konkordanz im strikten Sinne abbilden kann. Wie bei der Bewertung der Machbarkeit erörtert wurde, können und sollen AI-CDSS den Arzt allerdings nicht im Ganzen ersetzen (vgl. Kapitel 4.1.1). Wäre das tatsächlich das Ziel der betreffenden Systeme, würden diese ihre Begrenzungen nicht beachten, ihre Nutzung wäre daher ethisch nicht vertretbar.

Auch im Bereich der Therapieempfehlung besteht das übergeordnete Ziel der KI-Nutzung darin, die Qualität der Patientenversorgung zu verbessern. Um die Wirksamkeit eines AI-CDSS zu evaluieren, das Therapieempfehlungen abgibt, könnte man daher überprüfen, wie sich die Behandlungsergebnisse entwickeln, wenn Ärzte das betreffende System verwenden.

In jedem Fall ist es problematisch, wenn Anwender nicht bestimmen können, wozu ein System dienen soll und wie wirksam es ist. Für die Nutzer muss klar sein, worin das Ziel eines AI-CDSS besteht und inwieweit dieses durch die Systemnutzung erreicht werden kann. Aus diesen Überlegungen lässt sich die Forderung ableiten, Anwendungen nach einem klar definierten und evaluierbaren Ziel zu entwickeln (vgl. Kapitel 6.1.4).

4.1.3 Brauchbarkeit

Nachdem dargelegt wurde, welche Aufgaben AI-CDSS erfüllen können (Machbarkeit) und wie sie dabei abschneiden (Wirksamkeit), ist zu untersuchen, ob die Systeme einen ärztlichen Bedarf an Entscheidungsunterstützung erfüllen (Brauchbarkeit).

Der Frage nach der Brauchbarkeit von AI-CDSS kann man sich erstens auf abstrakte Weise nähern. Insofern es sich bei den Anwendungen um eine Form von KI-Systemen handelt, kann man analysieren, bei welchen kognitiven Aufgaben und Arten von Entscheidungen diese dem Menschen überlegen sind. Van Baalen et al. (2021) gehen etwa davon aus, dass KI-Anwendungen im gesamten Gebiet des deduktiven und induktiven Schlussfolgerns den Menschen überträfen. Ob dieses Urteil in dieser Allgemeinheit richtig ist, ist jedoch fragwürdig. D. W. Bates et al. (2021) differenzieren, dass KI-Systeme vor allem bei der Analyse und bei der Schlussfolgerung aus großen, unstrukturierten Datensätzen besser als Menschen abschnitten. Auch in Bezug auf die Erkennung hochspezifischer Muster werden KI-Systeme dem Menschen gegenüber als überlegen aufgefasst (Sonar & Weber, 2020).

Der Brauchbarkeit von AI-CDSS kann man nicht nur über eine Analyse der Stärken der Systeme, sondern auch mit Blick auf die Schwächen der Ärzte nachgehen: Inwiefern weisen diese einen Bedarf an Unterstützung durch die Anwendungen auf? Um darauf eine Antwort zu finden, ist es sinnvoll, sich im Folgenden mit dem Prozess auseinanderzusetzen, den die Systeme unterstützen sollen – der klinischen ärztlichen Entscheidungsfindung.

Mit Blick auf die Praxis lassen sich Aspekte identifizieren, die auf einen ärztlichen Bedarf an Unterstützung hinweisen. Ein Phänomen, das unbestrittenermaßen nicht dem Anspruch der ärztlichen Entscheidungsfindung entspricht und auf einen Unterstützungsbedarf aufmerksam macht, ist die hohe Zahl von Behandlungsfehlern in der klinischen Praxis. Unter einem Behandlungsfehler kann die „Verletzung des fachlichen Standards“ verstanden werden, „der zurzeit der Behandlung aus medizinischer Sicht anerkannt ist“ (Glanzmann & Schiltenwolf, 2017, S. 21).

Um zu analysieren, inwiefern man Behandlungsfehler durch die Nutzung von AI-CDSS verhindern kann, sind diese Fehler genauer zu untersuchen. So kann man feststellen, dass im Jahr 2019 3 688 mit einem Schaden verknüpfte Behandlungsfehler vorgekommen

sind. Diese werden zu 31,8 % der operativen Therapie, zu 24,0 % der Befunderhebung, zu 9,2 % der Pflege, zu 6,2 % dem Erheben und Beherrschen von Komplikationen, zu 5,9 % der interventionellen Therapie und zu 5,8 % der medikamentösen Therapie zugeordnet. Fehler bei der Diagnosestellung schlagen mit einem Anteil von 5,3 % zu Buche (Verband der Ersatzkassen [vdek], 2019).

Dass ein Bedarf an Unterstützung der Diagnostik besteht, lässt sich aus diesen Zahlen ableiten. Wie viele der Behandlungsfehler nicht auf die Diagnosestellung, sondern auf die Prognostik und auf die Therapieentscheidung entfallen, lässt sich anhand dieser Statistik nicht eindeutig beantworten. Klar ist aber, dass nicht alle Behandlungsfehler auf Fehler bei der ärztlichen Entscheidungsfindung zurückgehen. Diese können auch etwa durch eine fehlerhafte Durchführung von medizinischen Maßnahmen entstehen. Wie hoch deren Anteil an der Gesamtheit der Behandlungsfehler ist, lässt sich anhand der genannten Statistik nur unzureichend abschätzen. Dennoch dürfte unbestritten sein, dass jedenfalls ein nicht unbeträchtlicher Teil der Behandlungsfehler auf die ärztliche Entscheidungsfindung zurückgeht. Inwiefern diese Fehler durch die Nutzung von AI-CDSS verhindert werden könnten, soll nachfolgend untersucht werden. Zunächst wird erläutert, wie Fehler bei der klinischen ärztlichen Entscheidungsfindung entstehen.

Ein Faktor, den man für Fehler bei der klinischen Entscheidungsfindung verantwortlich macht, besteht in fehler- oder mangelhaftem Wissen der Ärzte (Norman et al., 2017). Das überrascht kaum – das medizinische Wissen ist nicht nur in seinen Einzelheiten komplex, sondern auch umfangreich. Es entwickelt sich außerdem dynamisch weiter: Das gesamte Wissen der Medizin wuchs bereits im Jahre 2010 derart schnell, dass von einer Verdopplung des Wissensumfangs alle dreieinhalb Jahre ausgegangen werden konnte (Densen, 2011). Das macht es verständlich, dass es für Ärzte herausfordernd ist, ihr Wissen kontinuierlich mit dem aktuellen Stand der Wissenschaft abzugleichen. Hier liegt ein Hinweis auf die Brauchbarkeit von AI-CDSS: Die Systeme können ihre Berechnungen jedenfalls theoretisch stets an den neuesten Stand der Wissenschaft anpassen, auf diese Weise Fehler im Wissen der Ärzte ausgleichen und somit zur Vermeidung von Behandlungsfehlern beitragen.

Ein weiterer Grund für die Entstehung von Behandlungsfehlern ist zu identifizieren, wenn man sich dem kognitiven Apparat des Menschen zuwendet: Dieser ist anfällig für Bias. Verschiedene Untersuchungen zeigen, welchen typischen ‚Denkfehlern' Ärzte aufsitzen (Lilienfeld & Lynn, 2014; Mamede et al., 2010). Das Potenzial von AI-CDSS zur

Vermeidung von Behandlungsfehlern, die auf Bias beruhen, liegt auf der Hand: Von *kognitiven* Verzerrungen ist nur menschliches Denken betroffen. Die Ergebnisse von KI-Systemen können zwar durch Verzerrungen verfälscht werden (vgl. Kapitel 3.4.2), von den typisch menschlichen kognitiven Bias ist die Funktionsweise der Systeme aber frei.

Die bisherigen Ausführungen konnten zeigen, inwiefern AI-CDSS brauchbar sein können. Von diesem allgemeinen Urteil aus soll nun differenzierter gefragt werden, inwiefern die Systeme in den jeweiligen Anwendungsbereichen – Diagnostik, Prognostik und Therapieempfehlung – einen Bedarf an ärztlicher Entscheidungsunterstützung decken können. Hierzu sind im Folgenden entsprechende Evaluationsstudien in den Blick zu nehmen.

Bezüglich der Wirksamkeit wurde bereits dargelegt, dass diagnostische und prognostische Anwendungen oft eine hohe Genauigkeit aufweisen (vgl. Kapitel 4.1.2). Aus dem Nachweis einer hohen Wirksamkeit mag sich die Vermutung ableiten lassen, dass die Systeme in diesen Bereichen auch brauchbar sind. Eine höhere Aussagekraft bezüglich der Brauchbarkeit eines Systems besitzen aber Vergleichsstudien: Stellt sich heraus, dass ein AI-CDSS Diagnosen oder Prognosen mit einer höheren Genauigkeit als bestimmte Ärzte treffen kann, so spricht das für die Brauchbarkeit des Systems. Zu kritisieren ist, dass solche Vergleichsstudien nur selten durchgeführt werden: Einen Vergleich der Performanz des jeweiligen Systems mit der Leistung von geeigneten Fachleuten führten laut einer Übersichtsarbeit von Liu et al. (2019) nur 17 % der untersuchten Studien durch. In diesen 14 Arbeiten, die die diagnostische Performanz von Deep-Learning-Algorithmen einerseits und von Personen aus den Gesundheitsberufen andererseits verglichen, erreichten die Systeme gegenüber dem Gesundheitspersonal mit 87,0 % zu 86,4 % einen leicht besseren Wert bezüglich der gepoolten Sensibilität. Auch hinsichtlich der gepoolten Spezifität übertrafen die Systeme die Versuchspersonen (92,5 % zu 90,5 %) (Liu et al., 2019).

Die Leistungsfähigkeit von AI-CDSS und Ärzten soll hier nicht verglichen werden, ohne auf methodische Herausforderungen solcher Vergleichsstudien einzugehen. Zum einen stellt sich die Frage, welche Ärzte zum Vergleich herangezogen werden. Abhängig von der Qualifikation und der Berufserfahrung der Personen ist es unterschiedlich herausfordernd, diese zu übertreffen. Die Auswahl der Vergleichspersonen sollte darum

so erfolgen, dass einzelne potenzielle Nutzer beurteilen können, ob sie dem System bei einer bestimmten Aufgabe unterlegen sind (vgl. Kapitel 6.1.18.2).

Wenn ein AI-CDSS Diagnosen oder Prognosen nachweislich mit einer höheren Genauigkeit als Ärzte stellt, kann man daraus auf ein Potenzial zur Verbesserung der Entscheidungsfindung schließen. Das bedeutet aber nicht unbedingt, dass die diagnostische oder prognostische Genauigkeit eines Arztes durch die Nutzung dieses Systems gesteigert werden kann. Anhand von Evaluationsstudien lässt sich das verdeutlichen. Obwohl einige Systeme – wie eben dargelegt – dem Menschen bei der Diagnostik überlegen sind, kommt ein systematisches Review zu dem Schluss, dass es nur geringe Evidenz dafür gibt, dass die Verwendung von ML-basierten CDSS die diagnostische Genauigkeit von Ärzten verbessern kann (Vasey et al., 2021).

Teilweise können Ärzte durch die Nutzung eines KI-Systems ihre diagnostische Genauigkeit aber auch stärker steigern, als im Hinblick auf die Wirksamkeit der Anwendung zunächst zu erwarten wäre: Während Pathologen in einer Evaluationsstudie bei der Diagnostik von Brustkrebs eine Fehlerquote von 3,5 % aufwiesen und das AI-CDSS in 2,9 % der Fälle falsche Diagnosen stellte, erreichten Ärzte im Zusammenspiel mit dem System mit 0,5 % eine Fehlerquote, die besser als diejenige der KI-Anwendung war (D. Wang et al., 2016). Aus diesem Befund lässt sich schließen, dass man die Brauchbarkeit eines AI-CDSS nicht hinreichend durch einen Vergleich der Wirksamkeit des Systems und der Leistung von Ärzten evaluieren kann. Um diesen Parameter zu bestimmen, gilt es stattdessen die diagnostische oder prognostische Genauigkeit zu ermitteln, die Ärzte unter Zuhilfenahme der betreffenden Anwendung erreichen. Diese ist dann mit der Leistungsfähigkeit zu vergleichen, die die Ärzte erzielen, ohne das System zu verwenden (vgl. Kapitel 6.1.18.2).

Im Rahmen der vorliegenden Bewertung der Brauchbarkeit von AI-CDSS wurden bisher nur diagnostische und prognostische Systeme explizit behandelt. Untersucht man die Brauchbarkeit der Anwendungen im Bereich der Therapieempfehlung, so ist – wie hinsichtlich der Wirksamkeit in Kapitel 4.1.2 – wieder zwischen zwei Gruppen von Systemen zu unterscheiden. Die Wirksamkeit der ersten Klasse wird evaluiert, indem man die Auswirkungen der Systemnutzung auf bestimmte Zielwerte untersucht. Wie die Brauchbarkeit von solchen Anwendungen bestimmt werden kann, liegt auf der Hand: Das AI-CDSS kann man als brauchbar beschreiben, falls dessen Verwendung dazu führt, dass

bestimmte Werte oder Parameter etwa für eine längere Zeit oder zuverlässiger in einem bestimmten Zielbereich liegen als wenn Ärzte das entsprechende System nicht benutzen. Beispielhaft sei hier die Evaluation des Systems von Escandell-Montero et al. (2014) genannt. Es wird angegeben, dass durch die Nutzung dieser Anwendung im Vergleich zur Therapie gemäß dem herkömmlichen Protokoll die Hämoglobinwerte von einem 27,6 % größeren Anteil der Patienten im Zielbereich lagen (Escandell-Montero et al., 2014).

Bei der Untersuchung der Wirksamkeit von AI-CDSS im therapeutischen Bereich wurde von dieser Systemklasse eine zweite Gruppe unterschieden, die anhand von Konkordanz evaluiert wird (vgl. Kapitel 4.1.2). Der Parameter der Konkordanz erweist sich auch in Bezug auf die Beurteilung der Brauchbarkeit als problematisch: Selbst wenn nachgewiesen würde, dass die Übereinstimmungsrate der Therapieempfehlungen bestimmter Ärzte mit den Empfehlungen von Experten durch die Systemnutzung steigt, wäre es aus folgenden Gründen problematisch, daraus die Brauchbarkeit des Systems abzuleiten:

Geht man davon aus, dass sich die Experten tatsächlich durch besondere Expertise und Kompetenzen in ihrem Fachgebiet auszeichnen, ist zu erwarten, dass deren Therapieempfehlungen von hoher Qualität sind. Angesichts dessen mag man erwarten, dass eine Behandlungsempfehlung, die von derjenigen der Experten abweicht, in vielen Fällen von geringerer Qualität ist. Im strengen Sinne bedeutet eine Abweichung von der Empfehlung der Experten aber nicht immer, dass die abweichende Therapieempfehlung ‚schlechter' ist. Hier wird das Problem deutlich, das sich bereits im vorangegangenen Kapitel gezeigt hat: Konkordanzwerte geben nur Aufschluss darüber, inwiefern die Problemlösefähigkeit bestimmter Experten imitiert werden kann.

AI-CDSS sollen den Arzt aber nicht nachahmen oder ersetzen, sondern Ärzte unterstützen und somit einen Mehrwert für die klinische Entscheidungsfindung schaffen (vgl. Kapitel 4.1.1). Ein solcher besteht dann, wenn die Behandlungsergebnisse unter Nutzung des AI-CDSS besser sind, als wenn ein Arzt das System nicht verwendet. Da die Qualität der klinischen Entscheidungsfindung zwischen verschiedenen Ärzten differiert, ist es sinnvoll, bei der Evaluation mehrere Gruppen von Ärzten heranzuziehen, die bei der klinischen Entscheidungsfindung unterschiedliche Qualitätsniveaus erreichen (vgl. Kapitel 6.1.18.2). So könnte sich etwa zeigen, dass die Nutzung eines Systems die Behandlungsqualität von bestimmten Ärzten verbessern kann (etwa von solchen, die

unerfahren sind), andere Ärzte durch den Gebrauch des AI-CDSS Patienten aber nicht besser behandeln können.

Damit man eruieren kann, ob die Verwendung eines Systems die klinische Entscheidungsfindung eines Arztes verbessern kann, muss klar sein, worin diese Optimierung bestehen soll. Schließlich können AI-CDSS nicht die gesamte klinische Entscheidungsfindung, sondern nur einzelne Teile davon unterstützen (vgl. Kapitel 4.1.1). Darum sollte die Zielsetzung einer KI-Anwendung machbar und klar formuliert sein. Außerdem muss man evaluieren können, inwiefern deren Nutzung zur Erreichung des intendierten Ziels führt. Das Ziel eines AI-CDSS sollte deswegen nicht nur machbar und klar formuliert, sondern auch evaluierbar sein (vgl. Kapitel 6.1.4).

4.2 Mögliche Alternativen

Wie im vorangegangenen Kapitel deutlich wurde, setzt man AI-CDSS zur Erreichung eines oder mehrerer bestimmter Ziele ein. Wenn diese nicht nur durch die Verwendung eines KI-Systems erreicht werden können, ist vor dessen Einsatz im Sinne der Zweck-Mittel-Rationalität zu überprüfen, ob es ein Mittel gibt, das den Zweck des Einsatzes besser oder möglicherweise mit geringerem Aufwand erfüllt (vgl. Kapitel 6.3.2). Im Hinblick auf die Heterogenität der Anwendungsbereiche und Funktionen von AI-CDSS erscheint es kaum möglich und auch wenig sinnvoll, an dieser Stelle Alternativen zu allen KI-Systemen zur Unterstützung der ärztlichen Entscheidungsfindung darzulegen. Stattdessen soll nachfolgend beispielhaft eine Alternative zu einer Gruppe von AI-CDSS aus einem bestimmten Anwendungsfeld – der Diagnostik von Brustkrebs – aufgeführt werden und mit KI-Anwendungen verglichen werden.

Brustkrebs ist eine der häufigsten Tumorerkrankungen der Frau (Bray et al., 2018). Zur Früherkennung wird ein Mammographie-Screening für Patientinnen ab dem Alter von 50 Jahren bis zum Ende des 70. Lebensjahres empfohlen (DGGG & DKG, 2021). Bei der Beurteilung von Mammographie-Aufnahmen besteht eine hohe Variabilität (Elmore et al., 2009): Werden 1000 derartige Bilder befundet, so werden 1,5 falsch negative und 121,2 falsch positive Diagnosen gestellt (Nelson et al., 2016).

Sowohl falsch positive als auch falsch negative Befunde gehen jedoch mit einem beträchtlichen Schadenspotenzial einher: Wird eine vorhandene Brustkrebserkrankung nicht erkannt, werden vorerst keine Maßnahmen gegen das Tumorwachstum ergriffen,

wodurch der Tumor sich weiter vergrößern kann und zu gesundheitlichen Einschränkungen, sogar zum Tod der Patientin führen kann (Given-Wilson et al., 1997). Neben falsch negativen sind auch falsch positive Ergebnisse schädlich für die Patientin, wenn deshalb unnötige Eingriffe mit Schadenspotenzial durchgeführt werden. Außerdem kann eine falsch positive Diagnose und die damit einhergehende Belastung, vielleicht oder sogar wahrscheinlich an einer potenziell tödlichen Krebserkrankung zu leiden, langfristig negative psychosoziale Folgen nach sich ziehen (Brodersen & Siersma, 2013).

Angesichts der nicht unbeträchtlichen Zahl an Fehldiagnosen bei der mammographischen Diagnostik und der negativen Auswirkungen dieser falschen Diagnosen dürfte es folglich unbestritten sein, dass bei der Befundung von Mammographie-Aufnahmen der Bedarf besteht, die diagnostische Genauigkeit zu verbessern.

Zur Deckung dieses Bedarfs kann man AI-CDSS verwenden. Im Gebiet der mammographischen Befundung wurden – wie in Kapitel 3.4.1 dargelegt – einige KI-Systeme entwickelt. Dass diese Anwendungen dabei helfen können, die Genauigkeit bei der mammographischen Diagnostik zu verbessern, geht aus den recht hohen Genauigkeitswerten jedenfalls einiger dieser Systeme hervor: In der Übersichtsarbeit von Houssami et al. (2019) werden die *accuracy*-Werte von KI-Anwendungen zur Detektion von Brustkrebs mit 69,2 % bis 97,8 % angegeben.

Um die diagnostische Genauigkeit von Ärzten bei der mammographischen Befundung zu erhöhen, kann man auch edukative Maßnahmen durchführen. Geller et al. (2014) konnten zeigen, dass die Teilnahme an einer DVD-gestützten Schulung die diagnostische Sensitivität der Schulungsteilnehmer bei der Befundung von Mammographie-Aufnahmen verbesserte.[8] Da diese Ergebnisse unter Testbedingungen erzielt wurden, ist nicht sicher zu sagen, ob die Durchführung dieser edukativen Maßnahme auch in der Praxis die diagnostische Sensitivität von Ärzten erhöhen kann. Angesichts der Testergebnisse erscheint das aber nicht abwegig.

Geller et al. (2014) geben zwar nicht an, wie hoch die Kosten dieser Schulung sind. Im Vergleich zu den Kosten des Kaufs, der Inbetriebnahme und der Wartung von AI-CDSS

[8] Das Odds Ratio betrug 1,40 relativ zur Diagnose von Experten und 1,34 relativ zum Krebsstatus (Geller et al., 2014).

(vgl. Kapitel 4.11) ist es aber durchaus möglich, dass diese edukative Maßnahme aus ökonomischer Perspektive vorteilhaft ist. Als weiterer Vorteil gegenüber der Verwendung von AI-CDSS lässt sich feststellen, dass die Schulung die Fähigkeiten der Nutzer verbessert. Der Gebrauch von AI-CDSS kann diese hingegen verschlechtern (*deskilling*) (vgl. Kapitel 4.7).

Es mag überraschen, dass durch die edukative Maßnahme zwar die Sensitivität, aber nicht die Spezifität der Diagnostik verbessert werden konnte (Geller et al., 2014). Die Schulung bietet sich also nur zur Erhöhung der Sensitivität bei der Befundung von Mammographie-Aufnahmen an. Hieran wird deutlich, dass das Ziel einer Intervention zur Verbesserung der klinischen ärztlichen Entscheidungsfindung eindeutig festgelegt werden sollte (vgl. Kapitel 6.3.1): Im Rahmen der Verbesserung der Diagnostik sollte klar sein, ob etwa nur die Sensitivität, die Spezifität oder *accuracy* verbessert werden soll. Der ärztliche Bedarf an Entscheidungsunterstützung sollte also möglichst genau spezifiziert werden. Um die verschiedenen Alternativen miteinander vergleichen zu können, muss bestimmbar sein, inwiefern diese jeweils den spezifischen Bedarf an Entscheidungsunterstützung decken (vgl. Kapitel 6.3.2).

In diesem Zusammenhang ist zu kritisieren, dass die Genauigkeit mancher AI-CDSS im Bereich der mammographischen Befundung nur mit *accuracy*–Werten beschrieben wird (Houssami et al., 2019). Eine solche Verkürzung der Evaluation kann problematisch sein: Es ist möglich, dass ein Bedarf an Entscheidungsunterstützung beispielsweise nicht hinsichtlich der Sensitivität, aber bezüglich der Spezifität der mammographischen Diagnostik besteht. Werden verschiedene Maßnahmen zur Deckung dieses Bedarfs an Entscheidungsunterstützung verglichen, sollte man beurteilen können, wie sich die Alternativen auf die Spezifität der Diagnostik auswirken. Daraus lässt sich die Forderung ableiten, dass AI-CDSS nicht nur anhand eines Parameters, sondern mittels verschiedener relevanter Größen evaluiert werden sollten (vgl. Kapitel 6.1.18.1).

4.3 Nutzenpotenzial für die Patientinnen und Patienten

Es liegt in der Natur von Technik, dass deren Nutzung im Allgemeinen keinen Selbstzweck darstellt, sondern zur Erreichung eines erstrebenswerten außertechnischen Ziels beiträgt. In diesem Zusammenhang wird auch vom Nutzen(-potenzial) und von Nützlichkeit gesprochen (Marckmann, 2003).

Wie bereits im Rahmen der Zieldefinition von AI-CDSS dargelegt wurde (vgl. Kapitel 3.3), können mit dem Einsatz der KI-Systeme verschiedene außertechnische Ziele erreicht werden. So kann deren Verwendung etwa den Ärzten nützen, indem diese von bestimmten Aufgaben entlastet werden und das ärztliche Tun effizienter wird (vgl. Kapitel 4.11). Insofern die KI-Nutzung die zeitliche und die ökonomische Effizienz der Behandlung erhöht, mag diese auch für die Versicherungsgemeinschaft beziehungsweise die Gesellschaft als Ganze nützlich sein. Hiermit seien nur einige der unterschiedlichen mit AI-CDSS verbundenen Nutzenpotenziale aufgeführt.

In der EBEA wird unter den verschiedenen Nutzenpotenzialen eine bestimmte Form desselben herausgestellt: Da eHealth-Anwendungen in der Medizin verwendet werden, sollte deren Gebrauch vor allem denjenigen nutzen, die in der Medizin im Mittelpunkt stehen – den Patienten. Was die Bestimmung des Nutzens für diese angeht, so schlägt die EBEA die Evaluation anhand der Größen Morbidität, Mortalität und Lebensqualität vor.

In der vorliegenden Arbeit werden das Nutzenpotenzial für die Patienten einerseits und die verschiedenen Potenziale andererseits, die ab Kapitel 4.5 behandelt werden, getrennt. Das ist wohl nur artifiziell möglich. Die später in dieser Arbeit behandelten Chancen von AI-CDSS stellen nicht selten auch ein Nutzenpotenzial für die Patienten dar.

Beispielsweise sei hier die in Kapitel 4.9 diskutierte Chance der Systeme zur Verbesserung der Arzt-Patient-Beziehung angeführt: Dass sich deren Qualität auf die Therapieergebnisse auswirken kann und somit deren potenzielle Verbesserung auch für das Nutzenpotenzial von AI-CDSS für die Patienten relevant ist, zeigen empirische Ergebnisse (Derksen et al., 2013; Riedl & Schüßler, 2017).

Darüber hinaus sind andere ethische Implikationen, die in der vorliegenden Arbeit an späterer Stelle behandelt werden, nicht ohne Auswirkung auf Überleben und Lebensqualität der Patienten. Es gibt etwa Hinweise darauf, dass die Länge der Zeit, die Ärzte mit Patienten verbringen, sich auf die Behandlungsergebnisse auswirkt (Andreyeva et al., 2018). Das Potenzial von AI-CDSS, das in Kapitel 4.11 diskutiert wird, zeitliche Ressourcen von Ärzten freizusetzen und es diesen somit zu ermöglichen, mehr Zeit mit den Patienten zu verbringen, erscheint in diesem Licht als ein Nutzenpotenzial für die Patienten.

Ein weiteres derartiges Potenzial von AI-CDSS besteht in der Individualisierung der Behandlung (vgl. Kapitel 3.4.4): Indem die klinische Entscheidungsfindung an individuelle Charakteristika des Patienten angepasst wird, kann die Genauigkeit von Prognosen und Diagnosen erhöht werden. Außerdem können so das therapeutische Ansprechen und die Therapieergebnisse verbessert werden.

Auch wenn sich die Auswirkungen der Verwendung von AI-CDSS auf die Behandlungsergebnisse theoretisch-analytisch abschätzen lassen, kann man diese Folgen letztlich nur durch empirische Evaluationsstudien bestimmen. Trotz der großen ethischen Bedeutung des Nutzens von AI-CDSS für die Patienten ist aber ein Mangel an derartigen Arbeiten festzustellen. Yin et al. (2021) stellen fest, dass sich nur 14 von 51 untersuchten Evaluationsstudien von medizinischen KI-Systemen überhaupt auf Therapieergebnisse beziehen. In 11 Arbeiten werden zur Evaluation klinische Prozess- und Outcome-Parameter herangezogen. Hierunter wird unter anderem die Mortalität der Patienten gefasst, aber auch beispielsweise die Aufenthaltsdauer in der Klinik oder die Frequenz von Überweisungen auf die Intensivstation und die Rate von Wiedereinweisungen werden als klinische Prozess- und Outcome-Parameter verwendet. Diese Parameter lassen allerdings nur bedingt Schlüsse auf das Überleben und die Lebensqualität der Patienten zu. Die Feststellung dieser Übersichtsarbeit, dass 8 dieser 11 Evaluationsstudien positive Wirkungen auf die klinischen Prozess- und Outcome-Parameter beschreiben, hat daher nur einen eingeschränkten Aussagewert bezüglich der Frage nach dem Nutzen von AI-CDSS für die Patienten (Yin et al., 2021).

Auch die Lebensqualität der Patienten sollte man bei der Evaluation des Nutzens für die Patienten als relevante Größe beachten. Dass die Verwendung von AI-CDSS diese positiv beeinflussen kann, wurde bereits gezeigt: Van Leeuwen et al. (2021) geben zum Beispiel an, dass die Nutzung eines KI-Systems zur Detektion von Großgefäßokklusionen bei Patienten mit Schlaganfall im Vereinigen Königreich mit einer Zunahme von *quality adjusted life years* einherging.

Zum Abschluss der Bewertung des Nutzenpotenzials von AI-CDSS kann als wesentlicher Befund der Ausführungen die niedrige Zahl an Evaluationsstudien festgehalten werden. Das dürfte unter anderem darin begründet sein, dass es als herausfordernd gilt, den Nutzen von KI-Systemen zu evaluieren (Zhou et al., 2021). Das kann aber die niedrige Zahl solcher Studien nicht rechtfertigen. Der Nutzen für die Patienten ist die zentrale ethische Rechtfertigung für die Verwendung von AI-CDSS. Es gilt daher die Bemühungen zu

intensivieren, um die Auswirkungen der Systemnutzung auf das Überleben und die Lebensqualität der Patienten zu erforschen (Wolff et al., 2021) (vgl. Kapitel 6.1.18.3).

4.4 Schadenspotenzial für die Patientinnen und Patienten

Das Schadenspotenzial der Nutzung von AI-CDSS für die Patienten kann anhand der Auswirkungen auf deren Überleben und auf deren Lebensqualität erhoben werden – also mithilfe derjenigen Größen, die auch der Bestimmung des Nutzenpotenzials dienen (vgl. Kapitel 4.3). Auch in Bezug auf das Schadenspotenzial lässt sich darum die Forderung, mit der das letzte Kapitel beschlossen wurde, wiederholen: Die Auswirkungen der Nutzung von AI-CDSS auf Morbidität, Mortalität und Lebensqualität der Patienten müssen empirisch evaluiert werden. Da die diesbezügliche Studienlage bereits im vorangegangenen Kapitel erläutert wurde, soll der Frage nach dem Schadenspotenzial der Systeme im vorliegenden Kapitel nicht mit dem Verweis auf empirische Ergebnisse, sondern theoretisch-analytisch begegnet werden.

Um zu eruieren, wie die Nutzung von AI-CDSS Patienten schaden kann, ist es sinnvoll, sich vor Augen zu führen, dass die Anwendungen keine Maßnahmen durchführen, sondern Ärzte bei der klinischen Entscheidungsfindung unterstützen (vgl. Kapitel 4.1.1). Diese zielt ihrerseits auf die Durchführung bestimmter diagnostischer und therapeutischer Maßnahmen ab, die dem Patienten nutzen wie auch schaden können. Anders als ein Röntgenapparat, dessen Nutzung mit Strahlenbelastung verbunden ist, können AI-CDSS als solche aber keinen Schaden bewirken. Ein Zusammenhang zwischen einer Anwendung und einer Schädigung des Patienten kann erst dann entstehen, wenn ein Arzt eine Systemempfehlung in seine Entscheidungsfindung einfließen lässt.

Mit AI-CDSS verbundene Schädigungen des Patienten gehen daher grundsätzlich entweder auf falsche Systemergebnisse oder auf Fehler bei der Systembedienung zurück. Dementsprechend wird nachfolgend eine Unterscheidung zwischen *falschen Systemergebnissen* und *Anwendungsfehlern* getroffen. Ausdrücklich soll noch darauf verwiesen werden, dass es nicht möglich ist, jede mit KI verbundene Schädigung des Patienten eindeutig *entweder* auf ein falsches Ergebnis *oder* auf einen Anwendungsfehler zurückzuführen. Schaden für den Patienten entsteht nämlich nicht selten dadurch, dass das System falsche Ergebnisse ausgibt und der Nutzer Anwendungsfehler begeht. Um in konstruktiver Absicht zu versuchen, einen Beitrag zur Minimierung dieser Fehler zu leisten, ist es dennoch sinnvoll, dem Schadenspotenzial getrennt hinsichtlich falscher

Systemergebnisse einerseits und Anwendungsfehler andererseits auf den Grund zu gehen.[9]

4.4.1 Schadenspotenzial durch die Ausgabe falscher Systemergebnisse

Die Ursachen für die Ausgabe falscher Systemergebnisse liegen vor allem in der Entwicklung der Systeme begründet. Deshalb ist hier auf verschiedene Formen von ML zu rekurrieren, die in Kapitel 3.1.2.1 erläutert wurden.

Um die Entstehungsweise eines bestimmten Fehlers zu erklären, der beim überwachten Lernen geschehen kann, soll nochmals diese Form des Maschinellen Lernens rekapituliert werden. Zur Veranschaulichung soll dieser Prozess im Anschluss an acatech (2020) anhand der fiktiven Entwicklung eines Systems erläutert werden, das die Aufgabe hat, Bildaufnahmen von malignen Melanomen und von gutartigen Hauttumoren zu unterscheiden.

Für die Entwicklung eines solchen Systems mit überwachtem Lernen benötigt man einen Datensatz, der verschiedene Aufnahmen von malignen Melanomen und von benignen Hauttumoren beinhaltet. Zur Erstellung dieses Datensatzes müssen die Bilder nach den Kategorien ‚maligne' und ‚benigne' klassifiziert werden. Bei dieser Zuordnung können Fehler geschehen: Eine Aufnahme einer bösartigen Läsion kann etwa fälschlicherweise als gutartig klassifiziert werden. Ebenso kann auch ein Bild eines benignen Tumors als maligne eingeordnet werden. Durch solche fehlerhaften Bilddiagnosen verfälscht man die sogenannte Grundwahrheit des Trainingsdatensatzes (*ground truth*) (DSK, 2019). Wird ein ML-System zur Detektion maligner Melanome an einem fehlerhaften Datensatz trainiert, kann es bei der Nutzung zu falschen Ergebnissen kommen. Um dieses Schadenspotenzial zu minimieren, sollte die Richtigkeit der Grundwahrheit in Trainingsdatensätzen gewährleistet werden (acatech, 2020) (vgl. Kapitel 6.1.17).

Nicht nur beim überwachten Lernen, sondern auch beim bestärkenden Lernen können Fehler geschehen, die die Qualität des entstehenden Systems beeinträchtigen können. Ein Phänomen, das beim bestärkenden Lernen dazu führen kann, dass das System falsche Ergebnisse ausgibt, ist *reward hacking*. Dieses Phänomen ist anhand des Lernprozesses zu erklären: Beim bestärkenden Lernen probiert das KI-Modell eigenständig

[9] Aufgrund der besonderen Bedeutung von ML für AI-CDSS wird hier explizit auch auf Fehler eingegangen, die im Zusammenhang mit ML stehen.

verschiedene Strategien aus, um ein von den Entwicklern vorgegebenes Ziel zu erreichen. Als *reward hacking* wird bezeichnet, wenn das KI-Modell dabei einen Weg findet, dieses Ziel auf eine von den Entwicklern nicht intendierte Weise zu erreichen – es findet gewissermaßen ein Schlupfloch (Challen et al., 2019).

Fehlerhafte Outputs von AI-CDSS können auch in Verzerrungen (Bias) der Trainingsdatenbasis begründet sein (vgl. Kapitel 3.4.2). Um derlei Fehler zu verhindern, kann man die Entwicklung des Systems am Ideal der Generalisierbarkeit ausrichten und etwa einen möglichst diversen Trainingsdatensatz verwenden. Problematisch ist allerdings, dass nicht in allen Bereichen hinreichend heterogene Datensätze vorhanden sind. Der Versuch, ein AI-CDSS möglichst generalisierbar zu gestalten, ist außerdem in der Hinsicht limitiert, dass universelle Generalisierbarkeit zumindest in manchen Fällen prinzipiell unerreichbar ist (Beede et al., 2020; Mitchell et al., 2021).

Das Ziel der universellen Generalisierbarkeit ist aber nicht nur in pragmatischer Hinsicht infrage zu stellen: Je generalisierbarer ein System ist, desto stärker sinkt üblicherweise die Leistungsfähigkeit des Systems in einzelnen Populationen (Futoma et al., 2020). Daraus kann sich folgendes Dilemma ergeben:

Entweder ein AI-CDSS wird mit dem Ziel einer möglichst hohen Performanz in einer bestimmten Population entwickelt, worunter die Generalisierbarkeit leiden kann. Das System kann dadurch in manchen Gruppen übermäßig häufig zu falschen Ergebnissen kommen (vgl. Kapitel 3.4.2). In diesem Fall sind für bestimmte Populationen die Nutzenpotenziale des AI-CDSS groß, für andere Gruppen birgt das System hingegen gewichtige Schadenspotenziale. Hier werden Fragen der Gerechtigkeit aufgeworfen (vgl. Kapitel 4.12).

Oder die Systementwicklung wird am Ideal der maximalen, universellen Generalisierbarkeit ausgerichtet. Dadurch können die lokalen Unterschiede der Leistungsfähigkeit des Systems nivelliert werden. Als potenzieller Nachteil dieses Vorgehens ergibt sich aber, dass die maximale Leistungsfähigkeit des Systems sinken kann. Insofern das Nutzenpotenzial der Anwendung für die Patienten von deren Leistungsfähigkeit abhängt, kann hier auch die Nützlichkeit des AI-CDSS beeinträchtigt werden. Das ist aus ethischer Sicht besonders kritisch: Der Nutzen eines Systems stellt schließlich dessen zentrale ethische Legitimation dar (vgl. Kapitel 4.3).

Hat man sich zu entscheiden, ob eine Anwendung am Ziel der universalen Generalisierbarkeit ausgerichtet wird, muss letztlich das Nutzenpotenzial des AI-CDSS gegenüber gerechtigkeitsethischen Implikationen abgewogen werden. Aus ethischer Perspektive erscheint ein vermittelnder Weg attraktiv: Falsche Systemergebnisse, die auf Verzerrungen des Trainingsdatensatzes beruhen, kann man auch verhindern, indem man Systeme lokal validiert. Dadurch können Verzerrungen des Trainingsdatensatzes identifiziert werden, die die Ergebnisse des Systems bei der Anwendung an einem bestimmten Ort verfälschen können. Gegebenenfalls können die Systeme dann für die lokale Anwendung rekalibriert werden (vgl. Kapitel 6.1.7) (Mitchell et al., 2021). Somit kann man Schadenspotenziale minimieren, die auf Verzerrungen des Trainingsdatensatzes beruhen – ohne dabei das zugehörige Nutzenpotenzial einzuschränken.

Bei der bisherigen Bewertung des Schadenspotenzials der Anwendungen wurde deutlich, dass die Menge und die Tragweite falscher Outputs nicht unbeträchtlich sind. Brisant ist, dass eine vollständige Fehlerfreiheit der Systeme generell nicht erreichbar ist. An dieser Stelle ist jedoch zu bedenken, dass auch Menschen nicht fehlerfrei arbeiten. Allein der Umstand, dass AI-CDSS teilweise falsche Ergebnisse ausgeben, sollte darum nicht zu einer kategorischen Ablehnung dieser Systeme führen. Zentral sollte vielmehr sein, ob der Arzt durch die Nutzung einer bestimmten Anwendung mehr oder weniger Fehler bei der klinischen Entscheidungsfindung begeht und ob folglich die Überlebenswahrscheinlichkeit und die Lebensqualität der Patienten von dessen Verwendung profitieren oder nicht (vgl. Kapitel 4.1.3). Hier wird erneut die Bedeutung von empirischen Studien augenfällig, welche die Auswirkungen der Systemnutzung auf Morbidität, Mortalität und Lebensqualität der Patienten untersuchen (vgl. Kapitel 6.1.18.3).

Damit die Verwendung von AI-CDSS, die den Patienten potenziell gefährden, ethisch vertretbar sein kann, ist noch eine weitere Bedingung zu nennen: Die Ausgabe fehlerhafter Systemergebnisse muss nach Möglichkeit minimiert werden. Der Fehlervermeidung kommt darum aus ethischer Sicht große Bedeutung zu (vgl. Kapitel 6.1.17). Dabei sollte beachtet werden, dass nicht jedes falsche Systemergebnis für den Patienten in gleicher Weise schädlich ist. Aus ethischer Sicht ist

zu fordern, dass man vor allem die Ausgabe von denjenigen falschen Outputs verhindert, die zu besonders schwerem Schaden für den Patienten führen können.

Man kann versuchen, den Schädlichkeitsgrad falscher Systemergebnisse anhand der drei Anwendungsbereiche von AI-CDSS – Diagnostik, Prognostik und Therapieempfehlung – zu differenzieren. In diesem Zusammenhang ist etwa bedeutend, dass die Therapieempfehlung enger als die Diagnostik oder die Prognostik mit der Durchführung von Maßnahmen und daher mit der potenziellen Schädigung des Patienten verknüpft ist. Somit mag man die Ausgabe falscher Ergebnisse im Bereich der Therapieempfehlung für besonders schädlich halten. Gleichwohl ist zu bedenken, dass auch mit falschen Diagnosen oder Prognosen schwerwiegende Folgen einhergehen können. Die Schwere der Konsequenzen von fehlerhaften Outputs kann man darum nur bedingt anhand des Anwendungsbereichs eines AI-CDSS bestimmen. Entscheidend sind vielmehr die spezifischen Eigenschaften eines Systems und die Beschaffenheit des jeweiligen Anwendungsbereichs. Daher sollte man in Bezug auf jedes System einzeln bewerten, wie schwer die Folgen von falschen Ergebnissen sind (vgl. Kapitel 6.1.18.4). In diesem Zusammenhang erscheint eine Einteilung von AI-CDSS in Risikoklassen sinnvoll, weil Ärzte somit gegebenenfalls darüber informiert werden können, dass bei der Verwendung eines Systems ein besonderes Maß an Aufmerksamkeit und Sorgfalt notwendig ist (vgl. Kapitel 6.2.2).

