Situativität, Funktionalität und Vertrauen: Ergebnisse einer szenariobasierten Interviewstudie zur Erklärbarkeit von KI in der Medizin

Marquardt, Manuela; Graf, Philipp; Jansen, Eva; Hillmann, Stefan; Voigt-Antons, Jan-Niklas

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Marquardt, M., Graf, P., Jansen, E., Hillmann, S., & Voigt-Antons, J.-N. (2024). Situativität, Funktionalität und Vertrauen: Ergebnisse einer szenariobasierten Interviewstudie zur Erklärbarkeit von KI in der Medizin. *TATuP - Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis / Journal for Technology Assessment in Theory and Practice*, *33*(1), 41-47. https://doi.org/10.14512/tatup.33.1.41

TATuP Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis
Journal for Technology Assessment in Theory and Practice

oekom science

RESEARCH ARTICLE

Situativität, Funktionalität und Vertrauen: Ergebnisse einer szenariobasierten Interviewstudie zur Erklärbarkeit von KI in der Medizin

Manuela Marquardt[1], Philipp Graf*[,2], Eva Jansen[1], Stefan Hillmann[3], Jan-Niklas Voigt-Antons[4]

Zusammenfassung • Eine zentrale Anforderung an den Einsatz von künstlicher Intelligenz (KI) in der Medizin ist ihre Erklärbarkeit, also die Bereitstellung von adressat*innengerechten Informationen über ihre Funktionsweise. Dies führt zu der Frage, wie eine sozial adäquate Erklärbarkeit gestaltet werden kann. Um Bewertungsfaktoren zu identifizieren, befragten wir Akteur*innen des Gesundheitswesens zu zwei Szenarien: Diagnostik und Dokumentation. Die Szenarien variieren den Einfluss, den ein KI-System durch das Interaktionsdesign und die Menge der verarbeiteten Daten auf die Entscheidung hat. Wir stellen zentrale Bewertungsfaktoren für Erklärbarkeit auf interaktionaler und prozessualer Ebene dar. Erklärbarkeit darf im Behandlungsgespräch situativ nicht interferieren und die professionelle Rolle infrage stellen. Zugleich legitimiert Erklärbarkeit ein KI-System funktional als Zweitmeinung und ist zentral für den Aufbau von Vertrauen. Eine virtuelle Verkörperung des KI-Systems ist vorteilhaft für sprachbasierte Erklärungen.

* *Corresponding author: pgraf@hm.edu*
[1] *Institut für Medizinische Soziologie und Rehabilitationswissenschaften, Charité – Universitätsmedizin Berlin, Berlin, DE*
[2] *Fakultät für angewandte Sozialwissenschaften, Hochschule München, München, DE*
[3] *Quality & Usability Lab, Technische Universität Berlin, Berlin, DE*
[4] *Immersive Reality Lab, Hochschule Hamm-Lippstadt, Hamm-Lippstadt, DE*

https://doi.org/10.14512/tatup.33.1.41
Received: 22. 08. 2023; revised version accepted: 15. 12. 2023;
published online: 15. 03. 2024 (peer review)

***Situativity, functionality and trust**: Results of a scenario-based interview study on the explainability of AI in medicine*

Abstract • *A central requirement for the use of artificial intelligence (AI) in medicine is its explainability, i. e., the provision of addressee-oriented information about its functioning. This leads to the question of how socially adequate explainability can be designed. To identify evaluation factors, we interviewed healthcare stakeholders about two scenarios: diagnostics and documentation. The scenarios vary the influence that an AI system has on decision-making through the interaction design and the amount of data processed. We present key evaluation factors for explainability at the interactional and procedural levels. Explainability must not interfere situationally in the doctor-patient conversation and question the professional role. At the same time, explainability functionally legitimizes an AI system as a second opinion and is central to building trust. A virtual embodiment of the AI system is advantageous for language-based explanations.*

Keywords • *explainability, XAI, AI in healthcare, embodied AI, voice dialog system*

This article is part of the Special topic "AI for decision support: What are possible futures, social impacts, regulatory options, ethical conundrums and agency constellations?," *edited by D. Schneider and K. Weber. https://doi.org/10.14512/tatup.33.1.08*

Zielsetzung

In den letzten Jahren erfolgte eine stete Zunahme von Systemen künstlicher Intelligenz (KI) im medizinischen und therapeutischen Bereich: von Sprachassistenzsystemen zur Pflegedokumentation über Chatbots zur Förderung der mentalen Gesund-

heit bis hin zu diagnostischen Systemen wird KI in vielfältigen Formen angewandt (Becker 2019). Dabei rücken nicht nur die Verarbeitung größerer Datensätze, sondern auch die Diskussion um sensible Entscheidungsprozesse als mögliche Anwendungsfälle in den Fokus (Heil et al. 2021). Auch ‚weiche' KI-Systeme, wie dialogbasierte Assistenzsysteme, die keine expliziten Entscheidungen treffen, üben durch die Selektion und Darstellung von Informationen (Gupta et al. 2022), insbesondere durch deren Wortwahl (Mahmood und Huang 2022), einen impliziten Einfluss auf das Entscheidungsverhalten der mit ihnen interagierenden Personen aus.

