Zeitschrift für Praktische Philosophie
Band 11, Heft 1, 2024, S. 419–448
www.praktische-philosophie.org
https://doi.org/10.22613/zfpp/11.1.17

Künstliche Intelligenz und Vertrauen im medizinischen Kontext

Artificial Intelligence and Trust in the Medical Context

Christian Budnik, Zürich

Zusammenfassung: KI-Anwendungen stellen bereits heute einen wichtigen Bestandteil der medizinischen Praxis dar. Ihr Einsatz verbindet sich allerdings mit einer Reihe von Problemen. In dem Aufsatz wird ein spezifisches dieser Probleme diskutiert – die Frage, wie Vertrauen im medizinischen Kontext durch den Einsatz von KI-Technologien betroffen ist. In einem ersten Schritt wird hierzu rekonstruiert, worin die philosophische Herausforderung besteht, die sich mit Vertrauen verbindet, und es wird zweitens erläutert, welche Rolle ein plausibel verstandenes Vertrauen im Verhältnis zwischen Patient:innen und Ärzt:innen spielt. In diesem Zusammenhang wird dafür plädiert, Vertrauen konsequent von bloßem Sich-Verlassen abzugrenzen und es als eine genuin personale Beziehungsform zu interpretieren, in der Personen sich auf eine spezifische Weise aufeinander beziehen. Für das Verhältnis zwischen Ärzt:innen und Patient:innen bedeutet dies, dass Patient:innen, die Vertrauen als Wert zu realisieren versuchen, nicht nur daran interessiert sind, dass Ärzt:innen sie auf eine medizinisch kompetente Weise behandeln, sondern dass sie von ihnen als individuelle Personen betrachtet werden möchten und dabei erwarten, dass sie so gut wie möglich ihre Perspektive einnehmen. Im Hauptteil des Textes wird die Frage diskutiert, ob KI-Technologien vertraut werden kann. Diese Frage ist zentral: Beantwortet man sie positiv, könnte das Ärzt:innen-Patient:innen-Verhältnis als Vertrauensbeziehung überflüssig werden; beantwortet man sie negativ, könnte dies die Vertrauensbeziehung zwischen Ärzt:innen und Patient:innen direkt schädigen, weil wir davon ausgehen müssten, dass unsere Ärzt:innen sich einer Technologie bedienen, der man nicht vertrauen kann. In diesem Zusammenhang werden zwei wichtige Positionen diskutiert, die in der jüngsten Zeit von der These ausgehen, dass man KI-Anwendungen im medizinischen Bereich vertrauen kann – die Ansätze von Ferrario et al. (2021) und von Philip Nickel (2022). Es wird dafür argumentiert, dass diese Ansätze scheitern, weil sie unterstellen, dass Vertrauen überhaupt die richtige

Kategorie im Hinblick auf eine Technologie wie KI ist. Abschließend wird angedeutet, inwiefern die Kategorie der Verlässlichkeit angemessener für unseren Umgang mit KI in der Medizin ist und uns dabei helfen könnte, Vertrauen zwischen Ärzt:innen und Patient:innen zu stärken.

Schlagwörter: Vertrauen, Künstliche Intelligenz, Medizin, Verlässlichkeit, Erwartungen

Abstract: Already today, AI applications are an important part of medical practice. Their employment, however, is connected to a wide array of problems. The paper discusses one of them – the question how trust in the physician-patient relationship is affected by the usage of AI algorithms. As a first step, the paper reconstructs the philosophical challenge relating to trust, before explaining what role a properly understood concept of trust plays in the relationship between physicians and patients. In this context, it is argued that trust has to be systematically distinguished from reliance, and that the best way to do this is to analyze it as a type of personal relationship in which the participants of this relationship treat each other in specific ways. For the physician-patient relationship this means that patients who want to realize the good of trust in their relationship to physicians are not merely interested in being treated in a competent manner, but that they want to be regarded as individual persons with a normative perspective that the physicians must attempt to understand. In its central section, the paper discusses the question whether it is possible to trust AI technologies. This question is central: Answered in the affirmative, the employment of AI technologies threatens to make the physician-patient relationship superfluous; answered in the negative, it could harm the physician-patient relationship since we would have to suppose that our physicians use a technology that cannot be trusted itself. The paper discusses two important accounts that recently defended the claim that we can trust medical AI – the account of Ferrario et al. 2021 and the account of Philip Nickel 2022. It is argued that both accounts are bound to fail, since they assume that trust is the appropriate category when it comes to a technology like AI. By way of a conclusion, it is argued that the category of reliability is more appropriate for our attitude towards AI, and that it has the potential to strengthen the physician-patient relationship.

Keywords: trust, artificial intelligence, medicine, reliability, expectations

1 Einleitung

Der Einsatz von Algorithmen, denen ‚künstliche Intelligenz'[1] zugeschrieben wird, hat in den letzten Jahren in nahezu allen Bereichen unseres Lebens rasant zugenommen. Egal ob es um autonomes Fahren, Spracherkennung, Partnersuche, Klimaschutz, Kunstproduktion demokratische Partizipation oder um mehr oder weniger banale Anwendungen auf unseren Smartphones geht, KI ist überall und verspricht, unser Leben einfacher, effizienter oder nachhaltiger zu machen. Der Trend wird von einer kritischen Reflexion durch verschiedene wissenschaftliche Disziplinen begleitet, und auch innerhalb der Angewandten Ethik hat sich das Thema KI schnell zu einem der prominentesten Forschungsgegenstände entwickelt.

Im Bereich der Medizin ist diese Entwicklung ebenfalls zu beobachten. Der Einsatz von KI-Technologien scheint hier sogar noch dramatischere Folgen zu haben, und es steht zu erwarten, dass sich in nächster Zukunft die Gesundheitsversorgung, wie wir sie kennen, durch den Einsatz dieser Technologien radikal ändern wird. In der medizinischen Forschung spielen KI-Anwendungen bereits eine wichtige Rolle. Man denke etwa im epidemiologischen Kontext an den Einsatz von KI-Algorithmen zur Prognose der Entwicklung der Corona-Pandemie oder an den Durchbruch in der Proteinstrukturvorhersage, der von Deepminds *AlphaFold* erzielt wurde. Noch deutlicher ist dieser Trend in der medizinischen Praxis ausgebildet, wo KI zur Diagnosefindung und immer mehr auch für Therapievorschläge eingesetzt wird.

---

1 In der Folge werde ich auf die Anführungsstriche verzichten und den Terminus mit ‚KI' abkürzen. An dieser Stelle markiere ich lediglich meine Skepsis gegenüber der Auffassung, dass es sich bei dem, was als künstliche Intelligenz bezeichnet wird, tatsächlich um Intelligenz handelt. Diese Skepsis ist inzwischen weit verbreitet. Sie bezieht sich nicht darauf, dass es einen Sinn von ‚intelligent' gibt, in dem man von künstlicher Intelligenz reden könnte. Diesen Sinn gibt es, und er umfasst im Wesentlichen das, was man als instrumentelle Rationalität bezeichnen könnte. Selbst für komplexe Ziele können die Algorithmen von heute dieses Kriterium bereits erfüllen. Mit ‚künstlicher Intelligenz' kann aber auch gemeint sein, dass es digitale Technologien gibt, die denken und handeln können wie menschliche Akteure. Der Terminus der Intelligenz steht dann für etwas, auf dessen Grundlage einem Algorithmus genuin personale Eigenschaften zugeschrieben werden können. Hier ist sehr wohl Skepsis angebracht, und der vorliegende Text kann als ein Versuch gelesen werden, diese Skepsis im spezifischen Kontext der Medizin zu erhärten.

Besonders betroffen von dieser Entwicklung sind diejenigen Medizinbereiche, in denen bildgebende Diagnostik zentral ist. KI-Algorithmen haben sich hier zum Teil bereits als leistungsfähiger in der Auswertung von medizinischen Bilddaten erwiesen als medizinisches Fachpersonal mit langjähriger Erfahrung, und ihr Einsatz hat in verschiedenen Medizinbereichen den Arbeitsaufwand beim Interpretieren großer Mengen an Bilddaten verringert. Konkrete Beispiele für den Einsatz von KI in der bildgebenden Diagnostik sind die Erkennung von Brustkrebs nach Mammographie oder Tomosynthese, die Diagnose von diabetischer Retinopathie, die Erkennung von Hautkrebs-Melanomen an dermatoskopischen Bildern oder die Frühdiagnose von Alzheimer anhand von MRT-Scans des Gehirns. Jenseits der Analyse medizinischer Bilddaten werden KI-Algorithmen etwa eingesetzt, um im Rahmen von Ganganalysen Daten aus tragbaren Sensoren (‚wearables') auszuwerten und auf dieser Grundlage Erkrankungen oder den Krankheitsverlauf vorherzusagen. Besondere Fortschritte erhofft man sich im Rahmen der Präzisionsmedizin, wo KI-Algorithmen große Datenmengen verschiedener Art analysieren sollen, um hochindividuelle Diagnosen zu formulieren und auf dieser Grundlage maßgenaue Behandlungsmethoden vorzuschlagen.

Wenn man nach den Ursachen für die explosionsartige Verbreitung von KI in der Medizin fragt, sind zwei miteinander verbundene Faktoren von entscheidender Bedeutung. Zum einen handelt es sich bei den KI-Formen, die für eine Großzahl der angesprochenen technologischen Fortschritte in verschiedenen Bereichen unseres Lebens gesorgt haben, um Algorithmen, die auf maschinellem Lernen beruhen. Dabei werden Programmen Daten zur Verfügung gestellt, auf deren Basis sie entweder eigenständig ‚lernen', in einem Datensatz Muster zu erkennen und verschiedene Datensätze miteinander auf Ähnlichkeiten zu überprüfen (‚unsupervised machine learning'), oder sie werden mithilfe von vorklassifizierten Daten dazu trainiert, selbständig neue Klassifikationen vorzunehmen und Prognosen aufzustellen (‚supervised machine learning').[2] In den letzten Jahren sind zudem besonders leistungsfähige Formen des maschinellen Lernens entwickelt worden,

2 Alternativ können Algorithmen dazu trainiert werden, in einer unbekannten Umgebung Entscheidungen zu treffen, indem sie positives oder negatives Feedback nach spezifischen Interaktionen erhalten und anhand der Konsequenzen ihres Verhaltens zukünftige Entscheidungen optimieren (‚reinforcement learning'). In dem Maße, in dem wir in Zukunft selbständig handelnde medizinische KI entwickeln sollten, die Anamnesegespräche führen, Diagnosen stellen sowie Therapien vorschlagen und begleiten, dürften solche Formen

bei denen Algorithmen programmiert werden, die zwischen ihrer Eingabe- und ihrer Ausgabeschicht mehrere Zwischenschichten aufweisen und die Struktur der neuronalen Netzwerke des Gehirns nachahmen (‚deep neural networks').

