Ethik in der Medizin
https://doi.org/10.1007/s00481-025-00878-1

ORIGINALARBEIT

KI in der bildgebenden Diagnostik verantwortet vertrauen. Erfahrungen aus Radiologie und Pathologie ethisch diskutiert

Wiebke Brandt · Alexis Fritz · Angelika Kießig · Philipp Lerch

Eingegangen: 25. März 2025 / Angenommen: 16. Juli 2025

Zusammenfassung In der medizinischen bildgebenden Diagnostik kommen zunehmend Assistenzsysteme zum Einsatz, die auf maschinellem Lernen basieren. Können Ärzt:innen jedoch noch Verantwortung für eine Diagnose übernehmen, deren Zustandekommen sie nicht bis ins Letzte verstehen? Wer blind auf die Technik vertraut, kann wohl schwerlich die epistemische Bedingung von Verantwortung erfüllen.

Qualitative Interviews mit Radiolog:innen und Patholog:innen rund um die Themen „Verantwortung" und „Vertrauen" zeigen, dass Ärzt:innen nur dann auf KI-Ergebnisse vertrauen, wenn sie die Möglichkeit zur Kontrolle haben. Kontrolle kann die Grundlage für ein *begründetes* (im Gegensatz zu *blindem*) Vertrauen schaffen. Es lohnt, darüber nachzudenken, ob ein solches begründetes Vertrauen auch die epistemische Bedingung von Verantwortung erfüllen kann: Ärzt:innen müssten dann nicht mehr ein bestimmtes Einzelergebnis der KI überprüfen (können), sondern es wäre ausreichend, dass sie *vorher* die Expertise von KI für diese Aufgabe überprüft haben und fortlaufend reevaluieren.

Es zeigt sich, dass die skizzierten Herausforderungen keineswegs KI-spezifisch sind: Verantwortungsübernahme trotz mangelndem Wissen begegnet u. a. auch im Bereich der Delegation. Dass innere Prozesse nur schwer nachvollzogen werden können, gilt zudem auch für Menschen und andere technische Geräte.

Um KI verantwortet vertrauen zu können, brauchen Ärzt:innen ein informationstechnologisches Grundverständnis sowie ein geschärftes Bewusstsein für die jeweili-

✉ Wiebke Brandt
Theologische Fakultät, Lehrstuhl für Moraltheologie, Katholische Universität Eichstätt-Ingolstadt, 85072 Eichstätt, Deutschland
E-Mail: Wiebke.Brandt@ku.de

Prof. Dr. Alexis Fritz
Albert-Ludwigs-Universität Freiburg, Freiburg, Deutschland

Angelika Kießig · Philipp Lerch
Katholische Universität Eichstätt-Ingolstadt, 85072 Eichstätt, Deutschland

Published online: 02 September 2025

gen Stärken von Mensch und Maschine. Sie sollten generell die Möglichkeit haben, KI-Ergebnisse nachzuvollziehen, allerdings sollte differenziert evaluiert werden, für welche Prozesse Nachvollziehbarkeit notwendig und hilfreich ist. Entwickler:innen von KI könnten durch eine Produkthaftung stärker in die Verantwortung genommen werden.

Schlüsselwörter Mensch-Maschine-Interaktion · Künstliche Intelligenz/ Maschinelles Lernen · Medizinische Diagnostik · Verantwortung · Vertrauen

Responsibly trusting artificial intelligence in medical imaging diagnostics. Ethical discussion of experiences in radiology and pathology

Abstract

Definition of the problem Assistance systems based on machine learning are increasingly being used in medical imaging diagnostics. But can physicians still take responsibility for a diagnosis if they do not fully understand how it came about? Those who blindly trust in technology can hardly fulfill the epistemic condition of responsibility.

Arguments Qualitative interviews with radiologists and pathologists on the topics of "responsibility" and "trust" show that physicians only trust artificial intelligence (AI) results if they have the opportunity to control them. Control can create the basis for *justified* (as opposed to *blind*) trust. It is worth considering whether such justified trust can also fulfill the epistemic condition of responsibility: Physicians would then no longer have to (be able to) check a specific individual result of the AI, but it would be sufficient that they have *previously* checked and continuously re-evaluate the expertise of AI for this task. It is clear that the challenges outlined above are by no means specific to AI: Assuming responsibility despite a lack of knowledge is also encountered, among others, in the area of delegation. The fact that internal processes are difficult to comprehend also applies to humans and other technical devices.

Conclusion In order to trust AI in a responsible way, physicians need a basic understanding of information technology and a heightened awareness of the respective strengths of humans and machines. They should generally have the opportunity to understand AI results, but there should be a differentiated evaluation of the processes for which traceability is necessary and helpful. AI developers could be held more accountable through product liability.

Keywords Human-machine interaction · Artificial intelligence/machine learning · Medical diagnostics · Responsibility · Trust

Einleitung

Kann eine Radiologin noch Verantwortung für eine Diagnose übernehmen, die mithilfe von KI – genauer mithilfe von Maschinellem Lernen (ML)[1] – erstellt wurde? Wenn sie das ML-System nicht durchschaut und also nicht weiß, wie das Ergebnis zustande gekommen ist, dürfte sie eine wichtige Bedingung für Verantwortung nicht mehr erfüllen. Gleichzeitig scheint es unverantwortlich, Patient:innen KI-Leistungen vorzuenthalten, welche die Diagnostik effizienter und besser machen, möglicherweise sogar medizinische Ressourcen in unterversorgten Gebieten zur Verfügung stellen – ‚nur' weil man die Funktionsweise der KI nicht versteht.

Es wird also ein verantwortbarer Mittelweg gesucht, wie ML-Systeme trotz ihrer augenscheinlichen Opazität zum Wohl der Patient:innen eingesetzt werden können. Dazu richtet sich der Blick auf das Phänomen des Vertrauens: Wenn Ärzt:innen ML nicht verstehen, es aber dennoch benutzen wollen, müssen sie wohl oder übel vertrauen. Doch werden sie ihrer Verantwortung damit noch gerecht? Unter welchen Bedingungen kann möglicherweise verantwortlich vertraut werden?

Vor dem Hintergrund dieser Fragestellungen wurden qualitative Interviews mit Radiolog:innen und Patholog:innen geführt, die mit bildgebenden Verfahren arbeiten.

Das Forschungsinteresse richtet sich dabei ausschließlich auf ML-basierte Assistenzsysteme, die die ärztliche Entscheidungsfindung unterstützen. Eigenständig diagnostizierende KI ist nicht im Blick.

Die Hauptachse der Interviews spannt sich auf zwischen den Themen „Verantwortung" und „Vertrauen". Außerdem wurde gefragt, wie ein verantwortetes Vertrauensverhältnis zu ML konkret gefördert werden könnte: Welche Kompetenzen brauchen Ärzt:innen im Umgang mit ML und welche Gestaltungsimpulse gibt es an die Adresse von Entwickler:innen eben solcher ML-basierter entscheidungsunterstützender Systeme?

Studiendesign und Untersuchungsmethode

Die Studienergebnisse basieren auf zehn ca. einstündigen, qualitativen Interviews mit Radiolog:innen und Patholog:innen, welche von der Autorin AK im Zeitraum Dezember 2023 bis Februar 2024 (bis auf ein telefonisches Interview) alle via Zoom durchgeführt wurden. Sechs der befragten Ärzt:innen stammen aus der Radiologie, je zwei aus der Humanpathologie bzw. der Veterinärpathologie[2] und arbeiteten zum Zeitpunkt des Interviews in unterschiedlichen Bereichen des deutschen Gesundheitssystems wie radiologischen Zentren, veterinärpathologischen Instituten oder (Universitäts)kliniken. Auch die Erfahrungsspanne der ärztlichen

[1] Innerhalb der Interviews wurde der Einfachheit halber von KI gesprochen. I. d. R waren damit ML-basierte Assistenzsysteme gemeint.

[2] Da Veterinärpathologie und Humanpathologie vergleichbar sind im Hinblick auf Diagnosekriterien, Technikeinsatz, Arbeitsprozesse sowie die Grundstruktur der Diagnose- und Entscheidungsfindung, wurden auch die Perspektiven von Veterinärpatholog:innen in die Studie mit einbezogen.

Tätigkeit als Radiolog:in bzw. Patholog:in variierte zwischen wenig erfahren (fünf vor bzw. in der Fachärzt:innenausbildung), moderat erfahren (zwei mit abgeschlossener Fachärzt:innenausbildung mit ersten Jahren Berufserfahrung), erfahren (drei Oberärzt:innen oder höher mit mehreren Jahren Berufserfahrung). Alle Proband:innen konnten auf praktische Erfahrung mit KI-Tools – auch ML-basiert – im Diagnosekontext bzw. im tatsächlichen Arbeitsalltag zurückgreifen; acht Proband:innen haben zusätzlich selbst zu KI-Anwendungen im jeweiligen Fachgebiet geforscht. Durchgeführt wurden die qualitativen Interviews semistrukturiert auf Grundlage eines Interviewleitfadens, wobei Audiodateien erstellt wurden. Der Kontakt zu den interviewten Ärzt:innen entstand einerseits über das Netzwerk des Forschungsprojektes „Responsibility Gaps in Human-Machine Interactions: The Ambivalence of Trust in AI“ (ReGInA). Andererseits wurden gezielt radiologische bzw. pathologische Bereiche deutscher Universitätskliniken mit der Bitte um Verbreitung der Einladung zum Interview angeschrieben. Zusätzlich wurde eine Aufwandsentschädigung angeboten. Die Befragten gaben an, dass im Wesentlichen das Interesse am Thema des Interviews wie auch die eigene Forschung und Arbeit an bzw. mit ML-Systemen die Motivation an der Teilnahme begründeten.