Auch für falsche Systemergebnisse, die besonders schweren Schaden für die Patienten nach sich ziehen, gilt jedoch, dass auch diese sich nur bedingt vermeiden lassen. Eine komplette Fehlerfreiheit ist nicht zu gewährleisten. Um das Schadenspotenzial durch die Ausgabe falscher Outputs so weit wie möglich zu reduzieren, kommt darum nicht nur der Minimierung der Fehlerwahrscheinlichkeit große Relevanz zu. Es gilt zu bedenken, dass ein falsches Systemergebnis an sich noch nicht in einer Schädigung des Patienten resultieren muss: AI-CDSS treffen keine Entscheidungen, sondern unterstützen den Arzt bei der klinischen Entscheidungsfindung. Erst wenn der Arzt falsche Systemempfehlungen nicht erkennt und diese in seine Entscheidungsfindung aufnimmt, können fehlerhafte Ergebnisse dem Patienten schaden.

Um also das Schadenspotenzial zu minimieren, das mit der Ausgabe falscher Outputs verbunden ist, sollten die Ärzte bei deren Detektion unterstützt werden. Das kann geschehen, indem die Nutzer darüber informiert werden, in welchen Szenarien, bei welchen Aufgaben und in welcher Hinsicht eine bestimmte Anwendung typischerweise

zu falschen Ergebnissen kommen kann (vgl. Kapitel 6.1.17). So können die Ärzte sich bei der Überprüfung der Ergebnisse auf die Detektion besonders häufiger Fehler konzentrieren, wodurch die Zahl der detektierten falschen Systemergebnisse steigen dürfte. Darüber hinaus sollte den Ärzten mitgeteilt werden, in welchen Situationen das System typischerweise falsche Outputs angibt, die dem Patienten in besonders schwerer Weise schaden können (vgl. Kapitel 6.3.12). Auch wenn also das Fehlerpotenzial von AI-CDSS nicht eliminiert werden kann, kann auf diese Weise jedenfalls deren Schadenspotenzial reduziert werden.

4.4.2 Schadenspotenzial durch Anwendungsfehler

Eine Schädigung des Patienten bei der Nutzung von AI-CDSS muss nicht durch die eben behandelte Ausgabe falscher Systemergebnisse bedingt sein. Auch dem Nutzer können Fehler unterlaufen. Diese Anwendungsfehler werden nachfolgend untersucht.

Die fehlerhafte Nutzung von AI-CDSS kann in mindestens drei Formen unterteilt werden. Erstens kann ein System in einem Kontext oder in einer Situation angewandt werden, in der es nicht benutzt werden kann oder sollte (vgl. Kapitel 4.4.2.1). Eine zweite Art von Anwendungsfehlern ist eng mit falschen Systemergebnissen verbunden: Der Nutzer kann falsche Empfehlungen des Systems, die theoretisch erkannt werden können, aus verschiedenen Gründen übersehen oder nicht als Fehler identifizieren und diese in seine klinische Entscheidungsfindung übernehmen (vgl. Kapitel 4.4.2.2). Schließlich kann eine fehlerhafte Anwendung der Systeme in Überdiagnostik resultieren (vgl. Kapitel 4.4.2.3).

4.4.2.1 Anwendung eines AI-CDSS trotz fehlender Anwendbarkeit

Schon in Kapitel 4.1.1 wurde festgestellt, dass als Input eines spezifischen AI-CDSS nur bestimmte Daten dienen können. Zum Beispiel ist ein System, das zur Interpretation von CT-Bildern des Röntgen-Thorax entwickelt wurde, typischerweise nicht zur Befundung von MRT-Aufnahmen geeignet. Auch die Qualität der Daten darf ein gewisses Niveau nicht unterschreiten, damit das System zu richtigen Ergebnissen kommen kann. Falls Daten in ein AI-CDSS eingespeist werden, die nicht die richtige Form oder nicht ein hinreichendes Qualitätsniveau aufweisen, liegt ein Anwendungsfehler vor (van Baalen et al., 2021). Damit die Nutzer diesen Fehler vermeiden können, sollten die Entwickler klarstellen, welche Form und welche Qualität die Inputdaten des Systems aufweisen müssen (vgl. Kapitel 6.1.19).

In Kapitel 3.4.2 wurde bereits die Problematik der eingeschränkten Generalisierbarkeit erläutert: Manche AI-CDSS geben falsche Ergebnisse aus, wenn sie in bestimmten Kontexten oder Populationen angewandt werden. Falls ein System für die Verwendung in einem spezifischen Zusammenhang nicht geeignet ist, aber dennoch benutzt wird, könnte man auch von einem Anwendungsfehler sprechen. Da universelle Generalisierbarkeit zumindest in einigen Fällen nicht erreichbar ist, ist davon auszugehen, dass die Nutzung von AI-CDSS auch weiterhin mit der Herausforderung der eingeschränkten Generalisierbarkeit konfrontiert sein wird.

Eine Möglichkeit, um verzerrte Systemergebnisse zu verhindern, zeigt sich, wenn man sich nochmal das Phänomen der eingeschränkten Generalisierbarkeit und dessen Ursache vor Augen führt: Zu falschen Ergebnissen kann ein AI-CDSS kommen, wenn zwischen den Trainings- und den Inputdaten ein zu großer Unterschied besteht (vgl. Kapitel 3.4.2). Insofern die Daten, die bei der Nutzung eingegeben werden, niemals identisch sind mit den Trainingsdaten, ist zu bedenken, dass stets ein Unterschied zwischen Trainings- und Inputdaten besteht. Wann dieser so groß ist, dass das System zu unverhältnismäßig vielen falschen Ergebnissen kommt, ist schwierig zu beurteilen. In manchen Fällen mögen sich Trainings- und Inputdaten aber offenkundig so stark voneinander unterscheiden, dass auch Ärzte erkennen können, dass das AI-CDSS in einem bestimmten Fall nicht verwendet werden sollte. Wenn beispielsweise ein System zur Identifizierung von Hautläsionen nur anhand von Bildern von hellhäutigen Patienten trainiert wurde, dürfte man darauf schließen können, dass es bei der Behandlung von dunkelhäutigen Patienten ein unverhältnismäßig hohes Fehlerpotenzial besitzt. Es ist also sinnvoll, die Charakteristika von Trainingsdatensätzen zu explizieren, damit Anwender überprüfen können, ob das System im vorliegenden Fall angewandt werden kann (vgl. Kapitel 6.3.11).

Gleichzeitig ist kaum davon auszugehen, dass Ärzte in jedem Fall entscheiden können, ob der Unterschied zwischen den Trainings- und den Inputdaten so groß ist, dass mit (unverhältnismäßig vielen) fehlerhaften Ergebnissen zu rechnen ist. Daher ist es ratsam, AI-CDSS lokal zu validieren (vgl. Kapitel 6.1.7). Dabei kann man eruieren, ob das System bei der Anwendung in bestimmten Umgebungen oder in bestimmten Populationen besonders häufig zu falschen Ergebnissen kommt. Aus diesem Grund kommt der lokalen Validierung und der Rekalibrierung zur Fehlervermeidung große Bedeutung zu. Werden mehrere lokale Validierungen eines Systems durchgeführt, kann

man eingrenzen, in welchen Populationen und in welchen Anwendungsumgebungen eine Nutzung des Systems vertreten werden kann (vgl. Kapitel 6.1.17). Um die Sicherheit des Patienten so gut wie möglich zu gewährleisten, sollte an zwei Stellen bedacht werden, ob das AI-CDSS bei der Anwendung zu verzerrten, falschen Ergebnissen kommen wird: Bei der Anschaffung eines Systems (vgl. Kapitel 6.3.3) und unmittelbar vor dessen Nutzung zur Behandlung eines bestimmten Patienten (vgl. Kapitel 6.3.11).

4.4.2.2 Übernahme falscher Systemergebnisse

Die zweite Form von Anwendungsfehlern, die hier zu behandeln ist, weist eine enge Verbindung mit falschen Systemergebnissen auf. Zum Verständnis dieser Fehler soll nochmal die Aufgabe von AI-CDSS in Erinnerung gerufen werden: Sie sollen die Ärzte *unterstützen* (vgl. Kapitel 4.1.1). Die Outputs der Anwendungen sind darum nicht als schlechthin richtige Entscheidungen aufzufassen, sondern als Empfehlungen, welche die Ärzte grundsätzlich zumindest auf Plausibilität hin zu kontrollieren haben. Falls ein falsches Ergebnis vorliegt, sollte der Nutzer dieses möglichst erkennen. Wenn er dennoch ein falsches Systemergebnis nicht identifiziert und er dieses in die klinische Entscheidungsfindung aufnimmt, können hierfür mindestens zwei Gründe vorliegen. Zum einen könnte der Anwender es unterlassen haben, das Ergebnis hinreichend zu kontrollieren. Dieses Szenario soll im Folgenden zuerst beleuchtet werden, bevor untersucht wird, warum Ärzte – wenn sie die Ergebnisse kontrollieren – dennoch falsche Outputs übersehen können.

Wie kann es dazu kommen, dass Nutzer die Ergebnisse von AI-CDSS nicht (ausreichend) kontrollieren? Hierbei spielt übersteigertes Vertrauen auf die Systeme eine wichtige Rolle: Wer übermäßig auf die Richtigkeit eines Outputs vertraut, dürfte dazu tendieren, dieses nicht hinreichend zu prüfen. Übersteigertes Vertrauen auf automatisierte Systeme zur Entscheidungsunterstützung wird vor allem unter dem Schlagwort des Automation Bias behandelt (Bahner, 2008; Parasuraman & Manzey, 2010). Zur Verhinderung von dessen potenziell schädlichen Folgen kommt dessen Entstehungsfaktoren große Bedeutung zu. Diese lassen sich zum einen in interne, den Nutzer betreffende und zum anderen in externe Faktoren unterteilen. Zu den **internen** Faktoren sind folgende zu zählen:

1. Ein hohes Verantwortungsbewusstsein kann die Entstehung des Automation Bias teilweise verhindern (Burdick et al., 1996; Mosier et al., 1998; Skitka et al., 2000).

2. Auch von individuellen persönlichen Eigenschaften der Nutzer scheint die Entstehung des Automation Bias abzuhängen (Bahner, 2008; Goddard et al., 2012).

3. Erfahrung bei der Benutzung des betreffenden Systems kann die Gefahr erhöhen, dass der Automation Bias eintritt: Wer ein System zu kennen meint, unterlässt eher die eingehende Prüfung der Outputs (Bailey, 2004).

4. Nutzer, die mit der Durchführung der automatisierten Aufgabe wenig erfahren sind, weisen ein höheres Risiko für die Entwicklung des Automation Bias auf (Bailey, 2004; Dreiseitl & Binder, 2005; Gaube et al., 2021; Goddard et al., 2012; Moray et al., 2000).

Darüber hinaus gibt es **externe** Faktoren für die Entstehung des Automation Bias:

1. Eine Übersichtsarbeit bestimmt als grundsätzlichen externen Auslöser des Automation Bias hohe kognitive Belastung – etwa durch Multitasking (Lyell & Coiera, 2017). Somit ist beispielsweise auch zu erklären, dass zeitlicher Druck die Entwicklung des Automation Bias begünstigt (Sarter & Schroeder, 2001).

2. Kognitive Herausforderungen können aber auch durch eine Vigilanzsteigerung das Risiko für die Entstehung des Automation Bias verringern (Goddard et al., 2012; Xu et al., 2007).

In diesem Zusammenhang bringt folgender Umstand ein ethisches Dilemma mit sich: Empfehlungen verbessern die Entscheidungsqualität von fachlich unerfahrenen Anwendern mehr als diejenigen von erfahrenen Nutzern. Gleichzeitig weisen aber Anwender mit wenig Erfahrung hinsichtlich der automatisierten Aufgabe eine besonders hohe Anfälligkeit für den Automation Bias auf (Bailey, 2004; Goddard et al., 2012; Moray et al., 2000). Um die Gefahr der Übernahme inkorrekter Empfehlungen zu minimieren, könnte man die Systeme so konstruieren, dass unerfahrenen Nutzern nur Empfehlungen mit einem besonders hohen Konfidenzgrad angezeigt werden (vgl. Kapitel 6.1.12). Die Auseinandersetzung mit weniger sicheren Empfehlungen wäre dann fachlich erfahrenen Nutzern vorbehalten, die die Richtigkeit der Outputs besser einschätzen können.

Alleine, dass der Arzt seiner Verpflichtung zur Kontrolle der Systemergebnisse nachkommt, reicht noch nicht aus, um zu verhindern, dass er falsche Outputs übernimmt.

Auch beim Prozess der Ergebniskontrolle können Fehler geschehen. Hier ist besonders der Fall relevant, dass sich das Output des AI-CDSS und das Ergebnis, zu dem der Anwender gekommen ist, widersprechen. Soll sich ein Arzt entscheiden, ob er ein Systemergebnis übernehmen soll, ist es hilfreich, wenn das System das Output erklärt. Geben die Erklärungen Hinweise darauf, dass die Berechnungen des AI-CDSS in dem vorliegenden Fall etwa auf falschen Prämissen aufbauen, kann der Arzt sich gegen die Übernahme des falschen Systemergebnisses entscheiden.

Zur Veranschaulichung kann hier ein fiktives System dienen, das einen Lungentumor in einer CT-Aufnahme detektiert. Wird als Begründung für diese Diagnose ein bestimmtes Bildareal hervorgehoben, das der Arzt sicher als Artefakt identifizieren kann, spricht das gegen die Richtigkeit der Diagnose des AI-CDSS. Insofern – wie in diesem Fall – Erklärbarkeit dabei helfen kann, Schaden zu vermeiden, ist diese ethisch geboten (vgl. Kapitel 6.1.16).

Wird ein Systemergebnis ohne Erklärung angegeben, so steht der Arzt in dieser Situation vor der Wahl zwischen ‚blinder' Akzeptanz oder Ablehnung des Systemergebnisses. Hier sollte entscheidend sein, welches Ergebnis mit höherer Wahrscheinlichkeit richtig ist – das, zu dem der Arzt ohne Zuhilfenahme des AI-CDSS kommt, oder das Output des Systems.

Um im Bereich der Diagnostik und der Prognostik herauszufinden, ob das Ergebnis des Arztes oder dasjenige des Systems mit höherer Wahrscheinlichkeit richtig ist, kann man sich auf die Genauigkeit des Systems beziehen und die (geschätzte) ärztliche Treffsicherheit damit vergleichen. Kann man davon ausgehen, dass das System bei der betreffenden Aufgabe eine höhere Leistungsfähigkeit als der Arzt aufweist, spricht das für die Übernahme des Systemergebnisses. Ein solcher Vergleich zwischen Mensch und Maschine ist im Bereich der Therapieempfehlung jedoch problembehaftet: Da Konkordanzwerte im strengen Sinne nicht die Wirksamkeit eines AI-CDSS angeben (vgl. Kapitel 4.1.2), ist die Evaluierung der Systeme anhand von Konkordanz auch an dieser Stelle kritisch zu bewerten. Hier zeigt sich erneut die Bedeutung einer klaren und evaluierbaren Zieldefinition von AI-CDSS (vgl. Kapitel 6.1.4).

Da eine rigorose Wirksamkeitsbestimmung dafür nötig ist, dass man sich für oder gegen die Übernahme eines umstrittenen Systemergebnisses entscheiden kann, erscheint in diesem Licht das sogenannte *overfitting* besonders problematisch: Das KI-Modell passt sich zu stark an den Trainingsdatensatz an, weshalb es bei der Eingabe von

Trainingsdaten typischerweise eine hohe Leistungsfähigkeit aufweist (Henzel, 2019). Die Performanz kann aber in der Praxis deutlich geringer ausfallen, wenn andere Daten als Input verwendet werden (Beil et al., 2019; Winkler et al., 2020). Somit können die trügerisch guten Evaluationsergebnisse die Nutzer hinsichtlich der Wirksamkeit des Systems in der Praxis täuschen. Damit man *overfitting* detektieren und somit möglichen Schaden von den Patienten abwenden kann, sollte ein System anhand von Daten evaluiert werden, die nicht für dessen Training verwendet wurden (Collins et al., 2014) (vgl. Kapitel 6.1.18.1).

4.4.2.3 Überdiagnostik

Anwendungsbedingten Schaden für den Patienten kann außerdem Überdiagnostik hervorrufen. Bevor untersucht wird, inwiefern die Verwendung von AI-CDSS Überdiagnostik und -therapie verstärken kann, soll zuerst dieses Phänomen erläutert werden. Hierbei ist zunächst zu bedenken, dass diagnostische wie auch therapeutische Maßnahmen neben Nutzen- auch Schadenspotenziale aufweisen. Damit deren Durchführung ethisch vertretbar sein kann, muss das zugehörige Nutzenpotenzial das Schadenspotenzial überwiegen. Wenn hingegen bei einer diagnostischen Maßnahme Letzteres Ersteres übertrifft, spricht man von Überdiagnostik (Brodersen et al., 2018).

Die Verwendung von AI-CDSS kann Überdiagnostik aus verschiedenen Gründen verstärken. Ein erster Faktor, der dieses Phänomen potenziell fördert, besteht in der oftmals hohen Genauigkeit und Sensitivität der KI-basierten Diagnostik (Carter et al., 2020). Dass man mithilfe von KI-Anwendungen Krankheiten anhand von Ausprägungen erkennen kann, die man ansonsten kaum als pathologisch identifizieren könnte, stellt zunächst eine große Chance dar: Patienten, deren Erkrankungen man ohne die Nutzung von AI-CDSS nicht detektieren könnte, können somit einer geeigneten Behandlung zugeführt werden.

Beispielsweise im Bereich der KI-basierten Tumordiagnostik entsteht durch die hohe Sensitivität aber auch eine Gefahr: Nicht alle Ausprägungen, die theoretisch als pathologisch gelten können, weisen auf eine *behandlungsbedürftige* Erkrankung hin. Die hohe Sensitivität diagnostischer AI-CDSS verstärkt somit etwa das Risiko, dass Tumore identifiziert und behandelt werden, die auch ohne Therapie klinisch unauffällig blieben und daher keiner Behandlung bedürften. Wenn eine Erkrankung diagnostiziert wird, die nicht behandlungsbedürftig ist, kann einer solchen diagnostischen Maßnahme kein

Nutzenpotenzial zugeschrieben werden. Weil das Schadenspotenzial der Maßnahme also nicht durch ein entsprechendes Nutzenpotenzial aufgewogen werden kann, handelt es sich um eine Form von Überdiagnostik.

Ein weiterer Grund dafür, dass die Verwendung von AI-CDSS Überdiagnostik verstärken kann, lässt sich identifizieren, wenn man Faktoren bedenkt, die zur Verhinderung dieses Phänomens beitragen. Grundsätzlich kann man der Überdiagnostik vorbeugen, indem man Nutzen- und Schadenspotenziale einer diagnostischen Maßnahme sorgfältig gegeneinander abwägt. Man darf davon ausgehen, dass hohe finanzielle Kosten einer medizinischen Maßnahme dazu beitragen, dass deren Erforderlichkeit genau überprüft wird und zugehörige Nutzen- und Schadenspotenziale bedacht werden. Falls die Nutzung von AI-CDSS die Kosten der Diagnostik erheblich senken kann (vgl. Kapitel 4.11), könnte das dazu führen, dass Nutzen- und Schadenspotenziale einer Maßnahme nicht mehr hinreichend gegeneinander abgewogen werden. Somit könnte die Nutzung von KI-Systemen die Gefahr von Überdiagnostik verstärken.

Nachdem dieses Risiko von AI-CDSS diskutiert wurde, ist es wichtig sich zu vergegenwärtigen, dass deren Nutzung nicht zwangsläufig mit Überdiagnostik einhergeht. Diese kann man verhindern, indem Nutzen- und Schadenspotenziale der Systemverwendung in jedem Einzelfall sorgfältig gegeneinander abgewogen werden (vgl. Kapitel 6.3.10).

4.5 Wahrung und Förderung der Patientenautonomie

Zu den ethischen Prinzipien, die innerhalb der Medizinethik in letzter Zeit am meisten an Bedeutung gewonnen haben, gehört der Respekt der Patientenautonomie (Schmietow & Marckmann, 2019). Wie sich die Verwendung von AI-CDSS auf dieses Prinzip auswirkt, ist für die ethische Bewertung der Systeme daher von großer Relevanz. In diesem Zusammenhang werden im vorliegenden Kapitel drei Aspekte beleuchtet: die mögliche Pflicht zur informierten Einwilligung (vgl. Kapitel 4.5.1), die Auswirkungen auf die Gesundheitsmündigkeit des Patienten (vgl. Kapitel 4.5.2) und auf die Flexibilität der Therapieentscheidung für Patientenpräferenzen (vgl. Kapitel 4.5.3). Implikationen, die mit der informationellen Selbstbestimmung verbunden sind, werden insbesondere im Zusammenhang mit dem Datenschutz und der Datenverfügbarkeit behandelt (vgl. Kapitel 4.10).

4.5.1 Informierte Einwilligung der Patientinnen und Patienten

Das Prinzip des Respekts der Patientenautonomie gebietet, dass Patienten nicht bevormundet werden, sondern selbstbestimmt über ihre Behandlung entscheiden können (Beauchamp & Childress, 2019). Vor der Durchführung einer Maßnahme die informierte Einwilligung des Patienten (Informed Consent) einzuholen, ist darum grundsätzlich ethisch geboten. Sollte auch für die Verwendung von AI-CDSS die Pflicht zur informierten Einwilligung des Patienten gelten? Dieser Frage kann man sich aus juristischer Perspektive nähern.[10] Hier soll diese hingegen aus ethischer Sicht behandelt werden.

Ein erstes Argument dafür verweist auf das Schadenspotenzial dieser Anwendungen für den Patienten: Insofern er derjenige ist, der unter deren falschen Outputs leidet, sollte über die Benutzung der Systeme nicht an ihm vorbei entschieden werden (Ploug & Holm, 2020b).

Weiterhin spricht der Schutz der Privatsphäre dafür, verpflichtend die informierte Einwilligung des Patienten vor der Verwendung eines AI-CDSS einzuholen. Diese kann nämlich unterminiert werden, wenn das KI-System patientenbezogene Daten verwendet (vgl. Kapitel 4.10).

Gegenüber diesen Überlegungen gibt es vor allem zwei Argumente *gegen* die verpflichtende Einwilligung des Patienten zur Nutzung eines AI-CDSS. Ein erster relevanter Gedanke ist vor dem Hintergrund zu verstehen, dass die Verwendung von entscheidungsunterstützenden KI-Systemen die Behandlungskosten senken kann (vgl. Kapitel 4.11). Wäre diese von der Zustimmung des Patienten abhängig, so könnte die Ablehnung eines AI-CDSS eine Zunahme der Behandlungskosten bedeuten (de Miguel Beriain, 2020). Da die Ressourcen im Gesundheitswesen begrenzt sind, könnten diese finanziellen Mittel an anderer Stelle fehlen und die Behandlungsqualität weiterer Patienten negativ beeinflussen. Falls die Verwendung des AI-CDSS die Behandlungskosten senken kann und dem Patienten im Vergleich zu den Alternativen einen größeren oder mindestens einen gleichwertigen Nutzen bietet, kann die ethische

[10] Einen rechtlichen Beitrag zu dieser Frage hat etwa die ZEKO (2021) geleistet.

Verpflichtung, dieses System zu nutzen, vom Prinzip der Gerechtigkeit abgeleitet werden (de Miguel Beriain, 2020).

Gegen eine Verpflichtung zur informierten Einwilligung des Patienten zur Benutzung eines AI-CDSS wird außerdem mit dem Verweis auf die Verpflichtung des Arztes zum Wohltun argumentiert: Falls die Verwendung einer KI-Anwendung einem Patienten nachgewiesenermaßen nutzen würde, spräche das Wohltunsprinzip für dessen Nutzung (de Miguel Beriain, 2020).

Für die Pflicht, dass Patienten der Verwendung von AI-CDSS zustimmen müssen, lässt sich folglich mit Bezugnahme auf das Prinzip des Respekts der Patientenautonomie argumentieren. Gegen diese Verpflichtung kann das Gebot des Wohltuns und die Verpflichtung zur Wahrung und Förderung der Gerechtigkeit sprechen. Betrachtet man diese Argumente zusammen, so stellt sich die Frage nach deren Abwägung.

Da das Gewicht der Argumente und der ethischen Verpflichtungen zum einen von der Beschaffenheit des jeweiligen AI-CDSS und zum anderen von der spezifischen Anwendungssituation abhängt, stellt sich heraus, dass man diese Frage letztlich nur im Einzelfall beantworten kann. Bei den vielen Systemen, die patientenbezogene Daten nutzen und bei denen daher stets die Gefahr des Datenmissbrauchs besteht, ist es bereits aus Gründen der informationellen Selbstbestimmung erforderlich, dass der Patient deren Verwendung zustimmt (vgl. Kapitel 4.10).

Wichtig ist bei dieser Abwägung der Grad der Autonomie des jeweiligen Systems. Falls es wenig bis gar nicht autonom agiert, quasi ein reines Hilfsmittel des Arztes darstellt und es in keiner Weise dazu neigt, Entscheidungen zu treffen und somit in gewisser Weise ein digitales Analogon zu herkömmlichen Hilfsmitteln zur Informierung von Ärzten (etwa Fachliteratur) ist, liegt es nahe, ein solches System auch wie die analogen Hilfsmittel zu beurteilen. Niemand dürfte ernsthaft fordern, dass Ärzte erst dann – stets möglicherweise fehlerhafte – Fachliteratur verwenden dürfen, wenn die Patienten dem zugestimmt haben. Im Analogieschluss bedeutet das für AI-CDSS: Je eher ein System als ‚reines Hilfsmittel' gelten kann, je weniger autonom ein System agiert, je weniger es dazu tendiert, Entscheidungen zu treffen, desto weniger ist eine verpflichtende informierte Einwilligung des Patienten ethisch geboten.

Im Umkehrschluss lässt sich feststellen: Je autonomer ein System agiert, je weniger der Arzt es kontrollieren kann und je mehr es die Entscheidungsfindung beeinflusst, desto eher sollte der Patient über das AI-CDSS informiert werden und dessen Nutzung explizit zustimmen.

Ein weiterer Aspekt, der im Zusammenhang mit Informed Consent wichtig erscheint, ist die Tragweite und die Bedeutung der klinischen Entscheidung für den Patienten, die das System unterstützt. Ein Beispiel für eine Entscheidung, die weitreichende Konsequenzen für den Patienten hat, ist die Wahl der Therapie einer onkologischen Erkrankung. Je größer die Tragweite und die Bedeutung der vom AI-CDSS unterstützten Entscheidung für den Patienten ist, desto eher sollte dieses System nur nach dessen informierter Einwilligung verwendet werden.

Aus ethischer Perspektive fragt es sich nicht nur, *ob* eine Aufklärung des Patienten nötig ist, sondern auch, *wie* diese durchgeführt werden soll. Problematisch ist hierbei vor allem, dass die Funktionsweise von AI-CDSS komplex, teilweise gar opak ist und deren Nutzen- und Schadenspotenzial nicht immer bekannt ist. Das Ziel muss es sein, den Patienten in laiengerechter Form über die wesentlichen Charakteristika des betreffenden Systems zu informieren (vgl. Kapitel 6.3.8). Dass die Aufklärung gelingt, ist nicht nur von den Bemühungen des Arztes abhängig. Damit der Patient die Aufklärung verstehen kann, ist ein gewisses Grundverständnis von KI erforderlich. Insofern jeder Bürger ein potenzieller Patient ist, sollte die gesamte Bevölkerung ein solches Verständnis von KI-Systemen entwickeln (vgl. Kapitel 6.2.10).

4.5.2 Förderung der Gesundheitsmündigkeit der Patientinnen und Patienten

Der Respekt der Patientenautonomie fordert nicht nur, dass der Patient über eine Maßnahme informiert wird und deren Durchführung zustimmt. Aus dem Autonomieprinzip geht des Weiteren hervor, dass der Behandlungsprozess nicht vom Arzt vorgegeben werden darf (Paternalismus) (Emanuel & Emanuel, 1992). Der Patient soll stattdessen mündig über seine Behandlung (mit-)entscheiden (Thomas et al., 2021). Förderlich dafür ist es, wenn dieser über seinen gesundheitlichen Zustand informiert ist. Daher gebietet es das Prinzip des Respekts der Patientenautonomie, die Informiertheit des Patienten zu fördern.

Fragt man sich, wie die Nutzung eines AI-CDSS die Informiertheit der Patienten beeinflusst, zeigt sich die Ambivalenz der Systeme. Zunächst kann die Verwendung eines KI-Systems die Informiertheit des Behandelten steigern, wenn dieses nicht nur dem Arzt, sondern auch dem Patienten Informationen und Ergebnisse anzeigt. Das System *DreaMed Advisor Pro* bietet dem Patienten beispielsweise die Möglichkeit, über das Smartphone den Behandlungsplan einzusehen, der mithilfe von KI entwickelt wurde (Hale, 2018). Auf diese Weise wird die Gesundheitsmündigkeit des Patienten und dessen Autonomie gestärkt. Im Allgemeinen ist darum zu fordern, dass Systemergebnisse auch an die Patienten ausgegeben werden sollen (vgl. Kapitel 6.1.13).

Von diesen Überlegungen ausgehend ist es zu kritisieren, dass viele AI-CDSS den Patienten nicht informieren. Darunter leidet seine Gesundheitsmündigkeit: Es wird ihm zum Beispiel erschwert, sich selbständig über Alternativen zu den Vorschlägen des Arztes zu erkundigen. So kann der Patient in eine Abhängigkeit vom Behandelnden gelangen, die der Patientenautonomie zuwiderläuft. In diesen Fällen kann der Arzt aber die Autonomie des Patienten fördern, indem er diesem die Systemergebnisse erklärt (vgl. Kapitel 6.3.12).

Im Zusammenhang der Informiertheit des Patienten ist auch die Erklärbarkeit des Systems relevant. Wenn es für bestimmte Ergebnisse keine Erklärungen angibt, ist es möglich, dass der Arzt die Informiertheit des Patienten mithilfe des Systems nur bedingt steigern kann (Bjerring & Busch, 2021). Ist ein AI-CDSS hingegen erklärbar, so kann dessen Nutzung aus folgendem Grund die Informiertheit des Patienten in besonderer Weise fördern: Ein großer Teil des ärztlichen Wissens ist impliziter Natur. Für Ärzte ist es daher oft schwierig zu explizieren, warum sie eine bestimmte Diagnose stellen oder wieso sie eine bestimmte therapeutische Maßnahme empfehlen. Diese oftmals impliziten Prozesse der Diagnostik, der Prognostik und der Therapieempfehlung zu formalisieren, somit für den Patienten verständlich zu gestalten und auf diese Weise die Gesundheitsmündigkeit des Patienten zu steigern, stellt ein Potenzial von erklärbaren AI-CDSS dar (van Baalen et al., 2021).

4.5.3 AI-CDSS und die Beachtung von Patientenpräferenzen

Nachdem erörtert wurde, inwiefern AI-CDSS im Allgemeinen die Gesundheitsmündigkeit des Patienten beeinflussen, soll ein Strukturelement der klinischen Entscheidungsfindung in den Blick genommen werden, hinsichtlich dessen

das Prinzip des Respekts der Patientenautonomie besonders relevant ist. Wie bereits in Kapitel 4.1.1 erläutert, sind bei der Therapieentscheidung Patientenpräferenzen zu beachten. Wenn AI-CDSS therapeutische Maßnahmen empfehlen, berühren sie den Therapieentscheidungsprozess. Es fragt sich daher, inwiefern es die Flexibilität der Behandlungsentscheidung für Patientenpräferenzen beeinflusst, wenn die Systeme therapeutische Empfehlungen abgeben. Weil AI-CDSS dazu (bisher) nicht in der Lage sind, ist es die Aufgabe der Nutzer, die Therapieentscheidung an den Wünschen und Werten des Patienten zu orientieren (McDougall, 2019) (vgl. Kapitel 4.1.1).

Bevor ein Arzt seiner Verpflichtung zum Respekt der Patientenautonomie nachkommen kann und er die Therapieempfehlung eines AI-CDSS an die Patientenpräferenzen anpassen kann, muss er sich zuerst über die Erforderlichkeit dessen bewusst sein. Dieses Bewusstsein dürfte besonders bedroht sein, wenn das betreffende System eine aus ärztlicher Sicht hohe epistemische Autorität aufweist. Benutzt der Arzt ein besonders leistungsfähiges AI-CDSS, könnte er dazu neigen, dessen Empfehlungen schlichtweg als ‚richtig' aufzufassen, sodass er übersehen mag, dass die Outputs des Systems an den Patientenpräferenzen orientiert sind. Diese potenzielle ‚Individualisierungsvergessenheit' birgt das Risiko in sich, Paternalismus zu entwickeln oder zu verstärken (McDougall, 2019).

Es ist wichtig darauf hinzuweisen, dass diese Gefahr nicht erst durch die Einführung von AI-CDSS in die Medizin einzieht. Eine illegitime Beschränkung der Patientenautonomie kann sich auch ergeben, wenn Ärzte bei klinischen Entscheidungen klinische Leitlinien zurate ziehen. Auch die daraus hervorgehenden Empfehlungen sind fast nie an die Präferenzen von Patienten angepasst (Fox et al., 2009).

Dennoch ist die Gefahr der ‚Individualisierungsvergessenheit', die von AI-CDSS auf der einen Seite und von klinischen Leitlinien auf der anderen Seite ausgeht, kaum als gleichwertig einzuschätzen. Dass Letztere *nur* Empfehlungen geben, die in ihrer Einfachheit individuelle Patientenpräferenzen grundsätzlich nicht einbeziehen, dürfte Ärzten bewusst sein. Demgegenüber ist es als unwahrscheinlich einzuschätzen, dass sie genau wissen, wie technisch hochkomplexe AI-CDSS entwickelt werden und wo die Grenzen der Systeme hinsichtlich der Beachtung von Patientenpräferenzen liegen.

Abhilfe gegenüber der ‚Individualisierungsvergessenheit' könnte man schaffen, indem AI-CDSS explizit darauf hinweisen, dass die Outputs nicht an Patientenpräferenzen angepasst sind und der Arzt derjenige ist, der diese Individualisierung durchführen muss

(McDougall, 2019) (vgl. Kapitel 6.1.8 und 6.3.12). Auch Edukation könnte für die Gefahr der ‚Individualisierungsvergessenheit' sensibilisieren (vgl. Kapitel 6.3.7). Um zu verhindern, dass der Arzt die vom AI-CDSS empfohlene therapeutische Maßnahme als schlechthin richtig missversteht, sollte das System außerdem mehrere Ergebnisse ausgeben (Rajput et al., 2020) (vgl. Kapitel 6.1.12).

Ist sich der Arzt darüber im Klaren, dass er die Systemempfehlungen an die Patientenpräferenzen anzupassen hat, ist eine erste Anforderung der Flexibilität für dieselben erfüllt. Was muss noch gegeben sein, damit der Arzt die KI-basierte Therapieempfehlung an die Patientenpräferenzen anpassen kann? An dieser Stelle ist zu bedenken, dass auch die Empfehlungen des AI-CDSS nicht werturteilsfrei sind. Um die werturteilsbehafteten Therapieempfehlungen des Systems an die Patientenpräferenzen anpassen zu können, muss der Arzt die Werturteile identifizieren können, die den Systemempfehlungen zugrunde liegen.

Hier sind mindestens zwei Möglichkeiten denkbar: In manchen Fällen ist davon auszugehen, dass Ärzte auf die Werturteile schließen können, die eine Therapieempfehlung impliziert. Beispielhaft kann man sich die fiktive Empfehlung vor Augen führen, einem bestimmten Krebspatienten eine Strahlentherapie zukommen zu lassen. Diesem Output liegt das Werturteil zugrunde, dass die Chance auf Heilung die mit der Radiotherapie verbundenen Schadenspotenziale – etwa die Kanzerogenität der Bestrahlung – überwiegen. Insofern man voraussetzen kann, dass Ärzten diese Nebenwirkung der Strahlentherapie bekannt ist, kann man davon ausgehen, dass für sie das Werturteil ersichtlich ist, auf dem diese Systemempfehlung basiert.

Manchmal ist es jedoch kaum möglich zu bestimmen, welche Werturteile eine Therapieempfehlung impliziert. In diesen Fällen können es vor allem zwei Maßnahmen erleichtern, zugrundeliegende Werturteile zu identifizieren.

Erstens könnten die Entwickler die Werturteile, auf der eine Empfehlung basiert, explizieren und dem Arzt verständlich darstellen (Andrews et al., 2013). Ein Beispiel dafür gibt ein Auszug aus einer klinischen Leitlinie: „Our recommendations reflect a belief that most women will place a low value on avoiding the pain, cost, and inconvenience of heparin therapy in order to avoid the small risk of even a minor abnormality in their child." (S. M. Bates et al., 2008, S. 848)

Es ist jedoch fraglich, ob die Entwickler von AI-CDSS immer angeben können, an welchen Zielen und Werten die einzelnen Therapieempfehlungen des Systems ausgerichtet sind. Wenn das System auf ML basiert, gehen darin schließlich nicht nur die Werturteile der Entwickler, sondern auch solche ein, die den Trainingsdatensätzen zugrunde liegen. Manche dieser Urteile mögen anhand von Charakteristika des Datensatzes erklärbar sein: Dessen geographische Herkunft mag etwa in begrenztem Ausmaß Hinweise auf lokale evaluative Präferenzen geben, die sich in der Architektur eines AI-CDSS und somit auch in dessen Therapieempfehlungen wiederfinden können.

Letztendlich ist es aber kaum möglich, *alle* Faktoren zu explizieren, die eine KI-basierte Therapieempfehlung gelenkt und beeinflusst haben. Das muss indessen nicht zur Kapitulation vor der Intransparenz der Werturteile führen, die einer Therapieempfehlung zugrunde liegen. Die Angabe einiger relevanter Informationen dürfte dafür ausreichen, dass der Arzt einschätzen kann, inwiefern sich die Ziele und Werte, welche die Therapieempfehlung impliziert, von den evaluativen Präferenzen des Patienten unterscheiden. Die Werturteile zu explizieren, die einer KI-basierten Therapieempfehlung zugrunde liegen, ist darum aus ethischer Sicht erstrebenswert (vgl. Kapitel 6.1.8). Somit wird der Arzt in die Lage versetzt, die Therapieentscheidung mit dem Patienten an dessen Werte und Ziele anzupassen. Auf diese Weise kann er dem Prinzip des Respekts der Patientenautonomie gerecht werden.

Wird hingegen nicht explizit angegeben, welche Werte oder Ziele die Therapieempfehlung verfolgt, kann das KI-System den Arzt dabei unterstützen, die der Empfehlung zugrundeliegenden Werturteile zu identifizieren. Das kann geschehen, indem das System auf herkömmliche Methoden der Erklärbarkeit zurückgreift: Die Kenntnis von entscheidungsrelevanten Faktoren kann dem Arzt Hinweise für mögliche Werturteile geben, auf denen die Therapieempfehlung aufbaut (Nyrup et al., 2019). Erklärt das AI-CDSS beispielsweise, dass eine Therapie empfohlen wird, weil sie im Gegensatz zu Alternativen nicht kanzerogen ist, wird der Arzt auf ein ethisch relevantes Schadenspotenzial der anderen möglichen Maßnahmen aufmerksam gemacht. Freilich muss er hier noch auf das hinter der Empfehlung stehende Werturteil schließen. Insofern Erklärungen aber die Transparenz der Therapieempfehlung – auch hinsichtlich der Werturteile – erhöhen, kann Erklärbarkeit auch relevant sein, um die Therapieempfehlungen an die Patientenpräferenzen anpassen zu können und dem Gebot des Respekts der Patientenautonomie zu entsprechen (vgl. Kapitel 6.1.16).

4.6 Wahrung der ärztlichen Entscheidungsautonomie

Nachdem im vorangegangenen Kapitel die Auswirkungen der Nutzung von AI-CDSS auf die Patientenautonomie untersucht wurden, soll das vorliegende Kapitel eine weitere Form der Freiheit in den Blick nehmen. Autonomie kommt nämlich nicht nur dem Patienten, sondern auch – in anderer Form – dem Arzt zu, wobei man diese auch als ärztliche Entscheidungsautonomie bezeichnet. Sie zeigt sich darin, dass der Arzt prinzipiell frei entscheiden kann, welche der zur Auswahl stehenden Maßnahmen er zur Behandlung eines Patienten zu welchem Zeitpunkt dem Patienten vorschlägt beziehungsweise ergreift (Kienle, 2008). Es ist aber wichtig darauf hinzuweisen, dass diese Freiheit kein Selbstzweck ist: „Die ärztliche Freiheit ist immer eine bedingte Freiheit, eine, um zu helfen, und nicht eine Freiheit um der Freiheit willen“ (Wiesing, 2001, S. 164).

In Kapitel 3.4.3 wurde deutlich, dass AI-CDSS zwar keine Entscheidungen treffen, also auch im strengen Sinne nicht als autonom zu begreifen sind. Gleichzeitig hat sich aus diesen Untersuchungen ergeben, dass die Systeme zwischen den beiden Polen – Entscheidungsunterstützung und -übernahme – einzuordnen sind (ZEKO, 2021).