Zwar wären sowohl künstliche neuronale Netze als auch gewöhnlichere Formen maschinellen Lernens ohne Hardware mit einer hohen Rechenleistung nicht möglich, noch wichtiger ist aber ein anderer Faktor: Alle Formen des maschinellen Lernen setzen voraus, dass digitale Daten verfügbar sind, mit denen Algorithmen trainiert werden können. Ohne solche Daten kann ein noch so leistungsfähiger Rechner nicht viel lernen. Die immer weiter voranschreitende Digitalisierung weiter Teile unseres Alltags und speziell auch des Gesundheitswesens hat zur Entstehung von großen und komplexen Datenmengen beigetragen und auf diese Weise den Weg für die dramatischen Fortschritte des maschinellen Lernens im Bereich der Medizin geebnet. Es steht zu vermuten, dass die großen Durchbrüche der letzten Zeit gar nicht hätten gefeiert werden können, wenn wir in einer Welt leben würden, in der Ärzt:innen einander die Röntgen-Aufnahmen ihrer Patient:innen[3] immer noch per Post schickten.

Maschinelles Lernen und Big Data können auf diese Weise als die treibenden Faktoren bei der Entwicklung von KI-Technologien betrachtet werden. Bedenkt man zusätzlich, dass im medizinischen Bereich wichtigste Güter wie Leben und Gesundheit auf dem Spiel stehen, so dass besonders starke Anreize zur Entwicklung medizinischer KI-Technologien vorliegen, überrascht das Tempo, mit dem diese Entwicklung voranschreitet, nicht besonders. Angesichts dieser Situation wäre es naiv, wollte man die Frage, ob wir im medizinischen Bereich KI-Technologien einsetzen sollen, ernsthaft

---

maschinellen Lernens in der medizinischen Praxis wichtiger werden, als sie es derzeit sind.

3 Es ist mir bewusst, dass die gegenderten Formen ‚Ärzt:innen' und ‚Patient:innen' grammatikalisch problematisch sind. Ich möchte allerdings sowohl das generische Maskulinum als auch durchgehende Doppelnennungen vermeiden, und so entscheide ich mich für eine Konvention, die inklusiv und schlank ist, auch wenn sie die grammatikalisch richtigen Formen nur unzureichend abbildet. Dasselbe gilt für einige andere Substantive, die ich in der Folge verwende. Da die Bezugnahme auf die Singularform ‚Ärzt:in' (wegen des Umlauts) noch problematischer ist, verwende ich in der Folge allerdings den Ausdruck ‚Arzt-Patient-Beziehung', der sich in der medizinischen Debatte etabliert hat.

als eine offene Frage behandeln. Am Ende scheint die Situation ganz einfach: Wenn ein KI-Algorithmus tatsächlich zuverlässiger ein Melanom erkennen können sollte als erfahrene Ärzt:innen, dann wäre es schwer zu rechtfertigen, warum wir es im Rahmen onkologischer Diagnostik nicht einsetzen sollten. Und dennoch lässt sich von philosophischer Warte aus fragen, was der flächendeckende Einsatz von KI-Algorithmen in der Medizin für uns bedeuten wird, und wie wir ihn am besten gestalten sollten.

Diese Fragen müssen allerdings eingeschränkt werden. Einerseits muss im Folgenden von konkreten Entwicklungen und ihren technologischen Details abstrahiert werden. Wie angedeutet, gehen diese Entwicklungen so rasant vonstatten, dass es im Rahmen des vorliegenden Textes nicht sinnvoll wäre, die einzelnen Sparten der Medizin durchzugehen und nach möglichen Problemen in der Anwendung von bestimmten KI-Anwendungen zu fragen. Andererseits benötigt die philosophische Beschäftigung mit KI in der Medizin aber auch einen spezifischen Fokus. Als Patient:innen befinden wir uns oft in Situationen, die von Vulnerabilität und Unsicherheit gekennzeichnet sind, und wir sind dadurch auf ein bestimmtes Verhältnis zu den Ärzt:innen, die uns behandeln, angewiesen. Der zentrale Wert, der in diesem Zusammenhang realisiert werden kann, ist Vertrauen. Aus diesem Grund werde ich mich im vorliegenden Text auf die Frage konzentrieren, wie sich *das Vertrauen zwischen Ärzt:innen und Patient:innen* durch den Einsatz von KI-Technologien verändern wird. Dieser Fokus wird zur Folge haben, dass ich wichtige philosophische Fragestellungen, die mit dem Thema ‚KI in der Medizin' zusammenhängen, vernachlässigen werde. Entsprechend werde ich etwa nicht auf die Fragen eingehen können, wie der Einsatz von KI-Algorithmen mit dem Wert der Privatheit zu vereinbaren ist, wie sich mit dem der KI-Technologie inhärenten Problem von diskriminierenden Verzerrungen umgehen lässt, oder inwiefern der Einsatz solcher Technologien zu einem Verlust von medizinischen Fähigkeiten (‚de-skilling') auf Seiten von Ärzt:innen führen könnte.

Stattdessen wird es mir zunächst darum gehen, zu rekonstruieren, worin die philosophische Herausforderung besteht, die sich mit Vertrauen verbindet (Abschnitt 2) und zu erklären, welche Rolle ein plausibel verstandenes Vertrauen im Verhältnis zwischen Patient:innen und Ärzt:innen spielt (Abschnitt 3). Hier werde ich dafür plädieren, Vertrauen konsequent von Sich-Verlassen abzugrenzen und es im medizinischen Kontext im Sinne einer *Vertrauensbeziehung* zu verstehen. Im Hauptteil des Textes werde ich mich mit der für die vorliegenden Zwecke zentralen Frage beschäftigen, ob

KI-Anwendungen angemessene Objekte von Vertrauen sind (Abschnitt 4). Ich werde in diesem Zusammenhang zwei wichtige Positionen diskutieren, die in letzter Zeit die These vertreten haben, dass man KI-Anwendungen im medizinischen Bereich tatsächlich vertrauen kann, und dafür argumentieren, dass diese Ansätze scheitern, weil Vertrauen in diesem Zusammenhang nicht die richtige Kategorie ist. Abschließend werde ich andeuten, inwiefern die Kategorie der Verlässlichkeit angemessener für unseren Umgang mit KI in der Medizin ist (Abschnitt 5).

2 Der Begriff des Vertrauens

Versteht man Vertrauen als eine mentale Einstellung,[4] so handelt es sich dabei um eine Einstellung, die in Situationen der epistemischen Unsicherheit eingenommen wird, in Situationen also, in denen Personen etwas nicht genau wissen. Macht man an dieser Stelle die vorläufige Annahme, dass Vertrauen sich immer auf eine andere Person richtet,[5] dann bedeutet dies, dass wir vertrauen können, wann immer wir uns nicht ganz sicher sind, wie sich eine andere Person verhalten wird. Zudem muss es sich bei den in Frage stehenden Handlungen einer anderen Person um solche handeln, die uns auf irgendeine Weise am Herzen liegen oder wichtig sind. Es macht keinen Sinn zu vertrauen, wenn wir uns *ganz sicher* sind, dass jemand etwas tun wird, und es macht keinen Sinn zu vertrauen, dass jemand etwas tun wird, wenn uns *egal* ist, ob diese Person es auch tatsächlich tun wird.

Mit epistemischer Unsicherheit und praktischer Relevanz sind allerdings nur begriffliche Randbedingungen bestimmt, die Vertrauen überhaupt erst sinnvoll machen. In einem nächsten Schritt kann darauf hingewiesen werden, dass Zuschreibungen von Vertrauen eine wichtige Rolle in der Rechtfertigung unserer Handlungen zukommt: Jemandem zu vertrauen, kann ein Handlungsgrund sein – so etwa, wenn ich auf die Frage, warum ich mich noch nicht um ein Taxi zum Bahnhof gekümmert habe, die Antwort

4 Diese Annahme ist unausgesprochener Konsens der klassischen Vertrauenstheorien; für einen (sehr rudimentären) Versuch, sie zu begründen, vgl. Jones 1996, 7–8.

5 Die Vorläufigkeit ist selbstverständlich darin begründet, dass man sich an dieser Stelle davor hüten sollte, durch eine begriffliche Festlegung eine Vorentscheidung bezüglich der Frage, ob wir KI-Systemen vertrauen können, zu fällen. Ein anderer Sinn, in dem die vorliegende Formulierung vorläufig ist, betrifft die Tatsache, dass Personen auch sich selbst vertrauen können.

gebe, dass ich darauf vertraue, dass mein Vater mich hinfahren wird. ‚Weil ich darauf *vertraue*, dass er j-en wird' kann im Rahmen solcher Handlungsrationalisierungen allerdings ohne Sinnverlust ersetzt werden durch ‚Weil ich *glaube*, dass er j-en wird', d. h. durch die Zuschreibung einer Überzeugung. Diese Tatsache hat eine ganze Reihe von Vertreter:innen philosophischer Vertrauenstheorien dazu bewogen, eine kognitivistische Lesart von Vertrauen zu bevorzugen.

Gemäß einer solchen Lesart ist Vertrauen eine evidenzsensitive Einstellung, deren Angemessenheit davon abhängt, wie es sich mit der Welt verhält: Wenn ich behaupte, dass ich darauf vertraue, dass mein Vater mich zum Bahnhof fahren wird, dann bedeutet das entsprechend, dass es etwas an meinem Vater gibt, das mich in der Auffassung bestärkt, dass er, wenn es soweit ist, in sein Auto steigen und mit mir zum Bahnhof fahren wird. Üblicherweise redet man an dieser Stelle davon, dass eine Person, der man auf angemessene Weise vertraut, das Kriterium der Vertrauenswürdigkeit erfüllt. Unterschiedliche Theorien buchstabieren dann auf unterschiedliche Weise aus, worum es sich bei Vertrauenswürdigkeit genau handelt. So kann etwa dafür argumentiert werden, dass eine Person vertrauenswürdig ist, wenn sie in ihren Interaktionen mit mir Wohlwollen an den Tag legt; andere Theorien verorten Vertrauenswürdigkeit in Faktoren wie Kompetenz, Aufrichtigkeit oder Integrität.[6]

Das größte Problem für evidenzbasierte Vertrauenstheorien betrifft die Frage, wie Vertrauen von Sich-Verlassen abzugrenzen ist.[7] Die Idee hinter dieser Herausforderung ist, dass wir uns auf Personen auf ähnliche Weise wie auf Gegenstände verlassen können, ohne dass wir ihnen dadurch bereits vertrauen: Ich kann mich darauf verlassen, dass die Bremsen an meinem Fahrrad funktionieren werden, wir würden aber nicht sagen, dass ich in so einem Fall meinem Fahrrad vertraue. Ganz ähnlich kann ich mich darauf verlassen, dass mein Nachbar am Dienstagabend die Bio-Tonne an den Straßenrand stellen wird, z. B. weil ich beobachtet habe, wie er es seit zwanzig Jahren jede Woche getan hat; der springende Punkt ist aber, dass ich ihm in so einem Fall nicht im eigentlichen Sinne vertraue. Das generelle Problem

6 Für prominente Vertreter:innen einer evidenzbasierten Theorie vgl. Jones 1996, Hardin 1999, McLeod 2002, Hieronymi 2008, McMyler 2011, O'Neill 2018 oder Simpson 2018.