Der Fokus der Interviews richtete sich auf das Spannungsverhältnis von Verantwortung und Vertrauen im Umgang mit KI. Als Einstieg ins Thema wurde dafür zunächst die grundsätzliche Einstellung zur Nutzung von KI im diagnostischen Kontext erhoben. Wie die konkrete Zusammenarbeit mit KI im eigenen Arbeitsalltag erlebt wird und wie die Rollen bei der Diagnosefindung verteilt sind, wurde ebenso thematisiert wie explizite Reflexionen zu Verantwortung und Vertrauen. Neben der Erhebung des Ist-Zustands wurden auch Anregungen eingeholt, wie die Mensch-Maschine-Interaktion zukünftig gestaltet werden sollte, um das Spannungsverhältnis von Verantwortung und Vertrauen bestmöglich auszutarieren: etwa durch bestimmte Kompetenzen auf Seiten des Menschen, bestimmte Funktionen auf Seiten der Maschine und eine Aufgabenverteilung, die sich an den jeweiligen Stärken von Mensch und Maschine orientiert.

Die einzelnen Themenblöcke waren unterschiedlich ergiebig und variieren entsprechend in Umfang und Tiefe der Darstellung.

Die Fragestellungen für den Interview-Leitfaden waren innerhalb des Forschungsprojekts ReGInA im Team entwickelt und im Austausch mit anderen interdisziplinären Forschungsteams des bidt (Bayerisches Forschungsinstitut für Digitale Transformation) weiter geschärft worden.

Im Anschluss an die Realisierung der Interviews wurden die Audiodateien unter der Anwendung üblicher Protokollierungsregeln wissenschaftlich transkribiert. Die Analyse der verschriftlichten Interviews wurde anhand der qualitativen Inhaltsanalyse nach Philipp Mayring (2022) vorgenommen, wobei insbesondere auf die induktive Kategorienbildung zurückgegriffen wurde. Dabei wurde das Interviewmaterial zu jeder Fragestellung systematisch analysiert und in inhaltlich ähnliche Aussagen gruppiert, um das Antwortspektrum zu strukturieren. Dieser Prozess wurde zwei Mal von unterschiedlichen Personen durchgeführt und die Ergebnisse im Team diskutiert.

Zusammenfassung der Ergebnisse

Einstellung zum Einsatz von ML in der bildgebenden Diagnostik

Grundsätzlich ist bei allen Befragten die Bereitschaft da, mit KI zusammenzuarbeiten. Bedenken, von KI in den eigenen Arbeitsbereichen oder beruflichen Entwicklungschancen eingeschränkt zu werden, sind nicht zu spüren, im Gegenteil: *„[I]n der KI sehe ich eben gerade die Chance, diese Massenabfertigung bewältigen zu können. Ich glaube tatsächlich, dass wir KI brauchen werden auf lange Sicht. Also ich fühle mich überhaupt nicht davon irgendwie bedroht, ja. Das ist jetzt sehr häufig in den, in der Presse, ja? ‚KI schafft den Radiologen ab.' Also wirklich überhaupt nicht. Ich würde mir wünschen, dass es besser oder mehr funktionierende KI-Systeme gibt, weil ich sonst mit meiner Arbeit nicht mehr hinterherkommen kann" (I8).*[3]

Verantwortungszuschreibung beim Einsatz von ML im ärztlichen Diagnoseprozess

In allen Interviews wird das Thema „Verantwortung" vorrangig unter haftungsrechtlicher Perspektive betrachtet. Auf offene Fragen der Interviewerin nach der Verantwortung im Diagnosegeschehen werden in den meisten Fällen Antworten gegeben, die die rechtliche Ebene betreffen (Schadenshaftung). Eine explizite Unterscheidung zwischen moralischer und rechtlicher Verantwortung wird fast nie vorgenommen; auch auf Nachfrage tun sich die Befragten schwer, diese beiden Formen von Verantwortung auseinanderzuhalten. Im Kontext der Schadens-Verantwortung geht es in zwei Interviews allerdings auch um das Thema „Fehlerkultur", wodurch dann doch die Sprache auf moralische Verantwortung kommt: Fehler seien menschlich und dass Menschen Fehler machen, unvermeidbar. Entsprechend wird als Kriterium für (verantwortliches) moralisches Handeln genannt, *„im besten Wissen und Gewissen"* zu handeln (I4).

Als Bedingung für Verantwortung wird angeführt, dass KI-Ergebnisse kontrollierbar und somit auch bis zu einem gewissen Grade nachvollziehbar sein müssen. Dazu gehört z. B., dass neben der KI-Auswertung auch weiterhin entscheidungsrelevantes Material zur Verfügung steht, um ggf. die KI-Ergebnisse mit den eigenen Überlegungen abzugleichen: *„Also im Moment, finde ich, kann man das noch [V. wahrnehmen], weil wir kriegen ja die normalen CT-Bilder. Also alles, was man bisher immer hatte, hat man zum Anschauen da. Und kann dahingehend auch kontrollieren, hat die KI die richtigen drei Herde gefunden? Oder sehe ich noch fünf mehr, oder nicht?" (I1).* Wer sich blind auf KI-Ergebnisse verlässt, ohne sie zumindest in Zweifelsfällen nochmal überprüfen zu können, handelt demnach nicht verantwortlich.: *„Dann spuckt es mir einfach nur noch irgendwas aus und ich muss es glauben, oder ich glaube es halt nicht. Das macht es schwierig, dafür die Verantwortung zu übernehmen" (I1).*

Die persönliche Verantwortung des:r einzelnen Ärzt:in wird als eingebettet in ein Geflecht aus organisationseigenen Hierarchien und Kooperationen verschiedener

[3] Aus den Interviews wird mit der Angabe „I" für „Interview" bzw. „P" für „Proband:in" nebst der jeweiligen Interview-Nummer zitiert.

ärztlicher Fachrichtungen beschrieben. Dadurch wird die eigene Verantwortung als klar begrenzt auf den jeweiligen eigenen Aufgabenbereich wahrgenommen. Nach einhelliger Meinung der Befragten tragen sie in diesem ihren begrenzten Bereich auch dann weiterhin selbst die volle Verantwortung, wenn sie von KI unterstützt werden. Niemand spricht der KI (Schadens-)Verantwortung zu; die Möglichkeit, dass die KI verantwortlich sein könnte, wird entweder gar nicht thematisiert oder explizit verworfen.

Die Rolle der Entwicklungsungsfirmen wird dagegen durchaus reflektiert: Es sollte so sein, dass Hersteller:innen/Entwickler:innen in einem begrenzten Rahmen Verantwortung für ihre Produkte übernehmen, etwa so wie sie auch für andere (Medizin-)Produkte haftbar gemacht werden können, wenn der Schaden nicht auf eine Falschnutzung, sondern z. B. auf Materialmängel zurückgeht: *„Also wenn ich jetzt einen Katheter in den Körper hineinstecke und der bricht mir ab und dieses Plastikteil verstopft ein Gefäß und der Patient klagt, dann kann ich sagen: ‚Also tut mir leid, beim besten Willen, ich habe alles richtig gemacht. Das war ein Materialfehler. Das ist nicht in meiner Verantwortung.‘ Dann muss die Firma mit einem Schadenersatzprozess vielleicht rechnen […] Und genauso ist [es] auch bei Software“ (I6).* Zum Teil wird es als starkes Desiderat geäußert, Entwickler:innen stärker in die Verantwortung zu nehmen, teilweise aber auch im Sinne einer unsicheren Zukunftsprognose ohne normative Komponente: *„Ob das relativ nahe steht, dass KI da mal eigenständig irgendwelche Befunde herausgibt, die dann auch akzeptiert werden, das weiß ich tatsächlich noch nicht. Weil ich mir auch nicht sicher bin, ob da die KI-Firmen irgendeine Art von Verantwortung übernehmen wollen“ (I1).* Einig sind sich die Befragten, dass die Abgabe von Verantwortung an Entwicklungsfirmen derzeit überhaupt nur bei klar umrissenen, repetitiven Aufgaben der KI (wie messen, zählen etc.) infrage kommt, keinesfalls jedoch für das Erstellen einer Gesamtdiagnose.