Um eine sichere und ethisch vertretbare Anwendung eines KI-Systems im medizinischen Bereich zu ermöglichen, besteht die Forderung, die Funktionsweise der KI transparent und nachvollziehbar zu gestalten (Markus et al. 2021; Samhammer et al. 2023). Ein vielversprechender Ansatz ist die Entwicklung von Explainable Artifical Intelligence (XAI) Systemen. Erklärbarkeit wird in Anlehnung an Miller (2019, S. 5) als „an active characteristic of a model, denoting any action or procedure taken by a model with the intent of clarifying or detailing its internal functions" definiert. Um das Ziel der Nachvollziehbarkeit zu erreichen, müssen Inhalt und Form der Erklärung an den Adressat*innen orientiert sein (Barredo Arrieta et al. 2020). Eine sprachvermittelte Interaktion zwischen Nutzenden und dem KI-System bietet das Potenzial zur Bildung von Vertrauen mittels XAI (Hillmann et al. 2021). Auch wenn natürlichsprachliche Erklärungen im Rahmen von XAI noch nicht im Stand der Technik angekommen sind, zeigen erste Forschungsergebnisse ihre Machbarkeit in Form von Dialogen (Feldhus et al. 2023).

Neben den unterschiedlichen Adressat*innen der Erklärungen lassen sich KI-Systeme im medizinischen Bereich nach dem Zweck der Erklärungen differenzieren. Markus et al. (2021) schlagen drei Bereiche vor: Um Modellannahmen zu verifizieren oder zu verbessern, um neue Erkenntnisse zu gewinnen oder „[t]o manage social interaction" (Markus et al. 2021, S. 3). Allerdings besteht wenig Wissen darüber, wie Erklärbarkeit in KI-Systemen bei der Translation ins Gesundheitssystem konkret gestaltet werden soll und welche kontextuellen Bedarfe der Nutzenden relevant werden. Der Fokus dieses Beitrags liegt auf den sozialen Aspekten von Erklärbarkeit von KI in der Medizin, unabhängig von der technischen Ausprägung des zugrunde liegenden KI-Ansatzes (bspw. Klassifikation vs. Prädiktion, oder die Unterscheidung nach verschiedenen Methoden des maschinellen Lernens wie Support Vector Machines oder Deep Learning) oder den regulatorischen Implikationen (HEG-KI 2019).

Ein besonderer Forschungsbedarf besteht zu der Frage, welche soziale Rolle ein erklärbares KI-System zwischen Behandelnden und Patient*innen einnehmen könnte und welche Auswirkungen eine virtuelle (Avatar) oder real-weltliche (Roboter) Verkörperung des KI-Systems auf die Interaktion hat. Aus Sicht der Technikfolgenabschätzung ist es wünschenswert, die Diskussion der Umsetzung von XAI und ihrer sozialen Folgen mit empirischen Daten zu informieren. Ziel unserer Studie ist es, mithilfe von szenariobasierten Interviews Bewertungsfaktoren im Hinblick auf Erklärbarkeit von KI in der Medizin herauszuarbeiten und mit sozialen Deutungsmustern im medizinischen Kontext in Relation zu setzen.

Methode

Zur Identifikation relevanter Faktoren der Deutungsmuster eignen sich qualitative Interviews (Meuser und Sackman 1992). Diese setzen mittels vignettenförmiger Szenarien einen Stimulus, um Teilnehmende zu aktivieren und emotional bedingte, implizite Anteile an Bewertungen aufzudecken (Kosow und Gaßner 2008). Während das ‚Baseline-Szenario' den allgemeinen Rahmen steckt, variieren die Vignetten einzelne Aspekte. So können graduelle Bewertungen zu einem konkreten Thema abgefragt werden (Barter und Renold 1999). Vignetten sollten präzise, einfach und plausibel sein, um einen möglichst anschlussfähigen Stimulus zu bieten (Hughes und Huby 2004). Die vergleichend ausgestalteten Vignetten animieren Interviewte zu einer differenzierten oder widersprüchlichen Bewertung (Jenkins et al. 2010).

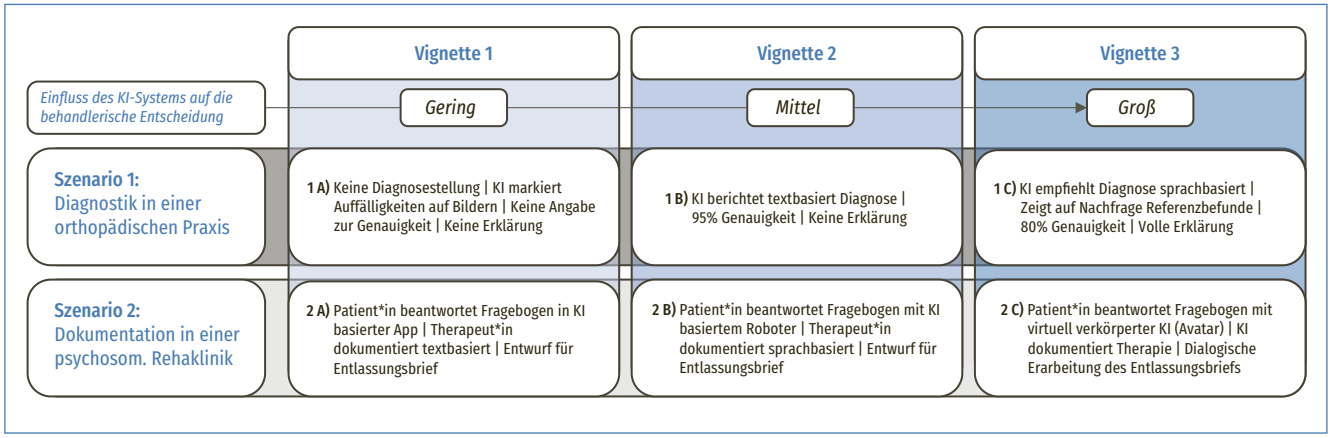

Abb. 1: Übersicht über die Konstruktion der Szenarien. *Quelle: eigene Darstellung*