7 Vgl. für diese für die Debatte um Vertrauen zentrale Unterscheidung Baier 1986.

für evidenzbasierte Ansätze besteht nun darin, dass sie Vertrauen mit einer Einstellung identifizieren, die auf einer evidenzgestützten Prognose über das Verhalten einer anderen Person beruht. Das scheint aber zu wenig zu sein.[8]

Nicht-evidenzbasierte Ansätze beschreiten deshalb einen anderen Weg. Sie identifizieren Vertrauen mit der Einstellung des Sich-Verlassens, formulieren aber Zusatzbedingungen, die erfüllt sein müssen, damit es sich bei einem speziellen Vorkommnis von Sich-Verlassen tatsächlich um Vertrauen handelt. Der meiner Ansicht nach ausgefeilteste solcher Ansätze bestimmt Vertrauen als ein Sich-Verlassen, das von einer normativen Erwartung begleitet wird.[9] Wenn ich mich darauf verlasse, dass mein Vater mich zum Bahnhof fahren wird, dann handelt es sich dabei nur dann um Vertrauen, wenn ich sein entsprechendes Verhalten nicht lediglich erwarte, so wie man erwartet, dass die Sonne am Morgen aufgehen wird; das wäre eine Erwartung, die prädiktiver Natur ist. Für Vertrauen ist im Gegensatz dazu zentral, dass ich im Sinne einer legitimen Forderung *von* meinem Vater erwarte, dass er mich zum Bahnhof fahren wird. Die Komponente einer normativen Erwartung scheint besser zu der normativ anspruchsvollen Rolle zu passen, die Vertrauen in unserem Alltag spielt. Es ist eben nicht nur eine Frage des ‚Funktionierens', wenn wir anderen Personen vertrauen. Genau deshalb glauben Vertreter:innen nicht-evidenzbasierter Ansätze, dass ihre jeweiligen Bestimmungen von Vertrauen besser mit dem Unterschied zwischen Vertrauen und bloßem Sich-Verlassen umgehen können.

Das Problem mit nicht-evidenzbasierten Ansätzen ist aber, dass sie der oben angesprochenen Rolle nicht gerecht werden können, die Vertrauen für uns als Akteur:innen spielt. Warum, ließe sich fragen, sollte die Tatsache, dass ich von meinem Vater berechtigterweise erwarten kann, dass er mich zum Bahnhof fährt, einen Grund für mich darstellen, mich nicht um alternative Transportmöglichkeiten zu kümmern? Wir sind im Hinblick auf

8 Dieser Einwand trifft kognitivistische Ansätze selbst dann, wenn das Kriterium für Vertrauenswürdigkeit im Sinne einer sehr ‚persönlichen' Eigenschaft verstanden wird, wie etwa im Fall der an Wohlwollen orientierten Ansätze: Es mag sein, dass ich davon ausgehe, dass mein Vater mich zum Bahnhof fahren wird, weil ich Anzeichen registriert habe, die dafür sprechen, dass er mir gegenüber wohlwollend eingestellt ist. Es ist aber nicht klar, warum diese Überzeugung einen anderen Status haben sollte, als jede andere Überzeugung, die auf Evidenzen beruht und eine Vorhersage begründet.

9 Vgl. die komplexe Theorie in Faulkner 2007; andere wichtige nicht-evidenzbasierte Ansätze finden sich in Holton 1994 und Hawley 2014.

sehr viele Menschen befugt, normative Erwartungen an sie zu haben. Sie können unter anderem auf der Grundlage moralischer Erwägungen entstehen. Es ist etwa moralisch falsch, Personen zu überfallen, und entsprechend kann ich von einer Person, die mir in einer dunklen Gasse entgegenkommt, erwarten, dass sie mich nicht überfällt, weil die Moral es verbietet. Aber das bedeutet klarerweise nicht, dass ich mich in so einer Situation nicht dennoch bedroht fühlen kann. Sich bedroht zu fühlen, scheint in so einer Situation aber genau das Gegenteil von Vertrauen zu sein.

Evidenzbasierte Ansätze sind also mit Problemen behaftet, mit denen nicht-evidenzbasierte Ansätze besser umgehen können, aber umgekehrt gilt ebenfalls, dass nicht-evidenzbasierte Ansätze mit Schwierigkeiten zu kämpfen haben, die für evidenzbasierte Ansätze nicht auftauchen. Angesichts dieser Situation liegt es nahe, die eingangs erwähnte Hintergrundannahme beider Theoriefamilien in Frage zu stellen: Vertrauen sollte möglicherweise nicht im Sinne einer Reduktion auf einen bestimmten Typus mentaler Einstellung analysiert werden. Wenn das Vertrauensprädikat aber nicht primär eine mentale Einstellung bezeichnet, bietet es sich an,[10] es im Sinne einer Charakterisierung einer bestimmten Beziehungsform zu verstehen. Wer eine Theorie dieser Art vertritt, ist keinesfalls darauf festgelegt zu bestreiten, dass Vertrauen mit mentalen Einstellungen zu tun hat. Es eröffnet sich damit vielmehr der begriffliche Raum, um zeigen zu können, wie verschiedene Typen von Einstellungen, seien sie evidenzbasiert oder nicht, innerhalb von Vertrauensbeziehungen miteinander zusammenhängen. Dieser Zusammenhang erlaubt wiederum, Vertrauen als ein dynamisches und temporal ausgedehntes Phänomen zu verstehen.

10 Es bietet sich dies vor allem angesichts einer weiteren nicht hinreichend begründeten Annahme vieler der gegenwärtig vertretenen Vertrauenstheorien an. Typischerweise wird nämlich davon ausgegangen, dass Vertrauen stets eine dreistellige Relation ist, die zwei Personen und eine Handlungserwartung miteinander verbindet. Eine zweistellige Lesart von Vertrauen, nach der ich immer einer anderen Person simpliciter vertraue, macht dagegen das von mir präferierte Verständnis von Vertrauen als einer Beziehung attraktiv; zur Zweistelligkeit von Vertrauen vgl. Faulkner 2015, Domenicucci und Holton 2017 und Budnik 2021, Kap. 2.

3 Die Vertrauensbeziehung zwischen Ärzt:innen und Patient:innen

Selbstverständlich muss eine ausgearbeitete Vertrauenstheorie mehr dazu sagen, wie solche Vertrauensbeziehungen genauer beschaffen sind. Ich beschränke mich an dieser Stelle auf die Skizze einer Idee.[11] In diesem Zusammenhang ist es hilfreich, sich zunächst an Freundschaft als dem paradigmatischen Fall einer Vertrauensbeziehung zu orientieren. Vertrauen besteht hier nicht darin, dass die miteinander befreundeten Personen voneinander etwas glauben oder sich aufeinander auf eine bestimmte Weise verlassen, wie es die beiden oben thematisierten Theorietypen fordern. Das ist zwar *auch* der Fall, aber eben auf eine besondere Weise.

Zunächst müssen solche Vertrauensbeziehungen über einen bestimmten Zeitraum wachsen, und sie haben eine gewisse temporale Stabilität. Das ist wichtig, weil Vertrauenspartner:innen einander nur in einem zeitlich ausgedehnten Prozess als individuelle Personen kennenlernen können, um auf diese Weise die Fähigkeit zu erwerben, den Gründen der jeweils anderen Person ein Gewicht beizumessen, das ähnlich groß ist wie das Gewicht der eigenen Gründe. Wenn ich in einer Vertrauensbeziehung davon ausgehe, dass jemand mich nicht hintergehen wird, dann liegt das nicht daran, dass ich diese Person lange genug beobachtet habe, um eine Prognose über ihr Verhalten treffen zu können, sondern daran, dass diese Person mir in normativer Hinsicht nahesteht, weil wir einander verstehen. Das bedeutet nicht zwangsläufig, dass wir dieselben Werte haben oder bezüglich bestimmter moralischer Gebote übereinstimmen, auch wenn das in einer Vertrauensbeziehung durchaus auch der Fall sein kann. Es bedeutet aber, dass wir uns hinreichend in die Perspektive der jeweils anderen Person hineinversetzen können, um ihre evaluativen und volitionalen Einstellungen nachvollziehen zu können und auf diese Weise Gründe zu erwerben, die sich nicht ausschließlich aus unserer eigenen normativen Perspektive speisen. Wenn ich Personen in normativer Hinsicht nahestehe, dann lassen mich ihre Bedürfnisse und Wünsche nicht kalt, auch wenn das selbstverständlich nicht bedeutet, dass ich deswegen automatisch meine eigenen Wünsche hintanstelle.[12]

11 Vgl. hierzu ausführlicher Budnik 2021.

12 Zum Spannungsfeld zwischen Vertrauen und Autonomie vgl. insbesondere Budnik 2021, Abschn. 5.1.5. Ich danke einer/einem anonymen Gutachter:in

An dieser Stelle wird man einwenden, dass dieses Vertrauensverständnis viel zu anspruchsvoll sei, was sich unter anderem daran zeige, dass es im Grunde nur innerhalb gewichtiger Beziehungen wie Freundschaften realisiert werden kann. Umgekehrt lässt sich aber geltend machen, dass Vertrauen tatsächlich ein weitaus anspruchsvolleres Phänomen ist, als es manche der vor allem in den Sozialwissenschaften verbreiteten Theorien annehmen. Besonders deutlich wird dies, wenn man sich den Kontrast zwischen Vertrauen und Sich-Verlassen vor Augen hält: Vieles von dem, wofür wir im Alltag relativ selbstverständlich das Vokabular des Vertrauens verwenden, gehört begrifflich in die Sphäre des Sich-Verlassens. Wir sagen zwar manchmal, dass wir Vertrauen in ein Auto, ein Pferd oder die uns unbekannte Person, die gerade den Bus lenkt, haben, tatsächlich sind das aber allesamt Kontexte, in denen wir uns lediglich darauf verlassen, dass etwas passieren oder eben nicht passieren wird, weil wir mehr oder weniger zuverlässige Prognosen darüber treffen können, wie sich Autos, Pferde oder Busfahrer:innen unter bestimmten Bedingungen verhalten werden.