Das Gespür für die eigene Verantwortung im Diagnosegeschehen könne geschwächt werden durch die Gewöhnung an gute Ergebnisse (*„wenn das Gerät neunmal gut vorgeschlagen hat und einmal nicht und man hat sich drauf verlassen, da kann man sich auch denken: ‚Ja, aber der hat einen Fehler gemacht, […] weil der macht [das] normalerweise gut und so‘“ (I3)).*

Gestärkt werden könne das Verantwortungsgefühl durch Aufklärung über die Arbeitsweise (und Fehlbarkeit) von KI sowie durch eine Benutzerschnittstelle, die nicht nur eine einzige Empfehlung der KI ausgibt, sondern mehrere Optionen (z. B. abgestuft durch Wahrscheinlichkeitsangaben) zur Auswahl stellt, wodurch den menschlichen Ärzt:innen die letztendliche Entscheidung obliegt.

Wahrnehmung des Entscheidungsgeschehens in der ärztlichen Diagnosesituation

KI wird nicht als Expertin auf Augenhöhe wahrgenommen, sondern eher als *„studentische Hilfskraft“* (I8). Um eine fundierte Zweitmeinung einzuholen, wenden sich die interviewten Ärzt:innen lieber an menschliche Expert:innen als an eine KI. Nur wenn der zur Verfügung stehende Mensch nicht im eigenen Fachbereich arbeitet und also kein einschlägiges Expertenwissen hat, ist KI für sie die bessere Ansprechpartnerin.

Sich auf die Ergebnisse der KI zu verlassen, kommt für die meisten nicht infrage. Die Kontrolle gehört dazu, weil beim derzeitigen Entwicklungsstand der KI auch immer mal wieder *„krasse Fehler“* (I6) passieren. Trotz dieser *„gesunden Skepsis“* (I6) wird die Arbeit mit KI als durchaus bereichernd erlebt, weil sie oft auf Dinge aufmerksam macht, die ein Mensch vielleicht übersehen hätte.

Die Einschätzung der Rollenverteilung in der Zusammenarbeit mit KI ist auch bestimmt von dem Bewusstsein, dass man selbst über mehr Kontext-Informationen verfügt (bzw. diese einholen kann) als die KI: *„Dann habe ich natürlich noch einen viel weiteren Horizont als die Maschine. Ich weiß, wie der Patient klinisch ist, ich weiß, wie die dynamische Entwicklung in der letzten halben Stunde war und so weiter. Das blendet die Maschine ja alles aus. Die wird ja nur mit einem kleinen Datenschnipsel-Ausschnitt bedient und ich habe dann natürlich eine breitere andere Datenbasis, mit der ich dann das Ergebnis der Maschine abgleichen kann“ (I6).*

Wenn Mensch und Maschine zum gleichen Ergebnis kommen, wirke das oft bestätigend auf den:die Ärzt:in und verleihe der eigenen Einschätzung besondere Sicherheit. Entsprechend könne Dissens verunsichernd wirken. Wer der KI nicht zustimmt und auch keinen offensichtlichen Fehler findet, könne versuchen, die eigene Einschätzung durch zusätzliche Evidenz oder Rücksprache mit Kolleg:innen zu überprüfen. Am Ende würden standardmäßig alle relevanten Überlegungen dokumentiert. Dazu gehöre, in der Rechtfertigung der eigenen (Diagnose-)Entscheidung die Vorschläge der KI entweder zu befürworten oder begründet abzuweisen.

Vertrauen auf ML-basierte Assistenzsysteme

Beim Thema „Vertrauen“ herrscht Einigkeit, dass man eben nicht *„blind vertrauen“* (I7, I8) darf, sondern die Ergebnisse der KI überprüfen muss. Als Ausnahme von dieser Regel werden überschaubare Aufgaben genannt, die den Stärken der KI besser angepasst sind als denen des Menschen: *„Also repetitive Aufgaben, meine Knochenmessungen, Winkelmessungen zum Beispiel. Da sehe ich absolut mich nicht mehr in der Verantwortung, das zu machen. Wenn ich das Tool habe, dann messe ich das auch nicht nach“ (I8).*

Ob Menschen einer KI eher (zu) viel oder eher (zu) wenig vertrauen, wird in Zusammenhang mit Lebensalter, Berufserfahrung und Selbstvertrauen gebracht: Menschen am Anfang ihrer beruflichen Laufbahn sowie unsichere Persönlichkeiten hätten vielleicht eher die Tendenz der KI viel Vertrauen entgegenzubringen. Dagegen seien gestandene Ärzt:innen und selbstbewusste Persönlichkeiten eher geneigt, ihrem eigenen Urteil mehr zu vertrauen als dem der KI.

Durch die Gewöhnung an verlässliche Ergebnisse der KI könne eventuell irgendwann eine gewisse unkritische Bequemlichkeit einsetzen, die ein leichtfertiges Vertrauen in KI begünstigt.

Demgegenüber werden als weitere Ursachen für (übermäßiges) Misstrauen fehlende Selbstreflexion und wenig Erfahrung im Umgang mit KI genannt.

Einer KI werde im Vergleich zu einem Menschen häufig auch deshalb weniger zugetraut, weil sie Kontext, Individualität und Komplexität des einzelnen Falls nicht abbilden kann. KI wird darauf trainiert, kranke Patient:innen mit gesunden Patient:innen zu vergleichen; dabei werden individuelle Merkmale wie Alter, Herkunft,

Geschlecht, Vorgeschichte, Vorerkrankungen, Medikamente etc. derzeit vom Algorithmus nicht einberechnet. *„Und ich kann mir ehrlicherweise nicht vorstellen, dass wir oder unsere Kinder und Urenkel das noch erleben werden, dass eine Intelligenz oder ein Code, ist es ja letztendlich, es schaffen kann, die individuelle Komplexität von Abermillionen von Patienten zu berücksichtigen" (I8).*

Im richtigen Maße zu vertrauen (also der KI weder leichtfertig zu vertrauen noch übermäßig zu misstrauen) gelingt nach Meinung der Befragten hauptsächlich durch Übung. Die KI selbst auszuprobieren, irgendwann alltäglich damit umzugehen und dabei immer besser ihre Schwächen und Stärken kennenzulernen, wird als sicherster Weg zu einem gerechtfertigten Vertrauen in die KI beschrieben.

Ist es auch maßgeblich für das Vertrauen, ob die Ergebnisse der KI nachvollziehbar sind? Das hängt offenbar sowohl von der Art der Aufgabe ab als auch davon, ob die KI-Ergebnisse von der eigenen Einschätzung abweichen. P9 etwa äußert wenig Interesse an einer genaueren Aufschlüsselung: *„Um mal bei dem Beispiel der Mustererkennung zu bleiben, da muss ich nicht im Detail wissen, warum er jetzt glaubt, dass da die Grenze ist, wenn ich dafür weiß, dass es ziemlich genau dem entspricht, was ich auch sagen würde, und wenn das dafür reproduzierbar ist. […] Und da muss ich eben nicht genau wissen, aufgrund welcher Schwellenwerte oder Grenzen er das jetzt entscheidet, wenn ich halbwegs mit dem Ergebnis dann am Ende einverstanden bin" (I9).*

Gegenteilig äußert sich P7: *„Ja, also das Wichtige bei eigentlich grundsätzlich allen KI-Empfehlungen ist für mich, dass die Empfehlung nachvollziehbar sein muss. […] Und erst, ja, dann würde ich einer KI-Diagnose vertrauen. Es muss also sozusagen irgendwie begründbar sein. Die muss mir zeigen können, wie ist sie darauf gekommen" (I7).*

Aus Sicht von P4 ist Nachvollziehbarkeit bei KI an den gleichen Stellen erforderlich wie bei Menschen: *„Es muss so lange, das AI, erklärbar sein, wie auch die menschliche Diagnosefindung erklärbar ist. Dann muss es auch erklärbar sein. Idealerweise muss es das reproduzieren, was der Mensch an Entscheidungen trifft. Sobald es Punkte gibt, in denen der Mensch auch nicht argumentieren kann, dann kann auch die Maschine nicht argumentieren. Und wenn es Parameter oder Diagnosen gibt, die durch eine Maschine definiert sind, dann muss sie nicht mehr erklärbar sein, wenn ich gezeigt habe, dass sie funktioniert. So. Und da geht es in beide Richtungen. Wir müssen vielleicht erst mal überprüfen, ob unsere Entscheidungsgrundlagen teilweise überhaupt stimmen. Also, das ist ja auch nicht ganz so klar" (I4).*

Ärztliche Kompetenzen bei der Nutzung von ML innerhalb der bildgebenden Diagnostik

Befragt nach der Art der Kompetenzen, die Ärzt:innen für den Umgang mit KI beherrschen sollten, werden drei verschiedene Arten von Kompetenzen genannt:

Zunächst müssen die medizinischen Kompetenzen so solide sein, dass man nicht darauf angewiesen ist, das Ergebnis der KI einfach zu glauben, sondern die Leistung der KI überprüfen, d. h. kritisch mit dem eigenen Wissens- und Erfahrungs-Horizont abgleichen kann. Zwei Befragte plädieren dafür, deshalb nicht zu früh in der Ausbildung KI zu verwenden, damit sich die eigenen Kompetenzen erst festigen

können: *„[...] das ist vielleicht so ein bisschen die Parallele wie so ein Taschenrechner in der Schule. Wenn man ab Tag eins mit dem Taschenrechner rechnet, wird man nie kopfrechnen können. Umgekehrt, wenn man erst ein paar Jahre kopfrechnet und sich dann eine Erleichterung mit dem Taschenrechner holt, [...] kann man trotzdem noch kopfrechnen" (I8).*

An typischen KI-Kompetenzen werden verschiedene Niveaustufen genannt: vom richtigen Anwenden des Programms über das richtige Interpretieren der Ergebnisse bis hin zum Verständnis der Funktionsweise von Algorithmen. Ein:e einzige:r Befragte:r sieht noch eine vierte Niveaustufe als wünschenswert an: Idealerweise sollten sogar Programmierkenntnisse vermittelt werden.