Die teilweise Autonomie der Systeme wirkt sich wiederum auf die Freiheit der Ärzte aus: In dem Maße, in dem KI-Systeme die klinische Entscheidungsfindung übernehmen oder beeinflussen, AI-CDSS also an Autonomie gewinnen, nimmt der ärztliche Einfluss auf die Entscheidungsfindung beziehungsweise die ärztliche Entscheidungsfreiheit ab. Darauf hinzuweisen, dass die Systeme den Behandelnden ‚nur‘ unterstützen, scheint deshalb nicht zu genügen, um die Bedenken zu zerstreuen, dass die Nutzung von AI-CDSS die ärztliche Autonomie beschränken könnte. Insofern die auch in Zukunft zu erwartende wachsende Leistungsfähigkeit der KI-Systeme eine zunehmend autonome Verwendung der Systeme ermöglichen mag, dürfte sich deren Einfluss auf die klinische Entscheidungsfindung weiter verstärken und die Frage nach der Entscheidungsübernahme durch AI-CDSS an Relevanz gewinnen. Wie also ist es ethisch zu bewerten, wenn die Systeme die ärztliche Entscheidungsautonomie einschränken?

Eine Verteidigung der ärztlichen Autonomie könnte auf das Prinzip des Nichtschadens verweisen: AI-CDSS besitzen ein Fehlerpotenzial, das sich nicht gänzlich eliminieren lässt (vgl. Kapitel 4.4.1). Die ärztliche Kontrolle der Outputs kann das Schadenspotenzial

zu einem bestimmten Grad minimieren (vgl. Kapitel 6.3.12). Damit Ärzte die Systemergebnisse überprüfen können, bedürfen sie aber einer gewissen Autonomie.

Ein weiteres Argument zum Schutz der ärztlichen Autonomie lässt sich aus dem Umstand gewinnen, dass AI-CDSS typischerweise die ‚*soft facts*' bei der klinischen Begegnung nicht hinreichend beachten können (Funer, 2021) (vgl. Kapitel 4.1.1). Wenn die ärztliche Entscheidungsautonomie beschränkt wird und kein jedenfalls potenziell mitfühlender, vernunftbegabter Mensch mehr die Oberhand über die klinische Entscheidungsfindung hat, könnte die reduktionistische Tendenz der Systeme die Qualität der Medizin negativ beeinflussen (vgl. Kapitel 4.9.2).

Problematisch ist außerdem, dass AI-CDSS grundsätzlich nicht flexibel für Patientenpräferenzen sind (vgl. Kapitel 4.1.1). Durch eine Beschränkung der ärztlichen Autonomie läuft die Medizin demnach Gefahr, durch die Etablierung eines neuen Paternalismus hinter bereits erreichte ethische Standards zurückzufallen (McDougall, 2019) (vgl. Kapitel 4.5.3).

Zur Verteidigung der ärztlichen Autonomie könnte man außerdem darauf verweisen, dass die Medizin am Wohl des Patienten orientiert ist und diese normative Orientierung durch die Autonomie des Arztes gewährleistet wird. Denn der Arzt hat sich mit dem Berufseid daran gebunden, sein Tun am Patientenwohl auszurichten (Kleeberg, 2015). Insofern AI-CDSS einen starken Einfluss auf die klinische Entscheidungsfindung ausüben, prägen diese Systeme die klinischen Entscheidungen auch in normativer Hinsicht. Führt man sich nun vor Augen, dass die Entwickler von AI-CDSS nicht an das ärztliche Ethos gebunden sind, wird die Befürchtung nachvollziehbar, dass die normative Orientierung der Medizin am Patientenwohl durch die Nutzung der Systeme untergraben werden könnte: Die Entwickler könnten die KI-Systeme an Interessen ausrichten, die nicht oder nur unzureichend mit dem Wohl des Patienten in Einklang stehen. AI-CDSS könnten etwa so programmiert werden, dass deren Verwendung vor allem den Absatz von Medikamenten einer bestimmten Firma erhöht oder die Kosten eines Krankenhauses senkt (Hodgkin, 2016).

Im vorliegenden Kapitel wurden nun bereits mehrere Einwände gegen die Beschränkung der ärztlichen Autonomie vorgetragen, die durch die Nutzung von AI-CDSS entstehen kann. Angesichts der Risiken, die mit einer Beschneidung der ärztlichen

Entscheidungsfreiheit einhergehen, wird die eingangs gestellte Frage in ein brisantes Licht gerückt: Wie kann eine Einschränkung der ärztlichen Autonomie durch die Nutzung von KI-Systemen ethisch vertretbar sein? In diesem Zusammenhang ist an die anfänglichen Überlegungen zum Wesen der ärztlichen Entscheidungsfreiheit zu erinnern: Diese ist daran gebunden, dem Patienten zu helfen (Wiesing, 2001). Wird sie durch den Einsatz von KI-Anwendungen beschränkt, ist das daher nicht per se negativ zu beurteilen. Um deren potenzielle Einschränkung durch die Nutzung der Systeme zu bewerten, sollte deren Auswirkung auf das Patientenwohl als Maßstab verwendet werden: Fördert die Beschränkung der ärztlichen Autonomie das Wohl des Patienten, ist diese ethisch geboten. Mindert diese hingegen das Patientenwohl, ist die Beschneidung der ärztlichen Autonomie ethisch nicht vertretbar.

Anhand dieses Maßstabs kann die potenzielle Einschränkung der ärztlichen Autonomie durch die Nutzung von AI-CDSS ethisch bewertet werden. Letztlich gilt es demnach für jedes System einzeln abzuwägen, ob die damit verbundenen Nutzenpotenziale für die Patienten die zugehörigen Schadenspotenziale aufwiegen.

Im Lichte dieser Überlegungen wird wieder die Bedeutung einer rigorosen Evaluation sichtbar: Der zusätzliche Nutzen für die Patienten, der eine Beschränkung der ärztlichen Autonomie begründen kann, ist anhand von Evaluationsstudien zu belegen (vgl. Kapitel 6.1.18.3). Durch eine rigorose Testung und Evaluation kann man auch den Bedenken begegnen, dass diejenigen Systeme die normative Orientierung der Medizin am Wohl des Patienten bedrohen, die innerhalb der Privatwirtschaft entwickelt wurden: Nutzt die Verwendung eines AI-CDSS dem Patienten nachweislich, ist es aus ethischer Sicht prima facie nicht unvertretbar, wenn bei der Entwicklung und beim Vertrieb des Systems gewinnorientiert gearbeitet wurde.

Besonders bei Hochrisiko-Anwendungen wäre es wünschenswert, auf eine Sicherheit bauen zu können, die über positive Evaluationsergebnisse hinausgeht. Könnte man auf andere Weise sicherstellen, dass sich AI-CDSS und deren Entwicklung am Ziel der Verbesserung des Patientenwohls orientieren? Hier kann eine Betrachtung des ärztlichen Tuns aufschlussreich sein: Wer als Arzt praktiziert, hat sich freiwillig an das ärztliche Ethos gebunden, welches das Patientenwohl als oberste Richtschnur vorgibt. Inspiriert davon könnte auch für die Entwickler von medizinischen KI-Systemen ein eigenes Berufsethos etabliert werden (Bundesverband Informationswirtschaft,

Telekommunikation und neue Medien [Bitkom] & Deutsches Forschungszentrum für Künstliche Intelligenz [DFKI], 2017) (vgl. Kapitel 6.2.9).

4.7 Auswirkung auf die ärztliche Entscheidungskompetenz

Die Nutzung von AI-CDSS kann nicht nur die Autonomie des Arztes, sondern auch seine Entscheidungskompetenz beeinflussen. Setzt man sich mit möglichen negativen Auswirkungen auf diese auseinander, ist zunächst Folgendes im Allgemeinen festzustellen: Wenn Prozesse automatisiert werden, können davon betroffene Personen bestimmte Fähigkeiten verlieren. Dieses Phänomen nennt man auch *deskilling* (Hoff, 2011; Manzey, 2008).

Deskilling kann auch bei der Nutzung von KI-Systemen auftreten. Beispielsweise kann man hier an das System *Moleanalyzer Pro* denken, das maligne Melanome detektiert (vgl. Kapitel 3.2.1). Angesichts der hohen Performanzwerte dieses Systems mag der Arzt dazu tendieren, verdächtige Hautläsionen nicht mehr eigenständig zu diagnostizieren, sondern für deren Detektion ausschließlich das KI-System zu verwenden (Haenssle et al., 2020). Somit kann *deskilling* auftreten – der Arzt verliert die Fähigkeit, maligne Melanome zu detektieren.

Dieses Phänomen kann im Zusammenhang mit AI-CDSS Befürchtungen aufkommen lassen: Ärzte sollen die Ergebnisse der Systeme nicht blind übernehmen, sondern grundsätzlich – wenigstens auf Plausibilität hin – überprüfen (vgl. Kapitel 6.3.12) (Arnold, 2021). Dazu ist es notwendig, dass die Nutzer unabhängig vom System fachlich qualitativ hochwertige Überlegungen anstellen können. Wenn Ärzte allerdings verlernt haben, selbst Prognosen und Diagnosen zu stellen oder Therapieempfehlungen zu geben, können sie die Ergebnisse von AI-CDSS nicht hinreichend überprüfen. Dadurch steigt die Wahrscheinlichkeit, dass sie falsche Systemergebnisse nicht detektieren und daher zu falschen Entscheidungen kommen, die dem Patienten schaden können (vgl. Kapitel 4.4.2.2). Aus diesen Überlegungen ergibt sich, wie wichtig es ist, *deskilling* etwa durch Schulungen der Ärzte möglichst zu vermeiden – besonders wenn schwerer Schaden droht (vgl. Kapitel 6.3.7).

Ein weiteres Phänomen, das die Entscheidungskompetenz der Nutzer von AI-CDSS untergraben kann, ist *alert fatigue*. Mit diesem Begriff beschreibt man Erschöpfung,

bedingt durch die zunehmende Zahl an Hinweisen und Nachrichten, die Nutzer von computerbasierten Systemen erhalten (Wasylewicz & Scheepers-Hoeks, 2019). Es ist wohl davon auszugehen, dass *alert fatigue* auch generell die Entscheidungskompetenz der Ärzte negativ beeinflussen kann.

Anhand dieser Beispiele – *deskilling* und *alert fatigue* – zeigt sich, dass die Verwendung von AI-CDSS für Ärzte negative Konsequenzen nach sich ziehen kann. Wie ist dieser Umstand ethisch zu bewerten? Bei Ärzten mögen diese Phänomene zu Unmut führen. Aus ethischer Sicht ist hier darauf hinzuweisen, dass die Unveränderlichkeit der ärztlichen Arbeitsbedingungen keinen Selbstzweck darstellt. Um die potenzielle Einschränkung der ärztlichen Entscheidungskompetenz zu bewerten, wird also ein Bewertungsmaßstab benötigt. Hierzu kann – wie im vorigen Kapitel – auf das Ziel der Medizin rekurriert werden: Diese ist auf die Wiederherstellung und die Förderung der Gesundheit der Patienten ausgerichtet.

Patientenorientiert ist auch die potenzielle Transformation des ärztlichen Berufs durch die Verwendung von AI-CDSS zu bewerten. Es mag für Ärzte ein Ärgernis darstellen, wenn sie bestimmte Fähigkeiten durch die Nutzung der Systeme verlieren oder sie in ihrem Tun durch eine steigende Zahl an Systemwarnungen einer zunehmenden Belastung ausgesetzt sind. Dass die Implementierung von AI-CDSS den ärztlichen Beruf verändern kann, bedeutet aber nicht, dass die Verwendung dieser Systeme ethisch unvertretbar wäre. Wenn sich jedoch herausstellen sollte, dass die Patientenversorgung durch diese Phänomene und Entwicklungen verschlechtert wird, sollte das alarmieren.

In diesem Zusammenhang kann man sich beispielhaft eine Studie zu *alert fatigue* vor Augen führen: Hier zeigte sich, dass Ärzte als Folge von *alert fatigue* 49 % bis 96 % der computerbasierten Hinweise übergehen (van der Sijs et al., 2006). Dieser Effekt ist aus ethischer Sicht heikel – AI-CDSS sollen den Arzt unterstützen und nicht seine Konzentrations- oder Arbeitsfähigkeit unterminieren. Um dessen Entscheidungskompetenz zu gewährleisten, sind darum entsprechende Maßnahmen zu ergreifen (vgl. Kapitel 6.1.11 und 6.3.9).

Angesichts von *deskilling* und *alert fatigue* ist noch Weiteres in den Blick zu nehmen: Belastet die Verwendung von AI-CDSS Ärzte unverhältnismäßig stark, ergeben sich Fragen der Gerechtigkeit (vgl. Kapitel 4.12). Man kann befürchten, dass diese versuchen, sich den Belastungen zu entziehen, indem sie die Nutzung der Systeme ablehnen. Vor

diesem Hintergrund kann es sinnvoll sein, Anreize für die Verwendung von AI-CDSS zu schaffen (vgl. Kapitel 6.2.7).

4.8 Zuschreibbarkeit von Verantwortung beim Einsatz von AI-CDSS

Verantwortung kann als einer der Schlüsselbegriffe der Ethik gelten. Insbesondere die Verwendung von Technologie wirft Fragen in Bezug auf die Verantwortung auf. Hinsichtlich der Nutzung von AI-CDSS lässt sich deren Bedeutung konsequentialistisch herleiten. Klinische ärztliche Entscheidungen haben Maßnahmen zum Gegenstand, die sich typischerweise direkt auf die Gesundheit des Patienten auswirken und somit zu erheblichem Schaden führen können (vgl. Kapitel 4.4). Dass jemand dafür Verantwortung übernimmt, ist für Patienten von großer Bedeutung. Wenn Schaden eintritt, gilt es diesen so weit wie möglich zu begleichen. Doch auch wenn es nicht dazu kommt, kann eine unklare Verantwortungsverteilung das Vertrauen der Patienten in die Behandlung und den Arzt unterminieren und somit der Qualität der Arzt-Patient-Beziehung schaden. Im Lichte dieser Überlegungen wird deutlich, dass der klaren Verantwortungszuschreibung bei der Verwendung von AI-CDSS große Relevanz zukommt. Dementsprechend konzentrieren sich die folgenden Untersuchungen besonders auf die Zuschreibbarkeit von Verantwortung.

Bezüglich der Verantwortung für KI-basierte Entscheidungen stellt sich zunächst die Frage, ob AI-CDSS Verantwortung übernehmen können. Dieser wurde in der vorliegenden Arbeit bereits nachgegangen (vgl. Kapitel 3.4.3). Dabei wurde klargestellt, dass KI-Systeme keine moralischen Agenten sind und keine Verantwortung übernehmen können. Das bedeutet aber nicht, dass Nutzer den Systemen in der Praxis nicht möglicherweise dennoch Verantwortungsfähigkeit zuschreiben. Schließlich zeigen einige Studien, dass Maschinen teilweise bestimmte kognitive und emotionale Fähigkeiten zugeordnet werden, die nur Menschen zukommen (Draude, 2011; Misselhorn, 2009; Slater et al., 2006). Weber und Zoglauer (2019) schlussfolgern daraus plausibel, dass Menschen – womöglich unterbewusst – bestimmten Systemen mit der Verantwortungsfähigkeit auch etwas weiteres genuin Menschliches zuschreiben könnten. Das ist vor allem im Hinblick auf besonders leistungsfähige KI-Systeme vorstellbar.

Als Zwischenfazit kann man demnach festhalten: Obwohl bei der Benutzung von AI-CDSS der klaren Verantwortungszuschreibung große Relevanz zukommt, besteht hier das Risiko der Verantwortungsdiffusion. Es gilt darum im Folgenden näher zu erörtern,

wem bei der Verwendung der Systeme Verantwortung zugeschrieben werden kann, welche Probleme dabei möglicherweise auftreten und wie mit diesen ethisch vertretbar umzugehen ist.

Hinsichtlich der Verantwortungszuschreibung für die Folgen der Techniknutzung unterscheidet man herkömmlicherweise zwei verschiedene Konzepte (Marckmann, 2003; Ropohl, 1993): Nach dem gebrauchsorientierten Konzept liegt die Verantwortung vor allem beim Nutzer der Technologie. Das herstellungsorientierte Konzept fokussiert hingegen die Verantwortung des Entwicklers. Dieses ist vor allem für die Fälle geeignet, in denen Technik automatisiert verwendet wird. Darunter fällt nicht der Einsatz von AI-CDSS, weil diese den Arzt nur unterstützen (vgl. Kapitel 4.1.1). Daher soll im Folgenden vor allem auf die Frage eingegangen werden, inwiefern man einem Anwender bei der Nutzung von entscheidungsunterstützenden KI-Systemen Verantwortung zuschreiben kann.

Es bietet sich an, dabei den Überlegungen von Liedtke und Langanke (2021) zu folgen, die sich mit der Zuschreibbarkeit von Verantwortung für KI-basierte medizinische Decision Support Systeme befassen. Sie formulieren dabei vier Vorbedingungen für die Zuschreibbarkeit von Verantwortung:[11]

> „1.) *Freiwilligkeit* – Handlungen müssen ohne „äußeren“ Zwang ausgeführt werden können.
>
> 2.) *Informiertheit* – Handelnde Subjekte müssen über die handlungsrelevanten Informationen verfügen. Handlungsrelevant sind dabei etwa Informationen, die sich auf Methoden, Abläufe oder Befunde beziehen können.
>
> 3.) *Selbstbestimmung und Zurechnungsfähigkeit* – Handelnde Subjekte müssen selbstbestimmt handeln können und zurechnungsfähig sein. […]
>
> 4.) *Existenz von Handlungsalternativen* – In einer Situation müssen mehrere Handlungsoptionen bestehen oder bestanden haben, andernfalls gerät die

[11] Der Verantwortungsbegriff, auf den sich die Autoren dabei stützen, wurde von Langanke et al. (2017) entwickelt.

> Verantwortungszuschreibung an die logische Grenze echter ‚Alternativlosigkeit'." (Liedtke & Langanke, 2021, S. 288–289)

Im Anschluss an Liedtke und Langanke (2021) soll nun analysiert werden, inwiefern diese Vorbedingungen der Zuschreibbarkeit von Verantwortung erfüllt sind, wenn ein Arzt ein AI-CDSS nutzt.

Vorbedingung (4) ist bei der Nutzung dieser Anwendungen stets erfüllt: Eine Handlungsalternative besteht immer darin, der Systemempfehlung nicht zu folgen. Da Ärzte grundsätzlich ohne äußeren Zwang und selbstbestimmt handeln, ist auch die Erfüllung der Vorbedingungen (1) und (3) wenig problematisch, auch wenn Liedtke und Langanke (2021) zurecht darauf hinweisen, dass der Automation Bias beide Vorbedingungen zu untergraben droht. Nachdem diese Problematik bereits behandelt wurde (vgl. Kapitel 4.4.2.2), soll darauf nicht mehr näher eingegangen werden. Dass der Automation Bias auch mit der Verantwortungszuschreibung verbunden ist, zeigt jedoch abermals die Relevanz davon, diesen zu verhindern (vgl. Kapitel 6.3.9).

Ein Kriterium der Verantwortungsübernahme, dessen Erfüllung bei der Nutzung von AI-CDSS jedenfalls fragwürdig ist, besteht in der Informiertheit (2). Wie bereits deutlich wurde, sind einige der Anwendungen als Black-Box-Systeme zu charakterisieren (vgl. Kapitel 3.4.1): Der Prozess, der zu der Ausgabe des Outputs führt, kann nicht sinnvoll nachvollzogen werden. Konfrontiert mit dem Ergebnis einer solchen opaken Anwendung kann der Arzt nicht überprüfen, ob die Faktoren, die zu diesem Output geführt haben, im vorliegenden Fall legitimerweise als entscheidungsrelevante Faktoren fungieren können. Es besteht hier nur die Wahl zwischen ‚blinder' Annahme oder Ablehnung der Systemempfehlung (Liedtke & Langanke, 2021).

Untergräbt die Opazität von AI-CDSS die Informiertheit der Nutzer und kann man ihnen somit keine Verantwortung zuschreiben, liegt es nahe, zur Lösung des Verantwortungsproblems bei der Auflösung der Opazität beziehungsweise bei der Gewährleistung von Erklärbarkeit anzusetzen: Wenn die Nutzer nachvollziehen können, wie und warum ein System zu einem bestimmten Ergebnis kommt, wird die Zuschreibung von Verantwortung nicht mehr dadurch unterminiert, dass die Vorbedingung (2) – Informiertheit – nicht gegeben ist. Wie in Kapitel 3.4.1 dargestellt, gibt es bereits einige Methoden zur Steigerung der Erklärbarkeit von ML-Systemen. Indem man diese verwendet, kann der Arzt bis zu einem gewissen Grad dazu befähigt werden, die Systemergebnisse nachzuvollziehen. Ob allerdings Erklärungen von Outputs

gewährleisten können, dass die Informiertheit als Vorbedingung (2) der Verantwortungsübernahme als erfüllt betrachtet werden kann, ist fragwürdig: Ein Black-Box-System ist schließlich dadurch gekennzeichnet, dass es nicht einmal theoretisch möglich ist, den Prozess sinnvoll nachzuvollziehen, der zur Ausgabe des Ergebnisses führt (Stein et al., 2019; ZEKO, 2021). Opazität wird sich folglich in manchen Fällen auch durch Erklärungen nicht gänzlich auflösen lassen.

Aufgrund der großen Bedeutung von klarer Verantwortungszuschreibung und der prinzipiellen Limitationen der Erklärbarkeit von Black-Box-Systemen könnte man dafür argumentieren, auf die Verwendung solcher Anwendungen zu verzichten (Shortliffe & Sepúlveda, 2018). Der Vorteil dieser Vorgehensweise liegt auf der Hand: Auf ebenso effektive wie einfach umzusetzende Weise könnte man verhindern, dass Informiertheit durch Opazität unterminiert wird. Somit könnte man den Ärzten Verantwortung für die Nutzung von AI-CDSS zuschreiben.

Auch wenn auf diese Weise die Informiertheit als Vorbedingung (2) der Verantwortungszuschreibung erfüllt wäre, ist es aus ethischer Sicht fragwürdig, ob es legitim sein kann, die Verantwortung für die Systemnutzung auf die Ärzte abzuschieben: Verbessert die Verwendung von AI-CDSS die Patientenversorgung, profitiert die Allgemeinheit. Mit der Verantwortung eine der zentralen Belastungen von Anwendungen alleine auf die Ärzte abzuschieben, erscheint in diesem Licht ethisch kaum vertretbar (ZEKO, 2021).

Kritisch kann man außerdem einwenden, dass Opazität nicht erst durch Black-Box-Systeme Einzug in die Medizin gehalten hat. Eine gewisse Form von Intransparenz ist vielen der medizinischen Instrumente und Pharmazeutika geradezu inhärent: Unzweifelhaft wissen Ärzte jedenfalls in manchen Fällen nicht, wie die von ihnen benutzten Instrumente und Pharmazeutika präzise funktionieren und wirken (Bjerring & Busch, 2021). Das liegt nicht nur an den Ärzten: Über die Wirkungsweise von einigen Pharmazeutika gibt es nur wenig wissenschaftliche Erkenntnis (Ghassemi et al., 2021). Ein Beispiel hierfür ist etwa das Schmerzmittel Acetaminophen (Kirkpatrick, 2005). Dass der Wirkmechanismus eines Arzneimittels nicht oder nur unzureichend bekannt ist, bedeutet jedoch nicht, dass dieses Pharmakon nicht verwendet werden dürfte. Ob ein Arzneimittel benutzt werden darf, hängt vielmehr davon ab, ob es sich in Evaluationsstudien als sicher und wirksam herausgestellt hat (Ghassemi et al., 2021).

Anhand dieser Ausführungen dürfte deutlich geworden sein, dass das Phänomen der Opazität keineswegs erst durch KI-Systeme in die Medizin eingetreten ist. Von daher ist es nicht angemessen, in Bezug auf Arzneimittel mit unbekanntem Wirkmechanismus zum einen und in Hinsicht auf opake AI-CDSS zum anderen mit zweierlei Maß zu messen und unbedingte Erklärbarkeit von Letzteren zu fordern.

Ein weiteres Argument dagegen, aus Gründen der Verantwortungszuschreibung absolute Erklärbarkeit zu beanspruchen, kann man gewinnen, indem man einen Blick auf den Zusammenhang der Opazität und der Leistungsfähigkeit von AI-CDSS wirft: Oftmals sind intransparente Systeme besonders leistungsfähig – ein radikaler Verzicht auf die Black-Box-Systeme würde darum bedeuten, dass man dem Gesundheitssystem und den Patienten die mit diesen Systemen verknüpften Potenziale vorenthalten würde (London, 2019).

Aus den obigen Überlegungen lässt sich schlussfolgern, dass es aus verschiedenen Gründen kaum angemessen ist, von KI-Anwendungen unbedingte Erklärbarkeit zu fordern. Ungeachtet dessen bleibt die ethische Relevanz der Zuschreibbarkeit von Verantwortung allerdings bestehen, die am Anfang dieses Kapitels kurz erläutert wurde. Wie kann also bei der Nutzung eines opaken AI-CDSS eine klare Verantwortungszuschreibung gewährleistet werden?

Anstatt in jedem Fall auf Erklärbarkeit zu beharren, könnte man evaluieren, ob die Nutzung eines bestimmten opaken AI-CDSS die Versorgungsqualität verbessert; gegebenenfalls könnte die Verwendung des Systems in bestimmten Fällen zum medizinischen Standard erklärt werden (Binkley, 2021; Ghassemi et al., 2021; Liedtke & Langanke, 2021). Die Verantwortung des Arztes könnte man so auf den sachgemäßen Systemgebrauch beschränken (vgl. Kapitel 6.2.3).

Da AI-CDSS – wie generell jede Technologie – ein nicht eliminierbares Fehlerpotenzial aufweisen (vgl. Kapitel 4.4.1), können bei deren Nutzung auch dann Schäden entstehen, wenn die Systeme sachgemäß verwendet werden und deren Entwicklungsprozess fehlerlos abgelaufen ist. Wenn man in einem solchen Fall also weder dem Arzt noch den Entwicklern Verantwortung zuschreiben kann, stellt sich die Frage, wer hier Verantwortung tragen soll. In diesen Fällen könnten neuartige Konstruktionen wie

diejenige der *e-person*[12] oder eine generelle Versicherungspflicht die Entstehung eines Verantwortungsvakuums verhindern (Braun et al., 2020) (vgl. Kapitel 6.2.3). Es fragt sich aber, inwiefern solche Lösungen erforderlich sind, um bei der Nutzung von AI-CDSS eine klare Verantwortungszuschreibung zu ermöglichen. Schließlich ist nicht nur die Verwendung von KI-Systemen mit einem nicht eliminierbaren Schadenspotenzial verbunden.

Dass medizinische Maßnahmen wie Operationen oder die Einnahme von Medikamenten ein gewisses Risiko aufweisen, kann als normal gelten. Mit diesen Risiken geht man innerhalb der Medizin grundsätzlich so um, dass man deren Wahrscheinlichkeit möglichst genau bestimmt und dem Patienten mitteilt. Auch in Bezug auf AI-CDSS bietet es sich an, diese hinsichtlich der Sicherheit systematisch zu evaluieren und die Patienten darüber aufzuklären (vgl. Kapitel 6.1.18.4 und 6.3.8).

4.9 Wahrung der Integrität der Arzt-Patient-Beziehung

Zu Beginn dieses Kapitels, in dem die Auswirkungen der Nutzung von AI-CDSS auf die Arzt-Patient-Beziehung beleuchtet werden, ist zunächst festzuhalten, dass einige der in der vorliegenden Arbeit schon behandelten Aspekte eng mit der Arzt-Patient-Beziehung verknüpft sind.

So stehen die Ausführungen zur Patientenautonomie und zur Zuschreibbarkeit von Verantwortung (vgl. Kapitel 4.5 und 4.8) in direktem Zusammenhang mit der Arzt-Patient-Beziehung. Im vorliegenden Kapitel könnte man auch die partizipative Entscheidungsfindung und den Paternalismus behandeln, die in dieser Arbeit im Zusammenhang mit der Patientenautonomie untersucht wurden (vgl. Kapitel 4.5.3). Schließlich gilt Paternalismus in der viel zitierten Arbeit von Emanuel und Emanuel (1992) als eine Form der Arzt-Patient-Beziehung. Die Auswirkungen der Nutzung von AI-CDSS auf die Arzt-Patient-Beziehung beschränken sich jedoch nicht darauf, dass bestimmte Formen derselben gefördert und andere Formen unterminiert werden.

Wenn Ärzte durch KI-Systeme unterstützt werden, wird das Fundament der Arzt-Patient-Beziehung – das Vertrauen des Patienten in den Arzt – berührt (Funer, 2021). Der Frage,

[12] Als *e-person* würde dem AI-CDSS der Status einer Rechtsperson zugesprochen werden, wodurch sich Fragen der Haftung und der Verantwortung leichter klären lassen sollen (Braun et al., 2020).

wie sich dieses und die Arzt-Patient-Beziehung im Allgemeinen verändern, wenn AI-CDSS verwendet werden, könnte man empirisch nachgehen. In Anbetracht der Tatsache, dass es sich bei den Systemen um eine neue Technologie handelt, deren Nutzung in der Breite noch stark eingeschränkt ist, verwundert es nicht, dass es (dem Wissen des Autors nach) bisher noch keine großen, aussagekräftigen empirische Studien zu dieser Frage gibt. Daher sollen die nachfolgenden Überlegungen analytisch ausloten, welche Entwicklungen wahrscheinlich sind. Wie sich die ärztliche Verwendung von AI-CDSS auf das Vertrauen des Patienten in den Arzt auswirken kann, soll als erstes diskutiert werden (vgl. Kapitel 4.9.1). Anschließend konzentrieren sich die Überlegungen auf die Folgen der KI-Nutzung für die Empathie und den nicht-reduktionistischen, bio-psycho-sozialen Fokus des Arztes (vgl. Kapitel 4.9.2).

4.9.1 Vertrauen in KI-unterstützte Ärztinnen und Ärzte

Bezüglich des Vertrauens des Patienten in den Arzt lassen sich mindestens zwei Dimensionen unterscheiden: eine epistemische und eine normative (Funer, 2021; Keren, 2007). Anhand dieser beiden Formen des Vertrauens analysieren die folgenden Überlegungen im Anschluss an Funer (2021), wie sich die Nutzung von AI-CDSS auf das Vertrauen des Patienten in den Arzt auswirkt.

Fasst man die epistemische Dimension des Vertrauens ins Auge, so kann der Arzt als Sprecher betrachtet werden. Das Vertrauen bezieht sich hierbei auf die Aussagen des Arztes: Einem Sprecher zu vertrauen ist eng mit der Einschätzung dessen epistemischer Autorität verknüpft (Keren, 2007). Der Patient vertraut dem Behandelnden in der Annahme, dass dieser als Arzt hinsichtlich fachlicher Fragen, die mit der Behandlung seiner Erkrankung verbunden sind, eine größere epistemische Autorität und fachliche Kompetenz als er, der Patient, besitzt (Funer, 2021).

Wie ist nun hier der Einfluss der KI-Nutzung einzuschätzen? An dieser Stelle ist zu bedenken, dass sich die KI-basierte Unterstützung des Arztes prinzipiell auf die kognitive Dimension klinischer Entscheidungen bezieht (vgl. Kapitel 4.1.1). Sofern ein AI-CDSS das Einsatzziel erreicht und es die Qualität der ärztlichen Entscheidungsfindung verbessert, kann man dem KI-Nutzer einen höheren Grad an fachlicher Kompetenz und epistemischer Autorität zusprechen. Damit Patienten aber überhaupt erkennen können, dass die KI-basierte Unterstützung die epistemische Basis der ärztlichen Entscheidung verbessert, muss die Leistungsfähigkeit des AI-CDSS für den Patienten *verständlich* sein.

Um das Vertrauen des Patienten in einen KI-unterstützten Arzt zu stärken, dürfte es darum förderlich sein, die Systemergebnisse – mitsamt zugehörigem Konfidenzgrad – auch an den Patienten auszugeben (vgl. Kapitel 6.1.13).

Wenn man den Arzt als Sprecher versteht, kann für das Vertrauen des Patienten in den KI-unterstützten Arzt auch noch ein weiterer Aspekt relevant sein. Um diesen Aspekt freizulegen, soll im Anschluss an Funer (2021) untersucht werden, wie Patienten in epistemischer Hinsicht zu Vertrauen in den Arzt gelangen. Hier ist zunächst festzustellen, dass die Patienten gegenüber dem Arzt in medizinisch-fachlicher Hinsicht gewöhnlicherweise einen epistemisch untergeordneten Standpunkt einnehmen. Von dort mag es den Patienten schwer fallen, die fachlichen Aussagen eines Arztes nach ihrer Qualität zu bewerten.

Dennoch müssen Patienten der Expertise eines Arztes nicht blind vertrauen. Stattdessen können sie nach Anhaltspunkten suchen, um die Expertise eines Arztes jedenfalls einzuschätzen. Dazu können etwa gezielte Nachfragen zu fachlichen Informationen dienen. Einem Arzt, der imstande ist, Fachwissen zu erklären, kann ein Patient einen höheren Grad an fachlicher Expertise zusprechen. Auch die Vertrauenswürdigkeit eines solchen Arztes dürfte ein Patient höher einschätzen. Man kann daraus schließen, dass die Vertrauenswürdigkeit eines KI-unterstützten Arztes in den Augen des Patienten gesteigert werden kann, wenn ein AI-CDSS Erklärungen für die Empfehlungen angibt (Funer, 2021). Auch von dieser Warte aus betrachtet ist es ethisch relevant, Erklärbarkeit zu gewährleisten (vgl. Kapitel 6.1.16).

Im oben beschriebenen Szenario wird davon ausgegangen, dass der Patient die epistemische Autorität und die Vertrauenswürdigkeit, die er dem AI-CDSS zuschreibt, der Expertise und der Vertrauenswürdigkeit des Arztes gewissermaßen hinzufügt. Dahinter steht das Paradigma der Unterstützung: Ist das KI-System vertrauenswürdig, so gewinnt auch der davon unterstützte Arzt in den Augen des Patienten an Vertrauenswürdigkeit. Der Patient könnte ein AI-CDSS allerdings (möglicherweise unbewusst) auch als Konkurrenz des Arztes auffassen. Dieser Fall könnte besonders dann eintreten, wenn sich die Diagnosen, Prognosen und Therapieempfehlungen des KI-Systems auf der einen Seite und des Arztes auf der anderen Seite nicht decken, sondern widersprechen.

Die hohe epistemische Autorität, die ein Patient einem System zuschreiben mag, könnte er dann gegen diejenige des Arztes abwägen. Nicht zuletzt weil etwa im populärwissenschaftlichen Kontext teilweise recht unkritisch über die Möglichkeiten von KI berichtet wird, ist die Befürchtung nicht abwegig, dass Patienten die Leistungsfähigkeit von AI-CDSS überschätzen könnten. So könnten sie davon ausgehen, dass diese Systeme bei klinischen Entscheidungen eine prominentere und autonomere Rolle einnehmen als das bisher in der Praxis tatsächlich der Fall ist (Triberti et al., 2020).

Solch eine Überschätzung der Rolle der Systeme kann gleichzeitig mit einer Unterschätzung der Aufgabe des Arztes bei der KI-unterstützten Entscheidungsfindung einhergehen. Hierdurch könnte die Position des Arztes untergraben werden und die Arzt-Patient-Beziehung nachteilig beeinflusst werden. Um das zu verhindern, sollten AI-CDSS jedenfalls relevante Outputs mit dem jeweiligen Konfidenzgrad auch an den Patienten ausgeben (vgl. Kapitel 6.1.13). Somit wird deutlich, dass die Systemempfehlungen nicht schlechthin richtig sind und AI-CDSS keine Entscheidungen treffen. Außerdem kann es sinnvoll sein, Patienten vor der Nutzung einer KI-Anwendung explizit über die Rollenverteilung zwischen Arzt und System aufzuklären (vgl. Kapitel 6.3.8).

Das Vertrauen gegenüber dem Arzt und dem AI-CDSS wurde bisher in seiner epistemischen Dimension behandelt. Vertrauen bezieht sich jedoch nicht nur auf die Einschätzung der epistemischen Autorität eines Sprechers. Es besitzt auch eine normative Komponente: Der Patient vertraut dem Arzt, weil er davon ausgeht, dass der Arzt dem ärztlichen Ethos entsprechend die Notlage des Patienten nicht zum eigenen Vorteil ausnutzt, sondern stattdessen das Patientenwohl als höchstes Ziel verfolgt und in diesem Sinne die Werte und Präferenzen sowie den individuellen Lebenskontext des Patienten bei der klinischen Entscheidungsfindung hinreichend berücksichtigt (Funer, 2021).

Falls der Patient zu der Überzeugung gelangt, dass das AI-CDSS die Interessen anderer – etwa die wirtschaftlichen Interessen bestimmter Gruppen – eine höhere Priorisierung als dem Patientenwohl gibt, dürfte dadurch das Vertrauen des Patienten in einen KI-unterstützten Arzt sinken. Von diesen Überlegungen her erhalten die Untersuchungen zur ärztlichen Autonomie (vgl. Kapitel 4.6) eine besondere Relevanz. Um das Vertrauen des Patienten in einen KI-unterstützten Arzt nicht zu unterminieren, müssen vor allem

wertsensitive Therapieempfehlungen an den Patienten angepasst werden (vgl. Kapitel 6.3.12).

4.9.2 Auswirkungen auf die Empathie und den nicht-reduktionistischen, bio-psycho-sozialen Fokus der Ärztinnen und Ärzte

In der Medizin geht es darum, Gesundheit wiederherzustellen und zu fördern. Diese geht über das hinaus, was sich rein biomedizinisch beschreiben lässt. Dass darunter umfassendes Wohlergehen zu verstehen ist, das neben biomedizinischen auch psychische und soziale Aspekte umfasst, geht auch aus der Gesundheitsdefinition der WHO (2022) hervor. Die Bedeutung des nicht reduktionistischen, bio-psycho-sozialen ärztlichen Fokus wird bei der Arzt-Patient-Beziehung deutlich: Diese dient nicht nur dem Austausch von Informationen. Die Zuwendung des Arztes zum Patienten und die Arzt-Patient-Beziehung als solche kann zur Heilung beitragen (Riedl & Schüßler, 2017).

Eine Vorbedingung dafür, dass ein Arzt die verschiedenen relevanten Dimensionen der Medizin bei klinischen Entscheidungen beachtet, liegt darin, dass er überhaupt die Notwendigkeit dessen erkennt: Der Fokus des Arztes darf nicht reduktionistisch, er muss bio-psycho-sozial geprägt sein. Inwiefern die Verwendung von AI-CDSS diese nicht-reduktionistische Orientierung des Arztes unterstützt oder behindert, gilt es im Folgenden zu erörtern.

Die Auswirkungen der KI-Nutzung auf den bio-psycho-sozialen Fokus des Arztes sind ambivalent. Es ist möglich, dass der Reduktionismus eines AI-CDSS den Arzt negativ beeinflusst. Das lässt sich anhand von Heideggers Technikkritik erklären (Karches, 2018). Heidegger bezeichnet Technik mit dem Begriff des Gestells: Sie ist eine Weise des Entbergens, des Zum-Vorschein-Bringens. Technik *ver-stellt* aber auch den Zugang zu bestimmten Aspekten der Realität (Heidegger, 1962).

AI-CDSS können dabei förderlich sein, Erkenntnisse über den Patienten zu gewinnen – aber nur in begrenzter Form (Karches, 2018). Die Systeme können nur dasjenige verwenden und in Berechnungen einbeziehen, was sich in computerisierter Form darstellen lässt (vgl. Kapitel 4.1.1). Hieraus ergibt sich eine Gefahr für die Arzt-Patient-Beziehung: Der den AI-CDSS inhärente Reduktionismus könnte den Arzt so beeinflussen, dass dieser den Patienten nicht mehr primär als menschliches Gegenüber betrachtet, sondern vor allem als Ansammlung von biomedizinischen Daten, gleichsam als Avatar, auffasst und behandelt (Arnold, 2021; Cabitza et al., 2017; Verghese, 2008).

Auch wenn diese Befürchtung begründet ist, soll hier darauf hingewiesen werden, dass diese Verkürzung nicht automatisch bei der Nutzung von AI-CDSS eintritt. Der Arzt kann die tendenziell reduktionistischen Systemempfehlungen in seine ihm als Menschen eigene Perspektive eingliedern, die auch das Psychische und das Soziale einbezieht (Funer, 2021; van Baalen et al., 2021). Bei der Systemnutzung sollte sich der Arzt folglich nicht der reduktionistischen Logik der KI-Anwendungen unterordnen und Systemergebnisse nicht blind übernehmen. Er sollte die Outputs hingegen – sofern nötig – an den Patienten individuell anpassen, um dessen Gesundheit nicht nur in ihrer biologischen, sondern auch in ihrer psychischen und sozialen Dimension zu erhalten und zu fördern (vgl. Kapitel 6.3.12).

Im Anschluss an diese Überlegungen lässt sich auch auf das Argument verweisen, dass die Auslagerung des Berechenbaren, des rein Biomedizinischen und der ‚*hard facts*' in das AI-CDSS den Arzt gerade dabei unterstützen könnte, sich besonders auf die genuin humane, holistische Perspektive zu konzentrieren (Topol, 2019). Da der Mensch jedenfalls auch ein biologisches Wesen ist, ist hier fraglich, inwiefern man noch von einer ganzheitlichen Behandlung sprechen kann, wenn das Biomedizinische im Extremfall komplett aus dem Aufgabenbereich des Arztes entfernt und in das AI-CDSS ausgelagert wird.

Um das Argument zu bekräftigen, dass die Nutzung von AI-CDSS potentiell eine empathische Behandlungsweise und den nicht-reduktionistischen, bio-psycho-sozialen Fokus des Arztes fördert, kann man zudem auf die Zeitersparnis hinweisen, die durch den KI-Gebrauch wohl erzielbar ist. Sich auf das Gegenüber und dessen Emotionen einzulassen sowie Empathie zu zeigen sind schließlich häufig mit einem erhöhten Zeitaufwand verbunden. Wenn AI-CDSS den Arzt etwa bei diagnostischen oder prognostischen Aufgaben entlasten (vgl. Kapitel 4.11), kann dieser die sich daraus ergebenden freien Zeiträume für eine Intensivierung der empathischen und genuin humanen Dimension des ärztlichen Tuns nutzen (Topol, 2019).