Geht es um das Verhältnis zwischen Ärzt:innen und Patient:innen, ist die Situation komplexer. Zwar können auch hier viele der Merkmale, die eine freundschaftliche Vertrauensbeziehung charakterisieren, nicht realisiert werden. Ärzt:innen sind eben keine Freundinnen oder Freunde; man kann sich nicht darauf verlassen, dass sie einem beim Umzug oder dem Schreiben eines Papers helfen werden. Und doch scheint es sich bei dem Arzt-Patient-Verhältnis um eines zu handeln, das nur unzureichend als ein bloßes Verlassensverhältnis bestimmt wäre. Das liegt daran, dass wir zu Ärzt:innen ein *persönliches* Verhältnis haben können: Wir können direkt miteinander interagieren und einander als individuelle Personen in den Blick nehmen. Das alles kann sich zudem über längere Zeiträume, in manchen Fällen sogar über Jahre, erstrecken, und bei den relevanten Interaktionen handelt es sich typischerweise um solche, die ein besonderes Gewicht haben, weil sie die zentralen Güter des Lebens und der Gesundheit betreffen.

Das alles spricht dafür, das Arzt-Patient-Verhältnis als eines zu verstehen, in dem das Gut des Vertrauens realisiert werden kann. Eine Theorie, die Vertrauen primär als eine Beziehung auffasst, kann uns dabei helfen, besser zu verstehen, wie dies am besten vonstattengehen sollte.[13] In einer

für Nachfragen, die mich an dieser Stelle gezwungen haben, ausführlicher auf den Begriff des normativen Nahestehens einzugehen.

13 Man könnte an dieser Stelle einwenden, es sei heutzutage illusorisch, über-

gelungenen Vertrauensbeziehung zwischen Ärzt:innen und Patient:innen wird es primär darauf ankommen, dass Ärzt:innen ihre Patient:innen als Individuen mit einer spezifischen Geschichte betrachten. Diese Forderung erklärt, warum Patient:innen, denen im Rahmen einer Anamnese viel Zeit und Aufmerksamkeit geschenkt wird, leichter ein Vertrauensverhältnis aufbauen. Sie erklärt ebenfalls, weshalb Patient:innen, die langjährige Arzt-Patient-Verhältnisse aufgeben mussten, dies oft als eine besonders schlimme Zäsur betrachten: Ärzt:innen sind eben nicht nur Personen mit einer bestimmten Expertise, die Patient:innen hilft, ein quasi-technisches Problem zu lösen, sondern wir schätzen in dem Verhältnis zu ihnen eine gewisse Nähe und Intimität, die sich darin manifestiert, dass sie uns wiedererkennen, sich an Gespräche mit uns erinnern, eine Sorge über unser weit gefasstes Wohlergehen zum Ausdruck bringen und oft mehr von uns wissen, als für eine spezifische Diagnose notwendig wäre.

In meiner Betonung des Werts eines spezifisch verstandenen Vertrauens für das Verhältnis zwischen Ärzt:innen und Patient:innen impliziere ich, dass es in diesem Kontext nicht nur darum geht, die Gesundheit von Personen wiederherzustellen. Den Jargon der Medizinethik aufgreifend, könnte man an dieser Stelle davon reden, dass für das Verhältnis zwischen Ärzt:innen und Patient:innen nicht nur die Prinzipien des Wohltuns und der Schadensvermeidung einschlägig sind,[14] sondern dass es hier auch um den Wert der Fürsorge gehen muss. Als Korrektiv zu den überkommenen Prinzipien der Medizinethik plädieren Vertreter:innen einer medizinischen Fürsorgeethik dafür, die Beziehung zwischen Ärzt:innen und Patient:innen in den Vordergrund zu rücken. Ärzt:innen sollen ihre Patient:innen nicht lediglich als zu lösende medizinische Fälle, sondern als individuelle Personen be-

haupt davon auszugehen, dass sich Vertrauensbeziehungen zwischen Ärzt:innen und Patient:innen etablieren können, weil im medizinischen Alltag oft keine Zeit für die relevanten Interaktionen vorhanden ist. Letzteres läßt sich schlecht bestreiten, gleichzeitig kann man aber daran festhalten, dass dies nicht dagegen spricht, dass wir solche Vertrauensbeziehungen anstreben sollten, z. B. indem wir Ärzt:innen wieder mehr zeitliche Ressourcen für die Kommunikation mit Patient:innen zur Verfügung stellen. In dieser Hinsicht betrachte ich die medizinische Vertrauensbeziehung eher im Sinne eines Ideals, dem wir uns in unterschiedlichem Ausmaß annähern können. Zudem muss bedacht werden, dass es nicht in jedem Teilbereich der Medizin notwendig oder auch nur wünschenswert ist, Vertrauensbeziehungen zu etablieren.

14 Vgl. Beauchamp und Childress 2001.

trachten, deren Perspektive sie durch den Einsatz empathischer Fähigkeiten einzunehmen versuchen, um in einem Dialog die für die Patient:innen besten Behandlungsstrategien zu ermitteln.[15]

Das Verständnis der Arzt-Patient-Beziehung, das ich im vorliegenden Kontext vorschlage, überschneidet sich in weiten Teilen mit den Forderungen einer ‚medical ethics of care‘,[16] es geht aber in mindestens zwei Hinsichten darüber hinaus. Zum einen ist Fürsorge, zumindest im Standardfall, ein asymmetrisches Phänomen: Ärzt:innen sollen Patient:innen fürsorglich behandeln, aber nicht umgekehrt. Eine Vertrauensbeziehung impliziert dagegen – wiederum zumindest im Standardfall – eine reziproke Einstellung der Teilnehmenden an dieser Beziehung zueinander.[17] Das mag überraschen: In welcher Hinsicht sollten denn Ärzt:innen Vertrauen in ihre Patient:innen setzen müssen? Tatsächlich ist der medizinische Kontext einer, in dem eine Partei *qua* ihrer medizinischen Kompetenzen aktiver ist als die andere Partei. Die Reziprozität in einer Vertrauensbeziehung erstreckt sich hier klarerweise nicht auf alles, was Patient:innen von Ärzt:innen erwarten können. Aber sie beinhaltet, dass Patient:innen ihren Ärzt:innen gegenüber ebenso respektvoll und aufrichtig sind, wie sie es umgekehrt von ihnen erwarten. In diesem Sinne können Ärzt:innen in einer Vertrauensbeziehung z. B. erwarten, dass ihre Patient:innen sie nicht über Symptome belügen, um an verschreibungspflichtige Medikamente zu kommen, und in bestimmten Rechtssystemen können sie beruhigt das Risiko eingehen, eine Therapie-

15 Vgl. Held 2006; für die Anwendung im medizinischen Kontext, vgl. etwa Nortvedt, Hem und Skribekk 2011; für einen Ansatz, der die Rolle der Empathie in der medizinischen Praxis betont, vgl. Halpern 2001.

16 Vgl. etwa die Beiträge in Cates und Lauritzen 2001.

17 Diese Reziprozitätsthese ist nicht selbstverständlich und bedürfte einer ausführlicheren Begründung, als ich sie an dieser Stelle vorlegen kann. Ich belasse es bei dem Hinweis darauf, dass unsere begrifflichen Intuitionen dagegen sprechen, eine interpersonelle Konstellation als Vertrauensbeziehung auszuzeichnen, bei der eine der beteiligten Personen eine Vertrauensbeziehung für sich reklamieren kann, während dies für die andere Person nicht möglich ist: ‚Ich stehe in einer Vertrauensbeziehung zu ihr, aber sie nicht zu mir‘ scheint aus begrifflichen Gründen eine falsche Aussage zu sein. Eine erschöpfende Argumentation für die Reziprozitätsthese müsste wiederum von einem Verständnis von Vertrauen als zweistelliger Relation ausgehen. Ich danke einer/einem anonymen Gutachter:in für die Aufforderung, diesen Punkt zu präzisieren.

empfehlung auszusprechen, weil sie erwarten können, dass sie bei Misserfolg nicht von ihren Patient:innen verklagt werden.

Die zweite wichtige Facette von Vertrauensbeziehungen, die sie von Fürsorge-Situationen unterscheidet, besteht darin, dass die Teilnehmenden an einer solchen Beziehung einander auf eine spezifische Weise interesselos behandeln. Der Aspekt der Interesselosigkeit wird in der philosophischen Debatte über Vertrauen zu wenig betont, aber er ist gerade für die medizinische Praxis von entscheidender Bedeutung. In einer Vertrauensbeziehung bedeutet ‚interesselos' keinesfalls, dass eine Person überhaupt keine eigenen Interessen verfolgt. Im medizinischen Kontext ist relativ klar, dass auch Ärzt:innen in der Regel eigene Interessen verfolgen, wie etwa Interessen, die mit der Erhaltung der eigenen Reputation oder mit finanziellen Faktoren zusammenhängen. Daran müssen sich nicht zwangsläufig Probleme knüpfen, und wollte man in der Arzt-Patient-Beziehung lediglich mit Mechanismen des Sich-Verlassens operieren, würden wir mit der Zuschreibung solcher Interessen kooperatives Verhalten etablieren können: Als Patient:innen könnten wir etwa davon ausgehen, dass Ärzt:innen uns angemessen behandeln, weil sie einer Kontrollaufsicht unterliegen, weil es etablierte Gruppen gibt, die Patient:innen-Interessen vertreten, und weil Ärzt:innen wissen, dass sie ihren Beruf nicht lange ausführen könnten, wenn sie Patient:innen-Interessen missachteten.

In einer Vertrauensbeziehung kommt aber mehr dazu. Wie oben angedeutet, versuchen die Teilnehmenden an so einer Beziehung, die Perspektive ihrer Vertrauenspartner:innen einzunehmen, um ihre Gründe nachvollziehen und die mit ihrer normativen Perspektive zusammenhängenden Emotionen nachfühlen zu können. Dadurch rücken ihre Perspektiven näher aneinander, so dass sie in konkreten Handlungssituationen Gründe dafür haben, sich auf die jeweils andere Person zu verlassen, die über die angesprochenen instrumentellen Erwägungen hinausgehen. Die Interesselosigkeit, die eine gelungene Arzt-Patient-Beziehung charakterisiert, hat deshalb auch nicht viel mit einem ungerichteten Altruismus zu tun: Ich kann mich auf die Ärzt:innen, zu denen ich Vertrauen habe, in vielen Hinsichten verlassen, aber nicht etwa, weil sie moralische Heilige wären, die keine eigenen Interessen befördern möchten, sondern weil sie in der Zeit, in der unsere Vertrauensbeziehung vorgelegen hat, gelernt haben, meine Interessen nicht als die Interessen einer völlig fremden Person zu betrachten.