Die Mehrheit der Befragten ist jedoch dafür, dass Mediziner:innen Kompetenzen auf Niveau 3 haben sollten. Zu einem solchen informationstechnologischen Grundverständnis gehört das Wissen darum, wie eine KI aufgebaut ist und wie sie trainiert wird. Erst damit wird ein kritisches Hinterfragen von KI-Ergebnissen möglich. Viele informationstechnologische Lai:innen sähen KI nämlich als eine Art *„höhere Macht" (I8)* an; es sei nicht hinreichend bewusst, dass KI auch Fehler machen kann.

Ein:e Befragte:r erachtet es in der Zusammenarbeit mit KI außerdem als wichtig, den eigenen Umgang mit Feedback bzw. Zweitmeinungen sowie die eigene Fähigkeit zur Selbstreflexion weiterzuentwickeln. Patholog:innen seien im Rahmen ihrer Ausbildung und dann auch ihrer beruflichen Tätigkeit selten in der Situation, eigene diagnostische Einschätzungen mit denjenigen von Kolleg:innen abzugleichen. Dazu komme, dass selten eindeutige Ergebnisse zu erwarten sind, *„[w]eil man einfach auch in der Pathologie keine wirkliche Wahrheit hat. Also gerade bei diesen quantitativen Sachen kriegt man keine Wahrheit und dann, was man sagt, ist halt die Wahrheit. Und sonst, man kann sich nie reflektieren und kriegt auch nie Feedback, ob da eine Wahrheit [...] zur richtigen Therapie und so weiter geführt hat" (I2).* Die fehlende Erfahrung mit Feedback und Zweitmeinungen könne den Umgang mit KI erschweren, weil die unzureichend eingeübte und nun plötzlich erlebte Diskrepanz von Meinungen zu einem Misstrauen in die KI führen kann.

Gestaltung der ML-basierten Assistenzsysteme und des Diagnoseprozesses

Bei aller Offenheit der Befragten, KI in die Arbeitsabläufe zu integrieren, besteht Diskussionsbedarf vor allem bzgl. der Verlässlichkeit der Ergebnisse und der Frage, an welchen Stellen, also für welche Aufgaben KI eigentlich eingesetzt werden soll. Da die Antworten zu diesem Punkt häufig recht konkrete Einblicke in den Arbeitsalltag der Ärzt:innen gewähren, fällt die zusammenfassende Darstellung entsprechend ausführlicher aus als bei den etwas abstrakteren vorangehenden Themenblöcken.

Die Arbeitsteilung zwischen Mensch und Maschine wird idealerweise gemäß den jeweiligen Stärken und Schwächen gestaltet. Im besten Fall ergänzen sich die Stärken von Mensch und Maschine und es kommt zu Effizienz- und Qualitätssteigerung der Arbeit.

Was sind nun aber die Stärken von KI? Die interviewten Ärzt:innen nannten insbesondere folgende Eigenschaften:

KI zählt und misst schneller und exakter als der Mensch. Wo ein Mensch schätzen muss, hat KI in Sekundenbruchteilen das genaue Ergebnis. Gerade wenn sehr viel zu

zählen ist und es dabei auf geringe Abweichungen ankommt, kann das von Vorteil sein: *„Multiple Sklerose, da kriegt man, wenn man ein MRT vom Kopf bekommt und man hat Multiple Sklerose, sind das so Herde. Das sind so weiße Punkte. Und bei fortgeschrittenen Erkrankungen, da sehen wir jetzt 10.000 weiße Punkte. Die kann kein Mensch mehr zählen. Die muss man auch nicht zählen. Aber wenn man jetzt eine Verlaufskontrolle hat und er hat irgendwie jetzt fünf Punkte mehr, aber/ne, das ist so, man sieht den Wald vor lauter Bäumen nicht. […] Und die KI kann dann natürlich Bilder im Hintergrund übereinanderlegen, und die Punkte markieren, die vorher nicht da waren. Und dann fällt das ins Auge" (I8).*

Ein weiteres Aufgabenfeld, in dem KI brilliert, ist der Bild-Vergleich. Anders als der Mensch arbeitet KI zudem mit gleichbleibender Qualität, was Reproduzierbarkeit und Objektivität verspricht: *„[D]as Produkt, was wir jetzt verwenden, […] macht Vergleiche mit einem Normen-Kollektiv. Dieses Normen-Kollektiv kann ein Mensch natürlich auch irgendwie mit viel Erfahrung auch haben, aber […] trotzdem bleibt dann noch so eine Subjektivität drinnen, ja? Also dann wiegen wahrscheinlich die letzten 20, die man gesehen hat, mehr als die 6000, die man insgesamt gesehen hat. Oder es ist morgens anders, als es abends ist, weil abends ist man müde oder am Ende der Schicht oder was auch immer. Würde man wahrscheinlich bestimmte Dinge anders bewerten oder vielleicht auch weniger Zeit investieren oder nicht mehr so gründlich sein. Und diese Varianz fällt weg." (I9).*

Stupide und repetitive Tätigkeiten fallen KI leichter als Menschen, die irgendwann mit Motivations- und Konzentrationsschwierigkeiten zu kämpfen haben. *„[Da]s sind so Aufgaben, wo man sagt: ‚Das ist so redundant, das ist so' – man langweilt sich schon, sozusagen, ja, wenn man nur daran denkt. Und das ist nicht die beste Startvoraussetzung, dann auch diese Aufgabe gut zu erfüllen" (I9).*

Wo KI das stumpfe Zählen übernimmt, werden Ressourcen frei, was *„dem Radiologen dann die Möglichkeit gibt, sich mit dieser freigewordenen Zeit mit anderen Dingen zu beschäftigen, die ihn dann letzten Endes auch besser machen, wieder, ja? Also Kommunikation mit Patient, Kommunikation mit anderen Klinikern, Vorbereitung von anderen Untersuchungen, logistische Fragen. Also Organisation von Untersuchungen, wer wann wohin? Das sind alles sozusagen Sachen, mit denen sich der Radiologe beschäftigen kann, wenn er nicht dabei ist, […] kleine Flecken zu zählen und zu vergleichen miteinander" (I9).*

Wenn der Mensch von Tätigkeiten entlastet ist, die seinem Leistungsprofil weniger entsprechen, kann er sich auf seine typisch menschlichen Stärken fokussieren. Neben dem Blick für das „große Ganze" und das „Drumherum" wird vor allem Kommunikation als eine typisch menschliche Stärke genannt. Kommunikation zwischen Menschen gehe über den reinen Informationsgehalt hinaus, lasse Raum für Zwischentöne, eröffne eine Beziehungsdimension. Zudem könne (gelegentliche) Interaktion auch die Arbeitszufriedenheit steigern. Diese menschliche Komponente wird als Trumpf erlebt: *„Das sind halt alles so Sachen, die/wo der Mensch wahrscheinlich, ja, auch über längere Zeit schwer zu schlagen sein wird. Und wo das Menschsein selbst wahrscheinlich auch da mit reinspielt, in den Arbeitsprozess. Ja? Also das geht ja nicht ganz weg und sollte vielleicht auch nicht ganz weg sein, dass man dann doch beim Arbeiten immer noch tatsächlich Mensch ist und nicht versucht,*

eine Maschine zu schlagen im Maschinesein. Das ist [...] auch nicht erstrebenswert, ja?" (I9).

Als weitere Stärke des Menschen, die konkret die diagnostische Tätigkeit betrifft, wird seine Fähigkeit gesehen, den Kontext eines Falls zu erfassen und flexibel auf Unvorhergesehenes zu reagieren: Während KI z. B. mit einem verwackelten Bild wenig anfangen kann, weil es aus dem bekannten Schema herausfällt, erkennt der Mensch den Störfaktor und ist u. U. sogar imstande, das Bild adäquat auszuwerten. Wo KI einen auffälligen Punkt als Krebs klassifiziert, erkennt ein Mensch, dass es sich nicht um Krebs, sondern um eine Narbe aus einer vorherigen Operation oder um verbliebenes Operationsmaterial handelt.

Während KI im Beantworten quantitativer Fragen besser sei, schlage der Mensch KI im Beurteilen und Diagnosenstellen. Auf dieser Ebene möchte sich (noch) niemand auf KI verlassen. Das sollten weiterhin Menschen tun, weil sie der Komplexität eines Falls besser Rechnung trügen als KI.