Konterkariert würde diese Chance zur Verbesserung der Behandlung, wenn frei gewordene Zeit dazu verwendet würde, ärztliche Arbeitsplätze abzubauen. In diesem Zusammenhang ist aus ethischer Sicht zu fordern, dass die Ressourcen, die durch die Nutzung von AI-CDSS freigesetzt werden, patientenorientiert genutzt werden (Topol, 2019) (vgl. Kapitel 6.3.14).

4.10 Datenschutz und Datenverfügbarkeit

Grundsätzlich handelt es sich bei AI-CDSS um eine Form von Informationstechnologie. Die vorliegende ethische Bewertung der Systeme hat daher auch die Aspekte in den Blick zu nehmen, die mit der Nutzung patientenbezogener Daten verbunden sind. Bei den folgenden Untersuchungen stellen sich besonders Datenschutz und -verfügbarkeit als relevant heraus. Bevor auf Letztere eingegangen wird, soll Ersterer fokussiert werden.

Zu Beginn dieser Ausführungen erscheint eine Klarstellung angebracht: Datenschutz kann nicht nur ethisch geboten, sondern auch rechtlich erforderlich sein. Auf juristische Aspekte kann hier nicht näher eingegangen werden.[13] Weil die vorliegende Arbeit aus dem Gebiet der Ethik stammt, soll zunächst die normative – die prinzipienethische und die konsequentialistische – Begründung des Datenschutzes im Anschluss an Marckmann (2003) dargelegt werden.

Prinzipienethisch lässt sich Datenschutz erstens vom Prinzip des Respekts der Patientenautonomie ableiten. Dieses verlangt, dass Patienten selbst bestimmen können, wer in welchem Ausmaß Zugang zu ihren Daten hat.

Die informationelle Selbstbestimmung kann beeinträchtigt werden, wenn etwa autorisierte Personen oder Institutionen die Datennutzungsrechte überdehnen, die ihnen zugesprochen wurden. Ein Beispiel hierfür kann anhand einer Contact-Tracing-App der singapurischen Regierung gegeben werden. Entgegen initialer Behauptungen der Regierung stellte sich heraus, dass die Nutzerdaten nicht nur zur Kontaktverfolgung von COVID-19-Infizierten, sondern auch zur Verfolgung von Straftaten benutzt werden konnten (Illmer, 2021). Solch eine Verwendung von Daten zu Zwecken, denen die Betroffenen nicht zugestimmt haben, wird auch als *repurposing* bezeichnet und läuft der Patientenautonomie zuwider (WHO, 2021).

Neben der informationellen Autonomie kann auch die allgemeine Selbstbestimmung über die eigene Lebensführung durch den Missbrauch von Gesundheitsdaten beschädigt werden. Denn der Zugriff auf Gesundheitsdaten ermöglicht es nicht nur, Einblicke in die Gegenwart und in die Vergangenheit der betreffenden Person zu gewinnen. Durch Extrapolation kann man auch Wahrscheinlichkeiten für bestimmte Verhaltensweisen in

[13] Rechtliche Aspekte des Datenschutzes im Zusammenhang mit KI-Systemen in der Medizin werden etwa von der ZEKO (2021) behandelt.

der Zukunft bestimmen. Die Lebensführung einer Person kann dadurch beeinflusst werden. Das kann im Interesse verschiedener Dritter liegen (etwa Versicherungen, Arbeitgeber oder in vielerlei Hinsicht konkurrierende Personen) (Sikma et al., 2020).

Prinzipienethisch lässt sich Datenschutz nicht nur vom Autonomie-, sondern auch vom Nichtschadensprinzip ableiten, weil personenbezogene Daten missbraucht werden können. In Finnland beispielsweise haben sich Kriminelle Zugang zu besonders sensiblen Daten aus digitalen psychotherapeutischen Sitzungen von mindestens 2 000 Patienten verschafft. Einige Betroffene gaben an, von den Hackern zur Überweisung von 200 Euro aufgefordert worden zu sein (aerzteblatt.de, 2020). Schweren Schaden kann auch der Missbrauch von Daten bewirken, die mit der medizinischen Identität des Betroffenen verbunden sind. Diese können in manchen Ländern etwa dazu benutzt werden, Gesundheitsleistungen zu beantragen oder zu bezahlen. Eine solche illegitime Verwendung von Daten hat in den USA Kosten in Höhe von beinahe 20 000 US-Dollar verursacht (CBS News, 2019; Steger, 2019).

Darüber hinaus lässt sich Gesundheitsdatenschutz auch mit konsequentialistischen Erwägungen begründen: Die Bedingung dafür, dass der Patient Informationen über sich preisgibt, besteht nämlich darin, dass er begründet davon ausgehen kann, dass diese Informationen nicht missbraucht werden. Ein unzureichendes Niveau an Datenschutz kann folglich dazu führen, dass der Patient Informationen zurückhält, die für die klinische Entscheidungsfindung relevant sind. Hierunter kann die Qualität der Patientenversorgung leiden (Marckmann, 2003).

Nachdem die normative Begründung des Datenschutzes im Bereich der klinischen Entscheidungsfindung erörtert wurde, ist zu untersuchen, worauf im Sinne des Datenschutzes bei der Entwicklung und Anwendung von AI-CDSS zu achten ist. Bevor der Umgang mit Trainingsdaten behandelt wird, sollen solche Daten in den Blick genommen werden, die bei der Anwendung von KI-Systemen genutzt werden.

In allgemeiner Form gibt man Daten in eine KI-Anwendung ein (Input), die in einem Rechenprozess verarbeitet werden. Daraufhin wird ein Ergebnis ausgegeben (Output) (vgl. Kapitel 3.1.1). Als Gemeinsamkeit weisen Input- sowie Outputdaten zunächst grundsätzlich einen Bezug zu demjenigen Patienten auf, für dessen Behandlung man das

AI-CDSS benutzt. Der Personenbezug von Daten geht jedoch mit einer Anfälligkeit für Datenmissbrauch einher (Weichert, 2021).

Relativierend ist darauf hinzuweisen, dass die mit der Nutzung von Patientendaten verbundenen Risiken in der Medizin nicht erst durch die Implementierung von AI-CDSS aufgekommen sind. Vielmehr ist grundsätzlich die Verwendung aller Systeme, die Daten von Patienten verwenden, mit diesen Gefahren verbunden. Wie etwa bei der digitalen Patientenakte, so ergibt sich auch bei der Nutzung von KI-Systemen das Dilemma, dass ein leichterer Zugang zu Patientendaten die Effektivität und die Qualität der medizinischen Versorgung verbessern kann, jedoch tendenziell auch deren Missbrauch ermöglicht und vereinfacht (L. Lee, 2017).

Wenn Patienten der Verwendung ihrer Daten nicht zustimmen, kann das Nutzenpotenzial von AI-CDSS auch deshalb eingeschränkt werden, weil dieses mit dem Phänomen Big Data und der Individualisierung der Entscheidungsfindung zusammenhängt (vgl. Kapitel 3.4.4 und 4.3). Folglich besteht die Gefahr, dass die mit den Systemen verbundene Hoffnung auf eine Individualisierung klinischer Entscheidungen sich nicht einlösen lässt, wenn Patienten bestimmte Daten nicht freigeben.

Im Rahmen des vorliegenden Kapitels ist der Umgang mit allen personengebundenen Daten zu behandeln, die mit KI-Systemen zur Unterstützung ärztlicher Entscheidungen in Verbindung stehen. Da zur Entwicklung von ML-basierten Anwendungen Trainingsdaten verwendet werden (vgl. Kapitel 3.1.2), die ebenfalls einen Patientenbezug aufweisen können, soll nachfolgend der Umgang mit diesen in den Fokus der Untersuchungen rücken. Weil Trainingsdaten beim Entwicklungsprozess eines AI-CDSS verwendet werden, kann deren Missbrauch darauf zurückgehen. Diese können aber auch gegebenenfalls danach identifiziert und missbraucht werden: Indem man eine ML-Anwendung oder deren Funktion analysiert, kann man teils recht genaue Rückschlüsse auf die Trainingsdaten ziehen, die bei der Entwicklung des KI-Systems verwendet wurden (Arshad & Strodthoff, 2021; Kaissis et al., 2020). Weil folglich patientenbezogene Trainingsdaten missbraucht werden können, ist auch in Bezug auf diese der Datenschutz zu beachten (vgl. Kapitel 6.1.14).

Im Zusammenhang des ethisch vertretbaren Umgangs mit patientenbezogenen Trainingsdaten ist nicht nur auf die Gefahr von deren Missbrauchs einzugehen. Wie in Kapitel 3.2 deutlich wurde, stellen diese eine Ressource dar: Zur Entwicklung von ML-Systemen sind umfassende Trainingsdatensätze nötig. Daher ist auch die

Realisierung der Chancen der Anwendungen an die Verfügbarkeit geeigneter Trainingsdaten geknüpft. Angesichts der Bedeutung von Daten als Ressource für die Entwicklung von KI-Systemen und für die Verbesserung der Patientenversorgung wird deutlich, dass für einen ethisch vertretbaren Umgang mit Trainingsdaten mehr als nur Datenschutz zu beachten ist.

Wie in Bezug auf die patientenbezogenen Daten, die bei der Nutzung von AI-CDSS verwendet werden, ist im Hinblick auf Trainingsdaten eine Abwägung erforderlich. Diese verfügbar zu machen gebietet das Prinzip des Wohltuns, insoweit die zu entwickelnden Systeme die Qualität der Patientenversorgung verbessern (vgl. Kapitel 4.3). Weil die Bereitstellung von Trainingsdaten aber auch mit dem Risiko des Datenmissbrauchs einhergeht, kann diese mit dem Prinzip des Nichtschadens konfligieren. Es ist daher das Wohltuns- gegen das Nichtschadensprinzip abzuwägen.

Auf den ersten Blick kann hier eine Parallele zu der vorhin erläuterten Abwägung bezüglich des Umgangs mit Daten bei der Nutzung eines KI-Systems in der Praxis sichtbar werden: Auch dass man Daten bei der KI-Nutzung verfügbar(er) macht, kann vom Prinzip des Wohltuns abgeleitet werden und dem Prinzip des Nichtschadens zuwiderlaufen.

Die Vergleichbarkeit zwischen der Anwendungs- und der Entwicklungssituation im Zusammenhang mit Datenschutz und -verfügbarkeit ist aber in einem wichtigen Punkt nicht gegeben: Geht es um die Verwendung von AI-CDSS in der Praxis, so ist es dieselbe Person, der die zunehmende Verfügbarkeit der Daten sowohl nutzen als auch schaden kann. Bei der Bereitstellung von Trainingsdaten sind es aber unterschiedliche Individuen, die zum einen vom damit verbundenen Nutzen profitieren und zum anderen unter dem mit der Datenverfügbarkeit verknüpften Schadenspotenzial leiden können. Alle Patienten, deren Behandlungsqualität durch die Verwendung eines bestimmten AI-CDSS verbessert werden kann, profitieren von der damit zusammenhängenden Bereitstellung von Trainingsdaten. Werden diese allerdings missbraucht, trifft der Schaden nicht die mit dem AI-CDSS Behandelten, sondern diejenigen, die ihre Daten für die Entwicklung des betreffenden Systems bereitgestellt haben. Man kann wohl davon ausgehen, dass es höchstens ausnahmsweise passieren dürfte, dass eine Person, die ihre Daten zur Entwicklung eines KI-Systems verfügbar gemacht hat, zu einem späteren Zeitpunkt mithilfe desselben behandelt wird. Daher lässt sich sagen, dass die Menge der Personen,

denen die Verfügbarkeit der Trainingsdaten nutzt, nicht zusammenfällt mit der Gruppe, der die Verfügbarkeit der Trainingsdaten schaden kann.

An diesem Punkt stellen sich Fragen: Wie kann man ethisch vertretbar damit umgehen, dass die Nutzen- und Schadenspotenziale der Bereitstellung von Trainingsdaten ungleich verteilt sind? Auf welche Weise ist zu verhindern, dass aufgrund des Missbrauchsrisikos keine Trainingsdaten bereitgestellt werden und somit die Entwicklung von – potenziell nützlichen – AI-CDSS verunmöglicht wird?

Insofern diejenigen, die ihre Daten zur Entwicklung von KI-Systemen freigeben, zwar unter Datenmissbrauch leiden können, von der Verfügbarkeit der Trainingsdaten aber nicht profitieren, ist es aus ethischer Sicht insbesondere geboten, den Missbrauch von Trainingsdaten zu unterbinden – sowohl bei der Entwicklung (vgl. Kapitel 6.1.14) als auch bei der Anwendung der Systeme (vgl. Kapitel 6.3.13).

Dennoch ist zu bedenken, dass man Datenschutz nur bedingt gewährleisten kann. Wollte man eine hundertprozentige Sicherheit vor dem Missbrauch von personenbezogenen Trainingsdaten erzielen, so dürfte man diese überhaupt nicht verwenden. Dadurch würde aber die Entwicklung von AI-CDSS erheblich erschwert, vielleicht sogar verunmöglicht. Doch auch wenn man die Bedeutung des Datenschutzes relativiert, ist zu beachten: Legt man den Fokus zu stark auf diesen, kann die Verfügbarkeit von Trainingsdaten dermaßen beeinträchtigt werden, dass die Entwicklung von KI-Systemen stark behindert wird. Somit würde allerdings – im Widerspruch zum Wohltunsprinzip – eine Chance zur Verbesserung der Patientenversorgung vergeben werden. Um im Zusammenhang mit der Verfügbarkeit von Trainingsdaten neben dem Datenschutz auch das Nutzenpotenzial der Daten angemessen zu beachten, kann man sich am Konzept der Datensouveränität orientieren. Dieses Konzept erklärt der Deutsche Ethikrat (2018a) wie folgt:

> „Unter Datensouveränität verstehen wir eine den Chancen und Risiken von Big Data angemessene verantwortliche informationelle Freiheitsgestaltung. Um dies zu gewährleisten, ist das traditionelle, primär auf die grundrechtlich geschützte informationelle Selbstbestimmung bezogene Datenschutzrecht weiterzuentwickeln und neu zu gestalten, indem inhaltlich umfassende grundlegende normative Vorgaben einbezogen und instrumentell neue Wege beschritten werden.“ (S. 262)

Die hier angesprochene „den Chancen und Risiken von Big Data angemessene verantwortliche informationelle Freiheitsgestaltung“ (Deutscher Ethikrat, 2018a, S. 262) erscheint aus ethischer Perspektive auch beim Umgang mit personenbezogenen Trainingsdaten attraktiv. Wenngleich die Idee der Datensouveränität in der Theorie sinnvoll ist, stellt sich dennoch die Frage nach deren praktischer Umsetzbarkeit. Der Begriff der Souveränität, den dieses Konzept in Anspruch nimmt, insinuiert ein recht hohes Maß an Kontrolle. Der Umgang mit gesundheitsbezogenen Daten ist jedoch von einer derartigen Komplexität gekennzeichnet, dass es fraglich ist, ob man den Anspruch der Souveränität hierbei sinnvoll erheben kann und sollte (Marckmann, 2020; Strech et al., 2020).

Praktikabler dürfte es jedenfalls sein, im Sinne von Datenspenden den Fokus weniger auf die kontinuierliche Kontrolle durch denjenigen zu legen, von dem die Daten stammen. Stattdessen kann man versuchen, eine angemessene Nutzung der gespendeten Daten sicherzustellen, indem man Transparenz schafft (Strech et al., 2020). Da das Ziel der Datensouveränität aus ethischer Sicht attraktiv ist, erscheint es empfehlenswert, Maßnahmen zur Gewährleistung von Datensouveränität in der Praxis zu erproben (vgl. Kapitel 6.1.14). Ob sich dieses Konzept als praxistauglich erweist, wird sich zeigen.

In jedem Fall dürfte sich in den obigen Untersuchungen die Bedeutung von Datenspenden gezeigt haben: Weil die Entwicklung von AI-CDSS auf die Bereitstellung von Trainingsdaten angewiesen ist und die Anwendung der Systeme die Patientenversorgung verbessern kann, dienen Datenspenden dem Wohl der Gemeinschaft (acatech, 2020). Um die Bereitschaft zur Datenspende zu erhöhen, sollte man – analog zur Blutspende etwa – das Bewusstsein in der Gesellschaft dafür stärken, dass diese die Qualität der Gesundheitsversorgung verbessern kann (vgl. Kapitel 6.2.10).

4.11 Effizienz

Besonders wenn eine medizinische Maßnahme nicht nur von dem betreffenden Patienten, sondern teilweise oder ganz von einem gemeinschaftlichen Gesundheitssystem finanziert wird, sollte diese ein positives Kosten-Nutzen-Verhältnis aufweisen. Schließlich sind die Ressourcen eines Gesundheitssystems begrenzt. Wenn bei der Behandlung eines Patienten finanzielle Mittel ineffizient verwendet werden, wird der Spielraum für die Behandlung anderer eingeschränkt. Weil bei der Entwicklung, der Implementierung und

dem Betrieb von AI-CDSS hohe Kosten anfallen, ist im Folgenden deren Effizienz zu untersuchen (Marckmann, 2003).

Dabei kann man zunächst deren zeitliche Dimension ins Auge fassen. Diese ist aus verschiedenen Gründen relevant. Die Chance der KI-Systeme, die in Kapitel 4.9.2 beschrieben wird, beruht etwa auf der Annahme, dass deren Nutzung den Ärzten zeitliche Freiräume schafft, wodurch diese sich vermehrt auf eine empathische und bio-psycho-soziale Behandlung des Patienten konzentrieren können. Die zeitliche Effizienz der KI-Nutzung hat auch eine ökonomische Dimension: Die Vergütung des Personals macht einen großen Anteil aller Kosten im Gesundheitswesen aus (vdek, 2024). Die Wirtschaftlichkeit der AI-CDSS ist daher eng mit deren Auswirkungen auf die zeitliche Belastung der Ärzte verknüpft. Grundsätzlich ist davon auszugehen, dass bestimmte ärztliche Aufgaben durch die Nutzung von KI-Systemen vereinfacht und beschleunigt werden können. Besonders bei kognitiv beanspruchenden Aufgaben wird deutlich, dass Rechenoperationen von KI-Anwendungen schneller als die menschliche Kognition sind. So dauert die Segmentierung einer Herzkammer auf MRT-Bildern durch eine ML-Anwendung nur 15 Sekunden, während Experten für diese Aufgabe etwa eine halbe Stunde benötigen (Marr, 2017).

Ärzte werden durch die Nutzung von AI-CDSS aber nicht nur Zeit gewinnen. Verwenden sie diese Systeme, so kommen auch einige zeitintensive Aufgaben auf sie zu. Bereits vor deren Benutzung gilt es, sich damit vertraut zu machen und in diesem Zusammenhang möglicherweise entsprechende Kurse zu besuchen, die deren Bedienung, aber auch darüber hinausgehende, etwa rechtliche oder organisatorische Fragen behandeln (vgl. Kapitel 6.3.7). Da sich die Systeme und die damit einhergehenden Anforderungen an die Benutzer kontinuierlich verändern, sind Schulungen für die Anwender wohl nicht nur einmal, sondern im Verlauf der Anwendung mehrmals vonnöten. Zeitliche Belastungen entstehen über Fortbildungen hinaus auch bei der Verwendung der Systeme. So erfordert etwa die Aufklärung des Patienten (vgl. Kapitel 6.3.8) oder die Überprüfung der Outputs (vgl. Kapitel 6.3.12) Zeitaufwand.

Ob der Gebrauch von AI-CDSS die Nutzer letztlich in zeitlicher Hinsicht be- oder entlasten wird, ist im Allgemeinen schwierig zu beurteilen. Besonders die Vielfältigkeit der zeitlichen Belastungen erschwert die Berechnung der Differenz zwischen zeitlichen Be- und Entlastungen. Daher sind Evaluationsstudien in der Praxis durchzuführen, um valide Aussagen zu den Auswirkungen der Benutzung der KI-Systeme auf die zeitliche

Effizienz von Ärzten treffen zu können. Um dennoch eine Tendenz auszumachen, kann man einen Blick auf die einschlägige Literatur werfen. Diese betont vor allem die Chance für den Arzt, mithilfe von KI-Systemen Zeit einzusparen (Liu et al., 2018; Topol, 2019). Obgleich das Ausmaß des teils hinter diesen Arbeiten stehenden Technikoptimismus skeptisch betrachtet werden kann, kann man doch die Annahme vorsichtig teilen, dass die Verwendung von AI-CDSS den Arzt in zeitlicher Hinsicht eher entlasten als zusätzlich belasten dürfte.

Ein weiterer Faktor, der bei der Verwendung von entscheidungsunterstützenden KI-Systemen zur Einsparung von Kosten beitragen könnte, ist die Reduktion von Behandlungsfehlern. Diese besitzen nämlich auch eine ökonomische Dimension: Alleine in Verbindung mit diagnostischen Fehlern wurden in den USA zwischen 1985 und 2010 Zahlungen in Höhe von 38,8 Milliarden US-Dollar fällig (Saber Tehrani et al., 2013). Soweit man Fehler bei der Diagnostik zumindest teilweise durch die Nutzung von AI-CDSS verhindern kann, wird in der Verwendung dieser Systeme ein beträchtliches Potenzial zur Einsparung von Kosten sichtbar (D. W. Bates et al., 2021).

Es ist klar, dass man das Kosten-Nutzen-Verhältnis der KI-Systeme angesichts ihrer Heterogenität nicht im Allgemeinen bestimmen kann. Ob die Kosten des Kaufs und der Wartung einer Anwendung durch den zugehörigen Nutzen gerechtfertigt sind, lässt sich letztendlich nur für jedes System einzeln beantworten. Dabei sollte man sämtliche Kosten dem Nutzen des jeweiligen AI-CDSS gegenüberstellen. Empirische Untersuchungen der Wirtschaftlichkeit von entscheidungsunterstützenden KI-Systemen in der Medizin sind aber rar. Ein Grund für diese schlechte Evidenzlage dürfte im noch jungen Alter der Systeme liegen. Jedoch sind auch Studien hinsichtlich der Wirtschaftlichkeit der schon länger im Betrieb befindlichen CDSS nur in geringer Zahl vorhanden (Sutton et al., 2020). In einem Review konnten nur sechs Studien zur Kosteneffizienz von KI-Systemen in der Medizin identifiziert werden, die die Autoren darüber hinaus als methodisch mangelhaft beschreiben (Wolff et al., 2020).

Als Beispiel für eine Studie zur Kosteneffizienz kann man eine Untersuchung eines KI-Systems zur Detektion von Großgefäßokklusionen bei akutem Schlaganfall anführen (van Leeuwen et al., 2021). Die Autoren gehen davon aus, dass Ärzte gewöhnlich 6 % der Großgefäßokklusionen übersähen und dass diese Rate durch die Nutzung des AI-CDSS um die Hälfte gesenkt werden könne. Den Kosten einer KI-basierten Detektion

von 40 US-Dollar stellen die Autoren unter anderem den Gewinn von *quality adjusted life years* entgegen und kommen so zu dem Ergebnis, dass die Verwendung des betreffenden Systems im Vereinigten Königreich jährlich zu einer Kostenersparnis von 11 Millionen US-Dollar führen würde.

Damit das Kosten-Nutzen-Verhältnis bei der Auswahl von AI-CDSS hinreichend Berücksichtigung erfahren kann, sollte die hierfür nötige Evidenz geschaffen werden (vgl. Kapitel 6.1.18.5). Wichtig ist in dieser Hinsicht auch der Vergleich verschiedener KI-Systeme und anderer Alternativen (Wolff et al., 2021) (vgl. Kapitel 6.3.2).

4.12 Gerechtigkeit

Bei der ethischen Bewertung von AI-CDSS sind auch gerechtigkeitsethische Implikationen zu berücksichtigen.

Als besonders bedeutsam kann im Gerechtigkeitsdiskurs John Rawls‘ Konzept von Gerechtigkeit als Fairness bezeichnet werden (Rawls, 2001). Ein Begriff, den man grundsätzlich als nicht vereinbar mit Fairness verstehen kann, ist ‚Diskriminierung‘. Zwar kann dieses Wort mit Bezug auf das zugrunde liegende lateinische Wort *discriminare* wertneutral ‚Unterscheidung‘ bedeuten (Schneider, 2021). Im gerechtigkeitsethisch relevanten Sinne ist Diskriminierung hingegen generell negativ konnotiert und kann etwa definiert werden als „ungerechtfertigte oder gesellschaftlich unerwünschte Benachteiligung […], die Personen aufgrund der Zuordnung bestimmter Merkmale und bzw. oder der Zugehörigkeit zu bestimmten Gruppen zuteil wird“ (Schneider, 2021, S. 328). Hier kann der Begriff der Gruppe „im Sinne einer sozial signifikanten Gruppe […] oder im Sinne einer Einteilung in verschiedene Kategorien, Klassen oder Segmente [verstanden werden]“ (Schneider, 2021, S. 328). Ob eine Unterscheidung gerechtfertigt ist oder als ungerechtfertigte Differenzierung eine Diskriminierung darstellt, hängt davon ab, ob ein *sachlicher* Grund für die Unterscheidung vorliegt (Orwat, 2019). Beispiele für ungerechtfertigte Ungleichbehandlung bestehen beispielsweise in der Bezugnahme auf rassistische oder sexistische Stereotype, rein persönliche Vorlieben oder in der Verfolgung eines egoistischen Kalküls.

Diskriminierende Praktiken wurden in Verbindung mit KI-Systemen bereits in verschiedenen Zusammenhängen geschildert. Ein klassisches Beispiel für ein

diskriminierendes KI-System ist das Softwareprogramm *Correctional Offender Management Profiling for Alternative Sanctions*. Dieses soll das Risiko dafür abschätzen, dass eine straffällig gewordene Person abermals ein Verbrechen begehen wird. Unfairerweise weist dieses System Personen afroamerikanischen Hintergrunds ein höheres Rückfallrisiko zu als Menschen mit kaukasischem Hintergrund (Angwin et al., 2016). Die Diskriminierung durch dieses System besteht primär darin, dass die Behandlung von einem Risikoscore abhängt, den der ethnische Hintergrund einer Person beeinflusst.

Lassen sich Erkenntnisse aus diesem Fall auf AI-CDSS übertragen? In der Medizin werden im Gegensatz zu Gerichtsprozessen keine Strafen verhängt. Es sind hier aber knappe Ressourcen zu verteilen. Die in der vorliegenden Arbeit untersuchten Systeme zur Unterstützung von Diagnostik, Prognostik und Therapieempfehlung dienen zwar primär nicht der Distribution von begrenzten Mitteln. Diese können die Anwendungen aber auf indirekte Weise beeinflussen (Paulus & Kent, 2020). Hier kann man etwa an die Entscheidung zwischen zwei therapeutischen Alternativen denken, die sich in ihrem finanziellen Aufwand voneinander unterscheiden. Vor diesem Hintergrund ist zu fordern, dass besonders solche KI-basierten Empfehlungen, die die Verteilung von knappen Ressourcen beeinflussen, sachlich begründet sein müssen und nicht zu Diskriminierung führen dürfen (vgl. Kapitel 6.1.15).

Fairness ist im Zusammenhang mit AI-CDSS nicht nur in Bezug auf einzelne Systemempfehlungen relevant. Insofern eine Anwendung die Qualität der Patientenversorgung verbessert, stellt deren bloße Verfügbarkeit ein Nutzenpotenzial dar. An dieser Stelle ist wiederum die Abhängigkeit der KI-Anwendungen von den Trainingsdaten von Belang (vgl. Kapitel 3.4.2): Sind die Unterschiede zu groß zwischen der Population, aus der die Trainingsdaten stammen und der Gruppe, in der ein AI-CDSS angewandt wird, kann es übermäßig häufig zu falschen Ergebnissen kommen (vgl. Kapitel 4.4.1). Zwar kann man die Anwendungen lokal validieren und gegebenenfalls rekalibrieren. Hierfür sind aber entsprechende Datensätze nötig (Mitchell et al., 2021). Wenn für eine bestimmte Patientengruppe keine passenden Daten zur Entwicklung oder Rekalibrierung von AI-CDSS verfügbar sind, kann diese gegenüber anderen hinsichtlich der Qualität der medizinischen Versorgung benachteiligt sein. Somit kann man hier von Diskriminierung sprechen.

Vor diesem Hintergrund ist es kritisch, dass in der Forschung als einem Bereich, in dem viele Daten gewonnen und verarbeitet werden, derzeit bestimmte Personen systematisch unterrepräsentiert sind – unter anderem Frauen und bestimmte ethnische Gruppen (Buolamwini & Gebru, 2018; Cirillo et al., 2020; Yoon et al., 2014). Damit die potenzielle Verbesserung der Patientenversorgung durch AI-CDSS nicht nur manchen, sondern idealerweise allen zuteil wird, ist es gerechtigkeitsethisch geboten, besonders für die bisher unterrepräsentierten und benachteiligten Gruppen eine entsprechende Basis an Trainingsdaten aufzubauen (vgl. Kapitel 6.2.5).

Eine weitere Form der möglichen Diskriminierung ist mit der Zustimmung zur Nutzung von AI-CDSS verbunden: Personen, die eine überdurchschnittlich negative und ablehnende Haltung gegenüber KI-Systemen aufweisen, könnten die Verwendung der Systeme im Vergleich zu anderen häufiger ablehnen und somit weniger von der potenziellen Verbesserung der Versorgungsqualität durch diese Systeme profitieren (WHO, 2021). Um eine möglichst gerechte Verteilung des Nutzenpotenzials der Anwendungen zu gewährleisten, sollte die Bevölkerung über KI-Systeme im Allgemeinen und insbesondere über deren Nutzen- und Schadenspotenziale aufgeklärt werden, sodass unbegründete Ängste vor den Systemen aufgelöst werden können und auf diese Weise ein gerechter Zugang zu den Potenzialen von AI-CDSS gewährleistet wird (vgl. Kapitel 6.2.10).

Bei der Auseinandersetzung mit Diskriminierung ist auch zu analysieren, ob bestimmte Personen durch die Nutzung der Anwendungen besonders zahlreichen oder schweren Belastungen ausgesetzt werden. Im Verlauf der vorliegenden Arbeit hat sich an mehreren Stellen gezeigt, dass der Gebrauch von AI-CDSS Ärzte mit Belastungen konfrontieren kann: Dieser erfordert durch Schulungen (vgl. Kapitel 6.3.7), die Kontrolle der Outputs (vgl. Kapitel 6.3.12) und die Aufklärung der Patienten (vgl. Kapitel 6.3.8) zeitlichen Aufwand der Ärzte. Weiterhin beanspruchen *alert fatigue* und *deskilling* die Ärzte (vgl. Kapitel 4.7). Gewiss mögen diese von der KI-Nutzung profitieren, wenn damit das Kosten-Nutzen-Verhältnis der Behandlung verbessert wird (vgl. Kapitel 4.11). Es ist aber fraglich, ob das Nutzenpotenzial der Systeme das zugehörige Schadenspotenzial für die Ärzte aufwiegen kann. Somit kann man befürchten, dass diese die Nutzung von AI-CDSS verweigern könnten, um sich den damit verbundenen Belastungen zu entziehen.

Wenn die Anwendungen allerdings keine Nutzer finden, kann auch deren Nutzenpotenzial für die Patienten nicht realisiert werden. Hier stellt sich die Frage, ob die Belastungen, die durch die Nutzung von AI-CDSS auf die Ärzte zukommen, zumutbar sind. Das hängt von deren Schwere und von der Anzahl und der Ausprägung der Nutzenpotenziale für die Patienten und für die Ärzte ab. Insofern es sich um zumutbare Belastungen handelt, sind die Ärzte anzuhalten, für die Patienten nützliche KI-Systeme zu verwenden. Darüber hinaus gilt es Maßnahmen zu treffen, damit KI-Nutzer nicht in gerechtigkeitsethisch unvertretbarer Weise unzumutbaren Belastungen ausgesetzt werden (vgl. Kapitel 6.2.7).

Die Nutzung von AI-CDSS bringt der Ärzteschaft aber wahrscheinlich nicht nur Belastungen. Wie in Kapitel 4.11 dargelegt wurde, kann der Gebrauch der Systeme die Effizienz der ärztlichen Entscheidungsfindung in zeitlicher Hinsicht verbessern. Die somit möglicherweise frei werdenden zeitlichen Ressourcen können die Ärzte dazu nutzen, sich künftig vermehrt der tendenziell zeitintensiven genuin humanen Dimension der Behandlung zu widmen (vgl. Kapitel 4.9.2). Der US-Amerikaner Topol (2019) sieht die Gefahr, dass die gesteigerte Effizienz auch dazu missbraucht werden könnte, die Medizin nur im ökonomischen Sinne zu optimieren, indem etwa ärztliche Arbeitsplätze abgebaut werden. In Deutschland und in anderen Ländern, in denen Ärztemangel herrscht, dürfte es durchaus eine Chance darstellen, wenn der Einsatz von AI-CDSS den Bedarf an Ärzten senkt. Gleichzeitig ist Topol aber in dem Sinne zuzustimmen, dass die KI-bedingte Effizienzsteigerung innerhalb der Medizin nicht nur wenigen innerhalb der Gesundheitsbranche zugutekommen darf. Stattdessen ist zu fordern: Die Ressourcen, die durch die Verwendung von AI-CDSS frei werden, sind patientenorientiert zu nutzen (vgl. Kapitel 6.3.14).

Dass dadurch die Qualität der Patientenversorgung beeinflusst werden kann, ist aber nicht der einzige Grund dafür, dass der potenzielle Abbau ärztlicher Arbeitsplätze ethisch relevant ist. Dieser betrifft auch die Lebensgrundlage der betreffenden Ärzte. Da sich die derzeit verfügbaren AI-CDSS nur auf bestimmte, abgegrenzte Aufgabenbereiche beziehen und nicht auf den gesamten klinischen ärztlichen Entscheidungsprozess (vgl. Kapitel 4.1.1), ist eine umfassende Ersetzung von Ärzten durch KI-Systeme – jedenfalls bisher – nicht möglich. Es ist aber anzunehmen, dass die Anwendungen in einigen Bereichen der klinischen Entscheidungsfindung wie etwa der Befundung von Bildaufnahmen in der Zukunft eine derart relevante Rolle spielen werden, dass sich die

Aufgaben des Arztes hier stark verändern werden. In diesem Zuge ist es möglich, dass der Bedarf an ärztlichem Personal in manchen Bereichen der Medizin sinken wird. Soweit das möglich ist, erscheint es sinnvoll, angehende Ärzte über für sie relevante, absehbare Transformationen zu informieren, sodass sie ihre Berufsplanung an den zukünftigen Bedarf in den einzelnen Bereichen der Medizin anpassen können (vgl. Kapitel 6.2.8).

Neben den gesundheitlichen Nutzenpotenzialen sollten auch die ökonomischen Potenziale von AI-CDSS gerecht verteilt werden. Der Markt von entscheidungsunterstützenden medizinischen Systemen weist schließlich ein großes Potenzial auf: Dieser soll der Prognose von BIS Research (2021) zufolge bis 2030 weltweit auf einen Wert von 3,739 Milliarden US-Dollar anwachsen. Angesichts der großen wirtschaftlichen Bedeutung von KI-Systemen ist es problematisch, dass deren Entwicklung mit einem erheblichen Aufwand an Ressourcen verbunden ist. Beispielsweise die Anschaffung von kostspieligen Computern mit ausreichender Rechenkapazität und die Beschäftigung von hinreichend ausgebildeten Entwicklern stellen ökonomische Barrieren für den Zugang zur Entwicklung von AI-CDSS dar. Somit besteht die Gefahr, dass die Partizipation an deren ökonomischen Potenzialen ungerecht verteilt wird (Rohde et al., 2021). Es wäre ethisch nicht vertretbar, wenn der Markt von entscheidungsunterstützenden KI-Systemen unter wenigen Firmen aufgeteilt würde, die über die entsprechenden Ressourcen verfügen. Es ist daher erforderlich, mit passenden Maßnahmen einen gerechten Zugang zu den ökonomischen Potenzialen der Systeme sicherzustellen (vgl. Kapitel 6.2.6).

Gerechtigkeit zeigt sich auch darin, dass sich die Behandlung eines Subjekts nach dessen individuellem Bedarf richtet. Eine Gruppe von Patienten, die bisher aufgrund einer tendenziell schlechteren gesundheitlichen Versorgung einen besonders großen Bedarf an Verbesserung der Versorgungsqualität aufweist, sind solche, die an seltenen Krankheiten leiden (Deutscher Ethikrat, 2018b). Vor diesem Hintergrund bergen AI-CDSS eine gerechtigkeitsethisch relevante Chance in sich: Die den Menschen übertreffenden Fähigkeiten von KI-Systemen zur Mustererkennung und Datenverarbeitung ermöglichen es Krankheiten zu detektieren, die im klinischen Alltag selten vorkommen und daher bei der Diagnostik oft nicht in Erwägung gezogen werden (Sachverständigenrat Gesundheit & Pflege, 2021; ZEKO, 2021).

Die Qualität der medizinischen Versorgung weist nicht nur zwischen Erkrankungen, sondern auch in verschiedenen Ländern und Weltregionen Unterschiede auf (Tosam et al., 2018). An dieser Stelle bieten AI-CDSS ein Potenzial: Der Einsatz von hochwertigen Systemen kann die medizinische Versorgungsqualität besonders dort verbessern, wo ein vergleichsweise großes Entwicklungspotenzial besteht (Panesar et al., 2019). Ein Beispiel in diesem Zusammenhang wurde in Kapitel 3.2.3 gegeben. Revell et al. (2018) entwickelten ein KI-System, das Ärzte bei der Auswahl von Medikamenten zur Behandlung von HIV unterstützt – ohne, dass dafür eine kostspielige genotypische Resistenztestung nötig wäre. Somit kann die Anwendung besonders in ressourcenarmen Regionen, in denen die Mittel für eine genotypische Resistenztestung nicht vorhanden sind, die Qualität der HIV-Behandlung verbessern.

Sofern es in unterprivilegierten Gegenden der Welt möglicherweise an kompetenten Nutzern von AI-CDSS mangelt, könnten auch dazu befähigte Ärzte an anderen Orten die Systeme synchron im Sinne der Telemedizin steuern. So könnte die Qualität der Patientenversorgung besonders in bisher unterprivilegierten Weltregionen verbessert werden.

In die Gerechtigkeitsdebatte ist in letzter Zeit vermehrt der Begriff der Umweltgerechtigkeit eingegangen (Bolte et al., 2012; Grafe, 2020). Auch die Entwicklung und Verwendung von AI-CDSS ist hinsichtlich deren Auswirkung auf die Umwelt zu analysieren. Besonders energieintensiv ist der Entwicklungsprozess von ML-Systemen: Eine einzige Trainingseinheit des Deep-Learning-Modells GPT-3, das zur Produktion von menschlicher Sprache entwickelt wurde, benötigte die Menge an Energie, die 126 dänische Haushalte in einem Jahr verbrauchen und führte zu einem CO2-Ausstoß, den Autofahrten einer Strecke von 700 000 km bedingen (DeWeerdt, 2020). An diesen Zahlen wird sichtbar, dass sich die Entwicklung von ML-Systemen in beträchtlichem Maß auf die Umwelt auswirkt und daher der umweltfreundlichen Entwicklung von AI-CDSS große Bedeutung zukommt (vgl. Kapitel 6.1.6). Auf Umweltgerechtigkeit ist außerdem bei der Auswahl und Nutzung eines AI-CDSS zu achten (vgl. Kapitel 6.3.2 und 6.3.6).

5. Übergreifende ethische Bewertung von AI-CDSS

Nachdem im vorangegangenen Kapitel einzelne Bewertungen anhand von verschiedenen Beurteilungskriterien vorgenommen wurden, sind diese zusammenzuführen und zu gewichten. Vor den Gefahren und Risiken von AI-CDSS sollen zuerst relevante Gründe für deren Einsatz zur Sprache kommen. Das bietet sich auch deshalb an, weil die Vertretbarkeit von Risiken davon abhängt, welche Nutzenpotenziale diesen gegenüberstehen.

Zentral für die ethische Legitimation der Systeme ist die Nützlichkeit für die Patienten: Der Gebrauch der KI-Anwendungen ist prima facie geboten, wenn man dadurch positive Auswirkungen auf die Überlebenswahrscheinlichkeit und die Lebensqualität der Behandelten erzielt. Nachweisen kann man diese Effekte nur anhand von empirischen Studien. Insofern entsprechende Forschungsergebnisse bisher nicht in hinreichender Menge und Qualität vorliegen, ist es schwierig, an dieser Stelle ein einheitliches übergreifendes Urteil über den Nutzen von AI-CDSS für die Patienten zu treffen (vgl. Kapitel 4.3). Tendenziell ist jedoch von einem großen Potenzial zur Verbesserung der medizinischen Versorgungsqualität auszugehen.

Ein Hinweis auf die Nützlichkeit der Systeme lässt sich im Hinblick auf deren oft hohe Wirksamkeit gewinnen: Da einige Anwendungen bei eng abgegrenzten Aufgaben der klinischen Entscheidungsfindung besser als Ärzte abschneiden, spricht das zunächst dafür, dass deren Einsatz die klinische Entscheidungsfindung optimiert und somit auch den Patienten nützt (vgl. Kapitel 4.1.2).

Die Untersuchung der Brauchbarkeit der Systeme kann diese Vermutung teilweise stützen: Evaluationsstudien zufolge kann der Einsatz eines AI-CDSS die Qualität klinischer Entscheidungen teils mehr, teils aber auch weniger stark verbessern, als man erwarten würde, wenn man die Leistungsfähigkeit des Systems und der Ärzte miteinander vergleicht (vgl. Kapitel 4.1.3). Für die Nützlichkeit der Anwendungen spricht auch, dass sie die Potenziale von Big Data heben können: Sie können diagnostische, prognostische und therapeutische Entscheidungen hochpräzise an individuelle Merkmale des Patienten anpassen. Da sie die Chance bieten, Lücken im ärztlichen Wissen auszugleichen und genuin humane Denkfehler aufzudecken, kann deren Nutzung außerdem die Zahl von Fehlern bei der Diagnostik, der Prognostik und der Therapieempfehlung minimieren (vgl. Kapitel 4.3).