Diese spezielle Interesselosigkeit, die in Vertrauensbeziehungen vorliegt, kann auch verstehen helfen, weshalb solche Beziehungen keine Be-

drohung für die Autonomie ihrer Teilnehmenden darstellen.[18] Wann immer Personen ihre Gesundheit in die Hände von Ärzt:innen legen, begeben sie sich in eine besonders verletzliche Position. Es ist zumindest auf den ersten Blick nicht fernliegend, diesen Vertrauensakt im Sinne einer Einschränkung der eigenen Autonomie zu verstehen, der Personen aus instrumentellen Gründen zustimmen, weil sie denken, dass sie auf diese Weise am besten ihre Gesundheit befördern können. Interpretiert man das Arzt-Patient-Verhältnis aber als eine Vertrauensbeziehung, ergibt sich diese Konsequenz, die die meisten von uns vermeiden wollten, nicht: Es ist ja keine völlig fremde Person, der ich in so einer Situation als Patient:in ausgesetzt bin, sondern jemand, der mir auf relevante Weise nahesteht und meine Perspektive einnehmen kann. Das macht einen Unterschied, durchaus auch auf einer emotionalen Ebene, und es führt dazu, dass ich meine Autonomie als Patient:in nicht nur nicht eingeschränkt sehe, sondern sie durch das für die Vertrauensbeziehung charakteristische Miteinander erweitere, weil ich in der vulnerablen Situation, in der ich mich befinde, nicht mehr auf mich allein gestellt bin. Diese Interpretation setzt nicht nur ein von den gängigen Theorien abweichendes Verständnis von Vertrauen voraus, sondern sie impliziert auch ein Autonomieverständnis, das sich von den ‚klassischen' Vorstellungen von Autonomie als Unabhängigkeit oder Kontrolle unterscheidet und Autonomie als ein relationales Phänomen versteht.[19]

4 Können wir medizinischen KI-Anwendungen vertrauen?

Der zunehmende Einsatz von KI-Technologien im medizinischen Bereich wird nicht ohne Folgen für die Vertrauensbeziehung zwischen Ärzt:innen und Patient:innen bleiben. Ob diese Beziehung gestärkt, geschwächt oder gar geschädigt wird, hängt wesentlich von der Frage ab, ob wir – und hier sind explizit sowohl Patient:innen als auch Ärzt:innen gemeint – KI-Technologien vertrauen können. Die Frage nach Vertrauen wird derzeit in den verschiedenen um KI geführten Debatten gestellt.[20] Es bieten sich an dieser

18 Zum Konflikt zwischen Autonomie und Vertrauen im Arzt-Patient-Verhältnis vgl. O'Neill 2002.

19 Vgl. ähnlich etwa McLeod 2002; für die Dynamik von Vertrauen und Autonomie im medizinischen Kontext vgl. Steinfath 2016 und Wiesemann 2016.

20 Alternativ wird nicht lediglich danach gefragt, ob wir KI-Algorithmen vertrauen können, sondern es wird gefordert, dass wir sie so programmieren sollten,

Stelle zunächst zwei Typen von Antworten an: Zum einen könnten Optimist:innen dafür argumentieren, dass wir KI-Systeme im Bereich der Medizin so programmieren können, dass ihnen vertraut werden kann. Das Problem für die Arzt-Patient-Beziehung ergibt sich bei dieser Antwort dadurch, dass nicht klar ist, ob Ärzt:innen und KI-Anwendungen als potentielle Vertrauenspartner:innen nicht langfristig in ein Konkurrenzverhältnis zueinander treten. Umgekehrt könnten Pessimist:innen argumentieren, dass KI-Technologien gefährlich sind, so dass wir ihnen gegenüber misstrauisch eingestellt sein sollten. Aus dieser Perspektive müssten Ärzt:innen auf ihren Einsatz verzichten, sobald sie das aber nicht tun, hätten Patient:innen Anlass, ihnen zu misstrauen.

Der Streit zwischen Optimist:innen und Pessimist:innen muss an dieser Stelle nicht entschieden werden. Die Frage, auf die beide antworten, impliziert nämlich, dass es überhaupt möglich ist, KI-Technologien zu vertrauen oder zu misstrauen. Von philosophischer Warte ist im Hinblick auf diese Frage Skepsis angemeldet worden. So argumentiert Mark Ryan, dass Vertrauen immer entweder eine affektive Komponente hat, etwa das Wohlwollen gegenüber der vertrauenden Person, oder aber mit normativen Erwartungen einhergeht. Weder die affektive noch die normative Komponente können seiner Ansicht nach im Fall von KI-Technologien realisiert sein, weil diese Technologien uns gegenüber nicht wohlwollend eingestellt sein können und wir als Vertrauende keine genuinen normativen, sondern lediglich prädiktive Erwartungen an sie haben (vgl. Ryan 2020). Joshua Hatherley macht einen ähnlichen Punkt stark, indem er betont, dass es zwar möglich ist, sich im medizinischen Kontext auf KI zu verlassen, dass wir diesen Technologien aber nie im eigentlichen Sinne vertrauen können. In seiner Begründung verweist er, analog zur Argumentation von Ryan, darauf, dass wir KI-Technologien keine Motive zuschreiben und keine normativen Erwartungen an sie haben können (vgl. Hatherley 2020).

Auf der Basis des von mir vorgeschlagenen Vertrauensverständnisses lässt sich die Skepsis von Ryan und Hatherley weiter erhärten. Die Frage, ob wir KI-Algorithmen vertrauen können, zielt meiner Ansicht nach ins Leere, weil es nicht möglich ist, in einer Vertrauensbeziehung zu solchen technischen Artefakten zu stehen. So eine Beziehung stellt eine komplexe, histo-

dass sie vertrauenswürdig sind. Vertrauenswürdigkeit ist etwa einer der Kernbegriffe des von der Europäischen Kommission erarbeiteten Weißbuchs zur KI; vgl. Europäische Kommission 2020.

risch gewachsene interpersonelle Konstellation dar, in der sich ihre Teilnehmenden auf verschiedene Weisen verstehend aufeinander beziehen; dafür müssen sie in der Lage sein, die Perspektive der anderen Person einzunehmen und empathisch ihre emotionalen Zustände nachzuvollziehen; dies setzt wiederum voraus, dass sie selbst eine normative Perspektive auf die Welt haben, sich eigene Zwecke setzen und Subjekte emotionaler Zustände sein können. Partner in einer Vertrauensbeziehung müssen *ein Leben führen*, so wie die meisten von uns menschlichen Personen es tun. Es mag übertrieben rührselig klingen, aber eine Vertrauensbeziehung zu führen, bedeutet nichts anderes, als ein Stück weit *ein Leben miteinander* zu führen.

Trotz aller Fortschritte in der KI-Forschung und selbst wenn man neuronale Netzwerke und verschiedene Formen des ‚deep learning' berücksichtigt, scheint die Annahme sehr fernliegend, dass KI-Systeme jemals auf diese Weise ein Leben werden führen können.[21] Damit will ich nicht bestreiten, dass es möglich ist, Programme zu entwickeln, die Personen *den Eindruck* vermitteln, dass sie mit ihnen interagieren, kommunizieren, sich in sie einfühlen oder ihre Perspektive einnehmen. Wir können uns allerdings darin täuschen, dass wir eine Vertrauensbeziehung führen, und genau das wäre auf eine ganz spezielle Weise der Fall, wenn ich mich von einem KI-Algorithmus ‚verstanden fühlen' würde, weil er ‚die richtigen Dinge' sagt. Die Fortschritte in der Entwicklung von KI sind beeindruckend. Das muss man nicht in Abrede stellen, wenn man gleichzeitig bemerkt, dass wir uns allzu oft von diesen Fortschritten zu sehr beeindrucken lassen, so etwa, wenn darüber berichtet wird, dass ein Algorithmus etwas ‚fast genauso gut' wie menschliche Personen macht, sei es Antworten auf komplexe Wissensfragen zu geben, Bilder zu malen oder, wie im Fall von sozialen Pflegerobotern, ‚vor Freude' zu ‚lachen'.

In der philosophischen Debattenlandschaft fehlt es allerdings nicht an Positionen, die dennoch daran festhalten, dass es möglich ist, KI-Algorithmen zu vertrauen. Diese Positionen reproduzieren dabei Probleme, die sich Theorien in der allgemeinen Debatte um die Begriffe des Vertrauens und der Vertrauenswürdigkeit stellen. Analog zu dieser Debatte sind es im Wesentlichen zwei Strategien, die verfolgt werden, um zu zeigen, was es heißt, einer KI zu vertrauen. Exemplarisch finden sich diese Strategien in den Theorien

21 Für Argumente für die These, dass KI-Systeme sich keine eigenen Zwecke setzen können und entsprechend keine Autonomie haben, vgl. Totschnig 2020; für die Argumentation, dass sie keine Empathie an den Tag legen können, vgl. Montemayor, Halpern und Fairweather 2022.

von Ferrario, Loi und Viganò auf der einen und Philip Nickel auf der anderen Seite.

Kritischer Ausgangspunkt für Ferrario et al. ist die erwähnte Position von Hatherley, nach der wir KI-Algorithmen im medizinischen Bereich nicht vertrauen können. Sie beanstanden an Hatherleys Strategie, dass „although convenient, the choice of applying human trust to human-AI interactions is not fully justified. This begs the question against AI." (Ferrario, Loi und Viganò 2021, 437). Es wird aber nicht ganz klar, worin diese einleitende Kritik der Autor:innen hinauslaufen soll. Hatherley macht darauf aufmerksam, dass Vertrauen bestimmte Aspekte beinhaltet, die es unmöglich machen, von Vertrauen in KI zu reden. Daran ist zunächst noch gar nichts ‚question begging'. Die Autor:innen schreiben an dieser Stelle weiter: „Rather, we shall strive, as much as possible, to identify a meaningful concept of trust that is applicable to human-human and human-AI relations" (ebd.). Es fragt sich aber, warum man gerade diesen Weg einschlagen sollte. Es scheint, dass hier mit der Annahme begonnen wird, dass Vertrauen in KI *möglich sein muss*, um dann in einem zweiten Schritt eine Definition vorzuschlagen, die eben dieses Desiderat erfüllt. Das wäre aber eine unzulässige Argumentation, zumindest, solange man keine guten Gründe vorbringt, weshalb wir einen solchen Begriff des Vertrauens benötigen.

Dass wir in dieser Weise *reden*, ist kein besonders guter Grund. Es ist unbestritten, dass wir in unserer Alltagssprache das Vertrauensvokabular viel großzügiger verwenden, als es eine Begriffsanalyse zulassen würde. ‚Können wir KI vertrauen?' oder ‚Sollen wir KI misstrauen?' sind Fragen, die kompetente Sprecher:innen des Deutschen gut verstehen. Das bedeutet aber nicht automatisch, dass ihre Implikation – nämlich, dass wir KI prinzipiell vertrauen könnten – richtig ist. Dass Vertrauen eine Reaktion auf Unsicherheit ist und dass der Einsatz von KI Verunsicherungen schafft, sind ebenfalls keine guten Gründe für die Annahme, dass wir KI vertrauen können müssen. Wie ich abschließend andeuten werde, könnte man dieser Verunsicherung nämlich auch durch Mechanismen der Verlässlichkeit zu begegnen versuchen. Es wäre vollkommen ausreichend, so meine Position, wenn wir uns auf KI-Systeme verlassen könnten.