Wie kann KI nun so eingesetzt werden, dass Mensch und Maschine ihre jeweiligen Stärken optimal entfalten können?

Auf die Frage, welche Einsatzbereiche im diagnostischen Geschehen sich für KI besonders eignen, werden hauptsächlich Vorarbeiten wie das Vorsortieren von Daten genannt. KI soll sowohl das zeitaufwändige Suchen nach den relevanten Arealen innerhalb eines Bildes ersparen (*„Die KI soll mir markieren, was sie auffällig findet. Zum Beispiel den Punkt soll die mir mit einem Pfeil markieren. [...] Dann muss ich nicht mehr vier Minuten lang den Punkt suchen, indem ich durch das Bild hoch- und runterscrolle" (I8)*) als auch das zu sichtende Datenmaterial insgesamt reduzieren: Aus der Fülle von Bildern kann KI schon mal die offensichtlich unauffälligen Fälle aussortieren und den Blick der Ärzt:innen damit direkt auf diejenigen Fälle lenken, die sie sich nochmal genauer anschauen müssen. KI eine Vorauswahl treffen zu lassen, erscheint auch als einzige Möglichkeit, der stetig steigenden Anzahl von Fällen noch Herr zu werden: *„[J]eder wird inzwischen hier, kriegt zum Beispiel eine Röntgenthorax-Untersuchung, der hier durch die Tür kommt, hat man manchmal den Eindruck. Ja, einfach, weil es immer mehr wird, immer mehr Ärzte wollen sich absichern, auch in der Notaufnahme, klar. Das sind/also teilweise pro Person sind das am Tag 200 Röntgenthoraxes. Und davon haben 180 einen unauffälligen Befund. Ja, da wäre es natürlich schön, wenn die KI zumindest 150 Unauffällige schon mal rausselektiert. Und ich kann mich mit den Übriggebliebenen vernünftig beschäftigen" (I8).*

Wenn die Aufgabe von KI in der Massenreduktion verortet wird, sei es wesentlich, dass sie nichts übersieht. KI müsse also übervorsichtig sein, sodass sie *„lieber einen rausfischt, der gar nicht erkrankt ist" (I8)*. Derzeit habe KI wohl genau diese Tendenz, zu vorsichtig zu sein und *„die Leute kränker [zu] machen, als sie [sind]" (I8)*. Im Zusammenspiel von maschineller Übervorsichtigkeit und menschlicher Nachkontrolle könnten Effizienz und Qualität der Arbeit gesteigert werden.

Eine weitere Chance wird darin gesehen, dass KI aus dem Datenmaterial von Patientenakten relevante Informationen bzw. Vergleichsbilder heraussucht. KI könnte dafür sowohl in der Vorgeschichte eines Patienten suchen als auch die Verknüpfung zu ähnlichen Krankheitsfällen herstellen und Informationen zu Diagnose, Therapie und Verlauf dieser Vergleichsfälle bereitstellen: *„Wenn da eine KI dahinterstehen*

würde, die die Patientenakten einfach drin hat und sieht: Okay, wir hatten vor fünf Jahren einen Patienten, der hatte den ziemlich gleichen Verlauf wie der aktuelle Patient. Und der hatte am Ende die Diagnose beziehungsweise der hat von der Therapie profitiert. Das könnte man vielleicht bei dem Patienten auch probieren" (I5).

Gestaltungswünsche für die KI-Entwicklung beziehen sich hauptsächlich auf das Thema „Nachvollziehbarkeit". Idealerweise soll die KI nicht nur ein Ergebnis ausgeben, sondern erklären, wie sie darauf gekommen ist. Dazu könnte sie offenlegen, welche Daten entscheidungsrelevant waren, indem sie etwa den Fundort (z. B. Punkte auf einem Bild) genau angibt oder Vergleichsmaterial zur Verfügung stellt, um eine Klassifikation zu begründen: *„Die KI schlägt mir vor, okay, ich sehe hier Lungenkrebs, dort und dort in der Lunge. Und dann soll mir die KI halt genau zeigen, wo sie den Lungenkrebs sieht und dann auch zeigen, wie sie darauf gekommen ist, dass das jetzt Krebs ist und nicht zum Beispiel eine gutartige Läsion. Indem sie zum Beispiel jetzt zeigt, okay, wie sieht diese Form dieser Läsion aus? Ist die jetzt glatt berandet, ist die/wächst die infiltrativ? Oder mir dann zum Beispiel auch Vergleichsbilder zeigt und sagt, so sieht Krebs aus und so sieht was Gesundes aus" (I7).* Eine noch höhere Stufe der Erklärbarkeit wäre durch Interaktion mit der KI erreicht: eine Art Gespräch, in dem man Rückfragen stellen bzw. die KI auf etwaige Fehler hinweisen und so gemeinsam das zunächst errechnete Ergebnis erhärten oder korrigieren könnte.

Allerdings sollten die Zusatzinformationen nicht so ausufernd sein, dass die Ärzt:innen in ihrem Arbeitsablauf gehemmt werden. *„Also so ein bisschen [...] ist gut. Jetzt aber noch irgendwie fünf Seiten mehr von wegen, das könnte und das könnte nicht, und vielleicht, und wahrscheinlich, glaube ich, führt das ganze wieder ad absurdum, weil Sie haben dann keine Zeit, um sich das durchzulesen" (I1).*

Ob Nachvollziehbarkeit gewünscht ist oder nicht, hängt letztlich wesentlich davon ab, welche Aufgabe der KI gestellt wird. Bei überschaubaren, konkret definierten Fragestellungen wird es selten als relevant empfunden, wie genau das Ergebnis zustande kam. Je komplexer die Fragestellung ist und je mehr Teilentscheidungen im Endergebnis enthalten sind, desto wichtiger ist es den Befragten, die maschinellen Abläufe erklärbar zu machen.

Wahrscheinlichkeitsangaben werden als weiterer wichtiger Gestaltungspunkt angeführt. Wenn KI verschiedene Vorschläge abgestuft nach Wahrscheinlichkeit mit Prozentangaben präsentiert, könne eine solche Rangordnung die ärztliche Entscheidungsfindung erleichtern. Ärzt:innen könnten so eine Liste an Diagnose-Optionen abarbeiten:

„Und eine Möglichkeit wäre, halt dann zu sagen, okay, liebe KI, du schlägst mir hier vor, das ist ein Lungenkarzinom, also Lungenkrebs. Das glaube ich jetzt nicht so. Was wäre denn das Zweitwahrscheinlichste?" (I7). Zugleich erhöhe es das Vertrauen in die KI, wenn sie Unsicherheiten „zugibt" bzw. eine Abstufung von Sicherheiten präsentiert.

Diskussion

Vertrauen

Ob der Begriff „Vertrauen“ in Bezug auf ML-basierte Assistenzsysteme überhaupt bemüht werden darf, darüber gibt es geteilte Meinungen.[4] Eigentlich ist er reserviert für zwischenmenschliche Beziehungen, da Vertrauen neben der rational-kognitiven Dimension, die eher eine Prognose erwartbaren Verhaltens anderer ist, eben auch eine affektive Seite hat: Vertrauen unterstellt anderen ein gewisses Wohlwollen uns gegenüber, hat nicht nur den reinen „Output“ im Blick, sondern auch die Motivation dahinter (vgl. Eschenbach 2021, S. 1611; Fritz 2021, S. 366f.). Von ML-Systemen kann jedoch kein Wohlwollen erwartet werden.

Die Ausweitung des Begriffs auf KI mag also irreführend und darüber hinaus auch riskant sein, da er die Grenzen zwischen moralischen Akteuren und Artefakten verwischt, was die Analyse komplexer Verantwortungsgefüge weiter erschwert (vgl. Fritz 2023, S. 163f.).[5]

Als Ausweg aus dieser Vokabelfrage bietet sich statt „vertrauen“ das Verb „sich verlassen“ an (vgl. Starke et al. 2022; Ryan 2020). In Anlehnung an die Formulierungen in den Interviews wird in der weiteren Diskussion jedoch nicht streng zwischen den Begriffen unterschieden.

In den Interviews ist auffällig, dass als Grundlage für Vertrauen in ML-Systeme die Nachvollziehbarkeit/Kontrollierbarkeit der ML-Ergebnisse genannt wird: Vertrauen ist nur in dem Maß gegeben, wie man das Ergebnis selbst noch einmal überprüfen kann.

Zunächst widersprechen sich jedoch Kontrolle und Vertrauen. Wie in dem Sprichwort „Vertrauen ist gut, Kontrolle ist besser“ anklingt, ist Kontrolle eigentlich ein Zeichen von Misstrauen bzw. ein Ersatz für Vertrauen: Wer kontrollieren kann, muss eben gar nicht mehr vertrauen (vgl. Hartmann 2003, S. 405).