Das Ziel der AI-CDSS, die Qualität der klinischen Entscheidungsfindung zu verbessern, kann man versuchen auf anderen Wegen zu erreichen. Eine Alternative, die allen KI-Systemen überlegen wäre, gibt es aber nicht. So kann etwa die Durchführung von edukativen Maßnahmen im Vergleich mit der Nutzung eines AI-CDSS in manchen Belangen überlegen, in anderer Hinsicht aber gleichzeitig unterlegen sein (vgl. Kapitel 4.2). Daher ist es wichtig, jedes KI-System vor dem Einsatz mit Alternativen in Bezug auf alle relevanten Aspekte zu vergleichen.

Über das Nutzenpotenzial hinaus bergen AI-CDSS einige weitere Chancen. Die Systeme können etwa die Gesundheitsmündigkeit der Patienten fördern, indem sie Outputs an diese ausgeben und erklären (vgl. Kapitel 4.5.2).

Auch auf die Qualität der Arzt-Patient-Beziehung kann sich die Verwendung der Systeme positiv auswirken: Wenn diese die ‚*hard facts*' der klinischen Entscheidungsfindung bearbeiten und zeitliche Ressourcen des Arztes freisetzen, kann dieser sich – so die Hoffnung – stärker auf den individuellen Patienten und dessen Bedürfnisse konzentrieren, woraus ein Mehr an Empathie und ein verstärkter Fokus auf die psychische und die soziale Dimension der Behandlung resultieren kann (vgl. Kapitel 4.9.2).

Chancen bestehen auch in einer Effizienzsteigerung: Die Verwendung von AI-CDSS kannn zeitliche Einsparungen bei der Behandlung ermöglichen. Außerdem dürfte die KI-Nutzung das Gesundheitssystem finanziell entlasten (vgl. Kapitel 4.11).

Darüber hinaus könnte sich die Verwendung von KI-Systemen in gerechtigkeitsethischer Hinsicht positiv auswirken und etwa dazu führen, dass Ungleichheiten in der Patientenversorgung – national wie international – ausgeglichen werden (vgl. Kapitel 4.12).

Diesen Chancen stehen Risiken gegenüber. Hinsichtlich der Patientenautonomie hat sich gezeigt, dass diese durch die Nutzung von AI-CDSS geschwächt werden kann. In manchen Fällen ist es strittig, ob der Patient über ein KI-System aufgeklärt werden muss und dessen Nutzung zustimmen muss. Besteht diese Pflicht zur informierten Einwilligung und wird diese nicht beachtet, so wird die Freiheit des Patienten beschränkt (vgl. Kapitel 4.5.1). Auch die Gesundheitsmündigkeit des Patienten kann durch die Verwendung von AI-CDSS unterminiert werden (vgl. Kapitel 4.5.2).

Während diese Risiken bei allen entscheidungsunterstützenden KI-Systemen vorhanden sind, sind mit Anwendungen im Bereich der Therapieempfehlung besondere Problemstellungen bezüglich der Patientenautonomie verbunden. In der Regel sind die Systemergebnisse nicht an individuelle Patientenpräferenzen angepasst. Unterlässt es der Arzt, die Therapieentscheidung am Patienten und seinen Werten auszurichten, kann die KI-Nutzung im therapeutischen Bereich daher Paternalismus befördern. Da die Patientenautonomie in der Medizin einen besonderen Stellenwert einnimmt, ist deren Einschränkung durchaus relevant. Die beschriebenen Risiken kann man durch eine entsprechende Entwicklungs- und Anwendungsweise der Systeme aber so weit vermeiden, dass sich aus der potenziellen Gefährdung der Patientenautonomie kein kategorisches Argument gegen die Nutzung von AI-CDSS ableiten lässt.

Im Gegensatz dazu ist die mögliche Beschränkung ärztlicher Freiheit für sich genommen aus ethischer Perspektive weniger relevant (vgl. Kapitel 4.6). Über der Wahrung der ärztlichen Entscheidungsautonomie steht nämlich die Ausrichtung der Medizin am Wohl des Patienten. Daher ist es sogar geboten, die ärztliche Entscheidungsautonomie einzuschränken, sofern man dadurch das Patientenwohl steigert.

Falls die Nutzung von AI-CDSS (etwa durch *deskilling* und *alert fatigue*) die ärztliche Entscheidungskompetenz beschneidet, können daraus kritisch zu bewertende Konsequenzen folgen (vgl. Kapitel 4.7). Weil die Systeme die ärztliche Entscheidungsfindung nur unterstützen, muss der Arzt die Outputs grundsätzlich wenigstens auf Plausibilität hin kontrollieren. Wenn dessen Entscheidungskompetenz so stark beeinträchtigt wird, dass er die Ergebnisse eines AI-CDSS nicht mehr hinreichend kontrollieren kann, ist dessen Verwendung aus ethischer Perspektive nicht zu rechtfertigen. Dass die Nutzung von KI-Anwendungen ärztliche Fähigkeiten einschränkt, ist zwar kaum gänzlich auszuschließen. Indem man entsprechende Maßnahmen ergreift, dürfte man aber verhindern können, dass die ärztliche Entscheidungskompetenz in einer Weise beeinträchtigt wird, die die Verwendung der Systeme ethisch unvertretbar erscheinen ließe.

Als eine ebenso ethisch relevante wie schwierig zu lösende Problemstellung hat sich – vor allem aufgrund der Opazität – die Verantwortungszuschreibung bei der Nutzung von AI-CDSS erwiesen (vgl. Kapitel 4.8). Sollte sich entgegen den Erwartungen in Zukunft zeigen, dass die Nutzung entscheidungsunterstützender KI-Systeme zu

unlösbarer Verantwortungsdiffusion führt, könnte hierin vor allem ein Argument gegen die Verwendung bestimmter Hochrisiko-Anwendungen liegen.

Außerdem ist neben einer Steigerung auch eine Abnahme des Vertrauens des Patienten in den Arzt möglich, wenn dieser ein AI-CDSS verwendet (vgl. Kapitel 4.9.1). Die Arzt-Patient-Beziehung könnte sich auch dann verschlechtern, wenn der Reduktionismus der KI-Systeme den Behandelnden beeinflusst und dieser im Kontakt mit dem Patienten das Soziale, das Psychische und die Empathie vernachlässigt (vgl. Kapitel 4.9.2).

Dieses Risiko kann zusammen mit einer anderen Gefahr von AI-CDSS bewertet werden: Die potenzielle Verschlechterung der Arzt-Patient-Beziehung und der drohende Datenmissbrauch (vgl. Kapitel 4.10) weisen die Gemeinsamkeit auf, dass sie nicht nur bei der Nutzung von KI-Anwendungen, sondern bei der Verwendung grundsätzlich aller digitaler Technologien in der Medizin auftreten können. Da auch diese anderen Systeme trotz dieser Risiken angewandt werden, wäre es nicht angemessen, die Nutzung der KI-Anwendungen aufgrund dieser Bedenken gänzlich abzulehnen.

Auch jenseits der Behandlungssituation und außerhalb der Arzt-Patient-Beziehung weist die Verwendung von AI-CDSS Gefahren auf. Ob dadurch die klinische Entscheidungsfindung mehr oder weniger effizient gestaltet wird, ist zwar angesichts der bisher spärlich ausgeprägten Studienlage kaum im Allgemeinen zu beantworten (vgl. Kapitel 4.11). Die Gefahr von KI-bedingten Effizienzverlusten kann man aber durch eine rigorose Evaluation des Kosten-Nutzen-Verhältnisses einzelner Anwendungen minimieren.

Darüber hinaus könnte die Verwendung von AI-CDSS auch in gerechtigkeitsethischer Hinsicht negative Konsequenzen nach sich ziehen (vgl. Kapitel 4.12). So besteht neben Diskriminierungsrisiken die Gefahr, dass die KI-Nutzung bestehende Ungleichheiten in der Patientenversorgung vertiefen könnte. Wenn diese Risiken durch eine entsprechende Adressierung entschärft werden können, gibt es aber auch aus gerechtigkeitsethischer Perspektive kein kategorisches Argument gegen die Verwendung von AI-CDSS.

Das schwerwiegendste Risiko bei der Nutzung der Systeme liegt in der potenziellen gesundheitlichen Schädigung der Patienten durch falsche KI-basierte Entscheidungen (vgl. Kapitel 4.4). Schließlich steht in der Medizin der Patient im Mittelpunkt. Nüchtern ist allerdings festzustellen, dass ein nicht eliminierbares Fehlerpotenzial nicht bloß eine Eigenschaft von AI-CDSS ist, sondern grundsätzlich ein Charakteristikum von

Technologie darstellt. Es gilt daher, das Schadenspotenzial der Anwendungen so weit wie möglich zu minimieren. Wenn das hinreichend erfolgreich geschieht, ist die Nutzung der Systeme auch mit Blick auf die mögliche Schädigung der Patienten aus ethischer Sicht vertretbar.

Da nun zum einen die Chancen und zum anderen die Risiken von AI-CDSS dargelegt und gegenübergestellt wurden, stellt sich die Frage nach der Synthese: Wie sind die Anwendungen aus ethischer Perspektive insgesamt zu bewerten?

Angesichts der Heterogenität der KI-Systeme bezüglich der Anwendungsgebiete, Zielsetzungen und der verwendeten Technologien ist deren allgemeine Bewertung nur bedingt möglich. Rekurrierend auf die Einteilung von AI-CDSS nach den drei Bereichen Diagnostik, Prognostik und Therapieempfehlung lässt sich tendenziell sagen, dass Systeme im letztgenannten Bereich ethisch besonders implikationsreich sind. Welche Implikationen eine Anwendung aufweist, hängt aber von deren individuellen Spezifika ab. Letztlich ist darum jedes System einzeln ethisch zu evaluieren.

Außerdem sind die Grenzen der Ethik zu bedenken: Sie mag zwar relevante ethische Aspekte von AI-CDSS herausstellen und Hinweise zu Abwägungen geben. Wie mit den Systemen umzugehen ist, kann aber nicht alleine innerhalb dieser Disziplin entschieden werden (Marckmann, 2003). Da es sich um eine Technologie handelt, die im ethisch sensiblen Feld der klinischen Entscheidungsfindung vielerlei Veränderungen mit sich bringt, sollte im Rahmen eines gesamtgesellschaftlichen Diskurses über den Umgang mit AI-CDSS deliberiert und entschieden werden.

An verschiedenen Stellen der vorliegenden Arbeit wurde deutlich, dass die ethischen Implikationen der Systeme wesentlich davon abhängen, wie man diese entwickelt und verwendet. So kann die Nutzung von inhärent reduktionistischen KI-Anwendungen Ärzte so beeinflussen, dass diese die soziale und psychische Dimension der Behandlung sowie die Empathie vernachlässigen. Wenn die Nutzer aber für dieses Risiko sensibilisiert sind und ihm entgegenwirken, kann eine Chance realisiert werden: Unterstützen KI-Systeme die Ärzte, können diese sich auf die genuin humanen Aspekte der Behandlung konzentrieren und somit die Qualität der Arzt-Patient-Beziehung verbessern.

Ein weiteres Beispiel für die Ambivalenz der Anwendungen sind deren Auswirkungen auf die Sensitivität der Therapieentscheidung für evaluative Präferenzen der Patienten.

Fasst ein Arzt eine – nicht an die Werte des Patienten angepasste – Systemempfehlung als schlechthin richtige Lösung auf, so unterminiert er die Autonomie des Patienten, wenn er die Therapieentscheidung nicht an dessen Präferenzen ausrichtet. Weist das System aber auf die Notwendigkeit einer solchen Anpassung hin und gibt es zudem gegebenenfalls Erklärungen für die einzelnen Empfehlungen ab, so kann die Nutzung einer solchen Anwendung die Flexibilität der Therapieentscheidung für evaluative Patientenpräferenzen steigern.

Da die ethische Bewertung von AI-CDSS also maßgeblich von deren Entwicklungs- und Nutzungsweise abhängt, werden im folgenden Kapitel ethisch begründete Empfehlungen zum Umgang mit den Systemen erarbeitet. Diese Empfehlungen verfolgen das Ziel, Potenziale und Chancen der Anwendungen zu maximieren und deren Risiken und Gefahren zu minimieren.

6. Empfehlungen

Die folgenden ethisch begründeten Empfehlungen beziehen sich auf die Entwicklung (vgl. Kapitel 6.1) und die Anwendung von AI-CDSS (vgl. Kapitel 6.3) sowie auf die Schaffung geeigneter Rahmenbedingungen für die Entwicklung und Nutzung der Systeme (vgl. Kapitel 6.2).

An dieser Stelle sei noch auf das Monitoring hingewiesen, das in der verwendeten Methodik als letzter Schritt vorgesehen ist. Dieses kann in der vorliegenden Arbeit noch nicht vorgenommen werden. Es soll hier aber dessen Relevanz betont werden: Angesichts des raschen technologischen Fortschritts gilt es, die zukünftigen Entwicklungen im Bereich von AI-CDSS zu verfolgen und den Umgang mit den Systemen entsprechend anzupassen.

6.1 Empfehlungen für die Entwicklung von AI-CDSS

6.1.1 Zusammensetzung des Entwicklungsteams

Bereits die Auswahl der Mitglieder des Entwicklungsteams ist ethisch relevant und sollte dementsprechend verantwortungsvoll geschehen. So sollte die fachliche Kompetenz der Programmierer den Anforderungen entsprechen, die mit der Entwicklung des jeweiligen Systems einhergehen. Besonders bei der Programmierung von Systemen, die ein hohes Missbrauchspotenzial und schwerwiegende Risiken aufweisen, sollten die KI-Entwickler unbedingt ethische Standards beachten.

Hinsichtlich der Zusammensetzung des Entwicklungsteams gilt es außerdem zu bedenken, dass die ethischen Implikationen, die mit einem AI-CDSS verbunden sind, bei dessen Entwicklung nur dann in angemessener Weise Beachtung finden können, wenn man diese frühzeitig identifiziert (Open Roboethics Institute, 2020). Falls derartige Gefahren erst bei der Anwendung des Systems entdeckt werden, kann der Zeitpunkt bereits verstrichen sein, zu dem eine Verhinderung negativer Auswirkungen möglich gewesen wäre. Dem *embedded ethics approach* gemäß sollten darum auch Ethiker einen Teil des Entwicklungsteams darstellen (McLennan et al., 2020). Mithilfe eines ethisch geschulten Blicks können etwa Aspekte als ethisch relevant identifiziert werden, die den Entwicklern möglicherweise nicht auffallen würden. Zum anderen können Ethiker dabei

unterstützen, bestimmte ethische Implikationen in ihrer Relevanz einzuschätzen und angemessen zu adressieren.

Darüber hinaus ist es sinnvoll, Stakeholder wie Anwender und Patientenvertreter in den Entwicklungsprozess zu integrieren (Open Roboethics Institute, 2020). Dadurch kann man gewährleisten, dass deren Belange bei der Entwicklung des AI-CDSS hinreichend Beachtung finden. Somit kann man dessen Akzeptanz unter Ärzten und Patienten erhöhen.

6.1.2 Identifizierung eines Bedarfs an Entscheidungsunterstützung

Die Verwendung von AI-CDSS soll die Qualität der Patientenversorgung verbessern (vgl. Kapitel 4.3). Es ist daher von großer Bedeutung, dass Systeme für Anwendungsbereiche entwickelt werden, in denen Ärzte einen Bedarf an Unterstützung aufweisen (vgl. Kapitel 4.1.3). Um solche zu identifizieren, bietet es sich für die Entwickler an, direkt bei Ärzten Informationen über deren Unterstützungsbedarf einzuholen.

Wenn bei den Anwendern ein Bedarf an Unterstützung in einem bestimmten Bereich festgestellt wurde, gilt es primär zu prüfen, ob es möglicherweise bereits ein System gibt, das die Anforderungen der Anwender erfüllt. Erleichtert wird das durch Datenbanken, die medizinische KI-Anwendungen verzeichnen.[14]

Stellt man als Ergebnis einer solchen Recherche fest, dass es kein derartiges AI-CDSS gibt, gilt es zu überprüfen, ob bereits Versuche angestellt wurden, ein System für den avisierten Anwendungsbereich zu entwickeln. Insofern diese scheiterten, kann das ein Indiz dafür sein, dass es in dem betreffenden Anwendungsgebiet Herausforderungen gibt, die bei der Entwicklung eines neuen AI-CDSS zu bedenken sind.

6.1.3 Vergleich des designierten AI-CDSS mit Alternativen

Wurde ein Bedarf an ärztlicher Entscheidungsunterstützung identifiziert, so sollte man überprüfen, welche anderen Mittel es gibt, um diesen Bedarf zu decken (vgl. Kapitel 4.2). Dabei sollten herkömmliche Technologien und nicht-technische Lösungen wie edukative Maßnahmen bedacht werden. Die verschiedenen Alternativen sind miteinander zu

[14] Beispielhaft kann hierfür die Website AIMe Registry (https://aime-registry.org) genannt werden.

vergleichen. Sollte sich abzeichnen, dass andere Mittel wesentlich besser als das designierte AI-CDSS dazu geeignet sind, den Bedarf an Entscheidungsunterstützung zu decken, sollte man von der Entwicklung eines solchen Systems absehen.

6.1.4 Machbare, klare und evaluierbare Zieldefinition

AI-CDSS zu nutzen sollte kein Selbstzweck sein. Stattdessen hat deren Verwendung dazu zu dienen, klar definierte Ziele zu erreichen. Diese sollten verschiedenen Anforderungen genügen:

1. Das Ziel eines AI-CDSS sollte **machbar** sein: Es sollte den prinzipiellen Limitationen von KI-Anwendungen nicht zuwiderlaufen. Schwierig abzubilden sind in solchen Systemen etwa psychische, soziale und spirituelle Aspekte (Funer, 2021). Auch Flexibilität für Patientenpräferenzen ist schwierig in die Systemarchitektur von AI-CDSS zu integrieren (McDougall, 2019) (vgl. Kapitel 4.1.1).

2. Des Weiteren sollte das Ziel eines AI-CDSS möglichst **klar** und konkret definiert sein. Schließlich kann man die Wirksamkeit eines Systems nur dann evaluieren, wenn klar ist, worin dessen Ziel besteht. Hier ist zu bedenken, dass manche Ziele (wie etwa die Individualisierung der Therapieentscheidung) so beschaffen sind, dass man kaum oder gar nicht bestimmen kann, inwiefern diese durch die Nutzung eines Systems erreicht werden. Wenn es nicht möglich ist, die Wirksamkeit eines AI-CDSS zu evaluieren, kann es problematisch sein, stattdessen Parameter wie die Konkordanz zu bestimmen (vgl. Kapitel 4.1.2 und 4.1.3).

3. Daher sollte die Zieldefinition eines AI-CDSS drittens **evaluierbar** sein. Dass man die Wirksamkeit eines KI-Systems bestimmen kann, ist nämlich aus mehreren Gründen relevant. Sofern das System Ärzte hinsichtlich der Erreichung des Ziels übertrifft, liegt hierin ein Hinweis auf die Brauchbarkeit der Anwendung (vgl. Kapitel 4.1.3). Außerdem kann man von der Wirksamkeit eines Systems auf die Wahrscheinlichkeit dafür schließen, dass ein Systemergebnis falsch ist (vgl. Kapitel 4.4.1 und 6.1.18.4). Je besser ein Arzt wiederum fehlerhafte Outputs detektieren kann, desto weniger wahrscheinlich ist es, dass er diese in seine Entscheidungsfindung übernimmt (vgl. Kapitel 4.4.2.2).

6.1.5 Auswahl eines geeigneten Trainingsdatensatzes

Bei der Auswahl eines Datensatzes zum Training des AI-CDSS sind verschiedene Gesichtspunkte zu beachten. Einige relevante Aspekte seien hier – primär im Anschluss an acatech (2020) – angeführt:

1. Der Datensatz sollte hinreichend groß sein. Deep Learning erfordert beispielsweise besonders große Datensätze (Müller-Quade et al., 2020; Vandewinckele et al., 2020).

2. Die Richtigkeit der Grundwahrheit (*ground truth*) ist sicherzustellen. Idealerweise sollten mehrere Experten zu demselben Ergebnis beziehungsweise zu derselben Klassifizierung kommen (vgl. Kapitel 4.4.1).

3. Die Charakteristika des Trainingsdatensatzes (z.B. Eigenschaften der Patienten, Umgebung der Datenerhebung) sind zu explizieren. Besteht ein zu großer Unterschied zwischen den Trainingsdaten auf der einen Seite und den Inputdaten in der Praxis auf der anderen Seite, kann hierin ein Hinweis auf Verzerrungen des Trainingsdatensatzes liegen (vgl. Kapitel 4.4.1).

4. Trainings-, Test- und Validierungsdatensatz sind zu trennen. Dadurch erleichtert man die externe Validierung (vgl. Kapitel 6.1.18.1).

Durch eine möglichst diverse Datenbasis kann man außerdem die Generalisierbarkeit des Systems erhöhen (vgl. Kapitel 4.4.1). Universelle Generalisierbarkeit ist zwar aus ethischer Perspektive wünschenswert, aber in der Praxis schwierig zu erreichen (Beede et al., 2020; Mitchell et al., 2021). Falls man dadurch die Performanz nicht unverhältnismäßig stark einschränkt, sollten AI-CDSS so universalisierbar wie möglich gestaltet werden und in möglichst vielen Populationen angewandt werden können. Darüber hinaus sollten die Systeme lokal validiert und rekalibriert werden (vgl. Kapitel 6.1.7).

6.1.6 Umweltfreundliche Entwicklung

Der Entwicklungsprozess von AI-CDSS kann energieintensiv sein (vgl. Kapitel 4.12). Um nachteilige Auswirkungen auf die Umwelt nach Möglichkeit zu minimieren, sollte man dabei auf Energieeffizienz achten (Rohde et al., 2021). Den Energiebedarf kann man

zum Beispiel reduzieren, indem man keine gänzlich neuen Systeme entwickelt, sondern bereits vorhandene KI-Modelle anpasst (Strubell et al., 2019).

6.1.7 Lokale Validierung und Rekalibrierung

Da Funktionsfähigkeit, Nutzen und Schaden von AI-CDSS stark von der Anwendungsumgebung und der Population abhängen, in der sie genutzt werden (vgl. Kapitel 3.4.2), sollten die Systeme lokal validiert werden (Mitchell et al., 2021) (vgl. Kapitel 4.12). Stellt sich heraus, dass eine Anwendung bei der Nutzung in einer bestimmten Gruppe oder in einer bestimmten Umgebung ein besonders hohes Fehlerpotenzial aufweist, sollte man das System rekalibrieren. Weil die meisten der AI-CDSS in vergleichsweise wohlhabenden Ländern entwickelt werden, kommt der lokalen Validierung und Rekalibrierung auch aus gerechtigkeitsethischer Perspektive große Relevanz zu (vgl. Kapitel 4.12). Die Datenbasis für die Rekalibrierung sollte ausreichend groß und so aktuell wie möglich sein (Mitchell et al., 2021).

6.1.8 Gewährleistung von Flexibilität für Patientenpräferenzen

Eine besondere ethische Herausforderung besteht bei KI-Systemen, die Therapieempfehlungen abgeben. Der Respekt der Patientenautonomie gebietet es, die Werturteile im Rahmen der Therapieentscheidung den Patientenpräferenzen entsprechend zu treffen (vgl. Kapitel 4.5.3). Die Therapieempfehlung des Systems an die Patientenpräferenzen anzupassen, ist zwar Aufgabe der Ärzte, doch die Entwickler sollten diese dazu befähigen.

Hier sind besonders zwei Maßnahmen erforderlich. Damit ein Arzt den Patientenpräferenzen hinreichend Beachtung schenken kann, muss er sich zunächst deren Relevanz bewusst sein (McDougall, 2019). Um einer potenziellen ‚Individualisierungsvergessenheit' vorzubeugen, wäre es zu begrüßen, wenn die Systeme die Ärzte darauf hinwiesen, dass – und hinsichtlich welcher Werturteile – die Therapieempfehlungen des Systems an den Patientenpräferenzen auszurichten sind (Kapitel 4.5.3). Um zu unterstreichen, dass die vom AI-CDSS empfohlene therapeutische Maßnahme individuell an den Patienten angepasst werden muss und nicht schlechthin ‚die richtige Therapie' ist, kann es besonders bei wertsensitiven Entscheidungen sinnvoll sein, dass KI-Systeme nicht nur eine, sondern mehrere Therapieempfehlungen abgeben (Rajput et al., 2020) (vgl. Kapitel 6.1.12).

Ist sich ein Arzt darüber im Klaren, dass eine Therapieempfehlung eines AI-CDSS am Patienten und seinen Präferenzen zu orientieren ist, ist es wichtig, dass er zusammen mit dem Patienten die jeweiligen Empfehlungen auch anpassen *kann*. Dem ist zuträglich, wenn der Arzt die Werturteile nachvollziehen kann, die der Systemempfehlung zugrunde liegen. Nicht immer kann der Behandelnde selbständig auf diese schließen. Es ist daher zu empfehlen, dass das System die Werturteile und Ziele transparent darstellt, auf denen eine Therapieempfehlung basiert (Andrews et al., 2013). Herkömmliche Erklärungen, die nur das Zustandekommen des Ergebnisses explizieren, können auch ohne spezielle Bezugnahme auf Werturteile dennoch auf diese hinweisen und somit dazu beitragen, dass die Therapieempfehlungen von AI-CDSS flexibel an Patientenpräferenzen angepasst werden können (Nyrup et al., 2019) (vgl. Kapitel 4.5.3).

6.1.9 Integrierbarkeit in den Workflow

Es ist bekannt, dass eine fehlende Einbindung von entscheidungsunterstützenden Anwendungen in den Workflow sowohl deren Verwendung als auch deren Nützlichkeit negativ beeinflussen kann (Kawamoto et al., 2005). Aus diesem Grund sollten Entwickler bereits bei der Programmierung darauf achten, dass sich das AI-CDSS in den Workflow der Nutzer integrieren lässt. Das System sollte sich möglichst nahtlos in die zukünftige Anwendungsumgebung einfügen lassen. In diesem Zusammenhang ist auf die große Bedeutung von Interoperabilität und von einheitlichen Datenstandards hinzuweisen (Kelly et al., 2019) (vgl. Kapitel 6.2.4).

6.1.10 Benutzerfreundlichkeit

AI-CDSS sind keine autonomen Akteure, sondern Instrumente der Ärzte. Die Funktionsfähigkeit der Systeme ist eng an deren Brauchbarkeit und damit an die Bedürfnisse der Nutzer gekoppelt (vgl. Kapitel 4.1.3 und 6.1.2). In diesem Zusammenhang ist darauf hinzuweisen, dass sich die Nutzerbedürfnisse interindividuell stark unterscheiden können. Eine individuelle Anpassung der Anwendungen an jeden einzelnen Nutzer ist gewiss kaum darstellbar. Grundsätzlich ist aber die Maxime zu verfolgen, die Systeme so individuell anpassungsfähig wie möglich zu gestalten. Einige mit der Benutzerfreundlichkeit verbundene Aspekte seien hier genannt:

1. **Ähnlichkeit der Bedienung im Vergleich zu anderen und älteren Systemen**: Sofern es keine relevanten Gründe dafür gibt, die Bedienung eines neuen

AI-CDSS im Vergleich zu älteren und anderen Modellen zu ändern, sollte die Bedienungsweise eines neuen Systems älteren oder anderen Anwendungen möglichst ähneln, mit denen die Zielgruppe vertraut ist. Auf diese Weise können Nutzer die Erfahrung und die Fähigkeiten, die sie in Bezug auf andere Entscheidungsunterstützungssysteme entwickelt haben, für die Nutzung des betreffenden AI-CDSS fruchtbar machen (Sutton et al., 2020).

2. **Beste Information zum richtigen Zeitpunkt**:

    a. **Idealer Zeitpunkt**: Die vom System gegebene Empfehlung oder Information sollte dem Arzt gegeben werden, wenn er in seiner Entscheidungsfindung am meisten davon profitiert.

    b. **Beste Information**: Die Information sollte für den Arzt neuartig sein und die gerade anstehende Entscheidung möglichst effektiv unterstützen.

    c. **Verfügbarkeit**: Das Output soll mit möglichst geringem Aufwand abrufbar sein (Wasylewicz & Scheepers-Hoeks, 2019).

6.1.11 Verhinderung von *alert fatigue*

Das in Kapitel 4.7 behandelte Phänomen *alert fatigue* stellt ein Risiko dar: Übergehen Ärzte relevante Hinweise von AI-CDSS, besteht die Gefahr, dass sie bestimmte Maßnahmen nicht ergreifen, die der Erhaltung oder der Förderung des Patientenwohls dienen (van der Sijs et al., 2006). Außerdem kann *alert fatigue* Burnouts bedingen (Jankovic & Chen, 2020). Insofern dieses Phänomen die Entscheidungskompetenz der Nutzer unterminiert, steigert es das Risiko, dass falsche Systemergebnisse ausgegeben werden. Dadurch wird das Schadenspotenzial von AI-CDSS für die Patienten verschärft. Das Prinzip des Nichtschadens gebietet darum, das Risiko von *alert fatigue* möglichst zu minimieren. Dazu sind unter anderem die in Tabelle 3 dargestellten Aspekte relevant.

Tabelle 3:

Nützliche und schädliche Faktoren in Bezug auf alert fatigue

Nützlich	Schädlich
Hohe Sensitivität und Spezifität sowie hoher positiver prädiktiver Wert der Hinweise	Unterbrechende Warnungen oder *hard stop alerts*
Anpassung der Rolle von Outputs	Fehlen der ‚richtigen Information'
Abstufung von Hinweisen	Fehlende Relevanz der Hinweise
Iteratives Design	Fehlende Integration der Outputs

Anmerkung. Übernommen von „Clinical decision support and implications for the clinician burnout crisis" von I. Jankovic und J. H. Chen, 2020. *Yearbook of Medical Informatics*, *29*(1), S. 151 (https://doi.org/10.1055/s-0040-1701986).

6.1.12 Ausgabe mehrerer Ergebnisse mit jeweiligem Konfidenzgrad

Eine Gefahr bei der Verwendung von AI-CDSS besteht darin, dass Ärzte deren Grenzen übersehen und diese nicht zur *Unterstützung* verwenden, sondern die klinische Entscheidungsfindung an die Systeme abgeben (vgl. Kapitel 4.1.1). Da diese Anwendungen wie grundsätzlich jede Technologie ein nicht gänzlich eliminierbares Fehlerrisiko aufweisen, sind Outputs wenigstens auf Plausibilität hin zu überprüfen. Wenn man Systemergebnisse – entgegen dem Prinzip des Nichtschadens – nicht hinreichend kontrolliert, können fehlerhafte Ergebnisse übersehen werden (vgl. Kapitel 4.4.2.2).

Damit Ärzte die Outputs hinreichend überprüfen, ist das Risiko des Automation Bias zu minimieren. Das kann etwa geschehen, indem zu den einzelnen Systemempfehlungen der jeweilige Konfidenzgrad angezeigt wird (McGuirl & Sarter, 2006). Outputs mit einem geringeren Konfidenzgrad könnten beispielsweise nur an Ärzte mit einem Wissens- und Erfahrungslevel ausgegeben werden, das die Beurteilung dieser Ergebnisse erlaubt (vgl. Kapitel 4.4.2.2). Um die Erfassung des Konfidenzgrads auch unter den Limitationen des klinischen Alltags zu ermöglichen, sollte man intuitive Darstellungsformen wählen.

Was die Angabe von Ergebnissen angeht, so ist außerdem zu empfehlen, dass dem Nutzer nicht nur ein Output, sondern mehrere Ergebnisse angezeigt werden. Die Ausgabe nur eines Outputs kann schließlich das Missverständnis stützen, dass es sich dabei um das ‚schlechthin richtige' Ergebnis handele. Im Anschluss daran kann es geschehen, dass ein

Systemergebnis nicht oder nur unzureichend kontrolliert wird (vgl. Kapitel 4.4.2.2). Um das zu verhindern, sollten AI-CDSS mehrere Outputs angeben. Auf diese Weise kann der Arzt das vom System präferierte Ergebnis mit alternativen Ergebnissen und deren jeweiligem Konfidenzgrad vergleichen und somit gegebenenfalls die Qualität seiner klinischen Entscheidung verbessern.

Die Angabe mehrerer Ergebnisse gebietet bei Systemen, die therapeutische Maßnahmen empfehlen, neben dem Nichtschadens- und dem Wohltuns- auch das Autonomieprinzip: Wird nur eine Therapieempfehlung ausgegeben, steigt die Gefahr der ‚Individualisierungsvergessenheit' und des daraus folgenden Paternalismus (vgl. Kapitel 4.5.3). Damit der Arzt nicht übersieht, dass er die Therapieentscheidung an den Patienten, seine individuelle Situation und seine evaluativen Präferenzen anzupassen hat, sollten mehrere Therapieempfehlungen ausgegeben werden (Rajput et al., 2020).

6.1.13 Ausgabe der Outputs an die Patientinnen und Patienten

AI-CDSS sollen den Arzt unterstützen – daher werden deren Ergebnisse grundsätzlich dem Arzt angezeigt. Aus verschiedenen Gründen ist es aber aus ethischer Perspektive wünschenswert, dass auch der Patient auf die Ergebnisse zugreifen kann.

Erstens kann die Gesundheitsmündigkeit des Patienten und somit seine Autonomie gefördert werden, wenn diesem die Ergebnisse des AI-CDSS angezeigt werden (vgl. Kapitel 4.5.2).

Zweitens kann die Ausgabe der Outputs an den Patienten fördern, dass dieser die Vertrauenswürdigkeit des Systems besser einschätzen kann. Stellen sich bei ihm Fragen in Bezug auf die mit den Ergebnissen verbundenen Unsicherheiten, können diese zwischen Patient und Arzt besprochen werden. So kann das Vertrauen in den KI-unterstützten Arzt gestärkt und die Qualität der Arzt-Patient-Beziehung verbessert werden (vgl. Kapitel 4.9.1).

Indem die Systemergebnisse dem Patienten mit jeweiligem Konfidenzgrad angegeben werden, kann man außerdem das Risiko dafür minimieren, dass der Patient die Fähigkeiten des Systems und dessen Einfluss bei der Entscheidungsfindung überschätzt und im Zuge dessen die Rolle des Arztes unterschätzt. Eine solche Konfusion über die Rollenverteilung zu verhindern, ist gefordert, um die Integrität der Arzt-Patient-Beziehung zu gewährleisten (vgl. Kapitel 4.9.1).

6.1.14 Datenschutzmaßnahmen

Beim Umgang mit Trainingsdaten sollte man bedenken, dass das Schadenspotenzial für die Datenspender nicht durch ein individuelles Nutzenpotenzial aufgewogen werden kann. Daher kommt dem Schutz dieser Daten vor Missbrauch eine besondere ethische Relevanz zu (vgl. Kapitel 4.10).

Um die Privatheit derjenigen zu schützen, von denen die Trainingsdaten stammen, ist zu gewährleisten, dass man mit diesen keinen Personenbezug herstellen kann. Hierfür bieten sich Techniken der Pseudonymisierung, der Anonymisierung und der Kryptifizierung an (Müller-Quade et al., 2020). Großes Potenzial weist in diesem Bereich die voll-homomorphe Verschlüsselung auf (Kaissis et al., 2020). Durch den Einsatz dieser Technik muss man die Daten bei Berechnungen nicht entschlüsseln. Somit sind sie in besonderer Weise vor Missbrauch geschützt. Vielversprechend ist außerdem das föderale Lernen. Dieser dezentrale Ansatz vermeidet die zentrale Speicherung von Daten und kann somit das Risiko von unautorisierten Zugriffen senken (Kaissis et al., 2020).

Das Konzept der Datensouveränität erscheint in der Praxis problematisch (Strech et al., 2020) (vgl. Kapitel 4.10). Zwar gibt es die Idee, diejenigen, von denen die Daten stammen, etwa über *Personal Information Managements Systems* dazu zu befähigen, die Nutzung ihrer Daten zu verfolgen und der zukünftigen Verwendung ihrer Daten zu bestimmten Zwecken zuzustimmen oder abzulehnen (Weichert, 2021). Dieses Vorgehen ist allerdings aufwändig. Es ist fraglich, ob man auf diese Weise zu der Menge an Daten kommen kann, die für die Entwicklung ML-basierter Systeme nötig ist.

Praxistauglicher erscheint das Konzept der Datenspende: Man beschränkt den Aufwand der potenziellen Datenspender darauf, einmalig die Vor- und Nachteile der Datenspende abzuwägen und anschließend ihr Einverständnis dazu zu erklären – oder dieses zu verweigern (Strech et al., 2020).

6.1.15 Sicherung von Fairness

In Kapitel 4.12 wurde die große ethische Bedeutung von Fairness deutlich. Um diese zu gewährleisten, ist sicherzustellen, dass insbesondere Outputs, die die Verteilung von Ressourcen beeinflussen, auf sachlichen Gründen beruhen und nicht bestimmte Gruppen diskriminieren. Hierzu bieten sich etwa folgende Maßnahmen an (Arshad & Strodthoff, 2021):

1. *Counter-factual fairness*: Im Sinne von Rawls‘ Fairnessbegriff wird überprüft, ob ein bestimmtes Output gleich ausgefallen wäre, wenn die betreffende Person einer anderen Gruppe angehört hätte.

2. *Post-processing* (Nachbearbeitungstechnik): Eine Regel, die Nichtdiskriminierung implizieren soll, wird im Nachhinein auf ein System angewandt.

6.1.16 Gewährleistung von Erklärbarkeit

Die ethische Bewertung von AI-CDSS hat gezeigt, dass Erklärbarkeit keine notwendige Bedingung von deren ethischer Legitimität darstellt. Für die Zuschreibbarkeit von Verantwortung etwa hat sich diese nicht als unbedingt notwendig erwiesen (vgl. Kapitel 4.8). Aus verschiedenen Gründen ist Erklärbarkeit aber aus ethischer Sicht wünschenswert: Diese kann den Arzt dabei unterstützen, falsche Outputs von AI-CDSS zu detektieren (vgl. Kapitel 4.4.2.2), die Gesundheitsmündigkeit des Patienten zu fördern (vgl. Kapitel 4.5.2), die Flexibilität der KI-basierten Therapieentscheidung für Patientenpräferenzen zu stärken (vgl. Kapitel 4.5.3) und das Vertrauen des Patienten in den Arzt fördern (vgl. Kapitel 4.9.1). Ein AI-CDSS sollte darum so erklärbar wie möglich gestaltet werden.

Um einen möglichst hohen Grad an Erklärbarkeit zu gewährleisten, sollten stets alle der einschlägigen Methoden in Erwägung gezogen werden (vgl. Kapitel 3.4.1). Welche der verschiedenen Strategien anzuwenden ist, hängt unter anderem vom Systemtyp, dessen Anwendungsgebiet und von den Nutzern ab. Idealerweise sollten Erklärungen möglichst individuell angepasst werden. So sollten die Erklärungen etwa an das jeweilige Wissensniveau und die Fähigkeiten – medizinischer wie informatischer Art – der einzelnen Nutzer(-gruppen) angepasst werden.

Die Erklärbarkeit eines AI-CDSS ist schließlich solide zu evaluieren, um gegebenenfalls einen Mangel an Erklärbarkeit identifizieren und beheben zu können. Die Evaluation der Erklärbarkeit ermöglicht außerdem, dass die Nutzer über diese für sie relevante Eigenschaft des Systems informiert werden können. Erklärbarkeit kann unter anderem anhand der folgenden Parameter beurteilt werden:

„**Genauigkeit:** Vorhersagegenauigkeit für unbekannte Daten.

Wiedergabetreue gibt an, inwieweit eine post-hoc Erklärung das Verhalten des Black-Box-Modells widerspiegelt. Die Wiedergabetreue ist eine besonders erstrebenswerte Eigenschaft, da nur bei korrekt wiedergegebenem Modellverhalten durch Analyse der Erklärungen sinnvolle Rückschlüsse gezogen werden können.

Konsistenz gibt an, ob für ein definiertes Modell ähnliche Datenpunkte ähnliche Erklärungen erhalten.

Verständlichkeit und Praxistauglichkeit gibt an, wie gut Nutzende die Erklärung verstehen und für ihre individuelle Aufgabe nutzen können. Die Verständlichkeit ist eine der wichtigsten und zugleich am schwersten zu formalisierenden Erklärungseigenschaften.“ (Schaaf et al., 2021, S. 20)

6.1.17 Vermeidung falscher Systemergebnisse

Wie generell alle technischen Geräte, so können auch AI-CDSS zu falschen Ergebnissen kommen. Auch wenn eine vollständige Eliminierung fehlerhafter Outputs illusorisch ist, gebietet es das Prinzip des Nichtschadens deren Häufigkeit so niedrig wie möglich zu halten. Insbesondere diejenigen falschen Systemergebnisse sind zu vermeiden, die mit besonders großem Schadenspotenzial für die Patienten verbunden sind.

Bereits bei der Entwicklung und bei der Auswahl des Trainingsdatensatzes ist anzusetzen: So sollte etwa die Richtigkeit des Trainingsdatensatzes von mehreren Experten überprüft werden (*ground truth*) (vgl. Kapitel 4.4.1 und 6.1.5). Falsche Ergebnisse, die durch Verzerrungen bedingt sind, können mithilfe einer lokalen Validierung und Rekalibrierung vermieden werden (Mitchell et al., 2021) (vgl. Kapitel 6.1.7).

Da es realistischerweise nicht erreichbar ist, dass das System immer richtige Outputs ausgibt, sollte man versuchen zu identifizieren, warum, bei welchen Aufgaben und in welchen Situationen typischerweise Fehler passieren (vgl. Kapitel 4.4.1). Möglicherweise lassen sich bestimmte Szenarien oder typische Fälle ausmachen, in denen ein AI-CDSS besonders häufig fehlerhafte Outputs anzeigt. Wissen die Nutzer, worauf sie sich bei der Kontrolle der Systemergebnisse besonders konzentrieren sollen (vgl. Kapitel 6.3.12), dürfte dadurch die Zahl der detektierten falschen Systemergebnisse steigen.