Ferrario et al. wollen aber weiter gehen. Dazu müssen sie einen Begriff von Vertrauen bestimmen, der auf KI anwendbar ist, gleichzeitig aber den Unterschied zwischen Vertrauen und Sich-Verlassen einfängt. Die Autor:innen operieren hier mit dem Begriff des ‚simple trust', der ihrer Ansicht nach immer dann vorliegt, wenn wir uns darauf verlassen, dass etwas passie-

ren wird *und* darauf verzichten, das in Frage stehende Vertrauensobjekt im Hinblick auf seine Vertrauenswürdigkeit zu überwachen: „It is this peculiar concession of the physician to economise on monitoring that characterizes the trust relation and distinguishes it from sheer reliance" (ebd., 438). Das Nicht-Überwachen wird auf diese Weise zur entscheidenden Bedingung, deren Erfüllung aus bloßem Sich-Verlassen genuines Vertrauen macht.

Ähnlich wie im Fall der evidenzbasierten Ansätze besteht das Problem mit diesem Vorschlag darin, dass man auf seiner Grundlage dem Unterschied zwischen Vertrauen und Sich-Verlassen nicht gerecht werden kann. Ein gutes epistemisches Kriterium für diesen Unterschied besteht darin, dass wir uns in einem emphatischen Sinne betrogen fühlen dürfen, wenn unser Vertrauen enttäuscht wurde, während wir lediglich enttäuscht sein können, wenn wir uns auf etwas verlassen haben, das dann nicht eingetreten ist. Es bleibt aber unverständlich, warum es diesen Unterschied in der emotional angebrachten Reaktion geben sollte, wenn Vertrauen lediglich mit Nicht-Überwachen zu tun hat: Wieso sollte ich mit ‚persönlichen' reaktiven Einstellungen wie Groll reagieren, wenn ich mich auf etwas verlassen habe, ohne es zu überwachen, während diese Einstellungen nicht angebracht wären, hätte ich mich auf dasselbe verlassen, ohne auf die Überwachung zu verzichten? Es kann nicht sein, dass der bloße Verzicht auf Überwachung diese Veränderung in der normativen Situation zur Folge hat.

Auch scheinen die Autor:innen zu übersehen, dass Situationen, in denen wir uns verlassen, ohne zu vertrauen, *sehr häufig* von einem Verzicht auf Überwachung begleitet sind.[22] Ich verlasse mich darauf, dass die Räder an meinem Fahrrad festgeschraubt sind, und verzichte auf eine tägliche Kontrolle. Ich wüsste nicht einmal, wie ich überprüfen sollte, ob die Sonne am Morgen aufgehen wird. In den Augen von Ferrario et al. müssten dies Fälle von Vertrauen sein. Die Gefahr bei einer solchen Proliferation von Vertrauen besteht darin, dass wir aufhören müssten, Vertrauen als ein normativ spezielles Phänomen zu betrachten, das sowohl für die Personen, die vertrauen,

22 Eine weitere Facette dieser Problematik besteht darin, dass die Entscheidung zwischen Monitoring und einem Verzicht darauf nicht binär ist; in jeder Situation, in der wir uns auf etwas oder jemanden verlassen, liegen vielmehr verschiedene Grade an Monitoring vor. Den Autor:innen scheint dies vor Augen zu stehen, denn sie verwenden an einer Stelle ihrer Argumentation die abschwächende Formulierung „no or little monitoring" (ebd.). Sie machen in der Folge aber nichts aus diesem ihren Ansatz potentiell unterminierenden Punkt.

als auch für die Personen, denen vertraut wird, besondere Gründe generiert und sie in einem Netzwerk reaktiver Einstellungen situiert. Einer Freundin zu vertrauen und das ‚Vertrauen', das man in eine Kaffeetasse setzt, die man zum Mund führt, würden sich dann nur dem Gegenstand nach unterscheiden, aber sie wären nicht Ausdruck einer grundsätzlich unterschiedenen Haltung.

Ferrario et al. könnten an dieser Stelle darauf beharren, dass Verlassen ohne Monitoring als eine Form des Vertrauens zählen sollte, die eben zu unterscheiden ist von dem Vertrauen, das wir in menschliche Akteur:innen setzen. Damit würden sie allerdings ihrem Vorhaben untreu, eine Vertrauensbestimmung vorzulegen, die sowohl auf zwischenmenschliche Konstellationen als auch auf solche zwischen Menschen und KI-Systemen anwendbar ist. Schlimmer noch: Ihr Vorschlag würde auf diese Weise als das Ergebnis eines banalen Streits um Worte gelesen werden müssen. Gegner dieser Sicht der Dinge wie Ryan, Hatherley oder auch ich würden dann sehr einfach entgegnen können, dass Ferrario et al. gerne von ‚simple trust' in KI-Systeme reden können, allerdings ohne dass dies in einem relevanten Sinne mit Vertrauen in KI zu tun hätte, weil es ja – wie sie dann selbst zugeben müssten – noch die andere genuin interpersonale Form von Vertrauen gibt, und in *dieser* Hinsicht würden wir KI-Systemen nicht vertrauen können.

Eine andere Weise, diese Kritik an Ferrario et al. zu formulieren, besteht darin, den Autor:innen vorzuwerfen, dass sie einen normativ leeren Vertrauensbegriff zu rekonstruieren versuchen, dessen Rolle sich durch nichts von der prädiktiven Rolle von bloßem Sich-Verlassen unterscheidet. Genau diese Kritik motiviert Philip Nickel, einen explizit normativen Ansatz zu formulieren, der aber gleichzeitig die Bezugnahme auf Vertrauen in medizinische KI-Systeme erlaubt. Nickels Ansatz stellt das andere Ende des Spektrums der optimistischen Positionen bezüglich der Frage nach Vertrauen in KI dar, aber sein Ansatz ist letzten Endes ähnlich problematisch wie die normativ allzu anspruchslose Position von Ferrario et al.

Wollte man Nickels Vertrauensbestimmung auf einen Satz verkürzen, so ließe sich sagen, dass Vertrauen für ihn darin besteht, dass man die Disposition hat, jemandem oder etwas in einem bestimmten Handlungskontext eine Entscheidungsbefugnis einzuräumen, und zwar nicht nur in prädiktiver Hinsicht, sondern auch, weil man normative Erwartungen an diese Person oder diesen Gegenstand hat: „In trust, one accords another entity discretionary authority because one has relevant normative and predictive expec-

tations toward it“ (Nickel 2022). Ich kann z. B. eine Buchhaltungssoftware bestimmte Beträge addieren lassen und aufgrund meiner bisherigen Erfahrungen damit annehmen, dass sie den richtigen Betrag ausrechnen wird. Bei dieser Vorhersage darüber, wie angemessen die Software die Aufgabe erfüllen wird, handelt es sich um bloßes Sich-Verlassen. Um aus einem Fall von Sich-Verlassen einen Fall von Vertrauen zu machen, muss zu dieser prädiktiven Komponente noch eine normative Komponente dazukommen. Nickel bedient sich hier der bereits im Zusammenhang mit nicht-evidenzbasierten Theorien thematisierten Strategie, eine explizit normative Erwartung ins Spiel zu bringen.[23]

Die unmittelbare Reaktion auf dieses Manöver besteht in der Nachfrage, wie es denn sein kann, dass wir normative Erwartungen an KI-Algorithmen haben. Der Standardauffassung zufolge können wir solche Erwartungen nur an Personen haben, weil wir nur Personen für verantwortlich für ihre Handlungen machen und ihnen gegenüber reaktive Einstellungen wie Groll, Lob oder Tadel an den Tag legen können.[24] Etwas von jemandem zu erwarten, ist Bestandteil dieses komplexen normativen Zusammenhangs, und das erklärt auch, warum wir nicht ernsthaft sagen würden, dass wir etwas von einem Gegenstand oder sogar von einem Tier erwarten. Gibt es irgendetwas an KI-Technologien, das sie dennoch prädestiniert dafür macht, Objekte von normativen Erwartungen zu sein?

Die Geschichte, die Nickel an dieser Stelle erzählt, bleibt systematisch unbefriedigend. Zum einen konstruiert er einen allzu engen begrifflichen

23 Nickel scheint dem oben formulierten Problem, das sich nicht-evidenzbasierten Ansätzen stellt, zu entgehen, indem er der normativen Erwartung eine prädiktive an die Seite stellt. Aber wie verhalten sich beide zueinander? Ergibt sich etwa die prädiktive Erwartung aus der normativen? Oder ist das Verhältnis umgekehrt? Beide Optionen scheinen unplausibel: Nur weil ich etwas fordern kann, spricht das nicht dafür, dass es auch tatsächlich passieren wird. Nur weil ich Grund zu der Annahme habe, dass etwas eintreten wird, bedeutet das nicht, das es normativ gefordert ist. Wenn beide Komponenten aber unabhängig voneinander sind, lässt sich dann nicht plausibel fordern, dass Vertrauen mit *einer* von beiden zu identifizieren wäre? Wenn es mit der prädiktiven Erwartung identifiziert wird, taucht erneut das Problem auf, dass sich Vertrauen nicht mehr von Sich-Verlassen unterscheiden lässt. Wenn es mit der normativen Erwartung identifiziert wird, stellt sich wiederum das Problem, dass nicht mehr klar ist, warum Vertrauen für uns als Handelnde wichtig ist.

24 Für den *locus classicus* vgl. Strawson 1962 und die Ausführungen in Wallace 1994.

Zusammenhang zwischen normativen Erwartungen und der *Funktion* von Gegenständen. Ich kann von einem KI-Algorithmus zur Diagnose von Hautkrebs erwarten, dass er auf angemessene Weise Melanome identifiziert, so die Auffassung von Nickel, weil es die Funktion des KI-Algorithmus ist, genau das zu tun. Es mag nun sein, dass Erwartungen, die wir auf diese Weise an einen medizinischen KI-Algorithmus haben, nicht rein prädiktiver Natur sind. Aus der Tatsache, dass eine Erwartung nicht nur prädiktiv ist, folgt aber nicht, dass sie in dem für Vertrauen relevanten Sinne normativ ist. Dass David das Geheimnis nicht ausplaudern sollte, das ich ihm anvertraut habe, ist deutlich unterschieden davon, dass ein KI-Algorithmus Melanome in dermatoskopischen Bildern identifizieren sollte. David kann ich Vorwürfe machen, sollte er mein Geheimnis nicht für sich behalten, während es absurd wäre, wollte man ein KI-System, das eine Fehldiagnose gemacht hat, mit Ausrufen wie ‚Wie konntest du nur?' konfrontieren.

An dieser Stelle drängt sich die folgende Frage auf: Es mag sein, dass wir dem KI-Algorithmus keine Vorwürfe machen würden, aber vielleicht würden wir in einem klaren Fall von Fehlfunktion den Programmierer:innen des Algorithmus oder den Personen, die das betreffende KI-Unternehmen leiten, Vorwürfe machen. Erstaunlicherweise behauptet Nickel genau das, wenn er darauf hinweist, dass sich das normative ‚commitment' in der Situation, in der wir einem KI-Algorithmus vertrauen, daraus speist, dass die KI-Hersteller:innen (in seiner Terminologie die ‚AI practitioners') dieses Vertrauen eingeladen haben: „By inviting clinicians' trust, AI practitioners create normative commitments, inviting the user to give discretionary authority to the application on the basis of normative expectations about its performance" (ebd.). Erstaunlich ist diese Argumentation, weil sie Nickels zuvor gewählte Strategie zu konterkarieren scheint: Er hatte ja behauptet, dass normative Erwartungen in dem bloßen Vorliegen der Funktion eines Artefakts verankert sind, und nun sollen die KI-Hersteller:innen die zentrale Urheberinstanz für normative Erwartungen sein.