https://doi.org/10.14512/tatup.33.1.41

Für diese Studie entwickelten wir zwei Baseline-Szenarien mit je drei Fallvignetten. Die Vignetten variieren im Einfluss auf die Entscheidung, die ein KI-System mittels des Interaktionsdesigns sowie des Umfangs der verarbeiteten Daten nimmt (siehe Abb. 1).

Das erste Szenario liegt im Bereich Diagnostik. Die Diagnostik ist eines der vielversprechendsten Einsatzgebiete für KI-Systeme in der Medizin (Samhammer et al. 2023), da hier große Mengen an Daten präzise und effizient analysiert und komplexe Muster erkannt werden können (Kinar et al. 2017). Wenn die Entscheidungsgrundlage nicht nachvollziehbar ist, kann der Einsatz von KI-Systemen zu erheblichen Vertrauensproblemen führen (Markus et al. 2021) und dadurch die traditionelle Beziehung zwischen Behandler*in und Patient*in untergraben (Čartolovni et al. 2022).

Das zweite Szenario adressiert den Bereich der Erhebung von Gesundheitsdaten sowie die Dokumentation einer Therapie und das Verfassen eines Entlassungsbriefes durch eine KI als Entlastung für Behandler*innen (Bossen und Pine 2023). Zentrale Konfliktlinien zeichnen sich in der Akzeptanz der Technologie (Čartolovni et al. 2022), im Umgang mit Datenschutz (Becker 2019) und in der Verstärkung von Ungleichheiten im Kontrast zur Arbeitserleichterung (Panch et al. 2019) ab.

Die Szenarien stellten wir in qualitativ-explorativen Online- und Präsenzinterviews insgesamt 16 Akteur*innen des Gesundheitswesens vor. Die Hälfte der Interviewten stammte aus dem ärztlichen Bereich (Tab. 1). Diese Auswahl wurde getroffen, da Ärzt*innen üblicherweise in den Gesundheitseinrichtungen, in denen sie arbeiten, am diagnostischen Prozess beteiligt sind. Zur Gewinnung einer vielfältigen Perspektive auf den KI-Einsatz in der Diagnostik und Dokumentation rekrutierten wir die Teilnehmenden aus verschiedenen Fachbereichen. Sie sollten die ihnen fremden Szenarien, die circa zehn Jahre in der Zukunft angesiedelt sind, auf ihren Bereich übertragen. Die Teilnehmenden wurden gebeten, die sozialen Folgen, den Nutzen, die Schwierigkeiten und die Funktionalität des jeweiligen KI-Systems und die gewünschte Form der Erklärung für Behandelnde und Patient*innen einzuschätzen.

Die Rekrutierung erfolgte über professionelle Netzwerke. Die Interviews fanden zwischen dem 12. 06. 23 und dem 14. 08. 23 statt und dauerten durchschnittlich 57 Minuten. Für die Auswertung der transkribierten Interviews verwendeten wir die Methode der Grounded Theory (Pentzold et al. 2018).

Inhaltliche Auswertung

Im Folgenden gliedern wir die zentralen Bewertungsfaktoren auf interaktionaler und prozessualer Ebene. Diese Differenzierung orientiert sich primär an raumzeitlichen Wirkungszusammenhängen: Die interaktionale Ebene thematisiert die konkrete Interaktionssituation zwischen Behandler*in und/oder Patient*in und dem KI-System. Sie fokussiert auf gleichzeitige und kopräsente Aspekte der Bewertung von Erklärbarkeit (situativ, kurzfristig), wobei die Form der Mitteilung einer Erklärung und die Verkörperung des KI-Systems besonders relevant werden. Die prozessuale Ebene umfasst längerfristige, translokale und organisationale Faktoren wie den Aufbau von Vertrauen, die Bedeutung von Standards sowie den Einfluss auf die fachliche Kompetenz.

Die von der KI abgegebene Erklärung oder gar das Aufdecken von ‚Fehlern' empfinden die Interviewten als unpassende Zweitmeinung, die die Fachkompetenz infrage stellt und das Vertrauensverhältnis gefährdet.