Darüber hinaus ist es auch sinnvoll, spätere Anwender zu schulen. Das Fehler- und Schadenspotenzial könnte man außerdem reduzieren, indem man fachliche Mindestanforderungen an die Nutzer stellt. Insofern die Anwender zur Kontrolle der Systemergebnisse möglicherweise nicht hinreichend befähigt sind, kann es außerdem sinnvoll sein, dass die Systemergebnisse eigens kontrolliert werden, bevor die Nutzer diese abrufen können. Die KI-basierten Bildanalysen des Systems *HeartFlow* werden beispielsweise von Angestellten der Herstellerfirma kontrolliert, ehe die Anwender Zugriff auf das Ergebnis haben (HeartFlow, o. D.).

6.1.18 Rigorose Evaluation und Testung

AI-CDSS werden wie alle technischen Instrumente zur Erreichung bestimmter Ziele und zur Erfüllung gewisser Zwecke eingesetzt. Wie die Systeme dabei abschneiden, sollten die Ärzte nachvollziehen können. Der Evaluation und Testung der Anwendungen kommt darum große Bedeutung zu. Im Hinblick auf deren Heterogenität und auf die Komplexität des Evaluationsprozesses kann im Rahmen der vorliegenden Arbeit nicht im Einzelnen dargestellt werden, wie ein AI-CDSS zu evaluieren ist. Es sollen aber einige Empfehlungen folgen, die aus ethischer Sicht besonders relevant sind.

Zunächst ist darauf hinzuweisen, dass Evaluationsergebnisse von verschiedenen KI-Systemen mit demselben Anwendungsbereich nicht selten stark divergieren (vgl. Kapitel 4.1.2). Angesichts dieser Variabilität ist die gemeinsame Evaluierung mehrerer Systeme in Gruppen nur eingeschränkt sinnvoll. Es ist wichtig, jede Anwendung einzeln zu evaluieren. Aus ethischer Perspektive ist des Weiteren im Allgemeinen die Transparenz der Evaluation zentral. Falls beispielsweise weniger gute Evaluationsergebnisse nicht veröffentlicht werden, kann darunter die Vertrauenswürdigkeit von KI-Systemen leiden. Wie ein Evaluationsprozess vonstatten ging und welche Ergebnisse dabei erzielt wurden, ist deshalb möglichst transparent zu kommunizieren. Außerdem sollte ein AI-CDSS nicht nur an dem Ort geprüft werden, an dem es entwickelt wurde, sondern auch dort, wo es angewandt wird, lokal validiert werden (vgl. Kapitel 6.1.7).

6.1.18.1 Wirksamkeit

Dass Nutzer über die Wirksamkeit eines AI-CDSS informiert werden, ist aus mehreren Gründen ethisch relevant. Zum einen kann von dieser Größe auf die Brauchbarkeit des Systems geschlossen werden (vgl. Kapitel 4.1.3). Zum anderen kann der Arzt im Hinblick auf die Wirksamkeit abschätzen, wie häufig ein System zu einem richtigen und wie oft es zu einem falschen Ergebnis kommt (vgl. Kapitel 4.4.1). Wenn der Behandelnde die Fähigkeiten des AI-CDSS realistisch einschätzen kann, kann er es eher vermeiden, falsche Systemergebnisse zu übernehmen (vgl. Kapitel 4.4.2.2). Damit die Wirksamkeit der Anwendungen bestimmt werden kann, sollten diese ein klares und evaluierbares Ziel verfolgen (vgl. Kapitel 6.1.3).

Bei der Wahl der Parameter, die zur Evaluation der Wirksamkeit verwendet werden, ist Verschiedenes zu beachten. Erstens sollten die verwendeten Größen so gebräuchlich sein, dass man das AI-CDSS mit anderen Systemen vergleichen kann (vgl. Kapitel 6.3.2). Darüber hinaus sollten die Parameter für die Ärzte verständlich sein (vgl. Kapitel 4.1.2): Falls die Ärzte diese nicht kennen, sollten sie erklärt werden. Schließlich ist in Bezug auf die Auswahl der Evaluationsparameter wichtig, dass man anhand der Evaluationsergebnisse bestimmen kann, inwieweit das System einen *spezifischen* Bedarf an Entscheidungsunterstützung deckt. Wenn beispielsweise die Sensitivität der Diagnostik erhöht werden soll, ist es wichtig, dass die Wirksamkeit der zur Auswahl stehenden Anwendungen anhand dieser Größe angegeben wird – und nicht etwa nur anhand von *accuracy*-Werten (vgl. Kapitel 4.2).

Im Rahmen der Wirksamkeitsevaluation eines AI-CDSS sollte nicht nur eine interne, sondern auch eine externe Validierung stattfinden. Dadurch kann man nämlich unter anderem *overfitting* detektieren: Wenn die Wirksamkeit des Systems bei der Eingabe von Trainingsdaten deutlich höher ist als bei der Einspeisung von externen Datensätzen, besteht ein Hinweis auf *overfitting* (Collins et al., 2014). Falls die externe Validierung unterlassen wird, kann jedoch ein Niveau an Wirksamkeit suggeriert werden, das in der Praxis nicht erreicht wird. Wenn die Nutzer von fälschlich hohen Wirksamkeitswerten ausgehen, vertrauen sie möglicherweise übermäßig auf deren Richtigkeit, weshalb sie falsche Outputs übersehen können (vgl. Kapitel 4.4.2.2). Das Prinzip des Nichtschadens gebietet darum die Durchführung einer externen Validierung.

Diese kann man in zwei Schritte aufteilen. Zuerst kann das System anhand eines Datensatzes validiert werden, der aus einer anderen Patientenpopulation stammt als

diejenige, die mit dem Trainingsdatensatz verbunden ist. Darüber hinaus ist es sinnvoll, das AI-CDSS in einem zweiten Schritt mit Daten zu evaluieren, die bezüglich ihrer Diversität denjenigen entsprechen, die in der Praxis in das System eingespeist werden (Hopkins et al., 2020).

6.1.18.2 Brauchbarkeit

Um die Brauchbarkeit eines AI-CDSS zu beurteilen ist es hilfreich, dessen Wirksamkeit mit den Fähigkeiten von Ärzten zur vergleichen (vgl. Kapitel 4.1.3). Das kann allerdings nur bedingt Rückschlüsse auf die Brauchbarkeit des Systems erlauben. Es gilt zu evaluieren, ob und inwieweit man die klinische ärztliche Entscheidungsfindung durch die Verwendung eines bestimmten Systems verbessern kann.

Angesichts der Heterogenität der Fähigkeiten einzelner Ärzte wäre es ideal, die Brauchbarkeit für jeden potenziellen Nutzer eines AI-CDSS einzeln zu evaluieren. Das könnte geschehen, indem die Qualität der klinischen Entscheidungsfindung eines bestimmten Arztes mit und ohne Verwendung des betreffenden Systems miteinander verglichen wird. Solche individuellen Vergleichsstudien wären mit großem Aufwand verbunden. Um diesen zu senken, kann man untersuchen, wie sich die Qualität der klinischen Entscheidungsfindung potenzieller Nutzergruppen durch die Verwendung eines bestimmten AI-CDSS verändert. Hier dürfte es sinnvoll sein, verschiedene Gruppen von Testpersonen zu evaluieren, die sich bezüglich der fachlichen Qualifikation unterscheiden (vgl. Kapitel 4.1.3). Damit einzelne potenzielle Nutzer darauf schließen können, inwieweit die Nutzung dieses Systems die Qualität ihrer individuellen klinischen Entscheidungsfindung verbessern könnte, sollten die Versuchspersonen möglichst repräsentativ für die designierte(n) Nutzergruppe(n) sein. Außerdem sollte deren Auswahl transparent erfolgen. Bezüglich der Vergleichsstudien ist darüber hinaus zu fordern, dass möglichst realistische Vergleichsbedingungen zwischen der Studien- und der Kontrollgruppe gewährleistet werden. Den Vergleichspersonen sollte man etwa Hilfsmittel zur Verfügung stellen, auf die Ärzte in der Praxis herkömmlicherweise zurückgreifen können (WHO, 2021).

6.1.18.3 Nutzenpotenzial für die Patientinnen und Patienten

Da das Nutzenpotenzial eines AI-CDSS für die Patienten zentral für die ethische Legitimation für die Verwendung desselben ist (vgl. Kapitel 4.2), sollte man die Auswirkungen der Systemnutzung auf Morbidität, Mortalität und Lebensqualität der

Patienten empirisch untersuchen. Das ist besonders im Hinblick auf die bisher geringe Zahl solcher Studien zu fordern (Wolff et al., 2021) (vgl. Kapitel 4.3).

Um das Nutzenpotenzial eines Systems für die Patienten zu bestimmen, bieten sich zweiarmige Evaluationsstudien an. In der Kontrollgruppe sollte die Behandlung ohne die Verwendung des betreffenden AI-CDSS erfolgen. Eine zweite Studiengruppe von Patienten sollten Ärzte unter Zuhilfenahme des jeweiligen Systems behandeln.

6.1.18.4 Fehler und Schadenspotenzial

Wie grundsätzlich jedes technische System, so arbeiten auch AI-CDSS nicht fehlerfrei. Die Wirksamkeit gibt an, wie häufig ein System zu einem richtigen Ergebnis kommt. Über diese Größe kann man daher auch errechnen, wie wahrscheinlich ein Output falsch ist. Damit ist das Schadenspotenzial des Systems verbunden: Je häufiger es falsche Ergebnisse ausgibt, desto mehr steigt auch grundsätzlich die Wahrscheinlichkeit dafür, dass ein Arzt bei der klinischen Entscheidungsfindung Fehler begeht und der Patient somit Schaden davonträgt.

Das Schadenspotenzial ist aber nicht gleichbedeutend mit dem Fehlerpotenzial eines Systems. Schließlich führt ein fehlerhaftes Ergebnis noch nicht direkt zu einem Fehler in der klinischen Entscheidungsfindung des Arztes: Wenn dieser das fehlerhafte Output erkennt, greift er es nicht auf. Ein weiterer Grund dafür, dass Schadens- und Fehlerpotenzial eines AI-CDSS nicht identisch sind, besteht darin, dass der Schaden für den Patienten, den die ärztliche Übernahme eines falschen Systemergebnisses nach sich ziehen kann, je nach System und Anwendungsgebiet differiert.

Dass sich das Fehlerpotenzial einer Anwendung anhand deren Wirksamkeit abschätzen lässt, reicht daher noch nicht aus, um deren Schadenspotenzial zu bestimmen. Dieses ist einzeln zu evaluieren. Hierbei ist zu prüfen, wie häufig Ärzte falsche Ergebnisse erkennen. Außerdem sollte man evaluieren, wie schwerwiegend die Auswirkungen von übernommenen falschen Empfehlungen für die Patienten sind.

6.1.18.5 Weitere Aspekte der Evaluation und Testung

Die Evaluation von AI-CDSS sollte sich nicht auf die Bestimmung der Wirksamkeit, der Brauchbarkeit und des Nutzenpotenzials der Systeme beschränken. Auch weitere ethische Implikationen einer Anwendung sind nach Möglichkeit empirisch zu untersuchen. Hierzu gehören etwa die Erklärbarkeit eines Systems (vgl. Kapitel 6.1.16), die Auswirkungen

der Verwendung eines AI-CDSS auf die Arzt-Patient-Beziehung (vgl. Kapitel 4.9), dessen Kosten-Nutzen-Verhältnis (vgl. Kapitel 4.11) und ökologische Aspekte wie der Energieverbrauch (vgl. Kapitel 4.12).

6.1.19 Klare Anforderungen an die Systemnutzung

Das Prinzip des Nichtschadens erfordert es, das Schadenspotenzial der KI-Nutzung für die Patienten möglichst zu minimieren. Da Anwendungsfehler durch eine unsachgemäße Verwendung von AI-CDSS entstehen können, sind klare Anforderungen an deren Nutzung zu stellen, um solche zu verhindern (vgl. Kapitel 4.4.2.1). Es sollte etwa klar sein, welche Form von Daten als Input zu verwenden ist und welche Qualitätsanforderungen erfüllt werden müssen. Außerdem sollten die Nutzer einsehen können, in welchen Populationen, Umgebungen und Situationen das System zu hinreichend guten Ergebnissen kommt. Auf diese Weise kann man die Ausgabe falscher Outputs aufgrund von Verzerrungen in den Trainingsdaten verhindern. Dass bestimmt wird, wie ein AI-CDSS sachgemäß zu benutzen ist, besitzt auch für die Zuschreibbarkeit von Verantwortung Relevanz (vgl. Kapitel 4.8 und 6.2.3).

6.1.20 Wartung und Aktualisierung der Systeme

Die vom Prinzip des Nichtschadens gebotene Reduzierung des Fehler- und Schadenpotenzials von AI-CDSS erstreckt sich nicht nur auf deren Entwicklung und ist mit der Implementierung eines Systems nicht beendet. Auch während des Nutzungsprozesses sollten die Entwickler kontinuierlich versuchen, das Fehler- und Schadenspotenzial weiter zu reduzieren. Schließlich ist davon auszugehen, dass manche Aspekte erst in der Praxis auffallen. Fehler sollten die Entwickler möglichst rasch und vollständig beheben. Außerdem ist kontinuierlich die Aktualität und die Richtigkeit der Annahmen zu kontrollieren, auf denen ein AI-CDSS beruht. So sollten etwa relevante medizinische Neuerungen wie Änderungen in einschlägigen klinischen Leitlinien möglichst schnell und umfassend in die Systeme implementiert werden. Im Rahmen der Wartung einer Anwendung ist auch zu überprüfen, ob eine Rekalibrierung erforderlich ist (vgl. Kapitel 6.1.7). Das kann der Fall sein, wenn sich etwa die Anwendungsumgebung oder die Patientenpopulation dermaßen verändert hat, dass das System vermehrt zu falschen Ergebnissen kommt (Antoniou & Mamdani, 2021).

6.2 Empfehlungen für die Schaffung geeigneter Rahmenbedingungen für die Entwicklung und Anwendung von AI-CDSS

Der ethisch vertretbare Umgang mit AI-CDSS bezieht sich nicht nur auf die Entwicklung und Anwendung der Systeme (vgl. Kapitel 6.1 und 6.3). Dafür gilt es auch geeignete Rahmenbedingungen zu schaffen. Diesbezügliche ethisch begründete Empfehlungen folgen in diesem Kapitel.

6.2.1 Transparente rechtliche Anforderungen für die Entwicklung und Nutzung von AI-CDSS

Bei Künstlicher Intelligenz handelt es sich um eine neue Technologie. Daher ist es verständlich, dass es ein zeitintensiver Prozess ist, auf die damit verbundenen rechtlichen Fragen Antworten zu finden. Gleichzeitig ist zu bedenken, dass eine (teils) diffuse Rechtslage ein Hindernis für die Entwicklung und Implementierung von AI-CDSS darstellt (Krumm et al., 2019). Aus ethischer Sicht ist es bedenklich, wenn das Nutzenpotenzial der Systeme nicht realisiert wird, weil bezüglich der Rechtslage Unklarheit besteht. Es sollten daher transparente und verständliche rechtliche Anforderungen für die Entwicklung und Anwendung von AI-CDSS geschaffen werden.

6.2.2 Einteilung von AI-CDSS in Risikoklassen

Verschiedene AI-CDSS sind mit unterschiedlich schwerwiegenden Risiken verbunden. Auch wenn bei der Nutzung der Systeme möglichst alle falschen Outputs zu vermeiden sind, so ist es aus ethischer Sicht bedeutend, besonders die Anzahl derjenigen Fehler zu reduzieren, die zu schwerem Schaden führen können (vgl. Kapitel 4.4.1).

Damit Entwickler und Anwender der Entstehung besonders schwerwiegender Schäden vorbeugen können, müssen diese sich des Schadenspotenzials eines bestimmten Systems bewusst sein (vgl. Kapitel 4.4.2.2). Zur Unterstützung ist es sinnvoll, die Anwendungen in verschiedene Risikoklassen einzuteilen.

6.2.3 Ermöglichung von klarer Verantwortungszuschreibung

Verantwortungsdiffusion stellt ein Risiko dar, das mit der Nutzung von entscheidungsunterstützenden KI-Systemen verknüpft ist (vgl. Kapitel 4.8). Nutzern für die Verwendung von AI-CDSS Verantwortung zuzuschreiben, wird insbesondere durch

Opazität erschwert: Wenn Informiertheit eine Vorbedingung der Verantwortungszuschreibung ist, können Ärzte für Entscheidungen keine Verantwortung übernehmen, die auf der Basis der Empfehlungen von Black-Box-Systemen getroffen wurden (Liedtke & Langanke, 2021).

Es wäre jedoch aus verschiedenen Gründen nicht angemessen, aufgrund von Herausforderungen bei der Verantwortungsübernahme auf die Nutzung von opaken Anwendungen vollständig zu verzichten (vgl. Kapitel 4.8). Es bietet sich stattdessen an, im Anschluss an eine rigorose Testung und Evaluierung klar zu bestimmen, in welchen Situationen und bei der Behandlung welcher Patienten ein bestimmtes AI-CDSS eingesetzt werden kann (Binkley, 2021; Ghassemi et al., 2021; Liedtke & Langanke, 2021). Es ist grundsätzlich ethisch vertretbar, ein KI-System zu verwenden, auch wenn man es nicht gänzlich versteht, sofern dessen Wirksamkeit und Sicherheit ausreichend evaluiert wurden und dessen Nutzenpotenzial das zugehörige Schadenspotenzial überwiegt. Die Verantwortung der Ärzte kann sich dann darauf beschränken, das AI-CDSS sachgemäß zu verwenden. Worin eine solche Nutzung besteht, haben unter anderem die Entwickler zu bestimmen (vgl. Kapitel 6.1.19).

Bezüglich der Zuschreibbarkeit von Verantwortung ist zu bedenken, dass AI-CDSS generell ein nicht eliminierbares Fehlerpotenzial besitzen (vgl. Kapitel 4.4.1). Es ist daher nicht auszuschließen, dass es auch im Rahmen einer sachgemäßen Systemnutzung zu Schadensfällen kommen kann. Für solche Szenarien ist durch geeignete Maßnahmen sicherzustellen, dass kein ‚Verantwortungsvakuum' entsteht. In diesem Zusammenhang erscheinen etwa das Konzept der *e-person* und eine allgemeine Versicherungspflicht für AI-CDSS attraktiv (Braun et al., 2020).

6.2.4 Standardisierung von AI-CDSS und der IT-Infrastruktur

Damit das Big-Data-assoziierte Potenzial von AI-CDSS zur Individualisierung der klinischen Entscheidungsfindung realisiert werden kann (vgl. Kapitel 3.4.4 und 4.3), gilt es die Systeme in die IT-Umgebung der Nutzer zu integrieren. Einheitliche Standards können das erleichtern. Da klinische IT-Dienste derzeit aber hinsichtlich der verwendeten Standards heterogen sind, sollte man auf deren Vereinheitlichung hinarbeiten (Krumm et al., 2019).

Das ist ethisch auch geboten, damit Anwender bei der Auswahl eines AI-CDSS nicht auf eine begrenzte Menge an Systemen beschränkt sind, die mit der IT-Infrastruktur vor Ort

kompatibel sind, sondern der Zweck-Mittel-Rationalität und dem Prinzip der Nutzenmaximierung entsprechend aus dem uneingeschränkten Angebot die für ihre Zwecke am besten geeignete Anwendung frei auswählen können.

6.2.5 Generierung und Bereitstellung von qualitativ hochwertigen Datensätzen

Die Entwicklung von ML-basierten AI-CDSS ist auf die Verfügbarkeit von passenden Trainingsdatensätzen angewiesen (vgl. Kapitel 3.1.2). Diese sind zu generieren und bereitzustellen. Das gilt besonders für Datensätze von Bevölkerungsgruppen, von denen es bisher unverhältnismäßig wenige oder nur qualitativ minderwertige Daten gibt (vgl. Kapitel 4.12).

Dabei ist zunächst bei der Sammlung von Daten in der klinischen Praxis anzusetzen. Aber nicht nur deren Menge ist relevant. Diese sollten so gesammelt werden, dass sie sich für die anschließende Entwicklung von AI-CDSS eignen. Dafür bietet sich die strukturierte Befundung an. Im Rahmen dessen werden Daten im klinischen Alltag geordnet erhoben (acatech, 2020; Hempel & Pinto dos Santos, 2021). Darüber hinaus sollte man sie in maschinenlesbarer Form sammeln und speichern (Krumm et al., 2019).

Man kann auch bereits bestehende Datensätze zusammenführen (Plattform Lernende Systeme, 2019). Weil mit Internationalität generell auch Vielfalt einhergeht, kann man durch die Zusammenführung von Datensätzen aus verschiedenen Ländern eine vielfältigere Datenbasis generieren. Dem kommt auch aus gerechtigkeitsethischer Perspektive Relevanz zu: Denn die generell aus Gründen der Fairness gebotene Generalisierbarkeit eines AI-CDSS kann durch die Vielfalt der Trainingsdaten gesteigert werden (Mitchell et al., 2021) (vgl. Kapitel 4.12).

6.2.6 Förderung der Entwicklung von AI-CDSS

Weil AI-CDSS das Potenzial bieten, die Patientenversorgung besser und effizienter zu gestalten, ist die Entwicklung derartiger Systeme vom Prinzip des Wohltuns prima facie geboten und daher zu fördern. Die allgemeine Verfügbarkeit entsprechender Computer und Datensätze kann es auch solchen Akteuren ermöglichen AI-CDSS zu entwickeln, welche die dazu nötigen Mittel nicht besitzen (Rohde et al., 2021). Auf diese Weise könnte man einer ethisch nicht vertretbaren Monopolbildung vorbeugen und eine gerechte Verteilung der ökonomischen Potenziale gewährleisten (vgl. Kapitel 4.12). Durch eine entsprechende Gestaltung der Rahmenbedingungen könnte man außerdem die

Entwicklung der Anwendungen im Sinne der Allgemeinheit beeinflussen (Krumm et al., 2019).

Die Unterstützung von KI-Entwicklern ist auch hinsichtlich regulatorischer und rechtlicher Fragen sinnvoll. Beispielhaft sei hier auf das Projekt *BAIM – Boost AI to Market* hingewiesen (Prinz, 2023).

6.2.7 Förderung der Nutzung von AI-CDSS

An verschiedenen Stellen der vorliegenden Arbeit wurden Belastungen für die Nutzer von AI-CDSS sichtbar: Diese können zunächst durch die Verwendung der Systeme bestimmte Fähigkeiten verlieren (*deskilling*) und *alert fatigue* erleiden (vgl. Kapitel 4.7). Da die Verantwortungszuschreibung für KI-basierte Entscheidungen nicht selten schwierig ist, werden KI-Nutzer auch in dieser Hinsicht beansprucht (vgl. Kapitel 4.8). Darüber hinaus erfordert die Verwendung von AI-CDSS etwa durch den Besuch von Fortbildungsveranstaltungen einen erheblichen zeitlichen Aufwand (vgl. Kapitel 6.3.7). Um diese Belastungen zu vermeiden, könnten Ärzte sich gegen die Nutzung von AI-CDSS entscheiden. Daraus würde jedoch resultieren, dass die Nutzenpotenziale der Systeme nicht realisiert werden können. Insofern die Anwender von AI-CDSS unverhältnismäßig großen Belastungen ausgesetzt sind, ist es aus Gründen der Gerechtigkeit geboten (vgl. Kapitel 4.12), die Verwendung der Systeme zu fördern.

6.2.8 Anpassung der ärztlichen Ausbildung

Durch die Implementierung von AI-CDSS – das hat sich im Laufe der vorliegenden Arbeit gezeigt – wird sich die klinische Praxis verändern (vgl. Kapitel 3.4.4). Es ist ethisch erforderlich, angehende Ärzte auf die Anforderungen vorzubereiten, die sich aus diesen Transformationen ergeben.

Damit Ärzte zur sachgemäßen Anwendung von KI-Systemen befähigt sind, benötigen sie Wissen und verschiedene Kompetenzen, die möglichst früh entwickelt und gestärkt werden sollten. So sollte die Vorbereitung der angehenden Ärzte auf die KI-Nutzung einen festen Platz im Medizinstudium erhalten. Beispielhaft ist hierfür das Projekt ‚Künstliche Intelligenz in der Lehre der AUgenheilkunde und der RAdiologie (KI-LAURA)‘ (Rheinische Friedrich-Wilhelms-Universität Bonn, 2021).

Die Anpassungen des Medizinstudiums sollten sich nicht auf die Etablierung von Lehrveranstaltungen beschränken, die den angehenden Ärzten die Nutzung der Systeme näherbringen. Diese sollten auf die Rolle vorbereitet werden, die ihnen im Zusammenspiel mit KI-Systemen zukommen wird. So sollte sich die Ausbildung auch auf diejenigen Aufgaben und Bereiche des ärztlichen Tuns konzentrieren, die von AI-CDSS nicht oder kaum unterstützt werden können und daher zukünftig innerhalb des ärztlichen Tuns an Relevanz zunehmen werden. Hierzu gehören etwa die Auseinandersetzung mit besonders komplexen Fällen, der Umgang mit ‚*soft facts*' oder die Beachtung von Patientenpräferenzen (Funer, 2021; Liu et al., 2018; Topol, 2019) (vgl. Kapitel 4.1.1).

Voraussichtlich wird der Einsatz von KI-Systemen in manchen Fachbereichen den Bedarf an Ärzten senken. Gleichzeitig ist anzunehmen, dass der derzeit bestehende Ärztemangel in einigen Bereichen in nächster Zeit nicht restlos aufgelöst werden kann. Insofern es sich absehen lässt, wie sich der Bedarf an Ärzten in einzelnen medizinischen Bereichen entwickeln wird, sollten angehende Ärzte darüber informiert werden, damit die KI-Transformation nicht in manchen Disziplinen einen Überhang, in anderen Bereichen aber einen Mangel an Ärzten schafft.

6.2.9 Ethische Sensibilisierung der Entwicklerinnen und Entwickler von AI-CDSS

Wie in Kapitel 6.1.1 dargelegt, sollten ethische Implikationen von AI-CDSS bei der Systementwicklung von Beginn an bedacht werden. Um das zu gewährleisten, wurde dafür argumentiert, dass Ethiker Teil des Entwicklungsteams sein sollten. Das darf jedoch nicht bedeuten, dass die Auseinandersetzung mit ethischen Fragen und Aspekten auf Ethiker abgeschoben wird. Damit diese Implikationen von AI-CDSS bei der Systementwicklung hinreichend Beachtung erfahren und vertretbar behandelt werden, müssen alle Mitglieder des Entwicklungsteams ethisch sensibilisiert sein. Dafür ist die Durchführung edukativer Maßnahmen sinnvoll. Schon im Rahmen der Ausbildung sollten die Entwickler auf die ethischen Aspekte der Entwicklung von KI-Systemen vorbereitet werden (Zweig, 2018). Idealerweise sollten in diesen Lehrveranstaltungen auch Implikationen Beachtung finden, die besonders mit der Verwendung von KI-Systemen in der Medizin einhergehen. Zur Gewährleistung dessen erscheint eine gesonderte Ausbildung für die Entwickler von medizinischen KI-Systemen sinnvoll.

So wichtig die Sensibilisierung der Entwickler für ethische Implikationen von AI-CDSS ist, gilt es dennoch zu bedenken, dass deren Kenntnis noch nicht bedeutet, dass diese angemessen adressiert werden. Zur Gewährleistung einer ethisch vertretbaren Entwicklung von AI-CDSS mag die Entwicklung eines Berufsethos dienlich sein (Bitkom & DFKI, 2017) (vgl. Kapitel 4.6) – auch wenn es fraglich ist, ob das kurzfristig möglich ist (Mittelstadt, 2019).

6.2.10 Gesellschaftliche Aufklärung über KI

Nicht nur Ärzte, auch Patienten sollten über die grundsätzliche Funktionsweise sowie über Nutzen- und Schadenspotenziale von AI-CDSS informiert sein. Das ist erforderlich, um diese in die Lage zu versetzen, mündig über die Nutzung der Systeme zu entscheiden (vgl. Kapitel 4.5.1). Indem die Gesellschaft über KI aufgeklärt wird, kann verhindert werden, dass sich Patienten aus unbegründeter Angst gegen die Verwendung dieser Technologie entscheiden (vgl. Kapitel 4.5.2).

Weil möglicherweise Vertreter mancher Gruppen überproportional häufig eine unbegründete Ablehnungshaltung gegenüber KI einnehmen und diese somit von der Partizipation an den Nutzenpotenzialen der Systeme ausgeschlossen sind, ist die gesellschaftliche Aufklärung über KI auch aus gerechtigkeitsethischen Gründen geboten (WHO, 2021) (vgl. Kapitel 4.12). Risiken, Möglichkeiten und Grenzen von KI-Systemen zu verstehen, ist außerdem erforderlich dafür, dass demokratisch und hinreichend informiert über die Regulation von AI-CDSS entschieden werden kann.

Gesellschaftliche Aufklärung über KI ist noch aus einem anderen Grund geboten. Da die Nutzung von AI-CDSS die Qualität der Patientenversorgung verbessern kann (vgl. Kapitel 4.3), zur Entwicklung dieser Systeme aber (große) Trainingsdatensätze erforderlich sind (vgl. Kapitel 3.1.2), sollten Spenden von Trainingsdaten gefördert werden (vgl. Kapitel 4.10). Die Bereitschaft zur Datenspende dürfte unter anderem davon abhängen, ob Bürger sich über deren Nutzenpotenzial im Klaren sind. Angesichts der Neuartigkeit von KI ist kaum davon auszugehen, dass ein solches Bewusstsein bisher hinreichend vorhanden ist. Deshalb ist die Gesellschaft über die Relevanz von Datenspenden aufzuklären.

6.3 Empfehlungen für die Nutzung von AI-CDSS

6.3.1 Überprüfung des Bedarfs an Entscheidungsunterstützung

Nicht nur die Entwickler (vgl. Kapitel 6.1.2), sondern auch die Nutzer sollten überprüfen, inwiefern ein Bedarf an Entscheidungsunterstützung besteht. Dieser ist möglichst genau zu spezifizieren, damit das am besten geeignete Mittel identifiziert werden kann (vgl. Kapitel 6.3.2). Gilt es zum Beispiel, die diagnostische Genauigkeit zu verbessern, sollte etwa klar sein, ob primär die Spezifität oder die Sensitivität der Diagnostik zu optimieren ist (vgl. Kapitel 4.2).

6.3.2 Vergleich mit Alternativen

Auch der Vergleich des Systems mit Alternativen ist sowohl für die Entwickler als auch für die Nutzer relevant (vgl. Kapitel 6.1.3): Ist ein Bedarf an ärztlicher Entscheidungsunterstützung festgestellt, sollte der Zweck-Mittel-Rationalität entsprechend ein Mittel zur Erfüllung dieses Bedarfs ausgewählt werden. Hierbei sollten die Anwender verschiedene Lösungsstrategien bedenken – KI-basierte, digitale wie auch analoge. Die Alternativen sind hinsichtlich ihrer Funktionsfähigkeit, ihrer Nutzen- und Schadenspotenziale miteinander zu vergleichen. Die ethischen Implikationen der einzelnen Lösungen sollten analysiert und abgewogen werden. Im Lichte der Ressourcenknappheit des Gesundheitssystems sind hier etwa ökonomische Gesichtspunkte und ökologische Aspekte zu berücksichtigen (vgl. Kapitel 4.11 und 4.12).

6.3.3 Überprüfung der Eignung des Systems für die spezifischen Anwendungsbedingungen

Vor der Implementierung eines AI-CDSS sollte überprüft werden, ob es sich für die Verwendung in der designierten Anwendungsumgebung eignet.

Dabei ist etwa sicherzustellen, dass die systemspezifischen Anforderungen an die Inputdaten erfüllt werden. So sollten diese in der richtigen Form und in hinreichender Qualität vorliegen (van Baalen et al., 2021) (vgl. Kapitel 4.4.2.1). Außerdem sind diese hinsichtlich ihrer Charakteristika mit dem Trainingsdatensatz zu vergleichen. Unterscheiden sich Trainings- und Anwendungsdaten zu stark, steigt das Risiko für die Ausgabe falscher Systemergebnisse (vgl. Kapitel 4.4.1).

Angesichts der Vielfalt verwendeter Standards im Bereich der medizinischen IT-Infrastruktur kann man nicht davon ausgehen, dass jedes System mit der vorliegenden IT-Umgebung kompatibel ist. Es ist darum auch die Kompatibilität des KI-Systems sicherzustellen.

6.3.4 Probatorische Nutzung vor der langfristigen Inbetriebnahme

Die in Kapitel 6.1.18 beschriebene Prüfung des AI-CDSS vor der Nutzung ist wichtig. Gleichzeitig ist zu bedenken, dass man vor der Inbetriebnahme eines Systems nicht alle relevanten Aspekte identifizieren kann. Es ist möglich, dass sich bestimmte Schwierigkeiten erst bei der Nutzung der Anwendung zeigen.

Ist das System allerdings bereits im Betrieb, kann sich die Entscheidung zu dessen Absetzung als schwierig erweisen: Schließlich ist die Implementierung einer Anwendung mit Kosten verbunden, die sich durch die Systemnutzung amortisieren sollen. Obwohl es manchmal erforderlich ist, ein AI-CDSS abzusetzen, dürfte die Entscheidung dazu schwer fallen, wenn das System bereits im Betrieb ist. Aus ethischer Sicht ist das bedenklich: Damit die Kosten der Implementierung ausgeglichen werden, könnte eine Anwendung weiter benutzt werden, auch wenn dadurch etwa das Patientenwohl gefährdet wird.

Um die Möglichkeit zu schaffen, ein AI-CDSS in der Praxis zu testen, ohne die Kosten für eine mögliche Absetzung in die Höhe zu treiben und somit die Entscheidungsfreiheit möglicherweise einzuschränken, ist es sinnvoll, ein System probatorisch zu implementieren. Indem etwa nicht eine ganze Klinik, sondern nur eine bestimmte Station mit der betreffenden Anwendung ausgestattet wird, kann man erste Erfahrungen damit in der spezifischen Umgebung gewinnen (Fulterer, 2021). Falls sich in der Praxis vor Ort etwa zeigt, dass das Schadenspotenzial das Nutzenpotenzial des Systems überwiegt, ist es mit vergleichsweise geringem Aufwand möglich, dessen Verwendung zu stoppen.

6.3.5 Etablierung einer passenden IT-Umgebung und Einbettung des AI-CDSS

Das Potenzial von AI-CDSS zur Individualisierung der Medizin ist nur dann realisierbar, wenn der Zugang zu Patientendaten gewährleistet wird (vgl. Kapitel 4.3). Darum sollten die Systeme mit der IT-Umgebung verbunden werden. Vor der Inbetriebnahme einer neuen Anwendung ist daher eine leistungsfähige IT-Infrastruktur zu etablieren, falls diese noch nicht vorhanden ist (Krumm et al., 2019). Dabei sollte man auf Kompatibilität mit

dem AI-CDSS achten. Um die anfallenden Daten nach der Zustimmung der Patienten als Ressource nutzen zu können – etwa für die Entwicklung weiterer Systeme – sollten innerhalb der IT-Infrastruktur Möglichkeiten zur strukturierten, maschinenlesbaren Sammlung der Daten geschaffen werden (vgl. Kapitel 6.2.5).

6.3.6 Umweltfreundliches Nutzungsverhalten

Bei der Verwendung von AI-CDSS ist wie bei anderen auf Stromzufuhr angewiesenen Systemen aus Gründen der Umweltgerechtigkeit auf eine möglichst ressourcen- und umweltschonenden Betriebsweise zu achten (vgl. Kapitel 4.12): Ungenutzte Systeme sollten von der Stromzufuhr genommen werden, nicht erforderliche, energetisch ineffiziente Funktionen sollten nicht verwendet werden.

6.3.7 Schulung und Weiterbildung der Nutzerinnen und Nutzer

Der Gebrauch eines AI-CDSS stellt vielerlei Anforderungen an die Nutzer. Um diesen gerecht werden zu können, sollten die Anwender im Rahmen von Schulungen und Weiterbildungen von kompetentem Personal zur sachgemäßen, ethisch vertretbaren Systemnutzung befähigt werden.

Zu den Aspekten, die von solchen Schulungen umfasst werden sollten, gehören zunächst die Grundlagen des AI-CDSS: Wie funktioniert es? Wo liegen dessen Möglichkeiten und Grenzen? Wie sollte es genutzt werden?

Das Prinzip des Nichtschadens gebietet es, hier besonders auf die Begrenzungen des Systems einzugehen. So sollte man die Nutzer darauf hinweisen, dass AI-CDSS bislang nur einzelne Aufgaben im Entscheidungsprozess, aber nicht den gesamten Ablauf übernehmen können (vgl. Kapitel 4.1.1), weshalb die Nutzer zur Kontrolle der Outputs – jedenfalls auf Plausibilität hin – angehalten sind (vgl. Kapitel 4.4.2.2 und 6.3.12). Dabei sollten die Ärzte auch darüber informiert werden, in welchen Situationen das jeweilige AI-CDSS typischerweise zu falschen Ergebnissen kommt (vgl. Kapitel 4.4.2.2). Somit können die Ärzte bei der Kontrolle der Outputs ihre Konzentration auf die besonders häufigen Fehler des Systems lenken und auf diese Weise die Zahl der detektierten falschen Ergebnisse erhöhen. Es gilt dabei insbesondere diejenigen falschen Ergebnisse zu erkennen, die zu besonders schwerem Schaden führen können.

Ärzte sollten darauf hingewiesen werden, dass AI-CDSS Patientenpräferenzen und ‚*soft facts*' grundsätzlich nicht in die Berechnungen einbeziehen und die hinreichende Beachtung dieser Aspekte daher in den ärztlichen Aufgabenbereich fällt (Funer, 2021) (vgl. Kapitel 4.1.1). Es sollte gewährleistet werden, dass der Arzt sein menschliches Gegenüber nicht naturwissenschaftlich verkürzt, sondern bio-psycho-sozial und empathisch behandelt (vgl. Kapitel 4.9.2).

Bezüglich der Zuschreibung von Verantwortung bei der Nutzung von AI-CDSS wurde offenbar, dass die – möglicherweise unbewusste – Übertragung der Verantwortung auf die Systeme eine Gefahr darstellt (vgl. Kapitel 4.8). Die Nutzer sollten über dieses Risiko informiert werden.

Man sollte außerdem versuchen, die Entstehung des Automation Bias mithilfe von edukativen Maßnahmen zu verhindern. Es gibt schließlich Hinweise darauf, dass man dessen interne Faktoren mithilfe von Training und Instruktionen modifizieren kann (Burdick et al., 1996; Skitka et al., 2000).

Ein weiterer Aspekt, der im Rahmen von Weiterbildungen adressiert werden sollte, ist die Gefahr von *deskilling*: Ärzte sollen die von AI-CDSS empfohlenen Diagnosen, Prognosen und Therapieempfehlungen nicht blind übernehmen, sondern kontrollieren (vgl. Kapitel 4.4.2.2). Um die Systemergebnisse aber überprüfen zu können, müssen Ärzte dazu in der Lage sein, qualitativ hochwertige Diagnosen, Prognosen und Therapieempfehlungen zu entwickeln (vgl. Kapitel 4.7). Damit Nutzer diese Fähigkeit nach einer gewissen Zeit der Verwendung von AI-CDSS nicht verlieren, sollten sie in dieser Hinsicht geschult werden. Dabei sind sie dafür zu sensibilisieren, welche Fähigkeiten besonders wichtig sind, um Patienten bei der KI-Nutzung vor Schaden zu schützen. Es gilt dafür zu sorgen, dass gerade diese Fähigkeiten nicht *deskilling* zum Opfer fallen.

Insofern durch die Nutzung von AI-CDSS manche Teile des ärztlichen Tuns und Denkens weniger relevant, andere Teile – etwa die Beachtung von *soft facts* – hingegen wichtiger werden, sollten vor allem diejenigen Kompetenzen von Ärzten gestärkt werden, die in der neuen Rolle des Arztes im Verbund mit KI-Systemen besonders an Bedeutung gewinnen.

6.3.8 Aufklärung und Einwilligung der Patientinnen und Patienten

Je nach der Funktion und der Beschaffenheit eines AI-CDSS kann es ethisch geboten sein, vor dessen Verwendung die informierte Einwilligung des Patienten einzuholen (vgl. Kapitel 4.5.1). Das lässt sich aus dem Prinzip des Respekts der Patientenautonomie ableiten. Dagegen kann einerseits das Prinzip der Gerechtigkeit, andererseits das Prinzip des Wohltuns sprechen.

Bei der Aufklärung des Patienten sollte darauf geachtet werden, dass dieser in laiengerechter Weise über relevante Charakteristika des Systems und über dessen Verwendungsweise informiert wird. Im Anschluss an Ploug und Holm (2020a) kann man in dieser Hinsicht auf vier relevante Gesichtspunkte hinweisen: Aus Gründen der informationellen Selbstbestimmung sollte der Patient über die Datennutzung des Systems und mögliche Risiken des Datenmissbrauches informiert werden (vgl. Kapitel 4.10). Darüber hinaus sollte er nachvollziehen können, welchen Zusatznutzen die Verwendung des AI-CDSS für den Patienten im vorliegenden Fall hat, wie hoch dessen Fehlerpotenzial ist und wie schwerwiegend mögliche fehlerhafte Outputs sind. Außerdem sollte dem Patienten die Aufgabenzuweisung zwischen dem KI-System und dem Arzt erklärt werden. Dadurch kann das Risiko reduziert werden, dass ein Patient diese missversteht – worunter die Qualität der Arzt-Patient-Beziehung leiden kann (vgl. Kapitel 4.9.1).

6.3.9 Verhinderung von Zeitdruck und von Ablenkung bei der Verwendung von AI-CDSS

Die Nutzung von AI-CDSS geht mit vielerlei – nicht zuletzt kognitiven – Anforderungen einher. Damit die Nutzer diesen gerecht werden können, reicht es wohl nicht aus, an Schulungen und Weiterbildungen teilzunehmen (vgl. Kapitel 6.3.7). Auch die Arbeitsbedingungen der Nutzer sind angemessen zu gestalten.