Andererseits spielt die Wissensdimension beim Vertrauen durchaus eine Rolle: Klassischerweise wird nämlich unterschieden zwischen „blindem“ und „begründetem“ Vertrauen. Letzteres basiert auf einer gewissen Kenntnis der Person, der ich mein Vertrauen entgegenbringe, einer Kenntnis, die z. B. auf der Erfahrung beruht, dass diese Person kompetent, zuverlässig, ehrlich und mir wohlgesonnen ist (vgl. Hartmann 2003, S. 395f.; Starke et al. 2022, S. 155; Fritz 2023, S. 166).

Letztlich kommt es also auf das richtige Mischungsverhältnis an: „Vertrauen [...] ist als Hypothese ein mittlerer Zustand zwischen Wissen und Nichtwissen um den

[4] S. z. B. Ryan (2020), Eschenbach (2021), Starke et al. (2022). Ryan verwehrt sich heftig gegen das Konzept des Vertrauens in Bezug auf KI und möchte die Rede von der „vertrauenswürdigen KI“ lieber durch „verlässliche KI“ ersetzt sehen. Eschenbach versucht das Problem zu umgehen, indem er von Vertrauen in das „soziotechnische System“ spricht, dem sowohl KI als auch menschliche Handelnde angehören. Einen ähnlichen Gedanken verfolgen Starke et al. mit Rückgriff auf Latour. Ihrer Meinung nach könne man KI aber durchaus vertrauen, schließlich vertraue man ja auch in Brücken und Gesundheitssysteme.

[5] Insbesondere auch die Ausweitung des *agency*-Begriffs auf KI ist mit den genannten Risiken behaftet. *Agency*/Handlungsfähigkeit impliziert Intentionen, also das Handeln aus Gründen und mithin die Fähigkeit zur Verantwortungsübernahme (vgl. Fritz und Brandt 2019).

Menschen. Der völlig Wissende braucht nicht zu vertrauen, der völlig Nichtwissende kann vernünftigerweise nicht einmal vertrauen“ (Simmel 1968, S. 263).

Anfängliche Kontrolle könnte so auch die Grundlage für späteres Vertrauen schaffen: Indem Ärzt:innen in einer Testphase die ML-Ergebnisse nachprüfen, machen sie sich *vertraut* mit der Arbeitsweise der ML-basierten Assistenzsysteme und können ihnen im Anschluss ggf. *begründet vertrauen*. Keinesfalls kann es nämlich wünschenswert sein, Vertrauen einfach pauschal zu erhöhen. Vertrauen muss vielmehr klug platziert werden, indem es sich an der Vertrauenswürdigkeit anderer orientiert (vgl. O’Neill 2020). Folglich muss es in einem ersten Schritt darum gehen, die Vertrauenswürdigkeit bzw. Verlässlichkeit der ML-Systeme adäquat einzuschätzen (bzw. auf Seiten der Entwickler:innen die Verlässlichkeit von ML zu erhöhen).

Wichtig scheint den Ärzt:innen jedoch zu sein, dass es immer die *Möglichkeit* zur Überprüfung der ML-Ergebnisse gibt, auch wenn sie nicht in jedem Fall davon Gebrauch machen würden.

Im Forschungsfeld *Explainable AI (XAI)* wird daran gearbeitet, opake ML-Systeme für deren Nutzer:innen überprüfbar, also nachvollziehbar(er) zu machen.

Grundsätzlich wird dabei zwischen zwei verschiedenen Arten der Nachvollziehbarkeit unterschieden: 1. innere Prozesse der Blackbox offenlegen („explaining what“) und 2. Entscheidungen verständlich machen, ohne dabei notwendigerweise die Prozesse zu rekonstruieren, die zu dieser Entscheidung geführt haben („explaining why“) (vgl. Eschenbach 2021, S. 1615).[6] Menschliches Handeln erklären wir üblicherweise auf die zweite Weise: Wir führen die Gründe an, die eine Person zu ihrem Handeln motiviert haben.

Um Ärzt:innen Ergebnisse eines ML-Systems verständlich zu machen, bietet sich ebenfalls die zweite Art der Erklärung an („explaining why“). Denn ebenso wenig wie man bei einem Menschen Einblick in seine neuronalen Prozesse einfordert, um sein Handeln verstehen zu können, hilft Ärzt:innen ein solcher Blick in die Blackbox der ML-Algorithmen weiter, wenn sie eine medizinische Einschätzung des ML-basierten Assistenzsystems nachvollziehen wollen (vgl. Eschenbach 2021, S. 1615f.; Durán und Jongsma 2021, S. 330; Zerilli et al. 2019, S. 669f.).

Die Interviews bestätigen diese Einschätzung: Die Befragten wünschen sich in einem konkreten Zweifelsfall keine informationstechnologischen Einblicke in die Funktionsweise des ML-Systems, sondern sie wollen unterstützt werden in ihrer eigenen medizinischen (!) Meinungsbildung. Damit besteht zumindest aus ärztlicher Sicht kein Interesse am Innenleben der Blackbox. Vielmehr interessiert, ob man die Blackbox „umgehen“ könnte, im Bedarfsfall also über menschliche Denkwege selbst zum gleichen Ergebnis käme wie das ML-System – ähnlich wie man sicherheitshalber noch eine Straßenkarte im Handschuhfach hat für den Fall, dass das Navigationssystem versagt.

Die nächstliegende und sinnvollste Form, wie XAI diesem Anliegen Rechnung tragen kann, wäre, dass sie die entscheidungsrelevanten Faktoren gewichtet präsentiert (vgl. Zerilli et al. 2019, S. 676).

[6] Zerilli et al. nutzen in diesem Zusammenhang auch Daniel Dennetts hilfreiche Unterscheidung in „design stance“ und „intentional stance“ (vgl. Zerilli et al. 2019, S. 668f.).

Dabei wird in den Interviews deutlich, dass nur eine gewisse Dosis an Erklärungen gewünscht wird: genug, damit man das Ergebnis einordnen kann, aber keine seitenlangen Ausführungen. Ein Übermaß an Transparenz würde – wie im Fall von AGBs – wohl eher bewirken, dass sich niemand mehr damit befasst.

Natürlich gibt es im wissenschaftlichen Diskurs auch Vorbehalte gegen XAI, z. B. dass das Problem nur auf eine höhere Ebene verschoben wird und völlig offenbleibt, inwiefern nun die Erklärung (ebenfalls auf opaken Algorithmen basierend) ihrerseits verlässlich ist (vgl. Durán und Jongsma 2021, S. 330). An dieser Stelle käme jedoch die medizinische Kompetenz des:der jeweiligen Ärzt:in als mögliches Korrektiv ins Spiel, womit aus rein medizinischer Sicht bewertet werden könnte, wie plausibel die Erklärung der XAI ist.

Verantwortung

„Verantwortung" scheint in den Interviews als rein rechtliches Konstrukt synonym zu „Haftung" verstanden zu werden. Diese Engführung könnte lediglich auf der Sprachebene angesiedelt sein, insofern, dass „Verantwortung" zwar nicht in der Bedeutung ‚moralische Verantwortung' verstanden wird, das Konzept von moralischer Verantwortung aber trotzdem vertraut ist und nur in anderen Vokabeln ausgedrückt wird. Eine andere Möglichkeit wäre, dass auch das Konzept moralischer Verantwortung im Diagnosegeschehen wenig Alltagsrelevanz hat und dementsprechend wenig reflektiert wird.

Tatsächlich geht der Begriff der Verantwortung ursprünglich auf den Kontext von Gerichtsprozessen zurück, wo er im Sinne der Schadenshaftung benutzt wird: „sich verantworten" im Sinne von „jemandem Rede und Antwort stehen" (vgl. Holderegger 1992, S. 203; Körtner 2001, S. 102). Moralische Verantwortung geht jedoch über rechtliche Verantwortung hinaus; sie ist nicht von der jeweiligen kontingenten Rechtsordnung einer Gesellschaft abhängig. Um verantwortlich handeln zu können, bedarf es zum einen eines Subjekts, das „antworten"/sich rechtfertigen kann, das Gründe für sein Handeln hat und diese nennen kann. Dieses Subjekt muss also in der Lage sein „frei, willentlich und bewusst" entscheiden und handeln zu können (Buddeberg 2011, S. 18). Zum anderen muss es in einer bestimmten Situation sowohl über das nötige Wissen für eine Handlungsentscheidung verfügen („epistemische Bedingung") als auch Handlungsmacht/Kontrolle über das eigene Handeln haben („Kontroll-Bedingung") (Rudy-Hiller 2022).

Wenn nun Ärzt:innen Diagnosen mithilfe eines opaken ML-basierten Assistenzsystems erstellen, steht die epistemische Bedingung von Verantwortung auf dem Spiel. Deshalb äußern die Interviewten immer wieder das Bedürfnis, Ergebnisse eines ML-Systems überprüfen zu können. In der Kontrolle des ML-gestützten Ergebnisses sehen sie die epistemische Bedingung von Verantwortung erfüllt.