Interaktionale Ebene

Der *erste* Faktor auf interaktionaler Ebene, der von den Teilnehmenden als Einflussfaktor auf die Erklärbarkeit von KI-Systemen referenziert wurde, ist das Behandlungsgespräch (BG). Hier wird die Institution der professionellen ärztlichen Rolle wirkmächtig, die mit einem Vertrauensvorschuss, der Entscheidungshoheit und der Haftbarkeit für Behandlungsfehler verknüpft ist. Obwohl ein KI-Einsatz im BG nicht Teil der Fallvignetten des ersten Szenarios war, reflektierte die Mehrheit der Interviewten in ihren Einschätzungen darüber. Innerhalb eines BG ist eine Interaktion mit einer ‚pro-aktiven' KI tendenziell nicht erwünscht – hier sollte die „KI nicht interferieren" (3). Während die passive Nutzung von KI als reines Werkzeug, das auf Zuruf arbeitet, in Einzelfällen akzeptiert wird, stellt das Erklären von Sachverhalten für die meisten Behandler*innen eine Grenze dar, die die ärztliche Integrität untergräbt: „Das ist wie wenn ein Kollege ungefragt bei mir reinquatscht" (3). Die von der KI abgegebene Erklärung oder gar das Aufdecken von „Fehlern" (13) empfinden die Interviewten als unpassende Zweitmeinung, die die Fachkompetenz infrage stellt und das Vertrauensverhältnis gefährdet (10). Zudem dürfen Diagnosen nicht in einem BG verhandelt werden, „um den Patienten nicht zu verunsichern" (14). Als vertrauensbildend wurde die Möglichkeit der Bezugnahme auf KI genannt – nämlich als „Zusatz, dass die KI das auch gesagt hat" (15). Die KI kann als Statussymbol für eine hohe technische Ausstattung (2) eine vertrauensfördernd positive Wirkung haben.

https://doi.org/10.14512/tatup.33.1.41

Die befragten Behandler*innen stellten allerdings die notwendige Kompetenz infrage, da KI „nicht empathisch genug“ sei und daher „nicht mitreden“ solle (3). Dies betrifft insbesondere das Erklären persönlicher Einzelfalldiagnosen, die bei schwerwiegenden Fällen emotionaler Arbeit bedürfen (15, 12). Eine Ausnahme davon stellt das Vermitteln von „Lehrbuchwissen“ (9) dar: Die direkte Interaktion zwischen KI und Patient*in könnte Entlastung schaffen und ein zusätzliches Informationsangebot sein (3).

Die Form der Mitteilung einer Erklärung ist der *zweite* relevante Faktor auf interaktionaler Ebene. Sie variierte zwischen den Fallvignetten beider Szenarien. Die Antworten auf die Frage nach der Präferenz einer sprachlichen und/oder visuellen Form divergieren situativ und mit der eigenen Erfahrung mit Sprachassistenzsystemen und deren Fehleranfälligkeit sowie den kontextuellen Faktoren der eigenen Berufspraxis. Viele Befragte wünschten sich eine gut funktionierende, korrigierbare Diktieroption (8, 16). Sprache ist für einige das effizientere Medium, um mit einer KI zu interagieren (2, 4, 8, 13, 14). Für andere ist ein Text und zusätzliches Bildmaterial unumgänglich zur Einschätzung der Erklärung. Einerseits, weil eine Hürde darin bestünde, in direkte, sprachliche Interaktion mit einer KI zu gehen: „Sprachbasiert funktioniert astrein, aber ich würde es lieber als Text haben, weil ich mich nicht mit einem Computer austauschen möchte“ (3). Andererseits ist für die Befragten die Dokumentation als Grundlage der Erklärung mit einer visuellen Darstellung der Vorbefunde und Statistiken leichter zugänglich (8) und damit nachvollziehbarer (6). Insbesondere im stationären Kontext, „wo eh schon immer so viel los ist“ (16), wird die textbasierte Form gegenüber Sprache aus Gründen des Datenschutzes bevorzugt. Neben der persönlichen Präferenz und Befähigung nimmt der räumlich-soziale Faktor die zentrale Rolle ein, ob sprachliche Interaktion mit einer KI als angemessen wahrgenommen wird oder nicht.

Nr.	Alter	Ausbildung	Position
1.	40	Psychologie	Wissenschaftliche Mitarbeiterin
2.	70	Zahnmedizin	Oberarzt
3.	35	FA Dermatologie	Funktionsoberarzt
4.	50	FA Psychosomatik	Oberarzt
5.	67	FA Pädiatrie & Neonatologie	Chefarzt
6.	45	FA Dermatologie	Fachärztin
7.	73	Orthopädie	Arzthelferin
8.	35	FA Radiologie	Fachärztin
9.	57	Krankenpflege	Hygienefachschwester
10.	52		Patient*innenvertretung
11.	53	Medizin	Gesundheitsmanagerin
12.	60		Patient*innenvertretung
13.	42	Psychosomatik	Psychotherapeutin
14.	42	FA Kardiologie	Fachärztin
15.	38	FA Gynäkologie	Fachärztin
16.	28		Medizinstudent
	ø = **49**	w = 10 | m = 6	