So ist sicherzustellen, dass Ärzte bei der Verwendung von AI-CDSS ausreichend Zeit für die sachgemäße Nutzung der Systeme haben und nicht unter zeitlichem Druck stehen. Das fordert das Prinzip des Nichtschadens: Schließlich konnte gezeigt werden, dass zeitlicher Druck die Entstehungsgefahr des Automation Bias erhöhen kann (Sarter & Schroeder, 2001) (Kapitel 4.4.2.2). Indem den Nutzern genügend Zeit für die Verwendung der AI-CDSS eingeräumt wird, können die Ärzte außerdem kontinuierlich ihre Systemnutzung evaluieren. Auf diese Weise können sie bei der Nutzung möglicherweise Verbesserungspotenzial identifizieren und anschließend realisieren.

Um Automation Bias zu vermeiden, sollte außerdem Multitasking unterbunden werden. Einen Zusammenhang zwischen beiden Phänomenen konnte nämlich ein Review aufzeigen: So trat Automation Bias in 10 von 11 Studien auf, bei denen die Probanden mehrere Aufgaben gleichzeitig zu bearbeiten hatten (Lyell & Coiera, 2017).

6.3.10 Verhinderung von Überdiagnostik

Das Prinzip des Nichtschadens gebietet es, Überdiagnostik zu vermeiden. Schließlich werden bei derartigen Maßnahmen die stets vorhandenen Schadenspotenziale nicht von entsprechenden Nutzenpotenzialen aufgewogen. Zur Vermeidung von Überdiagnostik gilt es zu bedenken, dass die Verwendung von AI-CDSS diese vor allem aus zwei Gründen fördern kann (vgl. Kapitel 4.4.2.3).

Erstens kann die mit KI-Nutzung potenziell einhergehende Reduzierung der Behandlungskosten (vgl. Kapitel 4.11) den Anreiz senken, die Erforderlichkeit einer diagnostischen Maßnahme hinreichend zu prüfen. Dieser Gefahr ist entgegenzuwirken: Ärzte sollten auch bei der Verwendung von AI-CDSS sicherstellen, dass die Schadenspotenziale nicht die Nutzenpotenziale einer medizinischen Maßnahme überwiegen.

Ein zweiter Faktor, der zur Förderung von Überdiagnostik führen kann, liegt in der oft hohen Genauigkeit und Sensitivität der diagnostischen AI-CDSS (vgl. Kapitel 4.1.2). Daher besteht die Gefahr, dass Pathologien identifiziert werden, die nicht behandelt werden können. Es muss auch bei der KI-basierten Diagnostik gewährleistet sein, dass diese jedenfalls einen potenziellen therapeutischen Nutzen aufweist. So sollte stets hinterfragt werden, ob die Erkrankung, die mithilfe eines AI-CDSS detektiert werden soll, behandlungsbedürftig ist.

6.3.11 Überprüfung der Eignung des AI-CDSS für den spezifischen Anwendungsfall

Inwieweit ein System für die Anwendung in einer bestimmten Umgebung und in einer Population geeignet ist, ist nicht nur für dessen Wirksamkeit, sondern auch für die Minimierung des Schadenspotenzials relevant. Daher sollte diese Eignung nicht nur vor der Implementierung des Systems überprüft werden (vgl. Kapitel 6.3.3). Diese ist vor jeder einzelnen Anwendungssituation zu kontrollieren. Hierbei ist unter anderem Folgendes sicherzustellen:

Es sollte zunächst gewährleistet sein, dass die für den Betrieb des Systems nötigen Daten in der richtigen Form vorhanden sind (zum Beispiel MRT- oder CT-Bild). Außerdem ist zu überprüfen, ob diese eine hinreichende Qualität aufweisen (vgl. Kapitel 4.4.2.1) (van Baalen et al., 2021).

Darüber hinaus gilt es auszuschließen, dass die Charakteristika der Trainingsdaten sich derart von den Inputdaten in der Praxis unterscheiden, dass die Anwendung verzerrte Ergebnisse ausgibt (vgl. Kapitel 4.4.2.1). Soll etwa untersucht werden, ob es sich bei einer verdächtigen Hautläsion eines Patienten afroamerikanischen Hintergrunds um ein malignes Melanom handelt, kann die Verwendung eines Systems, das nur mit Daten von Menschen kaukasischen Hintergrunds trainiert wurde, zu falschen Ergebnissen kommen. In diesem Fall darf das AI-CDSS nicht angewandt werden. Um Ärzte bei der Beantwortung der Frage zu unterstützen, ob ein System in einem bestimmten Fall verwendet werden kann, sollten die Entwickler die Anforderungen an die Inputdaten möglichst klar formulieren (vgl. Kapitel 6.1.19).

6.3.12 Kontrolle und Kontextualisierung der Outputs von AI-CDSS

AI-CDSS sollen die klinische Entscheidungsfindung des Arztes unterstützen, nicht übernehmen (vgl. Kapitel 4.1.1). Daher ist es geboten, dass der Behandelnde das Systemergebnis zumindest auf Plausibilität hin überprüft. Das Ausmaß der jeweils notwendigen Kontrolle hängt insbesondere vom jeweiligen System, von der konkreten Anwendungssituation sowie von der Schwere potenzieller Fehlentscheidungen ab.

Um die Häufigkeit der Schädigung von Patienten möglichst zu minimieren, sollten Ärzte darauf hingewiesen werden, in welchen Situationen das jeweilige AI-CDSS typischerweise fehlerhafte Ergebnisse ausgibt. Dabei sollte das Augenmerk speziell auf denjenigen falschen Outputs liegen, die besonders schweren Schaden für Patienten verursachen können.

Damit Ärzte potenzielle Fehler entdecken können, die mit den Begrenzungen von AI-CDSS verknüpft sind, sollten sich die Nutzer besonders auf die genuin humane Fähigkeit zur Entwicklung eines ganzheitlichen Bildes des Patienten stützen (van Baalen et al., 2021) (vgl. Kapitel 4.9.2). Die Outputs des AI-CDSS sollten auf Stimmigkeit mit diesem holistischen Bild überprüft werden. Auf diese Weise können Ärzte falsche Ergebnisse detektieren, die auf die Begrenztheit des Systems zurückzuführen sind.

Zur Kontrolle eines Ergebnisses sollte man – wenn vorhanden – auf Erklärungen der Anwendung rekurrieren. Lässt sich dadurch zeigen, dass das Output auf falschen Annahmen beruht, spricht das gegen dessen Richtigkeit (vgl. Kapitel 4.4.2.2). Lässt sich aus den Explikationen jedoch kein Hinweis auf ein falsches Ergebnis gewinnen – beziehungsweise gibt das System überhaupt keine Erklärung an – wird die Entscheidung für oder gegen die Übernahme des Outputs zunehmend schwierig. Letztlich sollte hier entscheiden, ob das AI-CDSS dem Arzt bei der Aufgabe über- oder unterlegen ist (vgl. Kapitel 4.4.2.2).

Hinsichtlich der Kontrolle und der Kontextualisierung von Systemergebnissen ist noch ein Weiteres zu beachten: Das Autonomieprinzip gebietet, dass Therapieentscheidungen an den Werten und Wünschen der Patienten auszurichten sind. Dies zu gewährleisten, ist Aufgabe des Arztes, weil AI-CDSS bislang für evaluative Präferenzen nicht flexibel sind (vgl. Kapitel 4.1.1 und 4.5.3).

Passt man die Therapieentscheidung nicht an die Werte und Wünsche des Patienten an, entsteht außerdem die Gefahr, dass dieser Vertrauen gegenüber dem KI-unterstützten Arzt verliert, wodurch die Arzt-Patient-Beziehung empfindlich gestört werden kann (vgl. Kapitel 4.9.1).

Nachdem der Arzt ein Ergebnis kontrolliert und kontextualisiert hat, sollte er dieses an den Patienten weitergeben, sofern nicht bereits das AI-CDSS das Ergebnis dem Patienten angezeigt hat (vgl. Kapitel 6.1.13). Das Prinzip des Respekts der Patientenautonomie fordert nämlich, den Patienten über seinen Gesundheitszustand zu informieren (vgl. Kapitel 4.5.2).

6.3.13 Umgang mit personenbezogenen Daten

Beim Umgang mit personenbezogenen Daten gilt es Datenschutz und -verfügbarkeit auszutarieren. Idealerweise sollten die Datenverfügbarkeit und das Nutzenpotenzial der Systeme nicht durch Datenschutzmaßnahmen eingeschränkt werden.

Zur Gewährleistung von Datenschutz sind Patientendaten vor unautorisiertem Zugriff zu schützen. Hierbei ist zu beachten, dass es verschiedene Gruppen von Personen gibt, welche die Daten unautorisiert und missbräuchlich verwenden könnten.

Erstens könnten Personen, die in einem begrenzten Maße Zugriffsrechte besitzen, diese überdehnen. Der Grund hierfür kann darin bestehen, dass die betreffenden Personen nicht

hinreichend über die Begrenzungen ihrer Zugriffsrechte informiert sind. Darüber aufzuklären, kann diese Form des Datenmissbrauchs verhindern.

Der illegitime Zugriff auf patientenbezogene Daten kann aber auch vorsätzlich geschehen. Edukative Maßnahmen sind hier nicht wirksam. Diese Form des Datenmissbrauchs fußt schließlich nicht auf mangelhafter Informiertheit und ist daher durch andere Maßnahmen zu verhindern. Der Zugang zu dem betreffenden System kann etwa mithilfe von Passwörtern geschützt werden. Daten können zum Schutz vor Missbrauch auch verschlüsselt werden.

Bei der Nutzung von AI-CDSS gilt es nicht nur die Daten derjenigen Patienten zu schützen, die mithilfe des Systems behandelt werden. Dem Schutz von Trainingsdaten kommt aus ethischer Perspektive besondere Bedeutung zu (vgl. Kapitel 4.10): Die Datenspender nehmen nämlich das Risiko des Datenmissbrauchs in Kauf, um die Entwicklung von KI-Systemen zu ermöglichen, mit denen später andere Patienten behandelt werden können. Illegitime Zugriffe auf Trainingsdaten sind daher in besonderer Weise zu verhindern (Arshad et al., 2021; Kaissis et al., 2020).

6.3.14 Patientenorientierte Nutzung der frei gewordenen Ressourcen

Die Nutzung von AI-CDSS kann die zeitliche Effizienz der Behandlung steigern (vgl. Kapitel 4.11). Eine Entlastung der Ärzte kann zum einen dazu dienen, dass diese vom notorischen Zeitdruck der ärztlichen Praxis befreit werden, mehr Zeit mit den Patienten verbringen können und sich dadurch stärker den zeitintensiven genuin humanen Aspekten der klinischen Entscheidungsfindung widmen können (vgl. Kapitel 4.9.2). Zum anderen könnte die Verringerung des ärztlichen Arbeitsaufwands auch zum Abbau ärztlicher Arbeitsplätze verwendet werden (Topol, 2019). Das erscheint besonders im Hinblick auf den finanziellen Druck in der Medizin und auf die hohen personalen Kosten durch Ärzte nicht unrealistisch. Der Abbau ärztlicher Arbeitsplätze kann in einem System, das von Ärztemangel geprägt ist, auch zur Besetzung bisher vakanter Arbeitsstellen führen. Aus ethischer Sicht ist jedenfalls zu fordern, dass die Ressourcen, die durch die Nutzung von AI-CDSS frei werden, *patientenorientiert* eingesetzt werden.

6.3.15 Kontinuierliche Überprüfung des Systems und der sachgemäßen Nutzung

Bereits in Kapitel 6.1.20 wurde darauf hingewiesen, dass auch nach der Implementierung eine kontinuierliche Überprüfung von AI-CDSS angezeigt ist. Diese sollten nicht nur die Entwickler, sondern auch die Anwender durchführen.

Im Rahmen dessen ist etwa die Aktualität der Software sicherzustellen. Weil Updates ein Einfallstor für Schadsoftware und Manipulation darstellen können, sollten diese auf ihre Autorisierung hin überprüft werden (Müller-Quade et al., 2020).

Darüber hinaus sollten die Anwender die Entwickler zur Optimierung der Systeme befähigen, indem sie diese über möglichen Verbesserungsbedarf informieren. Dass die Nutzer derartige Informationen an die Entwickler weitergeben, gebietet erstens das Nichtschadensprinzip: Erst wenn Letztere über Fehler eines Systems informiert werden, können sie diese beheben. Daneben erfordert auch das Prinzip des Wohltuns, mit den Entwicklern Informationen zu teilen, die diese zur Verbesserung der Systeme befähigen. Denn durch die Optimierung von AI-CDSS kann sich auch die Qualität der Versorgung zukünftiger Patienten verbessern.

Ein besonderes Augenmerk sollte bei der kontinuierlichen Überprüfung auf die Frage gerichtet werden, ob sich die Charakteristika der Inputdaten zwischenzeitlich möglicherweise derart verändert haben, dass das System unvertretbar häufig zu falschen Ergebnissen kommt (Futoma et al., 2020). So kann es – falls nötig – rekalibriert werden (vgl. Kapitel 6.1.7).

Im Hinblick auf die große Bedeutung der sachgemäßen Nutzung von AI-CDSS ist es angezeigt, neben den Systemen auch deren Verwendungsweise kontinuierlich zu überprüfen. Eine unabhängige Institution (‚Algorithmen-TÜV') könnte kontrollieren, ob ein Arzt eine Anwendung sachgemäß nutzt, oder ob er möglicherweise nicht (mehr) zu deren Verwendung befähigt ist (Zweig, 2018). Mit dieser Aufgabe könnten beispielsweise auch Mitarbeiter der jeweiligen Klinik oder des betreffenden medizinischen Versorgungszentrums betraut werden. Wichtig ist jedenfalls, dass die sachgemäße Nutzung des Systems dauerhaft sichergestellt ist. Schließlich gibt es Hinweise darauf, dass das Risiko des Automation Bias zunimmt, je erfahrener ein Nutzer mit der Verwendung eines Systems ist (Bailey, 2004). Durch die kontinuierliche Überprüfung der sachgemäßen Nutzung können mögliche Defizite identifiziert und gegebenenfalls durch Nachschulungen behoben werden (vgl. Kapitel 6.3.7).

7. Literaturverzeichnis

Abbas, H., Garberson, F., Liu-Mayo, S., Glover, E. & Wall, D. P. (2020). Multi-modular AI approach to streamline autism diagnosis in young children. *Scientific Reports*, *10*(1), 1–8. https://doi.org/10.1038/s41598-020-61213-w

Acatech. (2020). *Machine Learning in der Medizintechnik: Analyse und Handlungsempfehlungen* [Publikationsreihe Acatech POSITION]. https://www.acatech.de/publikation/machine-learning-in-der-medizintechnik/

Adler, R. (2019, 06. September). *Autonom oder vielleicht doch nur hochautomatisiert? Was ist z.B. der Unterschied zwischen autonomem Fahren und hochautomatisiertem Fahren?* Fraunhofer IESE. https://www.iese.fraunhofer.de/blog/autonom-oder-vielleicht-doch-nur-hochautomatisiert-was-ist-eigentlich-der-unterschied/

Aerzteblatt.de. (2020, 27. Oktober). *Vertrauliche Psychotherapiedaten in Finnland gehackt.* https://www.aerzteblatt.de/nachrichten/117742/Vertrauliche-Psychotherapiedaten-in-Finnland-gehackt

Al Meslamani, A. Z. (2023). Beyond implementation: The long-term economic impact of AI in healthcare. *Journal of Medical Economics*, *26*(1), 1566–1569. https://doi.org/10.1080/13696998.2023.2285186

Alt, R. & Zimmermann, H. D. (2021). The digital transformation of healthcare: An interview with Werner Dorfmeister. *Electronic Markets*, *31*(4), 895–899. https://doi.org/10.1007/s12525-021-00476-1

Andrews, J., Guyatt, G., Oxman, A. D., Alderson, P., Dahm, P., Falck-Ytter, Y., Nasser, M., Meerpohl, J., Post, P. N., Kunz, R., Brozek, J., Vist, G., Rind, D., Akl, E. A. & Schünemann, H. J. (2013). GRADE Guidelines: 14. Going from evidence to recommendations: The significance and presentation of recommendations. *Journal of Clinical Epidemiology*, *66*(7), 719–725. https://doi.org/10.1016/j.jclinepi.2012.03.013

Andreyeva, E., David, G. & Song, H. (2018, April). *The effects of home health visit length on hospital readmission* (NBER Working Paper Nummer 24566). National Bureau of Economic Research. https://econpapers.repec.org/RePEc:nbr:nberwo:24566

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016). *Machine bias*. Propublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A. & Mooney, C. (2021). Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review. *Applied Sciences (Switzerland)*, *11*(11), 1–23. https://doi.org/10.3390/app11115088

Antoniou, T. & Mamdani, M. (2021). Evaluation of machine learning solutions in medicine. *Canadian Medical Association Journal*, *193*(36), E1425–E1429. https://doi.org/10.1503/cmaj.210036

Arnold, M. H. (2021). Teasing out artificial intelligence in medicine: An ethical critique of artificial intelligence and machine learning in medicine. *Journal of Bioethical Inquiry*, *18*(1), 121–139. https://doi.org/10.1007/s11673-020-10080-1

Arshad, K., Schreiweis, B., Strodthoff, C. & Bergh, B. (2021). Überblick über ethische und soziale Herausforderungen im Kontext der klinischen Nutzung von KI-Systemen. *Zeitschrift für medizinische Ethik*, *67*(3), 297–308. https://doi.org/10.14623/zfme.2021.3.297-308

Asch, F. M., Mor-Avi, V., Rubenson, D., Goldstein, S., Saric, M., Mikati, I., Surette, S., Chaudhry, A., Poilvert, N., Hong, H., Horowitz, R., Park, D., Diaz-Gomez, J. L., Boesch, B., Nikravan, S., Liu, R. B., Philips, C., Thomas, J. D., Martin, R. P. & Lang, R. M. (2021). Deep learning–based automated echocardiographic quantification of left ventricular ejection fraction. *Circulation: Cardiovascular Imaging*, *June*, 528–537. https://doi.org/10.1161/CIRCIMAGING.120.012293

Bahner, J. E. (2008). *Übersteigertes Vertrauen in Automation: Der Einfluss von Fehlererfahrungen auf Complacency und Automation Bias* [Dissertation, Technische Universität Berlin]. Depositonce.tu-berlin.de. https://api-depositonce.tu-berlin.de/server/api/core/bitstreams/83dc4652-ffaf-477e-b3f2-6551491ec92a/content

Bailey, N. R. (2004). *The effects of operator trust, complacency potential, and task complexity on monitoring a highly reliable automated system* [Dissertation, Old Dominion University]. ODU Digital Commons. https://doi.org/10.25777/wfmq-tv11

Baio, J., Wiggins, L., Christensen, D. L., Maenner, M. J., Daniels, J., Warren, Z., Kurzius-Spencer, M., Zahorodny, W., Rosenberg, C. R., White, T., Durkin, M. S., Imm, P., Nikolaou, L., Yeargin-Allsopp, M., Lee, L. C., Harrington, R., Lopez, M., Fitzgerald, R. T., Hewitt, A., … Dowling, N. F. (2018). Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 Sites, United States, 2014. *MMWR Surveillance Summaries*, *67*(6). https://doi.org/10.15585/mmwr.ss6706a1

Baltzer, P. A. T. (2021). Künstliche Intelligenz in der Mammadiagnostik. *Der Radiologe*, *61*(2), 192–198. https://doi.org/10.1007/s00117-020-00802-2

Bao, J., Gilbertson, H. R., Gray, A. R., Munns, D., Howard, G., Petocz, P., Colagiuri, S. & Brand-Miller, J. C. (2011). Improving the estimation of mealtime insulin dose in adults with type 1 diabetes: The normal insulin demand for dose adjustment (NIDDA) study. *Diabetes Care*, *34*(10), 2146–2151. https://doi.org/10.2337/dc11-0567

Baratloo, A., Hosseini, M., Negida, A. & El Ashal, G. (2015). Part 1: Simple definition and calculation of accuracy, sensitivity and specificity. *Emergency (Tehran, Iran)*, *3*(2), 48–49. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4614595/pdf/emerg-3-48.pdf

Bates, D. W., Levine, D., Syrowatka, A., Kuznetsova, M., Craig, K. J. T., Rui, A., Jackson, G. P. & Rhee, K. (2021). The potential of artificial intelligence to improve patient safety: A scoping review. *Npj Digital Medicine*, *4*(1), 54. https://doi.org/10.1038/s41746-021-00423-6

Bates, S. M., Greer, I. A., Pabinger, I., Sofaer, S. & Hirsh, J. (2008). Venous thromboembolism, thrombophilia, antithrombotic therapy, and pregnancy: American College of Chest Physicians evidence-based clinical practice guidelines (8th edition). *Chest*, *133*(6 SUPPL. 6), 844S-886S. https://doi.org/10.1378/chest.08-0761

Beauchamp, T. L. & Childress, J. (2019). *Principles of biomedical ethics* (8. Aufl.). Oxford University Press.

Beede, E., Baylor, E., Hersch, F., Lurchenko, A., Wilcox, L., Ruamviboonsuk, P. & Vardoulakis, L. M. (2020). A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In R. Bernhaupt, F. Mueller, J. McGrenere, A. Cockburn, B. Pernille, Z. Shengdong (Hrsg.), *Conference*

on human factors in computing systems – proceedings (S. 1–12). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376718

Behrens, J., Wüller, H. & Remmers, H. (2021). Kontextspezifische Berücksichtigung ethischer Fragestellungen in der Entwicklung digitaler Lösungen. In M. Wiesche, I. M. Welpe, H. Remmers & H. Krcmar (Hrsg.), *Systematische Entwicklung von Dienstleistungsinnovationen* (S. 555–575). Springer Link. https://doi.org/10.1007/978-3-658-31768-3_27

Beil, M., Proft, I., van Heerden, D., Sviri, S. & van Heerden, P. V. (2019). Ethical considerations about artificial intelligence for prognostication in intensive care. *Intensive Care Medicine Experimental*, *7*(1), 70. https://doi.org/10.1186/s40635-019-0286-6

Berliner, D., Hänselmann, A. & Bauersachs, J. (2020). Therapie der Herzinsuffizienz mit reduzierter Ejektionsfraktion. *Deutsches Ärzteblatt International*, *117*(21), 376–385. https://www.aerzteblatt.de/int/article.asp?id=214094

Biecek, P. & Burzykowski, T. (2021). *Explanatory model analysis: Explore, explain and examine predictive models*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429027192

Binkley, C. E. (2021, 8. Februar). *The physician's conundrum: Assigning moral responsibility for medical artificial intelligence and machine learning*. Verdict. https://verdict.justia.com/2021/02/08/the-physicians-conundrum

BIS Research. (2021). Global clinical decision support systems (CDSS) market: Analysis and forecast, 2021-2030. *BIS Research*. https://bisresearch.com/industry-report/global-clinical-decision-support-systems-market.html

Bitterman, D. S., Aerts, H. J. W. L. & Mak, R. H. (2020). Approaching autonomy in medical artificial intelligence. *The Lancet Digital Health*, *2*(9), e447–e449. https://doi.org/10.1016/S2589-7500(20)30187-4

Bjerring, J. & Busch, J. (2021). Artificial intelligence and patient-centered decision-making. *Philosophy & Technology*, *34*, 349–371. https://doi.org/10.1007/s13347-019-00391-6

Bologheanu, R., Kapral, L., Laxar, D., Maleczek, M., Dibiasi, C., Zeiner, S., Agibetov, A., Ercole, A., Thoral, P., Elbers, P., Heitzinger, C. & Kimberger, O. (2023). Development of a reinforcement learning algorithm to optimize corticosteroid therapy in critically ill patients with sepsis. *Journal of Clinical Medicine*, *12*(4), 1513. https://doi.org/10.3390/jcm12041513

Bolte, G., Bunge, C., Hornberg, C., Köckler, H. & Mielck, A. (2012). Umweltgerechtigkeit durch Chancengleichheit bei Umwelt und Gesundheit: Eine Einführung in die Thematik und Zielsetzung dieses Buches. In G. Bolte, C. Bunge, C. Hornberg, H. Köckler & A. Mielck (Hrsg.), *Umweltgerechtigkeit. Chancengleichheit bei Umwelt und Gesundheit: Konzepte, Datenlage und Handlungsperspektiven* (S. 15–37). Hogrefe Verlag.

Braun, M., Hummel, P., Beck, S. & Dabrock, P. (2021). Primer on an ethics of AI-based decision support systems in the clinic. *Journal of Medical Ethics*, *47*(12), e3. https://doi.org/10.1136/medethics-2019-105860

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A. & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *68*(6), 394–424. https://doi.org/10.3322/caac.21492

Brodersen, J., Schwartz, L. M., Heneghan, C., O'Sullivan, J. W., Aronson, J. K. & Woloshin, S. (2018). Overdiagnosis: What it is and what it isn't. *BMJ Evidence-based Medicine 23*(1), 1–3. https://doi.org/10.1136/ebmed-2017-110886

Brodersen, J. & Siersma, V. D. (2013). Long-term psychosocial consequences of false-positive screening mammography. *Annals of Family Medicine*, *11*(2), 106–115. https://doi.org/10.1370/afm.1466

Bundesamt für Sicherheit in der Informationstechnik. (2022). Künstliche Intelligenz – wir bringen Ihnen die Technologie näher. *Bundesamt für Sicherheit in der Informationstechnik.* https://www.bsi.bund.de/DE/Themen/Verbraucherinnen-und-Verbraucher/Informationen-und-Empfehlungen/Technologien_sicher_gestalten/Kuenstliche-Intelligenz/kuenstliche-intelligenz_node.html

Bundesverband Informationswirtschaft, Telekommunikation und neue Medien & Deutsches Forschungszentrum für Künstliche Intelligenz.

(2017). *Künstliche Intelligenz: Wirtschaftliche Bedeutung, gesellschaftliche Herausforderungen, menschliche Verantwortung*. https://www.dfki.de/fileadmin/user_upload/import/9744_171012-KI-Gipfelpapier-online.pdf

Buolamwini, J. & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*, 1–15. https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Burdick, M. D., Skitka, L. J., Mosier, K. L. & Heers, S. (1996, August). The ameliorating effects of accountability on automation bias. In *Third annual symposium on human interaction with complex systems* (S. 142). IEEE Computer Society. https://doi.org/10.1109/HUICS.1996.549504

Cabitza, F., Rasoini, R. & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *Journal of the American Medical Association*, *318*(6), 517–518. https://doi.org/10.1001/jama.2017.7797

Carter, S. M., Rogers, W., Win, K. T., Frazer, H., Richards, B. & Houssami, N. (2020). The ethical, legal and social implications of using artificial intelligence systems in breast cancer care. *Breast*, *49(1)*, 25–32. https://doi.org/10.1016/j.breast.2019.10.001

Cavalot, F., Pagliarino, A., Valle, M., Di Martino, L., Bonomo, K., Massucco, P., Anfossi, G. & Trovati, M. (2011). Postprandial blood glucose predicts cardiovascular events and all-cause mortality in type 2 diabetes in a 14-year follow-up: Lessons from the San Luigi Gonzaga diabetes study. *Diabetes Care*, *34*(10), 2237–2243. https://doi.org/10.2337/dc10-2414

CBS News. (2019, 14. Februar). *Hackers are stealing millions of medical records – and selling them on the dark web.* https://www.cbsnews.com/news/hackers-steal-medical-records-sell-them-on-dark-web/

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T. & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, *28*(3), 231–237. https://doi.org/10.1136/bmjqs-2018-008370

Charles, C., Gafni, A. & Whelan, T. (1997). Shared decision-making in the medical encounter: What does it mean? (Or it takes, at least two to tango). *Social Science and Medicine*, *44*(5), 681–692. https://doi.org/10.1016/S0277-9536(96)00221-3

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

Collins, G. S., De Groot, J. A., Dutton, S., Omar, O., Shanyinde, M., Tajar, A., Voysey, M., Wharton, R., Yu, L. M., Moons, K. G. & Altman, D. G. (2014). External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. *BMC Medical Research Methodology*, *14*, 40. https://doi.org/10.1186/1471-2288-14-40

Datenschutzkonferenz. (2019). Positionspapier der DSK zu empfohlenen technischen und organisatorischen Maßnahmen bei der Entwicklung und dem Betrieb von KI-Systemen. https://www.datenschutzkonferenz-online.de/media/en/20191106_positionspapier_kuenstliche_intelligenz.pdf

Dawson, G. & Bernier, R. (2013). A quarter century of progress on the early detection and treatment of autism spectrum disorder. *Development and Psychopathology*, *25*(4pt2), 1455–1472. https://doi.org/10.1017/S0954579413000710

Dawson, G., Rogers, S., Munson, J., Smith, M., Winter, J., Greenson, J., Donaldson, A. & Varley, J. (2010). Randomized, controlled trial of an intervention for toddlers with autism: The Early Start Denver Model. *Pediatrics*, *125*(1), e17–e23. https://doi.org/10.1542/peds.2009-0958

De Miguel Beriain, I. (2020). Should we have a right to refuse diagnostics and treatment planning by artificial intelligence? *Medicine, Health Care and Philosophy*, *23*(2), 247–252. https://doi.org/10.1007/s11019-020-09939-2

Densen, P. (2011). Challenges and opportunities facing medical education. *Transactions of the American Clinical and Climatological Association*, *122*(319), 48–58. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3116346/

Derksen, F., Bensing, J. & Lagro-Janssen, A. (2013). Effectiveness of empathy in general practice: A systematic review. *British Journal of General Practice*, *63*(606), 76–84. https://doi.org/10.3399/bjgp13X660814

Deutsche AIDS-Gesellschaft (Hrsg.). (2020). *Deutsch-Österreichische Leitlinien zur antiretroviralen Therapie der HIV-1-Infektion* (9. Version).

https://daignet.de/media/filer_public/c7/2f/c72f0677-1677-4fc6-94ff-fb370a883811/deutsch_oesterreichische_leitlinien_zur_antiretroviralen_therapie_der_hiv_infektion.pdf

Deutsche Dermatologische Gesellschaft, Deutsche Krebsgesellschaft & Deutsche Krebshilfe. (2020). *S3-Leitlinie zur Diagnostik, Therapie und Nachsorge des Melanoms* (Version 3.3). https://www.leitlinienprogramm-onkologie.de/fileadmin/user_upload/Downloads/Leitlinien/Melanom/Melanom_Version_3/LL_Melanom_Langversion_3.3.pdf

Deutscher Ethikrat. (2018a). *Big Data und Gesundheit – Datensouveränität als informationelle Freiheitsgestaltung*. http://www.ethikrat.org/dateien/pdf/stellungnahme-big-data-und-gesundheit.pdf

Deutscher Ethikrat. (2018b). *Herausforderungen im Umgang mit seltenen Erkrankungen* [Ad-hoc-Empfehlung]. https://www.ethikrat.org/fileadmin/Publikationen/Ad-hoc-Empfehlungen/deutsch/herausforderungen-im-umgang-mit-seltenen-erkrankungen.pdf

Deutscher Ethikrat. (2023). *Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz*. https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf

Deutsche Gesellschaft für Gastroenterologie, Verdauungs- und Stoffwechselkrankheiten (Hrsg.). (2019). *S3-Leitlinie Kolorektales Karzinom – Kurzversion (Version 2.1)*. https://www.awmf.org/uploads/tx_szleitlinien/021-007OLk_S3_Kolorektales-Karzinom-KRK_2019-01.pdf

Deutsche Gesellschaft für Gynäkologie und Geburtshilfe & Deutsche Krebsgesellschaft (Hrsg.). (2021). Interdisziplinäre S3-Leitlinie für die Früherkennung, Diagnostik, Therapie und Nachsorge des Mammakarzinoms – Langversion (Version 4.4). https://www.leitlinienprogramm-onkologie.de/fileadmin/user_upload/Downloads/Leitlinien/Mammakarzinom_4_0/Version_4.4/LL_Mammakarzinom_Langversion_4.4.pdf

DeWeerdt, S. (2020, 10. November). It's time to talk about the carbon footprint of artificial intelligence. *Anthropocene*.

https://www.anthropocenemagazine.org/2020/11/time-to-talk-about-carbon-footprint-artificial-intelligence/

Dreiseitl, S. & Binder, M. (2005). Do physicians value decision support? A look at the effect of decision support systems on physician opinion. *Artificial Intelligence in Medicine*, *33*(1), 25–30. https://doi.org/10.1016/j.artmed.2004.07.007

Du-Harpur, X., Watt, F. M., Luscombe, N. M. & Lynch, M. D. (2020). What is AI? Applications of artificial intelligence to dermatology. *British Journal of Dermatology*, *183*(3), 423–430. https://doi.org/10.1111/bjd.18880

Eikelboom, J. & Hirsh, J. (2006). Monitoring unfractionated heparin with the aPTT: Time for a fresh look. *Thrombosis and Haemostasis*, *96(5)*, 547–552. https://doi.org/10.1160/TH06-05-0290

Elmore, J. G., Jackson, S. L., Abraham, L., Miglioretti, D. L., Carney, P. A., Geller, B. M., Yankaskas, B. C., Kerlikowske, K., Onega, T., Rosenberg, R. D., Sickles, E. A. & Buist, D. S. M. (2009). Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. *Radiology*, *253*(3), 641–651. https://doi.org/10.1148/radiol.2533082308

Elwyn, G., Frosch, D., Thomson, R., Joseph-Williams, N., Lloyd, A., Kinnersley, P., Cording, E., Tomson, D., Dodd, C., Rollnick, S., Edwards, A. & Barry, M. (2012). Shared decision making: A model for clinical practice. *Journal of General Internal Medicine*, *27*(10), 1361–1367. https://doi.org/10.1007/s11606-012-2077-6

Emanuel, E. J. & Emanuel, L. L. (1992). Four models of the physician-patient relationship. *Journal of the American Medical Association*, *267*(16), 2221–2226. https://doi.org/10.1001/jama.267.16.2221

Escandell-Montero, P., Chermisi, M., Martínez-Martínez, J. M., Gómez-Sanchis, J., Barbieri, C., Soria-Olivas, E., Mari, F., Vila-Francés, J., Stopper, A., Gatti, E. & Martín-Guerrero, J. D. (2014). Optimization of anemia treatment in hemodialysis patients via reinforcement learning. *Artificial Intelligence in Medicine*, *62*(1), 47–60. https://doi.org/10.1016/j.artmed.2014.07.004

Esmaeilzadeh, P. (2020). Use of AI-based tools for healthcare purposes: A survey study from consumers' perspectives. *BMC Medical Informatics and Decision Making*, *20*(1), 1–19. https://doi.org/10.1186/s12911-020-01191-1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P. & Vayena, E. (2018). AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Fox, J., Patkar, V., Chronakis, I. & Begent, R. (2009). From practice guidelines to clinical decision support: Closing the loop. *Journal of the Royal Society of Medicine*, *102*(11), 464–473. https://doi.org/10.1258/jrsm.2009.090010

Fulterer, R. (2021, 19. Januar). Künstliche Intelligenz in Medizinprodukten: Die Regulierung hinkt hinterher. *Neue Züricher Zeitung*. https://www.nzz.ch/technologie/kuenstliche-intelligenz-hokuenstliche-intelligenz-in-medizinprodukten-ki-hoert-covid-19ert-corona-und-andere-einsatzgebiete-in-der-medizin-ld.1596229

Funer, F. (2021). Arzt – Patient – Algorithmus: Ethische und kommunikative Erwägungen zu einer KI-gestützten Beziehung. *Zeitschrift für medizinische Ethik*, *67(3)*, 367–380. https://doi.org/10.14623/zfme.2021.3.367-380

Futoma, J., Simons, M., Panch, T., Doshi-Velez, F. & Celi, L. A. (2020). The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, *2*(9), e489–e492. https://doi.org/10.1016/S2589-7500(20)30186-2

Gartner. (o. D.). *Gartner Glossary: Big data*. https://www.gartner.com/en/information-technology/glossary/big-data

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E. & Ghassemi, M. (2021). Do as AI say: Susceptibility in deployment of clinical decision-aids. *Npj Digital Medicine*, *4*, 31. https://doi.org/10.1038/s41746-021-00385-9

Geller, B. M., Bogart, A., Carney, P. A., Sickles, E. A., Smith, R., Monsees, B., Bassett, L. W., Buist, D. M., Kerlikowske, K., Onega, T., Yankaskas, B. C., Haneuse, S., Hill, D., Wallis, M. G. & Miglioretti, D. (2014). Educational interventions to improve screening mammography interpretation: A randomized controlled trial. *American Journal of Roentgenology*, *202*(6), W586-W596. https://doi.org/10.2214/AJR.13.11147

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

Given-Wilson, R., Layer, G., Warren, M. & Gazet, J. C. (1997). False negative mammography: Causes and consequences. *Breast*, *6*(6), 361–366. https://doi.org/10.1016/S0960-9776(97)90693-7

Glanzmann, P. & Schiltenwolf, M. (2017). Haftung für Behandlungsfehler: Die Bedeutung des medizinischen Sachverständigengutachtens. *Deutsches Ärzteblatt International*, *114*(1–2), 21–23. https://www.aerzteblatt.de/int/article.asp?id=185397

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

Goff, D. C., Lloyd-Jones, D. M., Bennett, G., Coady, S., D'Agostino, R. B., Gibbons, R., Greenland, P., Lackland, D. T., Levy, D., O'Donnell, C. J., Robinson, J. G., Schwartz, J. S., Shero, S. T., Smith, S. C., Sorlie, P., Stone, N. J. & Wilson, P. W. F. (2014). 2013 ACC/AHA guideline on the assessment of cardiovascular risk. *Circulation*, *129*(25_suppl_2), S49–S73. https://doi.org/10.1161/01.cir.0000437741.48606.98

Goldhahn, J., Rampton, V. & Spinas, G. A. (2018). Could artificial intelligence make doctors obsolete? *BMJ*, *363*, k4563. https://doi.org/10.1136/bmj.k4563

Gómez-González, E., Gomez, E., Márquez-Rivas, J., Guerrero-Claro, M., Fernández-Lizaranzu, I., Relimpio-López, M. I., Dorado, M. E., Mayorga-Buiza, M. J., Izquierdo-Ayuso, G. & Capitán-Morales, L. (2020). *Artificial intelligence in medicine and healthcare: A review and classification of current and near-future applications and their ethical and social impact*. arXiv. http://arxiv.org/abs/2001.09778

Grafe, R. (2020). *Umweltgerechtigkeit: Aktualität und Zukunftsvision*. Springer Vieweg. https://link-springer-com.emedien.ub.uni-muenchen.de/book/10.1007/978-3-658-29083-2

Groß, D. & Schmidt, M. (2018). E-Health und Gesundheitsapps aus medizinethischer Sicht: Wollen wir alles, was wir können?

Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz, *61*(3), 349–357. https://doi.org/10.1007/s00103-018-2697-z

Grote, T. & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics*, *46*(3), 205–211. https://doi.org/10.1136/medethics-2019-105586

Haenssle, H. A., Fink, C., Toberer, F., Winkler, J., Stolz, W., Deinlein, T., Hofmann-Wellenhof, R., Lallas, A., Emmert, S., Buhl, T., Zutt, M., Blum, A., Abassi, M. S., Thomas, L., Tromme, I., Tschandl, P., Enk, A., Rosenberger, A., Alt, C., … Zukervar, P. (2020). Man against machine reloaded: Performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists under less artificial conditions. *Annals of Oncology*, *31*(1), 137–143. https://doi.org/10.1016/j.annonc.2019.10.013

Hale, C. (2018, 18. Juni). *FDA approves DreaMed's diabetes software for personalized insulin recommendations*. Fiercebiotech. https://www.fiercebiotech.com/medtech/fda-approves-dreamed-s-diabetes-software-for-personalized-insulin-recommendations

Hazlewood, G. S., Bombardier, C., Tomlinson, G. & Marshall, D. (2018). A Bayesian model that jointly considers comparative effectiveness research and patients' preferences may help inform GRADE recommendations: An application to rheumatoid arthritis treatment recommendations. *Journal of Clinical Epidemiology*, *93*, 56–65. https://doi.org/10.1016/j.jclinepi.2017.10.003

HeartFlow. (o. D.). *Our Technology Core*. https://www.heartflow.com/heartflow-ffrct-analysis/article/our-technology-core

Heidegger, M. (1962). *Die Technik und die Kehre*. Neske.