Die Idee, dass die Hersteller:innen von Artefakten die eigentlichen Adressat:innen von Vertrauen sind, ist zunächst nicht unplausibel. Auch wenn man an dieser Stelle darüber streiten könnte, ob unser Verhältnis zu uns unbekannten Firmenchef:innen, Entwickler:innen oder Ingenieur:innen am besten im Sinne einer Vertrauensbeziehung zu verstehen ist, wie ich es für das Verhältnis zwischen Patient:innen und Ärzt:innen fordere, so ist doch die Hypothese, dass es sich dabei um genuine Fälle von Vertrauen handeln kann, nicht unplausibel: Immerhin sind diese Akteur:innen Personen, die die für eine

Vertrauensbeziehung notwendigen Fähigkeiten haben können. Nickel möchte aber bei so einer Position nicht stehen bleiben. Seiner Ansicht nach können wir im eigentlichen Sinne des Wortes *einer medizinischen KI* vertrauen, wenn wir Vertrauen in die KI-Hersteller:innen setzen, die uns dazu einladen.

Es ist sehr schwer, sich diesen Prozess genauer vorzustellen. Behauptet Nickel etwa, dass wir, nachdem wir so eine Einladung angenommen haben, *beiden* Entitäten – den Hersteller:innen und ihren KI-Anwendungen – vertrauen? Aber in diesem Fall wäre es plausibler, davon auszugehen, dass sich unsere Erwartungen aufteilen, dass wir also normative Erwartungen an die Hersteller:innen und prädiktive Erwartungen an die KI-Anwendungen haben. Oder ist gemeint, dass die Hersteller:innen das in sie gesetzte Vertrauen auf die KI-Anwendungen *übertragen*? Aber das würde bedeuten, dass die Hersteller:innen aufhören, Adressat:innen unseres Vertrauens zu sein, sobald wir anfangen ihren Anwendungen zu vertrauen, – und das scheint doch sehr unplausibel. Zudem lauert bei dieser Interpretation eine weitere Gefahr: Ein solcher Vertrauenstransfer könnte zur Begründung für eine Verschiebung der Verantwortung für die auf der KI-Anwendung basierenden Handlungen verwendet werden. Die Hersteller:innen wären nicht mehr zuständig, wenn unsere Erwartungen nicht erfüllt werden, könnte dann argumentiert werden, weil unser Vertrauen sich ja nicht auf sie, sondern auf die KI-Anwendung selbst richtet.[25]

Tatsächlich möchte Nickel nicht bestreiten, dass wir in der Hauptsache den Hersteller:innen von KI-Anwendungen vertrauen, gleichzeitig scheint er aber auch daran festhalten zu wollen, dass auch die KI-Anwendungen selbst Adressat:innen unseres Vertrauens sind. Das ist der Grund für die Unschärfe in seiner Formulierung, dass „[t]he clinician trusts the practitioners *through* the application" (ebd., Herv. im Orig.). Ähnlich wie im Fall von Ferrario et al. scheint seine These, dass wir KI-Algorithmen im medizinischen Bereich vertrauen können, trotz der komplexen theoretischen Ressourcen, die er aufwendet, nicht hinreichend begründet zu sein. Nickels Position hat zudem ein weiteres Problem mit der Position von Ferrario et al. gemeinsam: Auch sie hat

25 Tatsächlich kann vermutet werden, dass die gegenwärtig ubiquitäre Bezugnahme auf ‚Vertrauen in KI' zumindest von Seiten der Industrie genau diesen Effekt haben soll: Die scheinbare Autonomie von KI-Algorithmen wird auf diese Weise dazu verwendet, um sich auf recht durchsichtige Weise der eigenen Verantwortung zu entledigen; vgl. hierzu Ryan 2020, 2762: „Referring to AI as trustworthy would inappropriately elevate AI, while disavowing the responsibility of those developing and implementing it."

eine unplausible Proliferation von Vertrauen zur Folge. Wann immer eine Person etwas herstellt, lädt sie Nickel zufolge implizit dazu ein, darauf zu vertrauen, dass das von ihr hergestellte Artefakt seine Funktion erfüllen wird; und in jedem dieser Fälle muss es sich seiner Ansicht nach um einen Fall handeln, in dem Personen dem betreffenden Artefakt vertrauen können. Wenn ich einen Schal stricke, und jemand legt sich ihn um den Hals, ist das auf der Basis der Theorie von Nickel ein Fall, in dem einem Schal vertraut wurde. Auch auf diese Weise droht der Vertrauensbegriff seinen Witz zu verlieren. Wir sollten diesen Preis nicht zahlen, nur weil es im medizinischen Kontext aufseiten der KI-Industrie, aber auch bei Krankenhausleitungen und Ärzt:innen ein Interesse daran geben mag, von vertrauenswürdiger KI zu reden, um Patient:innen etwaige Sorgen vor einer neuartigen Technologie zu nehmen.

5 Schluss

Ich habe dafür argumentiert, dass wir KI-Anwendungen im medizinischen Bereich nicht vertrauen können, weil Vertrauen als genuin interpersonelles Phänomen in diesem Kontext die falsche Kategorie darstellt. Ist die Forderung nach vertrauenswürdiger KI in der Medizin demnach fehlgeleitet? Ich denke, sie ist es nicht, wenn mit ‚vertrauenswürdig' etwas anderes gemeint ist, als KI-Systeme so zu programmieren, dass wir eine Vertrauensbeziehung zu ihnen aufbauen. Ich habe bereits darauf hingewiesen, dass in konkreten Situationen epistemischer Unsicherheit die Einstellung des bloßen Sich-Verlassens für uns als Akteur:innen ähnlich hilfreich sein kann, wie das, was in zwischenmenschlichen Konstellationen von einer Vertrauensbeziehung geleistet werden kann. Die Einstellung des Sich-Verlassens ist zudem flexibler, weil wir in verschiedenen Kontexten sehr viele Dinge tun können, um Verlässlichkeit herzustellen, während es oft nahezu aussichtslos erscheint, eine Vertrauensbeziehung zu forcieren.

An dieser Stelle kann ich selbstverständlich keine Empfehlungen dazu abgeben, wie wir KI-Anwendungen im medizinischen Bereich verlässlich machen sollten, weil das in der Hauptsache eine technische Frage darstellt. Man wird sich dabei primär am Wohl der Patient:innen orientieren müssen. Verlässlichkeit ist eine Relation, bei der mit Wahrscheinlichkeiten operiert werden kann. Sollte sich herausstellen, dass die Wahrscheinlichkeit, mit der eine KI-Anwendung eine korrekte Diagnose stellt, höher ist als im Fall der Diagnosen menschlicher Expert:innen, dann spricht aus der Perspektive des Wohlergehens von Patient:innen zunächst alles dafür, die KI-Anwendung

auch tatsächlich einzusetzen. KI-Anwendungen wären dann nichts anderes als Instrumente, die wir in der medizinischen Praxis verwenden sollten, wenn sie zu den besten Ergebnissen führen, genauso wie Ärzt:innen heutzutage ohne zu zögern Fieberthermometer verwenden, anstatt Patient:innen die Hand auf die Stirn zu legen.

Damit ist nicht behauptet, dass es neben dem Patient:innenwohl keine anderen Erwägungen gibt, die man beim Einsatz von KI-Technologien im medizinischen Bereich berücksichtigen sollte. Eingangs habe ich bereits erwähnt, dass sich mit KI-Technologien Probleme der Privatheit, des Datenschutzes oder der diskriminierenden Verzerrungen verbinden. Dabei handelt es sich aber lediglich um zusätzliche Aspekte der Verlässlichkeit von KI-Anwendungen, und wir sollten auch hier nach geeigneten Lösungen suchen und gegebenenfalls bestimmte Formen von KI-Technologie stärker regulieren. Neben der technischen Seite dieser Herausforderung werden dabei genuin normative Debatten geführt werden müssen, in denen die erwähnten Faktoren gegeneinander abgewogen werden. Es versteht sich etwa nicht von selbst, wie viel Privatheit aufs Spiel gesetzt werden kann, wenn es darum geht, Patient:innenleben zu retten. Fragen, die solche Wertekonflikte betreffen, sind aber kaum spezifisch für KI-Technologien. Wir führen die entsprechenden Diskussionen bereits seit Jahrzehnten. Vertrauen zwischen Ärzt:innen und Patient:innen ist ein Wert, der in solchen Abwägungen an zentraler Stelle mitberücksichtigt werden sollte. Welchen Einfluss haben aber verlässliche KI-Technologien auf dieses Vertrauen, wenn man es, wie ich es vorgeschlagen habe, im Sinne einer Vertrauensbeziehung versteht?

Eine mögliche Antwort fällt pessimistisch aus: Je mehr Aufgaben im medizinischen Bereich von KI-Anwendungen übernommen werden, desto schwächer wird das Vertrauen zwischen Ärzt:innen und Patient:innen, behauptet etwa Hatherley, weil dadurch die „epistemic authority of human clinicians“ (Hatherley 2020, 478) verdrängt werde. An dieser Stelle sieht man gut, inwiefern mein Ansatz sich von Hatherleys Ansatz unterscheidet. Hatherley versteht Vertrauen meiner Ansicht nach allzu instrumentell und strategisch, und so ist es kein Wunder, dass er so viel Gewicht auf die epistemische Autorität der Ärzt:innen legt, die tatsächlich von leistungsfähigen und verlässlichen KI-Systemen untergraben werden könnte. Aus der systematischen Perspektive des Vertrauensansatzes, den ich vorgeschlagen habe, ist das eine verzerrte Sicht der Dinge. In einer Vertrauensbeziehung kommt es nicht nur darauf an, dass jemand eine epistemische Autorität hat, sondern zentral auch darauf, dass die Beziehungspartner:innen einander verstehen

und die normative Perspektive der jeweils anderen Person einnehmen können. Für Ärzt:innen, die sich in ihrer Arbeit KI-Technologien bedienen, bedeutet dies, dass sie nicht befürchten müssen, dass sie dadurch die Vertrauensbeziehung zu ihren Patient:innen schwächen. Man beachte wiederum die Analogie zur Freundschaft: Dass ich jemanden kennenlerne, der besser Auto fahren kann, ist keine Bedrohung für die Vertrauensbeziehung zu einer Freundin, die mich sonst zum Bahnhof zu fahren pflegte.