Tab. 1: Übersicht über die Teilnehmenden. *Quelle: eigene Darstellung*

Der *dritte* und letzte Faktor ist die Verkörperung einer KI im Hinblick auf Erklärbarkeit, die in den Fallvignetten des zweiten Szenarios variiert. Einige der Befragten betrachteten eine virtuelle oder real-weltliche Verkörperung als angemessen für ihre Patient*innen, nicht jedoch für sich selbst, da hier „eine ganz normale Spracheingabe“ ausreichend sei (14). Andere würden hingegen bevorzugt mit einem Avatar (6) oder einem Roboter (3) interagieren, wobei eine kleine Minderheit der Interviewten eine virtuelle Verkörperung bevorzugt. Die Präferenz geht auf eine Differenz in der Bewertung des ontologischen Status beider Formen zurück: Einen Roboter nehmen die Befragten als Maschine wahr, die ein „eigenartiges Gefühl“ auslöst (15). Seine real-weltliche Verkörperung wird aufgrund der geringeren Kosteneffizienz (8) und der nicht vorhandenen Translokalität (8, 11) als ungeeignet wahrgenommen und implizit stärker mit dem Narrativ des ‚Ersetzt-Werdens‘ assoziiert (2, 6). Die Teilnehmenden perzipieren den ‚stärker‘ verkörperten Roboter als Entität mit graduellem Personenstatus. Dieser ist nicht mehr nur Werkzeug, sondern erlangt eine darüber hinaus gehende, nicht erwünschte, Präsenz (6). Im Kontakt mit Patient*innen erachteten sie eine real-weltliche Verkörperung hingegen aufgrund der „Plastizität“ (3) und Anonymität (9) als Chance. Den Avatar sehen die Befragten als passendere Form an, die dem ‚Werkzeugcharakter‘ der KI entspräche. Für den Zweck sprachlicher Erklärbarkeit wird er als angemessener empfunden, da er abschaltbar und dann „nicht so invasiv“ wie ein Roboter (16) sei.

Prozessuale Ebene

Als *ersten* Bewertungsfaktor auf prozessualer Ebene im Hinblick auf die Erklärbarkeit von KI-Systemen identifizierten wir den Aufbau von Vertrauen. Erklärbarkeit stellt eine Bedingung für eine „Vertrauensbasis“ dar, „die aufgebaut werden [muss]“ (9), wie folgendes Zitat verdeutlicht: „Feedback hätte ich schon gerne, warum das System auf die Diagnose kommt, ansonsten würde ich dem System nicht vertrauen“ (6, auch 9, 14). Dabei ist es wichtig, „aus dem Schwarz-Weiß [raus zu gehen] – es muss eine Graustufe haben und auf die reagieren können“ (9). Eine differenzierte Einschätzung mit Angabe des Grades der Unsicherheit

 https://doi.org/10.14512/tatup.33.1.41

rege zu einem eigenen Kontrollprozess an: „Gleichzeitig kritisiert es sich noch selber, was es glaubwürdiger macht, […] das heißt, es sagt dem Betrachter selber, ich bin nicht 100 % genau, das heißt, du musst selber noch mal gucken – das ist super“ (10). Als weitere „vertrauensbildende Maßnahme“ nannte eine Ärztin, „dass man am Anfang, vielleicht im ersten halben Jahr, viel kontrolliert – so wie wenn ich mit einem Assistenzarzt arbeite“ (6). Später reiche es, nur noch bei schweren Diagnosen zu kontrollieren. Der Aufbau von Vertrauen kann also den Bedarf an Erklärbarkeit zum Teil ersetzen, dennoch muss die Möglichkeit der Nachvollziehbarkeit immer gegeben bleiben. Eine besondere Stellung nimmt der Zugriff und der Bezug in den Erklärungen auf die Rohdaten (8, 11, 14) ein. Erklärbarkeit und Vertrauen in die Reliabilität dieser Erklärungen sind diejenigen Aspekte, die aus dem ‚Werkzeug KI‘ eine „Zweitmeinung“ (13) werden lassen und eine „Kommunikation auf Augenhöhe“ (6) ermöglichen: „Man hat die Möglichkeit nachzufragen und bekommt Infos. Das wäre, als ob sich zwei Ärzte unterhalten“ (13).

Vertrauen kann durch fachspezifische Standards, etwa im Sinne von „Datenbankstandards“ (10), sowie der Absicherung durch medizinische Fachgesellschaften oder Verbände mittels Zertifikaten externalisiert werden, „ähnlich wie bei einem neuen Medikament“ (5). Von besonderem Interesse für die Befragten war, wer in welchem Umfang welche Daten für das Training der KI berücksichtigt hat (5, 6, 9, 10, 16) und welche Verantwortlichkeit daraus resultiert (9). Sollten sich bestimmte KI-Systeme als „Goldstandard“ (11) durchsetzen, würde dies den Bedarf an Erklärbarkeit des allgemeinen Funktionierens reduzieren.

Der *zweite* Faktor zur Bewertung von Erklärbarkeit auf prozessualer Ebene resultiert aus verschiedenen Nutzen- und Problemvisionen und bewegt sich im Rahmen von Verbesserungen der Qualität medizinischer Versorgung. Erklärbarkeit kommt bei diesen Visionen ein ambivalenter Stellenwert zu. Der Einsatz von KI als „bessere Expertise“ (5) kann die Effizienz (3, 6, 12, 15), Genauigkeit und damit Qualität der medizinischen Versorgung verbessern. Erklärungen als zusätzliche Reflexionsangebote „zur Absicherung, Bestätigung [und] Hilfestellung“ (2) haben das Potenzial, die Genauigkeit zu steigern und zur Sicherstellung gleichbleibender Qualität der medizinischen Versorgung trotz hoher Arbeitsbelastung (8) beizutragen. Die Interviewten sahen es jedoch als Gefahr für die Qualität medizinischer Versorgung an, wenn Empfehlungen einer KI „unreflektiert übernommen“ (14, auch 5) werden oder „man sich als Therapeut zu sehr auf die Technik verlässt“ (4, auch 15). Erklärungen sind eine zentrale Stellschraube für die Vermeidung blinden Technikvertrauens, das zu Deskilling, also dem Abbau von Kompetenzen, führen könne. Bei einer schlechten Umsetzung bestünde zudem das Risiko eines Effizienzverlusts (5).