Hempel, J. M. & Pinto dos Santos, D. (2021). Structured reporting and artificial intelligence. *Der Radiologe*, *61*(11), 999–1004. https://doi.org/10.1007/s00117-021-00920-5

Henzel, M. (2019). *Analyse der Generalisierbarkeit von maschinell gelernten Algorithmen in Fahrerassistenzsystemen* [Dissertation, Technische Universität Darmstadt]. TUprints.

https://tuprints.ulb.tu-darmstadt.de/9246/1/20191103_Dissertation_veroeffentlichte_Version01.pdf

Hodgkin, P. K. (2016). The computer may be assessing you now, but who decided its values? *BMJ*, *355*, i6169. https://doi.org/10.1136/bmj.i6169

Hoff, T. (2011). Deskilling and adaptation among primary care physicians using two work innovations. *Health Care Management Review*, *36*(4), 338–348. https://doi.org/10.1097/HMR.0b013e31821826a1

Holzinger, A. (2018). Explainable AI (ex-AI). *Informatik-Spektrum*, *41*(2), 138–143. https://doi.org/10.1007/s00287-018-1102-5

Hopkins, J. J., Keane, P. A. & Balaskas, K. (2020). Delivering personalized medicine in retinal care: From artificial intelligence algorithms to clinical application. *Current Opinion in Ophthalmology*, *31*(5), 329–336. https://doi.org/10.1097/ICU.0000000000000677

Houssami, N., Kirkpatrick-Jones, G., Noguchi, N. & Lee, C. I. (2019). Artificial intelligence (AI) for the early detection of breast cancer: A scoping review to assess AI's potential in breast screening practice. *Expert Review of Medical Devices*, *16*(5), 351–362. https://doi.org/10.1080/17434440.2019.1610387

Huang, S., Yang, J., Fong, S. & Zhao, Q. (2020). Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Letters*, *471*, 61–71. https://doi.org/https://doi.org/10.1016/j.canlet.2019.12.007

Illmer, A. (2021, 5. Januar). *Singapore reveals Covid privacy data available to police*. BBC. https://www.bbc.com/news/world-asia-55541001

Jankovic, I. & Chen, J. H. (2020). Clinical decision support and implications for the clinician burnout crisis. *Yearbook of Medical Informatics*, *29*(1), 145–154. https://doi.org/10.1055/s-0040-1701986

Jenkins, D. J., Wolever, T. M., Taylor, R. H., Barker, H., Fielden, H., Baldwin, J. M., Bowling, A. C., Newman, H. C., Jenkins, A. L. & Goff, D. V. (1981). Glycemic index of foods: A physiological basis for carbohydrate exchange. *The American Journal of Clinical Nutrition*, *34*(3), 362–366. https://doi.org/10.1093/ajcn/34.3.362

Jie, Z., Zhiying, Z. & Li, L. (2021). A meta-analysis of Watson for Oncology in clinical application. *Scientific Reports*, *11*, 5792. https://doi.org/10.1038/s41598-021-84973-5

Kächele, M., Schels, M. & Schwenker, F. (2014). Inferring depression and affect from application dependent meta knowledge. In *AVEC 2014: Proceedings of the 4th international workshop on audio/visual emotion challenge* (S. 41–48). Association for Computing Machinery. https://doi.org/10.1145/2661806.2661813

Kaissis, G. A., Makowski, M. R., Rückert, D. & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, *2*(6), 305–311. https://doi.org/10.1038/s42256-020-0186-1

Karches, K. E. (2018). Against the iDoctor: Why artificial intelligence should not replace physician judgment. *Theoretical Medicine and Bioethics*, *39*(2), 91–110. https://doi.org/10.1007/s11017-018-9442-3

Kather, J. N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C. A., Gaiser, T., Marx, A., Valous, N. A., Ferber, D., Jansen, L., Reyes-Aldasoro, C. C., Zörnig, I., Jäger, D., Brenner, H., Chang-Claude, J., Hoffmeister, M. & Halama, N. (2019). Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine*, *16*(1), 1–22. https://doi.org/10.1371/journal.pmed.1002730

Kaul, V., Enslin, S. & Gross, S. A. (2020). History of artificial intelligence in medicine. *Gastrointestinal Endoscopy*, *92*(4), 807–812. https://doi.org/10.1016/j.gie.2020.06.040

Kawamoto, K., Houlihan, C. A., Balas, E. A. & Lobach, D. F. (2005). Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success. *British Medical Journal (Clinical Research Edition)*, *330*(7494), 765. https://doi.org/10.1136/bmj.38398.500764.8F

Kazzazi, F. (2021). The automation of doctors and machines: A classification for AI in medicine (ADAM framework). *Future Healthcare Journal*, *8*(2), e257–e262. https://doi.org/10.7861/fhj.2020-0189

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, *17*(1), 1–9. https://doi.org/10.1186/s12916-019-1426-2

Kendale, S., Kulkarni, P., Rosenberg, A. D. & Wang, J. (2018). Supervised machine-learning predictive analytics for prediction of postinduction hypotension. *Anesthesiology*, *129*(4), 675–688. https://doi.org/10.1097/ALN.0000000000002374

Keren, A. (2007). Epistemic authority, testimony and the transmission of knowledge. *Episteme*, *4*(3), 368–381. https://doi.org/10.3366/E1742360007000147

Kidholm, K., Ekeland, A. G., Jensen, L. K., Rasmussen, J., Pedersen, C. D., Bowes, A., Flottorp, S. A. & Bech, M. (2012). A model for assessment of telemedicine applications: MAST. *International Journal of Technology Assessment in Health Care*, *28*(1), 44–51. https://doi.org/10.1017/S0266462311000638

Kienle, G. S. (2008). Evidenzbasierte Medizin und ärztliche Therapiefreiheit: Vom Durchschnitt zum Individuum. *Deutsches Ärzteblatt*, *105*(25), 1381–1385. https://www.aerzteblatt.de/archiv/60581/Evidenzbasierte-Medizin-und-aerztliche-Therapiefreiheit-Vom-Durchschnitt-zum-Individuum

Kim, B., Khanna, R. & Koyejo, O. (2016). Examples are not enough, learn to criticize! Criticism for interpretability. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (Hrsg.), *Advances in neural information processing systems 29 (NIPS 2016)* (S. 2288–2296). Association for Computing Machinery. https://papers.nips.cc/paper_files/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf

Kirkpatrick, P. (2005). New clues in the acetaminophen mystery. *Nature Reviews Drug Discovery*, *4*(11), 883. https://doi.org/10.1038/nrd1887

Kleeberg, J. (2015). Eide Und Bekenntnisse in der Medizin. S. Karger.

Krumm, S., Frederking, A., Schaat, D. S. & Schürholz, D. M. (2019). *Anwendung Künstlicher Intelligenz in der Medizin.* Begleitforschung Smart Service Welt II. https://www.digitale-technologien.de/DT/Redaktion/DE/Downloads/Publikation/SSW_Policy_Paper_KI_Medizin.pdf?__blob=publicationFile&v=6

Kulikowski, C. A. (2019). Beginnings of artificial intelligence in medicine (AIM): Computational artifice assisting scientific inquiry and clinical art - with reflections on present AIM challenges. *Yearbook of Medical Informatics*, *28*(1), 249–256. https://doi.org/10.1055/s-0039-1677895

Kusunose, K., Shibayama, K., Iwano, H., Izumo, M., Kagiyama, N., Kurosawa, K., Mihara, H., Oe, H., Onishi, T., Onishi, T., Ota, M., Sasaki, S., Shiina, Y., Tsuruta, H. & Tanaka, H. (2018). Reduced variability of visual left ventricular ejection fraction assessment with reference images: The Japanese Association of Young Echocardiography Fellows multicenter study. *Journal of Cardiology*, *72*(1), 74–80. https://doi.org/10.1016/j.jjcc.2018.01.007

Kwan, J. L., Lo, L., Ferguson, J., Goldberg, H., Diaz-Martinez, J. P., Tomlinson, G., Grimshaw, J. M. & Shojania, K. G. (2020). Computerised clinical decision support systems and absolute improvements in care: Meta-analysis of controlled clinical trials. *BMJ*, *370*. https://doi.org/10.1136/bmj.m3216

Landefeld, C. S., Cook, E. F., Flatley, M., Weisberg, M. & Goldman, L. (1987). Identification and preliminary validation of predictors of major bleeding in hospitalized patients starting anticoagulant therapy. *The American Journal of Medicine*, *82*(4), 703–713. https://doi.org/10.1016/0002-9343(87)90004-0

Langanke, M., Liedtke, W. & Buyx, A. (2017). Patients' responsibility for their health. In T. Schramme & S. Edwards (Hrsg.), *Handbook of the philosophy of medicine* (S. 619–640). Springer Netherlands. http://link.springer.com/10.1007/978-94-017-8688-1_22

Lee, C. S., & Lee, A. Y. (2020). Clinical applications of continual learning machine learning. *The Lancet. Digital health*, *2*(6), e279–e281. https://doi.org/10.1016/S2589-7500(20)30102-3

Lee, L. (2017). Ethics and subsequent use of electronic health record data. *Journal of Biomedical Informatics*, *71*, 143-146 . https://doi.org/10.1016/j.jbi.2017.05.022

Li, J.-P. O., Liu, H., Ting, D. S. J., Jeon, S., Chan, R. V. P., Kim, J. E., Sim, D. A., Thomas, P. B. M., Lin, H., Chen, Y., Sakomoto, T., Loewenstein, A., Lam, D. S. C., Pasquale, L. R., Wong, T. Y., Lam, L. A. & Ting, D. S. W. (2021). Digital technology, tele-medicine and artificial intelligence in ophthalmology: A global perspective.

Progress in Retinal and Eye Research, *82*, 100900. https://doi.org/10.1016/j.preteyeres.2020.100900

Liedtke, W. & Langanke, M. (2021). Der Einsatz von IT-basierten Decision-Support-Systemen in der medizinischen Versorgung aus verantwortungsethischer Sicht. *Zeitschrift für medizinische Ethik 67(3)*, 279–296. https://doi.org/10.14623/zfme.2021.3.279-296

Lilienfeld, S. O. & Lynn, S. J. (2014). Errors/biases in clinical decision making. *The Encyclopedia of Clinical Psychology*. Abgerufen am 10. April 2024, von https://doi.org/10.1002/9781118625392.wbecp567

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, *61*(10), 35–43. https://doi.org/10.1145/3233231

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J. R., Schmid, M. K., Balaskas, K., Topol, E. J., Bachmann, L. M., Keane, P. A. & Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *The Lancet Digital Health*, *1*(6), e271–e297. https://doi.org/10.1016/S2589-7500(19)30123-2

Liu, X., Keane, P. A. & Denniston, A. K. (2018). Time to regenerate: The doctor in the age of artificial intelligence. *Journal of the Royal Society of Medicine*, *111*(4), 113–116. https://doi.org/10.1177/0141076818762648

Loder, J. & Nicholas, L. (2018). *Confronting Dr Robot: Creating a people-powered future for AI in health* [Forschungsbericht]. Nesta Health Lab. https://www.nesta.org.uk/report/confronting-dr-robot/

Lohr, S. (2022, 21. Januar). IBM is selling off Watson Health to a private equity firm. *New York Times*. https://www.nytimes.com/2022/01/21/business/ibm-watson-health.html

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

Lyell, D. & Coiera, E. (2017). Automation bias and verification complexity: A systematic review. *Journal of the American Medical Informatics Association*, *24*(2), 423–431. https://doi.org/10.1093/jamia/ocw105

Mai, N. D., Lee, B. G. & Chung, W. Y. (2021). Affective computing on machine learning-based emotion recognition using a self-made eeg device. *Sensors*, *21*(15), 1–19. https://doi.org/10.3390/s21155135

Mamede, S., van Gog, T., van den Berge, K., Rikers, R. M. J. P., van Saase, J. L. C. M., van Guldener, C. & Schmidt, H. G. (2010). Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. *Journal of the American Medical Association*, *304*(11), 1198–1203. https://doi.org/10.1001/jama.2010.1276

Mantelakis, A., Assael, Y., Sorooshian, P. & Khajuria, A. (2021). Machine learning demonstrates high accuracy for disease diagnosis and prognosis in plastic surgery. *Plastic and Reconstructive Surgery - Global Open*, *9*(6), e3638. https://doi.org/10.1097/GOX.0000000000003638

Manzeschke, A. (2015). MEESTAR: Ein Modell angewandter Ethik im Bereich assistiver Technologien. In K. Weber, D. Frommfeld, H. Fangerau & A. Manzeschke (Hrsg.), *Technisierung des Alltags: Beitrag für ein gutes Leben?* (S. 263–283). Franz Steiner Verlag.

Manzey, D. (2008). Systemgestaltung und Automatisierung. In P. Badke-Schaub, G. Hofinger & K. Lauche (Hrsg.), *Human factors* (S. 307–324). Springer Berlin Heidelberg. http://link.springer.com/10.1007/978-3-540-72321-9_19

Marckmann, G. (2003). *Diagnose per Computer? Eine ethische Bewertung medizinischer Expertensysteme.* Deutscher Ärzteverlag.

Marckmann, G. (2016). Ethische Aspekte von eHealth. In F. Fischer & A. Krämer (Hrsg.), *eHealth in Deutschland* (S. 83–99). Springer. http://link.springer.com/10.1007/978-3-662-49504-9_4

Marckmann, G. (2020). Ethical implications of digital public health. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, *63*(2), 199–205. https://doi.org/10.1007/s00103-019-03091-w

Markus, A. F., Kors, J. A. & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, *113*, 103655. https://doi.org/10.1016/j.jbi.2020.103655

Marr, B. (2017, 20. Januar). First FDA approval for clinical cloud-based deep learning in healthcare. *Forbes*. https://www.forbes.com/sites/bernardmarr/2017/01/20/first-fda-approval-for-clinical-cloud-based-deep-learning-in-healthcare/?sh=2f7ef2a4161c

Mbirimtengerenji, N. D. (2007). Is HIV/AIDS epidemic outcome of poverty in sub-saharan Africa? *Croatian Medical Journal*, *48*(5), 605–617. https://doi.org/10.1016/j.ajog.2007.03.032

McDougall, R. J. (2019). Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics*, *45*(3), 156–160. https://doi.org/10.1136/medethics-2018-105118

McGuirl, J. M. & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors*, *48*(4), 656–665. https://doi.org/10.1518/001872006779166334

McLennan, S., Fiske, A., Celi, L. A., Müller, R., Harder, J., Ritt, K., Haddadin, S. & Buyx, A. (2020). An embedded ethics approach for AI development. *Nature Machine Intelligence*, *2*(9), 488–490. https://doi.org/10.1038/s42256-020-0214-1

Mesko, B. & Győrffy, Z. (2019). The rise of the empowered physician in the digital health era: Viewpoint. *Journal of Medical Internet Research*, *21*(3), e12490–e12490. https://doi.org/10.2196/12490

Meyer, A., Zverinski, D., Pfahringer, B., Kempfert, J., Kuehne, T., Sündermann, S. H., Stamm, C., Hofmann, T., Falk, V. & Eickhoff, C. (2018). Machine learning for real-time prediction of complications in critical care: A retrospective study. *The Lancet Respiratory Medicine*, *6*(12), 905–914. https://doi.org/10.1016/S2213-2600(18)30300-X

Mitchell, W. G., Dee, E. C. & Celi, L. A. (2021). Generalisability through local validation: Overcoming barriers due to data disparity in healthcare. *BMC Ophthalmology*, *21*(1), 4–6. https://doi.org/10.1186/s12886-021-01992-6

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Moray, N., Inagaki, T. & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology: Applied, 6*(1), 44–58. https://doi.org/10.1037/1076-898X.6.1.44

Morley, J., Morton, C., Karpathakis, K. (2021). *Towards a framework for evaluating the safety, acceptability and efficacy of AI systems for health: An initial synthesis*. JMIR Preprints. https://doi.org/10.2196/preprints.31654

Mosier, K. L., Skitka, L. J., Heers, S. & Burdick, M. (1998). Automation bias: Decision making and performance in high-tech cockpits. *International Journal of Aviation Psychology*, *8*(1), 47–63. https://doi.org/10.1207/s15327108ijap0801_3

Moumjid, N., Gafni, A., Brémond, A. & Carrère, M. O. (2007). Shared decision making in the medical encounter: Are we all talking about the same thing? *Medical Decision Making*, *27*(5), 539–546. https://doi.org/10.1177/0272989X07306779

Müller-Quade, J., Damm, W., Holz, T., Houdeau, D., Schauf, T., Schindler, W., Neumuth, T. & Schapranow, M. (2020). *Sichere KI-Systeme für die Medizin* [Whitepaper]. Plattform Lernende Systeme. https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_6_Whitepaper_07042020.pdf

Navarrete-Welton, A. J. & Hashimoto, D. A. (2020). Current applications of artificial intelligence for intraoperative decision support in surgery. *Frontiers of Medicine*, *14*(4), 369–381. https://doi.org/10.1007/s11684-020-0784-7

Nelson, H. D., O'Meara, E. S., Kerlikowske, K., Balch, S. & Miglioretti, D. (2016). Factors associated with rates of false-positive and false-negative results from digital mammography screening: An analysis of registry data. *Annals of Internal Medicine*, *164*(4), 226-235. https://doi.org/10.7326/M15-0971

Nemati, S., Ghassemi, M. M. & Clifford, G. D. (2016). Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, *38*, 2978–2981. https://doi.org/10.1109/EMBC.2016.7591355

Neumuth, T. (2020). Künstliche Intelligenz – Anwendungsbereiche in der Onkologie. *Forum*, *35*(2), 104–108. https://doi.org/10.1007/s12312-019-00734-6

Norman, G. R., Monteiro, S. D., Sherbino, J., Ilgen, J. S., Schmidt, H. G. & Mamede, S. (2017). The causes of errors in clinical reasoning: Cognitive biases, knowledge deficits, and dual process thinking. *Academic Medicine*, *92*(1), 23–30. https://doi.org/10.1097/ACM.0000000000001421

Nsier, H. (2023). Ethical considerations of using artificial intelligence to drive clinical decision support in pediatric medical settings. *Pediatric Research*, *94*, 10–11. https://doi.org/10.1038/s41390-022-02452-7

Nyrup, R., Whittlestone, J. & Cave, S. (2019). *Why value judgements should not be automated.* Apollo - University of Cambridge Repository. https://doi.org/10.17863/CAM.41552

O'Leary, L. (2022, 31. Januar). *How IBM's Watson went from the future of health care to sold off for parts*. Slate. https://slate.com/technology/2022/01/ibm-watson-health-failure-artificial-intelligence.html

O'Sullivan, C. (2020, 17. September). *Interpretable vs explainable machine learning*. Towards Data Science. https://towardsdatascience.com/interperable-vs-explainable-machine-learning-1fa525e12f48

Open Roboethics Institute. (2020). *Foresight into AI ethics in healthcare (FAIE-H): A toolkit for creating an ethics roadmap for your healthcare AI project* [Digitale Broschüre]. https://openroboethics.org/wp-content/uploads/2021/07/ORI-Foresight-into-Artificial-Intelligence-Ethics-Launch-V1.pdf

Orwat, C. (2019). *Diskriminierungsrisiken durch Verwendung von Algorithmen.* Antidiskriminierungsstelle des Bundes. https://www.antidiskriminierungsstelle.de/SharedDocs/downloads/DE/publikationen/Expertisen/studie_diskriminierungsrisiken_durch_verwendung_von_algorithmen.pdf?__blob=publicationFile&v=3

Ott, K. (1997). *Ipso facto: Zur ethischen Begründung normativer Implikate wissenschaftlicher Praxis*. Suhrkamp.

Pan, H., Tao, J., Qian, M., Zhou, W., Qian, Y., Xie, H., Jing, S., Xu, T., Zhang, X., Dai, Z., You, M., Liu, Y., Liu, X. & Wang, S. (2019). Concordance assessment of Watson for Oncology in breast cancer chemotherapy: First China experience. *Translational Cancer Research*, *8*(2), 389–401. https://doi.org/10.21037/tcr.2019.01.34

Panesar, S., Cagle, Y., Chander, D., Morey, J., Fernandez-Miranda, J. & Kliot, M. (2019). Artificial intelligence and the future of surgical robotics. *Annals of Surgery*, *270*(2), 223–226. https://doi.org/10.1097/SLA.0000000000003262

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

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

Paulus, J. K. & Kent, D. M. (2020). Predictably unequal: Understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *Npj Digital Medicine*, *3*(1), 1–8. https://doi.org/10.1038/s41746-020-0304-9

Peiffer-Smadja, N., Rawson, T. M., Ahmad, R., Buchard, A., Pantelis, G., Lescure, F. X., Birgand, G. & Holmes, A. H. (2020). Machine learning for clinical decision support in infectious diseases: A narrative review of current applications. *Clinical Microbiology and Infection*, *26*(5), 584–595. https://doi.org/10.1016/j.cmi.2019.09.009

Plattform Lernende Systeme (2019). *Lernende Systeme im Gesundheitswesen: Bericht der Arbeitsgruppe Gesundheit, Medizintechnik, Pflege*. https://www.plattform-lernende-systeme.de/publikationen-details/lernende-systeme-im-gesundheitswesen.html

Ploug, T. & Holm, S. (2020a). The four dimensions of contestable AI diagnostics - a patient-centric approach to explainable AI. *Artificial Intelligence in Medicine*, *107*, 101901. https://doi.org/10.1016/j.artmed.2020.101901

Ploug, T. & Holm, S. (2020b). The right to refuse diagnostics and treatment planning by artificial intelligence. *Medicine, Health Care and Philosophy*, *23*(1), 107–114. https://doi.org/10.1007/s11019-019-09912-8

Poretschkin, M., Schmitz, A., Akila, M., Adilova, L., Becker, D., Cremers, A. B., ... & Wrobel, S. (2021). *Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz: KI-Prüfkatalog*. Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS. https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkatalog/202107_KI-Pruefkatalog.pdf

Prinz, T. (2021, 20. Oktober). *Software als Medizinprodukt: Bias bei Künstliche Intelligenz (KI)-basierter Software berücksichtigen.* VDE Health. https://www.vde.com/topics-de/health/beratung/software-als-medizinprodukt

Prinz, T. (2023, 05. Juni). *CE conformity for AI-based software in medicine with BAIM.* VDE Health. https://www.vde.com/topics-en/health/consulting/ce-conformity-for-ai-based-software-in-medicine-with-baim

Qiu, J., Wu, Q., Ding, G., Xu, Y. & Feng, S. (2016). A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*, 2016, 67. https://doi.org/10.1186/s13634-016-0355-x

Rajput, V. K., Dowie, J. & Kaltoft, M. K. (2020). Are clinical decision support systems compatible with patient-centred care? *Studies in Health Technology and Informatics*, *270*, 532–536. https://doi.org/10.3233/SHTI200217

Rawls, J. (2001). *Justice as fairness: A restatement* (E. Kelly (Hrsg.)). Harvard University Press. https://doi.org/10.2307/j.ctv31xf5v0

Reinhardt, K., Scriba, P. C., Dietel, M. & Kroemer, H. K. (2020). Präzisionsmedizin: Bewertung unter medizinisch-wissenschaftlichen und ökonomischen Aspekten [Stellungnahme des Wissenschaftlichen Beirats der Bundesärztekammer]. *Deutsches Ärzteblatt*, *117*(22-23). https://doi.org/10.3238/baek_sn_praezision_2020

Revell, A. D., Wang, D., Perez-Elias, M. J., Wood, R., Cogill, D., Tempelman, H., Hamers, R. L., Reiss, P., Van Sighem, A. I., Rehm, C. A., Pozniak, A., Montaner, J. S. G., Clifford Lane, H. & Larder, B. A. (2018). 2018 update to the HIV-TRePS system: The development of new computational models to predict HIV treatment

outcomes, with or without a genotype, with enhanced usability for low-income settings. *Journal of Antimicrobial Chemotherapy*, *73*(8), 2186–2196. https://doi.org/10.1093/jac/dky179

Riedl, D. & Schüßler, G. (2017). The influence of doctor-patient communication on health outcomes: A systematic review. *Zeitschrift für Psychosomatische Medizin und Psychotherapie*, *63*(2), 131–150. https://doi.org/10.13109/zptm.2017.63.2.131

Rohde, F., Wagner, J., Reinhard, P., Petschow, U., Meyer, A., Voß, M. & Mollen, A. (2021). Nachhaltigkeitskriterien für künstliche Intelligenz: Entwicklung eines Kriterien- und Indikatorensets für die Nachhaltigkeitsbewertung von KI-Systemen entlang des Lebenszyklus (IÖW-Schriftenreihe 220/21). Institut für ökologische Wirtschaftsforschung. https://www.ioew.de/publikation/nachhaltigkeitskriterien_fuer_kuenstliche_intelligenz

Rosengrün, S. (2021). *Künstliche Intelligenz zur Einführung*. Junius Verlag.

Ruck, A., Wagner-Bondorf, S. & Lowe, C. (2016). *First draft of guidelines: EU guidelines on assessment of the reliability of mobile health applications*. European Commission. https://www.twobirds.com/-/media/pdfs/news/articles/2016/firstdraftguidelinesandannexes.pdf?la=en&hash=196705F551B3B2AAE686F9672BD6C20BB7E221CD

Russell, S. J. & Norvig, P. (2012). *Künstliche Intelligenz: Ein moderner Ansatz*. Pearson.

Saber Tehrani, A. S., Lee, H., Mathews, S. C., Shore, A., Makary, M. A., Pronovost, P. J. & Newman-Toker, D. E. (2013). 25-year summary of US malpractice claims for diagnostic errors 1986–2010: An analysis from the National Practitioner Data Bank. *BMJ Quality & Safety*, *22*(8), 672 – 680. https://doi.org/10.1136/bmjqs-2012-001550

Sachverständigenrat Gesundheit & Pflege. (2021). *Digitalisierung für Gesundheit: Ziele und Rahmenbedingungen eines dynamisch lernenden Gesundheitssystems.* https://www.svr-gesundheit.de/fileadmin/Gutachten/Gutachten_2021/SVR_Gutachten_2021.pdf

Sarter, N. B. & Schroeder, B. (2001). Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. *Human Factors*, *43*(4), 573–583. https://doi.org/10.1518/001872001775870403

Schaaf, N., Wiedenroth, S. J. & Wagner, P. (2021). *Erklärbare KI in der Praxis* (T. Bauernhansl, M. Huber & P. Wagner (Hrsg.)). Fraunhofer IPA. https://doi.org/10.24406/publica-fhg-30084

Schmietow, B. & Marckmann, G. (2019). Mobile health ethics and the expanding role of autonomy. *Medicine, Health Care and Philosophy*, *22*(4), 623–630. https://doi.org/10.1007/s11019-019-09900-y

Schwab, K. (2016, 14. Januar). *The fourth industrial revolution: What it means, how to respond*. World Economic Forum. https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/

Sechopoulos, I., Teuwen, J. & Mann, R. (2021). Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art. *Seminars in Cancer Biology*, *72*, 214–225. https://doi.org/10.1016/j.semcancer.2020.06.002

Shaikh, F., Dehmeshki, J., Bisdas, S., Roettger-Dupont, D., Kubassova, O., Aziz, M. & Awan, O. (2021). Artificial intelligence-based clinical decision support systems using advanced medical imaging and radiomics. *Current Problems in Diagnostic Radiology*, *50*(2), 262–267. https://doi.org/10.1067/j.cpradiol.2020.05.006

Sheng, Y., Li, T., Yoo, S., Yin, F. F., Blitzblau, R., Horton, J. K., Ge, Y. & Wu, Q. J. (2019). Automatic planning of whole breast radiation therapy using machine learning models. *Frontiers in Oncology*, *9*, 750. https://doi.org/10.3389/fonc.2019.00750

Shortliffe, E. H. & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. *Journal of the American Medical Association*, *320*(21), 2199–2200. https://doi.org/10.1001/jama.2018.17163

Sikma, T., Edelenbosch, R. & Verhoef, P. (2020, November). *The use of AI in healthcare: a focus on clinical decision support systems* (Case Study 8). RECIPES. https://recipes-project.eu/sites/default/files/2020-11/D2_3_AI_In_Healthcare%28CDSS%29_HarvardStyle.pdf

Skitka, L. J., Mosier, K. & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human Computer Studies*, *52*(4), 701–717. https://doi.org/10.1006/ijhc.1999.0349

Smith, H., Birchley, G. & Ives, J. (2024). Artificial intelligence in clinical decision-making: Rethinking personal moral responsibility. *Bioethics*, *38*(1), 78–86. https://doi.org/10.1111/bioe.13222

Somashekhar, S. P., Sepúlveda, M. J., Puglielli, S., Norden, A. D., Shortliffe, E. H., Rohit Kumar, C., Rauthan, A., Arun Kumar, N., Patil, P., Rhee, K. & Ramya, Y. (2018). Watson for Oncology and breast cancer treatment recommendations: Agreement with an expert multidisciplinary tumor board. *Annals of Oncology*, *29*(2), 418–423. https://doi.org/10.1093/annonc/mdx781

Sonar, A. & Weber, K. (2020). KI gestern und heute: Einsichten aus der Frühgeschichte der KI für aktuelle ethische Überlegungen zum Einsatz von KI in der Medizin. *Arbeit*, *29*(2), 105-122. https://doi.org/10.1515/arbeit-2020-0009

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, *24*(1), 62–77. http://www.jstor.org.emedien.ub.uni-muenchen.de/stable/24355087

Steger, A. (2019, 30. Oktober). *What happens to stolen healthcare data?* HealthTech Magazine. https://healthtechmagazine.net/article/2019/10/what-happens-stolen-healthcare-data-perfcon

Stein, H., Fischer, S. & Pohlink, C. (2019). *Blick in die Blackbox: Nachvollziehbarkeit von KI-Algorithmen in der Praxis*. Bundesverband Informationswirtschaft, Telekommunikation und neue Medien. https://www.bitkom.org/Bitkom/Publikationen/Blick-in-die-Blackbox-Nachvollziehbarkeit-von-KI-Algorithmen-in-der-Praxis

Strech, D., Kielmansegg, S. G. von, Zenker, S., Krawczak, M. & Semler, S. (2020). *„Datenspende“ – Bedarf für die Forschung, ethische Bewertung, rechtliche, informationstechnologische und organisatorische Rahmenbedingungen. Bundesministerium für Gesundheit* [Wissenschaftliches Gutachten, erstellt für das Bundesministerium für Gesundheit]. https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/5_Publikationen/Ministerium/Berichte/Gutachten_Datenspende.pdf

Strickland, E. (2019). IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum*, *56*(4), 24–31. https://doi.org/10.1109/MSPEC.2019.8678513

Strubell, E., Ganesh, A. & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In A. Korhonen, D. Traum & L. Màrquez (Hrsg.), *Proceedings of the 57th annual meeting of the Association for Computational Linguistics* (S. 3645–3650). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1355

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N. & Kroeker, K. I. (2020). An overview of clinical decision support systems: Benefits, risks, and strategies for success. *Npj Digital Medicine*, *3*(1), 17. https://doi.org/10.1038/s41746-020-0221-y

Svoboda, E. (2020). Artificial intelligence is improving the detection of lung cancer. *Nature*, *587*(7834), S20–S22. https://doi.org/10.1038/d41586-020-03157-9

Thavendiranathan, P., Popović, Z. B., Flamm, S. D., Dahiya, A., Grimm, R. A. & Marwick, T. H. (2013). Improved interobserver variability and accuracy of echocardiographic visual left ventricular ejection fraction assessment through a self-directed learning program using cardiac magnetic resonance images. *Journal of the American Society of Echocardiography*, *26*(11), 1267–1273. https://doi.org/10.1016/j.echo.2013.07.017

Thomas, E. C., Bass, S. B. & Siminoff, L. A. (2021). Beyond rationality: Expanding the practice of shared decision making in modern medicine. *Social Science and Medicine*, *277*(März), 113900. https://doi.org/10.1016/j.socscimed.2021.113900

Tian, Z., Liu, L. & Fei, B. (2017). Deep convolutional neural network for prostate MR segmentation. In R. J. Webster & B. Fei (Hrsg.), *Medical imaging 2017: Image-guided procedures, robotic interventions, and modeling*. SPIE, the International Society for Optics and Photonics. https://doi.org/10.1117/12.2254621

Topol, E. J. (2019). *Deep medicine*. Basic Books.

Tosam, M. J., Chi, P. C., Munung, N. S., Oukem-Boyer, O. O. M. & Tangwa, G. B. (2018). Global health inequalities and the need for solidarity: A view from the global south. *Developing World Bioethics*, *18*(3), 241-249.

Triberti, S., Durosini, I. & Pravettoni, G. (2020). A “third wheel” effect in health decision making involving artificial entities: A psychological perspective. *Frontiers in Public Health*, *8*(April), 1–9. https://doi.org/10.3389/fpubh.2020.00117

Tupasela, A. & Di Nucci, E. (2020). Concordance as evidence in the Watson for Oncology decision-support system. *AI and Society*, *35*(4), 811–818. https://doi.org/10.1007/s00146-020-00945-9

Rheinische Friedrich-Wilhelms-Universität Bonn. (2021, 12. Juli). *Medizinstudierende lernen Umgang mit KI - Lehr-Projekt KI-LAURA wird vom BMBF gefördert*. Uni Bonn. https://www.uni-bonn.de/de/neues/170-2021

Ursin, F., Timmermann, C. & Steger, F. (2021). Explicability of artificial intelligence in radiology: Is a fifth bioethical principle conceptually necessary? *Bioethics*, *36*(2), 143–153. https://doi.org/10.1111/bioe.12918

Van Baalen, S., Boon, M. & Verhoef, P. (2021). From clinical decision support to clinical reasoning support systems. *Journal of Evaluation in Clinical Practice*, *27*(3), 520–528. https://doi.org/10.1111/jep.13541

Van der Sijs, H., Aarts, J., Vulto, A. & Berg, M. (2006). Overriding of drug safety alerts in computerized physician order entry. *Journal of the American Medical Informatics Association*, *13*(2), 138–147. https://doi.org/10.1197/jamia.M1809

Van Leeuwen, K. G., Meijer, F. J. A., Schalekamp, S., Rutten, M. J. C. M., van Dijk, E. J., van Ginneken, B., Govers, T. M. & de Rooij, M. (2021). Cost-effectiveness of artificial intelligence aided vessel occlusion detection in acute stroke: An early health technology assessment. *Insights into Imaging*, *12*(1), 133. https://doi.org/10.1186/s13244-021-01077-4

Vandewinckele, L., Claessens, M., Dinkla, A., Brouwer, C., Crijns, W., Verellen, D. & van Elmpt, W. (2020). Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. *Radiotherapy and Oncology*, *153*, 55–66. https://doi.org/10.1016/j.radonc.2020.09.008

Vasey, B., Ursprung, S., Beddoe, B., Taylor, E. H., Marlow, N., Bilbro, N., Watkinson, P. & McCulloch, P. (2021). Association of clinician diagnostic performance with machine learning-based decision support systems: A systematic review. *JAMA Network Open*, *4*(3), e211276. https://doi.org/10.1001/jamanetworkopen.2021.1276

Verghese, A. (2008). Culture shock — patient as icon, icon as patient. *New England Journal of Medicine*, *359*(26), 2748–2751. https://doi.org/10.1056/nejmp0807461

Verghese, A., Shah, N. H. & Harrington, R. A. (2018). What this computer needs is a physician: Humanism and artificial intelligence. *Journal of the American Medical Association*, *319*(1), 19–20. https://doi.org/10.1001/jama.2017.19198

Verband der Ersatzkassen (2024). *Gesundheitsausgaben nach Ausgabenträgern* [Diagramm]. Vdek. https://www.vdek.com/presse/daten/d_versorgung_leistungsausgaben.html

Vivanti, R., Joskowicz, L., Karaaslan, O. A. & Sosna, J. (2015). Automatic lung tumor segmentation with leaks removal in follow-up CT studies. *International Journal of Computer Assisted Radiology and Surgery*, *10*(9), 1505–1514. https://doi.org/10.1007/s11548-015-1150-0

Wang, D., Khosla, A., Gargeya, R., Irshad, H. & Beck, A. H. (2016). *Deep learning for identifying metastatic breast cancer.* ArXiv. http://arxiv.org/abs/1606.05718

Wang, Y., Yao, Q., Kwok, J. T. & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, *53*(3), 63. https://doi.org/10.1145/3386252

Wasylewicz, A. T. M. & Scheepers-Hoeks, A. M. J. W. (2019). Clinical decision support systems. In P. Kubben, M. Dumontier & A. Dekker (Hrsg.), *Fundamentals of clinical data science* (S. 153–169). Springer Open. https://doi.org/10.1007/978-3-319-99713-1_11

Weber, K. & Zoglauer, T. (2019). Maschinenethik und Technikethik. In O. Bendel (Hrsg.), *Handbuch Maschinenethik* (S. 145–163). Springer Fachmedien Wiesbaden. http://link.springer.com/10.1007/978-3-658-17483-5_10

Weichert, T. (2021). Datenschutz im Kontext der medizinischen Nutzung von KI-Systemen: Heute und morgen. *Zeitschrift für medizinische Ethik, 67(3)*, 351–365. https://doi.org/10.14623/zfme.2021.3.351-365

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M. & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLOS ONE*, *12*(4), e0174944. https://doi.org/10.1371/journal.pone.0174944

Whelton, P. K., Carey, R. M., Aronow, W. S., Casey, D. E., Collins, K. J., Himmelfarb, C. D., DePalma, S. M., Gidding, S., Jamerson, K. A., Jones, D. W., MacLaughlin, E.

J., Muntner, P., Ovbiagele, B., Smith, S. C., Spencer, C. C., Stafford, R. S., Taler, S. J., Thomas, R. J., Williams, K. A., … Wright, J. T. (2018). 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: Executive summary: a report of the American college of cardiology/American Heart Association task force on clinical practice guidelines. *Hypertension*, *71*(6), 1269–1324. https://doi.org/10.1161/HYP.0000000000000066

Wiesing, U. (2001). Evidenz-basierte Medizin aus ethischer Sicht. In J. Michaelis & H. Raspe (Hrsg.), *Die Evidenz-basierte Medizin im Lichte der Fakultäten* (S. 159–164). Schwabe Verlag.

Winkler, J. K., Sies, K., Fink, C., Toberer, F., Enk, A. & Haenssle, H. A. (2020). Digitalisierte Bildverarbeitung: Künstliche Intelligenz im diagnostischen Einsatz. *Forum*, *35*(2), 109–116. https://doi.org/10.1007/s12312-019-00729-3

Wolff, J., Pauling, J., Keck, A. & Baumbach, J. (2020). The economic impact of artificial intelligence in health care: Systematic review. *Journal of Medical Internet Research*, *22*(2), 1–8. https://doi.org/10.2196/16866

Wolff, J., Pauling, J., Keck, A. & Baumbach, J. (2021). Success factors of artificial intelligence implementation in healthcare. *Frontiers in Digital Health*, *3*(June), 594971. https://doi.org/10.3389/fdgth.2021.594971

World Health Organization. (2017). *HIV drug resistance report 2017*. https://iris.who.int/bitstream/handle/10665/255896/9789241512831-eng.pdf?sequence=1

World Health Organization. (2021). *Ethics and governance of artificial intelligence for health*. https://iris.who.int/bitstream/handle/10665/341996/9789240029200-eng.pdf?sequence=1

World Health Organization. (2022). *Constitution: WHO remains firmly committed to the principles set out in the preamble to the constitution*. https://www.who.int/about/governance/constitution

Wright, A. & Sittig, D. F. (2008). A four-phase model of the evolution of clinical decision support architectures. *International Journal of Medical Informatics*, *77*(10), 641–649. https://doi.org/10.1016/j.ijmedinf.2008.01.004

Xu, X., Wickens, C. D. & Rantanen, E. M. (2007). Effects of conflict alerting system reliability and task difficulty on pilots' conflict detection with cockpit display of traffic information. *Ergonomics*, *50*(1), 112–130. https://doi.org/10.1080/00140130601002658

Yang, J., Guo, F., Lyu, T., Yan, L. N., Wen, T. F., Yang, J. Y., Wu, H., Wang, W. T., Song, J. L., Xu, H. & Zhang, Q. H. (2020). [Research of artificial intelligence-based clinical decision support system for primary hepatocellular carcinoma]. *Zhonghua yi xue za zhi*, *100*(48), 3870–3873. https://doi.org/10.3760/cma.j.cn112137-20200905-02571

Yin, J., Ngiam, K. Y. & Teo, H. H. (2021). Role of artificial intelligence applications in real-life clinical practice: Systematic review. *Journal of Medical Internet Research*, *23*(4). https://doi.org/10.2196/25759

Yoo, S., Sheng, Y., Blitzblau, R., McDuff, S., Champ, C., Morrison, J., O'Neill, L., Catalano, S., Yin, F. F. & Wu, Q. J. (2021). Clinical experience with machine learning-based automated treatment planning for whole breast radiation therapy. *Advances in Radiation Oncology*, *6*(2), 100656. https://doi.org/10.1016/j.adro.2021.100656

Yoon, D. Y., Mansukhani, N. A., Stubbs, V. C., Helenowski, I. B., Woodruff, T. K. & Kibbe, M. R. (2014). Sex bias exists in basic science and translational surgical research. *Surgery*, *156*(3), 508–516. https://doi.org/10.1016/j.surg.2014.07.001

Zentrale Ethikkommission. (2021). Entscheidungsunterstützung ärztlicher Tätigkeit durch Künstliche Intelligenz [Stellungnahme]. *Deutsches Ärzteblatt*, *118*(33-34). https://doi.org/10.3238/arztebl.zeko_sn_cdss_2021

Zhou, Q., Chen, Z., Cao, Y. & Peng, S. (2021). Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: A systematic review. *Npj Digital Medicine*, *4*(1), 154. https://doi.org/10.1038/s41746-021-00524-2

Zweig, K. A. (2018). *Wo Maschinen irren können: Verantwortlichkeiten und Fehlerquellen in Prozessen algorithmischer Entscheidungsfindung*. Bertelsmann Stiftung. https://doi.org/10.11586/2018006

8. Affidavit

Eidesstattliche Versicherung

Englich, Florian Michael

Name, Vorname

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Dissertation mit dem Titel:

KI-Systeme zur Unterstützung ärztlicher Entscheidungen: eine ethische Bewertung

selbständig verfasst, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

Ich erkläre des Weiteren, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

Zürich, 02.03.2025 Florian Michael Englich

Ort, Datum Unterschrift Doktorandin bzw. Doktorand

9. Danksagung

Zunächst danke ich Herrn Prof. Dr. Georg Marckmann für die engagierte Betreuung meiner Promotion. Seine Hinweise und der Austausch mit ihm waren für das Gelingen dieser Arbeit von großer Bedeutung. Auch den weiteren Mitgliedern der Betreuungskommission, Frau Prof. Dr. Verina Wild und Herrn Prof. Dr. Martin Fischer, gilt mein Dank.

Danken möchte ich meinen Eltern und meinen Geschwistern Johannes, Mirjam und Christina. Sie haben mich unterstützt und motiviert. Dafür gilt mein Dank auch Bernhard, Benedikt, Aggrey, Eman, Tobias, André, Ihssan, Rahel, Jannik und Christopher.

Für wertvolle fachliche Impulse danke ich Orhan Önder, M.D., dem Netzwerk Junge Medizinethik und der Austauschgruppe *PhD Students in AI Ethics*. Von der Hanns-Seidel-Stiftung wurde ich dankenswerterweise finanziell unterstützt.

Viele andere, die hier nicht namentlich erwähnt sind, waren am Gelingen dieser Arbeit beteiligt. Auch ihnen gilt mein Dank.