Auf ähnliche Weise könnte sich auf der Grundlage meines Vorschlags das viel diskutierte Problem entschärfen lassen, das sich mit der Black-Box-Natur mancher KI-Algorithmen verbindet. Es stellt sich insbesondere im Zusammenhang mit neuronalen Netzwerken, deren komplexe interne Struktur sogenannte ‚hidden layers' enthält, die es zum Teil unmöglich machen, die Entstehung des Ergebnisses auf der Ausgabeschicht des Algorithmus nachzuvollziehen. Sollten Ärzt:innen sich eines KI-Algorithmus bedienen, ohne Patient:innen erklären zu können, wie dieser Algorithmus zu seinem Ergebnis gekommen ist, könnten Patient:innen sich die Frage stellen, warum sie den Ärzt:innen überhaupt vertrauen sollen.[26] Versteht man Vertrauen als eine Beziehung, stellt sich diese Frage allerdings nicht in dieser Schärfe, weil es in einer Vertrauensbeziehung nicht primär darum geht, dass etwas erklärt wird. Man könnte sogar dafür argumentieren, dass Vertrauensbeziehungen den Ort darstellen, an dem umständliche Erklärungen unnötig werden.

Das bedeutet nun keinesfalls, dass wir uns blind auf die Ergebnisse von medizinischen KI-Algorithmen verlassen sollten. Auch hier sollten wir Grade an Verlässlichkeit anstreben, die relativ zu den Gütern, die auf dem Spiel stehen, angemessen sind.[27] Auch wären Ärzt:innen gut beraten, ihre Patient:innen darüber aufzuklären, mit welcher Wahrscheinlichkeit ein KI-Algorithmus in seinem Diagnose- oder Therapievorschlag richtig liegt. Aber es würde keinen Vertrauensverlust darstellen, sollten sie nicht in der Lage sein, detailliert zu erläutern, wie der Algorithmus zu seinem Ergebnis gekommen ist. Erklärungen sind nie vollständig. Die wenigsten Ärzt:innen könnten alle medizinischen Faktoren erläutern, die in eine Diagnose an einem modernen Universitätskrankenhaus einfließen, und noch weniger Patient:innen könnten mit so einer Erklärung auch wirklich etwas anfangen. Es gibt wohl kaum Ärzt:innen, die mit ihren Patient:innen über die Struktur der

26 Vgl. Watson et al. 2019.

27 Vgl. die Ausführungen zum ‚computational reliabilism' in Durán und Jongsma 2021.

6-Aminopenicillansäure reden, wenn sie ihnen ein Antibiotikum verschreiben, und für eine gelungene Vertrauensbeziehung ist das auch unnötig.

In einer solchen Vertrauensbeziehung müssen Ärzt:innen und Patient:innen vielmehr einander auf eine spezifische Weise betrachten, habe ich argumentiert. Ich habe den historisch gewachsenen Charakter der Vertrauensbeziehung zwischen Ärzt:innen und Patient:innen betont und dabei immer wieder darauf aufmerksam gemacht, wie wichtig empathische Fähigkeiten, Fürsorge und das Bemühen um ein wechselseitiges Verstehen sind. Das alles sind Prozesse, die viel Zeit und Energie kosten – Zeit und Energie, die Ärzt:innen immer weniger zur Verfügung stehen, so dass Vertrauensbeziehungen oft gar nicht erst entstehen können. In genau diesem Zusammenhang ist die eigentliche Herausforderung von KI-Technologien im medizinischen Bereich zu suchen.

Die zentrale Frage lautet, ob wir es schaffen, diese Technologien so einzusetzen, dass sie Raum für Vertrauensbeziehungen zwischen Ärzt:innen und Patient:innen schaffen. Immerhin sind KI-Algorithmen wie dafür gemacht, Aufgaben zu übernehmen, die Ärzt:innen heutzutage routinemäßig davon abhalten, sich ihren Patient:innen zu widmen. Andererseits steht zu vermuten, dass Ärzt:innen die Arbeit von KI-Anwendungen zumindest in der nächsten Zukunft werden überwachen und kontrollieren müssen, was eher einen zusätzlichen Arbeitsaufwand als eine Entlastung für sie bedeuten könnte. Zudem ist es nicht so, dass die zum Teil dramatischen Verhältnisse im Gesundheitswesen, die heutzutage oft das Entstehen von Vertrauensbeziehungen verhindern, eine Naturnotwendigkeit darstellen: Sie sind das Ergebnis einer nutzenmaximierenden Haltung, die die Arbeit von Ärzt:innen auf das Wiederherstellen der Gesundheit von entindividualisierten Patient:innen reduziert. Sollten wir uns als Gesellschaft weiterhin mit diesem Trend abfinden, droht die Gefahr, dass überall dort, wo KI-Anwendungen eingesetzt werden können, ökonomischer Druck zu einer Verdrängung menschlicher Akteur:innen im Gesundheitswesen führt. Soweit muss es aber nicht kommen. In dieser Hinsicht stellen die rasanten Entwicklungen von KI-Technologien, wie wir sie zurzeit erleben, eine Chance dar, darauf zu reflektieren, was uns genuin zwischenmenschlicher Kontakt im medizinischen Bereich bedeutet, um auf diese Weise eine neue Wertschätzung von Vertrauensbeziehungen zwischen Ärzt:innen und Patient:innen zu gewinnen.[28]

28 Dieser Text ist im Rahmen des Nationalen Forschungsprogramms «Digitale Transformation» (NFP 77) des Schweizerischen Nationalfonds (SNF) ent-

Literatur

Baier, Annette. 1986. „Trust and Antitrust". *Ethics* 96: 231–260.

Beauchamp, Tom L., und James F. Childress. 2001. *Principles of Biomedical Ethics*. 5. Aufl. Oxford: Oxford University Press.

Budnik, Christian. 2021. *Vertrauensbeziehungen. Normativität und Dynamik eines interpersonellen Phänomens*. Berlin/Boston: De Gruyter.

Cates, Diana Fritz, und Paul Lauritzen. 2001. *Medicine and the Ethics of Care*. Washington, D.C.: Georgetown University Press.

Domenicucci, Jacopo, und Richard Holton. 2017. „Trust as a Two-Place Relation". In *The Philosophy of Trust*, herausgegeben von Paul Faulkner und Thomas Simpson, 149–160. Oxford: Oxford University Press.

Durán, Juan Manuel, und Karin Rolanda Jongsma. 2021. „Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI". *Journal of Medical Ethics* 47: 329–335.

Europäische Kommission. 2020. *Weissbuch. Zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen*. https://commission.europa.eu/document/d2ec4039-c5be-423a-81ef-b9e44e79825b_de.

Faulkner, Paul. 2007. „On Telling and Trusting". *Mind* 116: 875–902.

Faulkner, Paul. 2015. „The Attitude of Trust is Basic". *Analysis* 75: 424–429.

Ferrario, Andrea, Michele Loi und Eleonora Viganò. 2021. „Trust does not need to be human: it is possible to trust medical AI". *Journal of Medical Ethics* 47: 437–438.

Halpern, Jodi. 2001. *From Detached Concern to Empathy. Humanizing Medical Practice*. Oxford: Oxford University Press.

Hardin, Russell. 1999. „Do we want trust in government?". In *Democracy and Trust*, herausgegeben von Mark E. Warren, 22–41. Cambridge: Cambridge University Press.

Hatherley, Joshua James. 2020. „Limits of trust in medical AI". *Journal of Medical Ethics* 46 (7): 478–481.

Hawley, Katherine. 2014. „Trust, Distrust and Commitment". *NOÛS* 48 (1): 1–20.

Held, Virginia. 2006. *The Ethics of Care. Personal, Political, Global*. Oxford: Oxford University Press.

Hieronymi, Pamela. 2008. „The Reasons of Trust". *Australasian Journal of Philosophy* 86: 213–236.

standen (Projekt 187446). Für hilfreiche Kommentare danke ich Holger Baumann, Monika Betzler, Karoline Reinhardt, Peter Schaber, Johanna Sinn, den Teilnehmenden des Oberseminars «Aktuelle Forschungsthemen der Praktischen Philosophie» an der Ludwig-Maximilians-Universität München sowie zwei anonymen Gutachter:innen.

Holton, Richard. 1994. „Deciding to trust, coming to believe“. *Australasian Journal of Philosophy* 72: 63–76.

Jones, Karen. 1996. „Trust as an Affective Attitude“. *Ethics* 107: 4–25.

McLeod, Carolyn. 2002. *Self-Trust and Reproductive Autonomy*. Cambridge, MA: MIT Press.

McMyler, Benjamin. 2011. *Testimony, Trust, and Authority*. Oxford: Oxford University Press.

Montemayor, Carlos, Jodi Halpern und Abrol Fairweather. 2022. „In principle obstacles for empathic AI: why we can't replace human empathy in healthcare“. *AI & Society* 37: 1353–1359. https://doi.org/10.1007/s00146-021-01230-z.

Nortvedt, Per, Marit Helene Hem und Helge Skribekk. 2011. „The ethics of care: Role obligations and moderate partiality in health care“. *Nursing Ethics* 18 (2): 192–200.

O'Neill, Onora. 2002. *Autonomy and Trust in Bioethics*. Oxford: Oxford University Press.

O'Neill, Onora. 2018. „Linking Trust to Trustworthiness“. *International Journal of Philosophical Studies* 26: 293–300.

Ryan, Mark. 2020. „In AI We Trust: Ethics, Artificial Intelligence, and Reliability“. *Science and Engineering Ethics* 26, 2749–2767. https://doi.org/10.1007/s11948-020-00228-y.

Simpson, Thomas W. 2018. „Trust, Belief, and the Second-Personal“. *Australasian Journal of Philosophy* 96: 447–459.

Steinfath, Holmer. 2016. „Das Wechselspiel von Autonomie und Vertrauen – eine philosophische Einführung“. In *Autonomie und Vertrauen. Schlüsselbegriffe der modernen Medizin*, herausgegeben von Holmer Steinfath und Claudia Wiesemann, 11–68. Wiesbaden: Springer.

Strawson, Peter F. 1962. „Freedom and Resentment“. In: Peter F. Strawson. 2008. *Freedom and Resentment. And other Essays,* 1–28. New York: Routledge. [Erstveröffentlichung 1962.]

Totschnig, Wolfhart. 2020. „Fully Autonomous AI“. *Science and Engineering Ethics* 26: 2473–2485. https://doi.org/10.1007/s11948-020-00243-z.

Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

Watson, David S., Jenny Krutzinna, Ian N. Bruce et al. 2019. „Clinical applications of machine learning algorithms: beyond the black box“. The BMJ 364: l886.

Wiesemann, Claudia. 2016. „Vertrauen als moralische Praxis – Bedeutung für Medizin und Ethik“. In *Autonomie und Vertrauen. Schlüsselbegriffe der modernen Medizin*, herausgegeben von Holmer Steinfath und Claudia Wiesemann, 69–99. Wiesbaden: Springer.