Im Zusammenhang mit einer möglichen Kompetenzerweiterung durch den Einsatz von KI (4, 14) für den medizinischen Nachwuchs (7, 8) und das nicht-ärztliche Personal (7, 9) oder wenn kein Facharzt vor Ort ist (6, 11), existieren gegenläufige Tendenzen im Hinblick auf Erklärbarkeit. In Kombination mit der eigenen fachlichen Expertise schätzen die Behandelnden das Risiko für Fehlentscheidungen beim Einsatz von erklärbarer KI als gering ein (6). Erklärbarkeit sei nur dann sinnvoll, wenn die adressierte Person in der Lage ist, die zusätzlich zur Verfügung gestellten Informationen in die eigene Entscheidungspraxis zu integrieren und dafür die Verantwortung zu übernehmen. Ein kontrastierendes Beispiel aus unserem Interviewmaterial verdeutlicht dies anhand des Einsatzes eines Diagnostik-KI-Systems durch nicht-ärztliche Gesundheitsfachkräfte (7). Die Befragte äußerte ihre Präferenz für ein Blackbox-KI-System, „schon damit nicht so viele Fragen kommen“ (7) und keine eigenständige Entscheidung getroffen werden müsse. Die Übernahme der Verantwortung für die Diagnosestellung wird in diesem Fall aufgrund fehlender Fachkompetenz und geringerer Bezahlung abgelehnt (9).

Folgerungen

Die vorliegende Analyse widmete sich mithilfe szenariobasierter Interviews empirisch ermittelten Bewertungsfaktoren sozialadäquater Erklärbarkeit von KI-Systemen im medizinischen Bereich. Erklärbarkeit von KI ist durch den definitorischen Zuschnitt auf eine soziale Adressat*in (Miller 2019) der Erklärung eines erklärenden Agenten (Barredo Arrieta et al. 2020) ein sozialwissenschaftlich anschlussfähiges Thema. In der Auswertung zeigt sich ein differenziertes – teils widersprüchliches – Netzwerk an Faktoren, die bei der Bewertung von XAI in der Medizin auf interaktionaler und prozessualer Ebene wirksam werden. Im Folgenden möchten wir die Anforderungen an sozial-adäquate Erklärbarkeit unter den drei Aspekten Situativität, Funktionalität und Vertrauen zusammenfassen.

Situativität

Unter den Aspekt der Situativität fällt die Frage, wann und wie eine Erklärung angebracht erscheint. Die Befragten äußerten sich je nach persönlicher Erfahrung, Fachgebiet und Berufspraxis ambivalent zur Frage der sprachlichen Interaktion mit einer KI. Sie assoziierten spezifische Herausforderungen sprachlicher Kommunikation mit Technik (Dickel 2021). Häufig präferierten sie für die Berufspraxis text- bzw. datenbasierte Erklärungen, obwohl sie situative Vorteile in sprachbasierten und verkörperten Formen von KI sahen. Tendenziell empfanden sie real-weltliche Verkörperungen durch Roboter für den Zweck von Erklärbarkeit als weniger angemessen als eine virtuelle Form. Eine häufig formulierte Grenze für den Einsatz von Erklärbarkeit stellte das BG dar, was im Einklang mit internationaler Literatur steht (Aminololama-Shakeri und López 2019; Powell 2019). Diese Grenze schützt die Autorität und Institution der professionellen ärztlichen Rolle.

Funktionalität

Der Aspekt der Funktionalität umfasst die Bandbreite von Fähigkeiten und Aufgaben, die einer KI zugewiesen werden können, welche Erwartungen im medizinischen Alltag an sie gerichtet sind und inwiefern Erklärbarkeit hier interferiert. Während

https://doi.org/10.14512/tatup.33.1.41

einerseits akkurate Erklärungen die Genauigkeit und Reflexionsleistung steigern können, besteht zugleich Skepsis hinsichtlich der Frage, wie diese zusätzlichen Informationen konkret in bestehende Arbeitsabläufe integriert werden können (Stichwort Zweitmeinung und Kompetenzerweiterung) und welche Konsequenzen sich daraus mittelfristig ergeben (Stichwort Patient*innenkontakt und Deskilling). Entgegen der von den Befragten geäußerten Vision KI als Zweitmeinung zu nutzen, finden sich in der Literatur empirische Hinweise, die eine bloße Assistenzfunktion von KI plausibler erscheinen lassen. KI zeigt sich hier als „unreliable in predictable ways" (Bossen und Pine 2023, S. 16), die eine stetige Kontrolle nötig macht. Erklärbarkeit kann helfen, diese Lücke zu schließen, birgt aber ebenso das Risiko, zu Fehlern zu führen, die ein Mensch nicht machen würde. Eine einheitliche soziale Rolle für eine KI findet sich in unseren Ergebnissen noch nicht. Vielmehr oszilliert diese zwischen einer assistierenden Funktion und einem Agenten mit gleich- oder höherwertiger Kompetenz. Konkrete Rollenerwartungen an ein je spezifisches KI-System dürften sich erst durch wiederholt bewährte Interaktionen verfestigen.