In diesem Zusammenhang wäre allerdings auch zu überlegen, ob die epistemische Bedingung um eine Ebene vorverlagert werden kann: auf die generelle Überprüfung eines ML-basierten Assistenzsystems als Expertin für bestimmte Aufgaben, um ein *begründetes*, also gerechtfertigtes Vertrauen zu etablieren. Dieses Vertrauen gründet dann z. B. auf der Erfahrung, dass das Assistenzsystem in ähnlichen Fällen

meist richtige Ergebnisse liefert, also *zuverlässig kompetent* arbeitet.[7] Diese Sichtweise modifiziert möglicherweise das Verständnis von Verantwortung bzw. der Art des Wissens, welches für verantwortliches Handeln notwendig ist: Um der epistemischen Anforderung an Verantwortung Genüge zu tun, müssten Ärzt:innen eben nicht mehr ein bestimmtes Einzelergebnis der KI überprüfen (können). Es wäre ausreichend, dass sie *vorher* die Expertise von KI für diese Aufgabe überprüft haben und fortlaufend reevaluieren. Ein solchermaßen begründetes Vertrauen könnte die epistemische Bedingung für Verantwortung in ähnlicher Weise erfüllen wie die Ergebniskontrolle in jedem Einzelfall.

Als Verantwortliche sehen Ärzt:innen in erster Linie sich selbst, wenn sie auch teilweise Einschränkungen ihres Verantwortungsbereichs vornehmen: Kolleg:innen vor ihnen in der Arbeitskette sehen sie z. B. für die Qualität der Gewebeproben oder der Röntgenbilder in der Verantwortung. Beim Vorschlag einer begrenzten Produkthaftung seitens der Entwickler:innen klingt ebenfalls der Wunsch nach einer klaren Begrenzung des eigenen Verantwortungsbereichs durch.

Damit Verantwortung handlungswirksam werden kann und nicht in Überforderung mündet, ist es wichtig, den eigenen Verantwortungsbereich nicht endlos auszudehnen. Von daher sind (rechtliche) Überlegungen zu Verantwortungsaufteilung und damit -begrenzung, die auch die Entwicklungsfirmen in den Blick nehmen, essentiell.

Spezifische Herausforderungen durch ML?

Es lohnt sich, genauer zu prüfen, inwiefern die diskutierten Herausforderungen spezifisch für den Einsatz von ML sind. Dass Verantwortung und Vertrauen in einem Spannungsverhältnis zueinander stehen, ist schließlich kein Alleinstellungsmerkmal von Bereichen, in denen mit ML gearbeitet wird. Die Debatte um das rechte Maß an Verantwortung und Vertrauen im Umgang mit ML kann von einer vergleichenden Perspektive profitieren: Wo sich strukturelle Parallelen zeigen, lässt sich ein realistisches Problembewusstsein schärfen und bereits bewährte Lösungsansätze in anderen Bereichen können als Anregung dienen.

Inwieweit sind also die Herausforderungen rund um Verantwortung, Vertrauen und Nachvollziehbarkeit wirklich spezifisch auf ML-Systeme zurückzuführen?

Dass man für etwas Verantwortung übernimmt, ohne im strengen Sinn die epistemische Bedingung zu erfüllen, kennen wir aus mindestens zwei anderen Bereichen: Sowohl beim Autoritätsargument ist dies der Fall (man vertraut auf eine Autorität, die in einem bestimmten Bereich mehr Expertise besitzt als man selbst) als auch bei der Delegation (man vertraut auf Mitarbeitende, ohne ihre Arbeitsergebnisse in allen Fällen nochmals zu kontrollieren).

Hier gilt es eben nicht pauschal als verantwortungslos, sich auf das Urteil einer Autorität zu verlassen oder Aufgaben an andere zu delegieren.

[7] Durán und Jongsma nennen verschiedene weitere Kriterien für eine solche *Computational Reliability*: „verification and validation methods, robustness analysis, a history of (un)successful implementations, and expert knowledge“ (2021, S. 332).

Wie in einem Unternehmen mit guter Personalführung Aufgaben nach Stärken und Schwächen (und Entwicklungspotenzial) der Mitarbeitenden zugeteilt werden, könnte ML als eine Mitarbeiterin (vgl. *„studentische Hilfskraft“* (I8)) betrachtet werden. Die Mitarbeiterin ML hat gewisse Inselbegabungen: Sie ist ein mathematisches Genie, dafür tritt sie aber z. B. in Kommunikationssituationen sehr ungelenk auf. Ihr würde man nur ganz bestimmte Aufgaben anvertrauen, die sie erfahrungsgemäß sehr akkurat und zuverlässig löst. Verantwortungslücken ergäben sich nach diesem Verständnis genauso selten oder häufig wie bei Delegation an Menschen.

Was die Opazität innerer Prozesse betrifft, lohnt ebenfalls der Vergleich zwischen ML-System und Mensch. Oft werden vergleichsweise hohe Anforderungen an die Durchschaubarkeit von ML gestellt, die auch Menschen kaum erfüllen können. Das menschliche Gehirn ist gewissermaßen auch eine Blackbox, deren kognitive Prozesse verborgen sind. Menschliches Denken ist in vielfältiger Hinsicht mit kognitiven Verzerrungen behaftet (P9 erwähnt z. B., dass Menschen dazu neigen, frischere Eindrücke stärker zu gewichten als ältere). Besonders interessant in diesem Zusammenhang ist die menschliche Tendenz zu nachträglichen Rationalisierungen (vgl. Zerilli et al. 2019, S. 666). Zudem gibt es das Phänomen der ärztlichen Intuition, bei dem die Opazität der Entscheidungsfindung offenbar als weniger problematisch empfunden wird.

Vor diesem Hintergrund mutet es merkwürdig an, dass in der Diskussion um ML-basierte Assistenzsysteme ein so starker Fokus auf Nachvollziehbarkeit gelegt wird. Nicht *wer* etw. entscheidet, sondern *was* entschieden wird und wieviel von dieser Entscheidung abhängt, sollte ausschlaggebend dafür sein, ob Nachvollziehbarkeit gewährleistet sein muss. Je größer die Tragweite einer Entscheidung, desto wichtiger wird Nachvollziehbarkeit (vgl. Zerilli et al. 2019, S. 679). Statt pauschal darauf zu pochen, dass ML-Systeme immer nachvollziehbar sein müssen, lohnt es also, nach der Wichtigkeit der von den ML-basierten Assistenzsystemen zu treffenden Entscheidungen zu differenzieren.

P4 teilt offenbar die Einschätzung, dass es auf die Art der Entscheidung ankommt, weniger darauf, ob sie von Menschen oder ML-Systemen getroffen wird: *„Es muss so lange, das AI, erklärbar sein, wie auch die menschliche Diagnosefindung erklärbar ist“* (I4).

In I2 wird besonders deutlich, dass KI nicht in ihrer Eigenschaft als *künstliche* Intelligenz zur Herausforderung wird, sondern eher im Sinne einer *weiteren* Intelligenz, indem sie Ärzt:innen mit einer anderen Meinung konfrontiert. Weniger die Mensch-Maschine-Interaktion scheint hier das Problem zu sein, sondern eher generell Interaktion bzw. der adäquate Umgang mit Feedback. Auf den ersten Blick stehen diese Angaben aus I2 im Widerspruch zu den Aussagen der anderen Befragten, dass sie bei Unsicherheiten Unterstützung im Kollegium suchen bzw. suchen könnten. Da P2 systemische Ursachen für den von ihm:r erkannten Missstand nennt, liegt es nicht nahe, dass die unterschiedlichen Aussagen allein dem unterschiedlichen individuellen Erleben der Befragten geschuldet sind. Stattdessen könnte sich der vermeintliche Widerspruch dadurch auflösen, dass in Zweifelsfällen durchaus eine hohe Offenheit gegenüber kollegialem Feedback besteht, ein systematischer Routine-Abgleich mit einer Zweitmeinung dagegen nicht eingeübt ist und daher größere Widerstände hervorruft.

Anders als Menschen können ML-Systeme keine Verantwortung übernehmen. Im Vergleich mit anderen technischen Geräten ist das jedoch wiederum kein Alleinstellungsmerkmal: Auch ein Röntgenapparat kann keine Verantwortung übernehmen. Hier scheint diese Problematik jedoch weniger virulent, was damit zu tun haben mag, dass Schäden an „normalen" Geräten eindeutiger erkennbar sind als „schadhafte" Algorithmen, die unerkannt falsche Ergebnisse produzieren. In Bezug auf die Opazität schenken sich ML-basierte Assistenzsysteme und Röntgenapparat jedoch auch wenig: Wenige Ärzt:innen werden das Innenleben eines Röntgenapparats wirklich verstehen. Dennoch scheint es in diesem Fall kein Problem zu sein, Verantwortung zu übernehmen für das Ergebnis, dessen Zustandekommen man technisch nicht überblickt (vgl. Durán und Jongsma 2021, S. 333).