Vertrauen

Ein zentrales Leitmotiv ist Vertrauen, das aufgebaut, in Anspruch genommen, untergraben oder externalisiert wird und mit dem Bedarf an Erklärbarkeit eng verwoben ist. Durch Erklärungen kann das Vertrauen in eine KI gestärkt werden, was den Bedarf an Erklärbarkeit in der wiederholten Nutzung mit der Zeit verringert, wenngleich Kontrolle durch den Zugriff auf die ‚Rohdaten' gegeben sein sollte. Dieser Befund steht in Einklang mit der Unterscheidung von Vertrauen als Einstellung und ‚reliance' als Verhalten „[…] that follows the level of trust" (Scharowski et al. 2023, S. 4). Kloker et al. (2022, S. 1) unterstreichen die vermittelnde Rolle von Erklärbarkeit zwischen Vertrauen und Vorsicht: „[… M]aximizing trust is not the goal, rather to balance caution and trust to find the level of appropriate trust". Dies spiegelt sich in unseren Ergebnissen wider: Das Vertrauen in die Sicherheit einer KI geht dialektisch aus dem Umgang des Systems mit Unsicherheit hervor, weil dieser zu einer eigenständigen Reflexionsleistung anregt und einer ‚overreliance' vorbeugt.

Zusammenfassend möchten wir das disruptive Potenzial von XAI für mögliche Zukünfte im Kontext der Medizin betonen, das sich an den heterogenen Bewertungsfaktoren ablesen lässt. Durch den Zuschnitt der Szenarien und die Auswahl der Interviewten konnte nur ein Bereich möglicher Anwendungen von XAI beleuchtet werden. Weitere Forschung ist notwendig, um das Zusammenspiel der einzelnen Faktoren besser zu verstehen. Ein ethnographischer Ansatz, der auf die reale und langfristige Einbindung von XAI fokussiert, bietet sich zur Untersuchung der sozialen Dimension von Erklärbarkeit an.

Funding • This article received funding by the German federal ministry of education and research (BMBF) as part of the MIA-PROM project (Multimodal Interactive Assistance for the Collection of Patient Reported Outcome Measures).

Competing interests • The authors declare no competing interests.

Literatur

Aminololama-Shakeri, Shadi; López, Javier (2019): The doctor-patient relationship with artificial intelligence. In: American Journal of Roentgenology 212 (2), S. 308–310. https://doi.org/10.2214/AJR.18.20509

Barredo-Arrieta, Alejandro et al. (2020): Explainable artificial intelligence (XAI). Concepts, taxonomies, opportunities and challenges toward responsible AI. In: Information Fusion 58, S. 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Barter, Christine; Renold, Emma (1999): The use of vignettes in qualitative research. In: Social Research Update 25. Online verfügbar unter https://sru.soc.surrey.ac.uk/SRU25.html, zuletzt geprüft am 11. 01. 2024.

Becker, Aliza (2019): Artificial intelligence in medicine. What is it doing for us today? In: Health Policy and Technology 8 (2), S. 198–205. https://doi.org/10.1016/j.hlpt.2019.03.004

Bossen, Claus; Pine, Kathleen (2023): Batman and Robin in healthcare knowledge work. Human-AI collaboration by clinical documentation integrity specialists. In: ACM Transactions on Computer-Human Interaction 30 (2), S. 1–29. https://doi.org/10.1145/3569892

Čartolovni, Anto; Tomičić, Ana; Lazić Mosler, Elvira (2022): Ethical, legal, and social considerations of AI-based medical decision-support tools. A scoping review. In: International Journal of Medical Informatics 161, S. 104 738. https://doi.org/10.1016/j.ijmedinf.2022.104738

Dickel, Sascha (2021): Wenn die Technik sprechen lernt. Künstliche Kommunikation als kulturelle Herausforderung mediatisierter Gesellschaften. In: TATuP – Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis 30 (3), S. 23–29. https://doi.org/10.14512/tatup.30.3.23

Feldhus, Nils; Wang, Qianli; Anikina, Tatiana; Chopra, Sahil; Oguz, Cennet; Möller, Sebastian (2023): InterroLang. Exploring NLP models and datasets through dialogue-based explanations. In: Houda Bouamor, Juan Pino und Kalika Bali (Hg.): Findings of the Association for Computational Linguistics: EMNLP 2023. Singapur: Association for Computational Linguistics, S. 5399–5421. https://doi.org/10.18653/v1/2023.findings-emnlp.359

Gupta, Akshit; Basu, Debadeep; Ghantasala, Ramya; Qiu, Sihang; Gadiraju, Ujwal (2022): To trust or not to trust. How a conversational interface affects trust in a decision support system. In: Frédérique Laforest et al. (Hg.): Proceedings of the ACM Web Conference 2022. New York, NY: Association for Computing Machinery, S. 3531–3540. https://doi.org/10.1145/3485447.3512248

HEG-KI – Hochrangige Expertengruppe für KI (2019): Ethik-Leitlinien für eine vertrauenswürdige KI. Brüssel: Europäische Kommission. https://doi.org/10.2759/22710

Heil, Reinhard et al. (2021): Artificial intelligence in human genomics and biomedicine. Dynamics, potentials and challenges. In: TATuP – Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis 30 (3), S. 30–36. https://doi.org/10.14512/tatup.30.3.30