Für andere technische Geräte existiert eine Produkthaftung, in der Hersteller:innen bzw. Entwickler:innen bis zu einem gewissen Punkt Verantwortung für ihr Produkt übernehmen, indem sie für Herstellungs- und Materialfehler haften. Bei ML-basierten Assistenzsystemen könnte analog eine Produkthaftung eingeführt werden, die in begrenztem Maße ihre Rechenleistung umfasst. Dafür müsste im Vorfeld genau der Leistungsbereich von ML-Systemen definiert werden, für den Entwickler:innen bei richtiger Anwendung Verantwortung übernehmen können. Eine tatsächlich ML-spezifische Herausforderung, die sich bei anderen technischen Geräten so nicht stellt, liegt hier jedoch darin, dass ML-Systeme beständig „dazulernen" und sich dadurch weiterentwickeln. Der Algorithmus wird sich also nach einigen Monaten der Anwendung schon erheblich von dem anfänglich hergestellten und zugelassenen Produkt unterscheiden. Diesem Problem könnte z. B. durch ein System fortlaufender Folge-Zulassungsverfahren begegnet werden, in denen nach bestimmten Abständen der dann jeweils veränderte Algorithmus wieder neu für die Praxis freigegeben wird (vgl. Beck et al. 2023, S. 258f.).

Schlussfolgerungen

Die Expert:innen-Interviews haben Einblicke in die Herausforderungen und Chancen gegeben, die sich bei der Zusammenarbeit mit ML im diagnostischen Alltag zeigen. Dabei wurden verschiedene Überlegungen dazu geäußert, wie eine gelingende Mensch-Maschine-Interaktion im Spannungsfeld von Verantwortung und Vertrauen gestaltet werden könnte. Auch wenn der Stichprobenumfang mit zehn Personen begrenzt ist, lassen sich aus den Interviews und der systematisch-ethischen Reflexion darüber einige richtungsweisende Impulse ableiten:

- Ärzt:innen sollten ein informationstechnologisches Grundverständnis davon haben, wie Algorithmen funktionieren, wie und mit welchen Datensätzen sie trainiert worden sind, wo ihre Stärken und Schwächen liegen etc. Sodann sollten sie Zeit investieren (dürfen), um sich mit der spezifischen Funktionsweise des ML-basierten Assistenzsystems an ihrem jeweiligen Arbeitsplatz zunächst vertraut zu machen. Das ist notwendig, um ein Gespür dafür zu bekommen, welche typischen Fehler zu erwarten sind, und eine realistische Einschätzung davon, für welche Aufgaben das ML-System eine verlässliche Autorität darstellt. Nur so kann ein

begründetes Vertrauen auf ML-Systeme entstehen, das mit einer gesunden Skepsis gepaart ist. Es sollte überlegt werden, in welchen Fällen damit die epistemische Bedingung von Verantwortung bereits erfüllt ist, wenn Einzelergebnisse des ML-Systems nicht mehr nachgeprüft werden (können).

- Entwickler:innen von ML-basierten Assistenzsystemen sollten die Möglichkeit für geeignete Formen der Nachvollziehbarkeit gewährleisten, damit Ärzt:innen sich im Zweifelsfall selbst eine Meinung bilden können. Gleichwohl scheint es ratsam, genau zu überlegen, an welchen Stellen Nachvollziehbarkeit von ML-Systemen wirklich notwendig und hilfreich ist: Wann müssen Einzelergebnisse des ML-basierten Assistenzsystems nachgeprüft werden und wann ist es verantwortbar, sich darauf zu verlassen, weil das ML-System diese Aufgabe aller Erfahrung nach besser löst als Menschen? Je komplexer die Fragestellung und je gravierender die möglichen Folgen der jeweiligen KI-„Entscheidung“, desto eher mag Nachvollziehbarkeit angeraten sein.
- Mensch und Maschine sollten in ihren unterschiedlichen Stärken profiliert werden, damit Aufgaben optimal zugeteilt werden können. Es muss also das Bewusstsein dafür geschult werden, was Menschen besser können als ML-basierte Assistenzsysteme und umgekehrt. Auf dieser Grundlage kann dann in unterschiedlichen Arbeitsbereichen im Einzelnen ausbuchstabiert werden, welche Kompetenzen für welche Aufgabe erforderlich sind und ob sich Mensch oder Maschine besser dafür eignen. Mögliche Einsatzbereiche von ML-Systemen können vorab im Team identifiziert werden; dabei sollten die von dem ML-basierten Assistenzsystem zu beantwortenden Fragestellungen möglichst präzise beschrieben sein. Man muss nicht zwangsläufig alle informationstechnologischen Prozesse verstehen, sollte aber genau wissen, an welcher Stelle ML gewinnbringend eingesetzt werden kann. Anders formuliert: Viel wichtiger, als die „Blackbox“ zu öffnen, ist es, sie richtig zu platzieren.
- Damit Ärzt:innen Verantwortung wahrnehmen können, muss sie begrenzt sein. Deshalb sollten Überlegungen zur Produkthaftung für ML-basierte Assistenzsysteme vorangetrieben werden.

Förderung Dieser Beitrag ist ein Ergebnis des Forschungsvorhabens „Responsibility Gaps in Human-Machine Interactions: The Ambivalence of Trust in AI“. Dieses Forschungsprojekt wird gefördert durch das Bayerische Forschungsinstitut für Digitale Transformation (bidt), einem Institut der Bayerischen Akademie der Wissenschaften.

Funding Open Access funding enabled and organized by Projekt DEAL.

Einhaltung ethischer Richtlinien

Interessenkonflikt W. Brandt, A. Fritz, A. Kießig und P. Lerch geben an, dass kein Interessenkonflikt besteht.

Ethische Standards Ein Ethikvotum wurde nicht eingeholt, da ausschließlich Interviews mit Expert:innen durchgeführt wurden, deren Anonymität im Rahmen der Auswertung und Publikation gewahrt wurde. Die Teilnahme erfolgte freiwillig nach Information über Ziel und Ablauf der Studie. Alle Interview-Partner:innen haben ihre Einwilligung zur Publikation der Studienergebnisse gegeben. Das Projekt inklusive der Befragungen wurde im Einklang mit nationalem Recht durchgeführt.

Literatur

Beck S, Faber M, Gerndt S (2023) Rechtliche Aspekte des Einsatzes von KI und Robotik in Medizin und Pflege. Ethik Med 35:247–263

Buddeberg E (2011) Verantwortung im Diskurs. Grundlinien einer rekonstruktiv-hermeneutischen Konzeption moralischer Verantwortung im Anschluss an Hans Jonas, Karl-Otto Apel und Emmanuel Lévinas. De Gruyter, Berlin

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

Eschenbach W (2021) Transparency and the black box problem: why we do not trust AI. Philos Technol 34:1607–1622. https://doi.org/10.1007/s13347-021-00477-0

Fritz A (2021) Vertrauen. In: Frick E, Hilpert K (Hrsg) Spiritual Care von A bis Z. De Gruyter, Berlin, S 365–368

Fritz A (2023) Vertrauenswürdigkeit in der digitalen Arbeitswelt. Eine Relektüre von Onora O'Neills Reith Lectures: A Question of Trust. In: Fritz A, Karl K (Hrsg) Persönlichkeitsbildung interdisziplinär. Die Bedeutung von Anerkennung und das Spannungsverhältnis zur Professionalität. Nomos, Baden-Baden, S 155–177 https://doi.org/10.5771/9783748931652

Fritz A, Brandt W (2019) Verteilte Moral in Zeiten von KI? Über die moralische Bedeutung technischer Artefakte in der Mensch-Maschine-Interaktion. ThPh 94(4):526–555

Hartmann M (2003) Akzeptierte Verletzbarkeit. Elemente einer normativen Theorie des Vertrauens. DZPhil 51(3):395–412

Holderegger A (1992) Verantwortung. In: Wils J-P, Mieth D (Hrsg) Grundbegriffe der christlichen Ethik. Schöningh, Paderborn, S 199–208

Körtner U (2001) Freiheit und Verantwortung. Studien zur Grundlegung theologischer Ethik. Univ.-Verlag, Freiburg

Mayring P (2022) Qualitative Inhaltsanalyse. Grundlagen und Techniken, 13. Aufl. Beltz, Weinheim Basel

O'Neill O (2020) Trust and accountability in a digital age. Philosophy 95(1):3–17. https://doi.org/10.1017/S0031819119000457

Rudy-Hiller F (2022) The epistemic condition for moral responsibility. In: Zalta E, Nodelman U (Hrsg) The Stanford encyclopedia of philosophy (https://plato.stanford.edu/archives/win2022/entries/moral-responsibility-epistemic. Zugegriffen: 30. Nov. 2024)

Ryan M (2020) In AI we trust: ethics, artificial intelligence, and reliability. Sci Eng Ethics 26(5):2749–2767. https://doi.org/10.1007/s11948-020-00228-y

Simmel G (1968) Soziologie. Untersuchungen über die Formen der Vergesellschaftung, 5. Aufl. Gesammelte Werke, Bd. 2. Duncker & Humblot, Berlin

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

Zerilli J, Knott A, Maclaurin J, Gavaghan C (2019) Transparency in algorithmic and human decision-making: is there a double standard? Philos Technol 32:661–683. https://doi.org/10.1007/s13347-018-0330-6