Hillmann, Stefan; Möller, Sebastian; Michael, Thilo (2021): Towards speech-based interactive post hoc explanations in explainable AI. In: Astrid Carolus, Carolin Wienrich und Ingo Siegert (Hg.): Proceedings of the 1st AI-Debate Workshop. Magdeburg: Universität Magdeburg, pp. 13–15. http://dx.doi.org/10.25673/38473

Hughes, Rhidian; Huby, Meg (2004): The construction and interpretation of vignettes in social research. In: Social Work and Social Sciences Review 11 (1), S. 36–51. https://doi.org/10.1921/swssr.v11i1.428

Jenkins, Nicholas; Bloor, Michael; Fischer, Jan; Berney, Lee; Neale, Joanne (2010): Putting it in context. The use of vignettes in qualitative interviewing. In: Qualitative Research 10 (2), S. 175–198. https://doi.org/10.1177/1468794109356737

Kinar, Yaron et al. (2017): Performance analysis of a machine learning flagging system used to identify a group of individuals at a high risk for colorectal cancer. In: PLoS One 12 (2), S. e0171759. https://doi.org/10.1371/journal.pone.0171759

Kloker, Anika; Fleiß, Jürgen; Koeth, Christoph; Kloiber, Thomas; Ratheiser, Patrick; Thalmann, Stefan (2022): Caution or trust in AI? How to design XAI in sensitive use cases? In: AMCIS 2022 Proceedings 16.

Kosow, Hannah; Gaßner, Robert (2008): Methoden der Zukunfts- und Szenarioanalyse. Überblick, Bewertung und Auswahlkriterien. Berlin: Institut für Zukunftsstudien und Technologiebewertung. Online verfügbar unter https://www.researchgate.net/publication/262198781_Methoden_der_Zukunfts-und_Szenarioanalyse_Uberblick_Bewertung_und_Auswahlkriterien, zuletzt geprüft am 15. 01. 2024.

Mahmood, Amama; Huang, Chien-Ming (2022): Effects of rhetorical strategies and skin tones on agent persuasiveness in assisted decision-making. In: Carlos Martinho, João Dias, Joana Campos und Dirk Heylen (Hg.): Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents. New York, NY: Association for Computing Machinery, S. 1–8. https://doi.org/10.1145/3514197.3549628

Markus, Aniek; Kors, Jan; Rijnbeek, Peter (2021): The role of explainability in creating trustworthy artificial intelligence for health care. A comprehensive survey of the terminology, design choices, and evaluation strategies. In: Journal of Biomedical Informatics 113, S. 103 655. https://doi.org/10.1016/j.jbi.2020.103655

Meuser, Michael; Sackmann, Reinhold (1992): Zur Einführung. Deutungsmusteransatz und empirische Wissenssoziologie. In: Michael Meuser und Reinhold Sackmann (Hg.): Analyse sozialer Deutungsmuster. Beiträge zur empirischen Wissenssoziologie. Pfaffenweiler: Centaurus-Verlagsgesellschaft, S. 9–37.

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

Panch, Trishan; Mattie, Heather; Atun, Rifat (2019): Artificial intelligence and algorithmic bias. Implications for health systems. In: Journal of Global Health 9 (2), S. 010 318. https://doi.org/10.7189/jogh.09.020318

Pentzold, Christian; Bischof, Andreas; Heise, Nele (2018): Einleitung. Theoriegenerierendes empirisches Forschen in medienbezogenen Lebenswelten. In: Christian Pentzold, Andreas Bischof und Nele Heise (Hg.): Praxis grounded theory. Theoriegenerierendes empirisches Forschen in medienbezogenen Lebenswelten. Wiesbaden: Springer, S. 1–24. https://doi.org/10.1007/978-3-658-15999-3_1

Powell, John (2019): Trust me, I'm a chatbot. How artificial intelligence in health care fails the Turing Test. In: Journal of Medical Internet Research 21 (10), S. e16222. https://doi.org/10.2196/16222

Samhammer, David et al. (2023): Klinische Entscheidungsfindung mit Künstlicher Intelligenz. Ein interdisziplinärer Governance-Ansatz. Heidelberg: Springer. https://doi.org/10.1007/978-3-662-67008-8

Scharowski, Nicolas; Perrig, Sebastian; Svab, Melanie; Opwis, Klaus; Brühlmann, Florian (2023): Exploring the effects of human-centered AI explanations on trust and reliance. In: Frontiers in Computer Science 5, S. 1–15. https://doi.org/10.3389/fcomp.2023.1151150

MANUELA MARQUARDT
ist ausgebildete Techniksoziologin und arbeitet seit 2019 als wissenschaftliche Mitarbeiterin an der Charité zu rehabilitationswissenschaftlichen Themen. Seit 2022 promoviert sie im Projekt MIA-PROM.

PHILIPP GRAF
ist Techniksoziologe und Doktorand an der TU Chemnitz zum Thema sozialer Robotik. Seit 2022 arbeitet er für die Hochschule München im Projekt MIA-PROM.

DR. EVA JANSEN
ist Medizinethnologin und seit 2021 am Institut für medizinische Soziologie und Rehabilitationswissenschaften der Charité.

DR. STEFAN HILLMANN
forscht und arbeitet seit 2010 zu sprachgestützter Mensch-Maschine Interaktion. 2017 promovierte er zur Simulation von Nutzerverhalten bei Interaktionen mit multimodalen Dialogsystemen. Aktuell forscht er am Einsatz von KI-Methoden zum Trainieren und Evaluieren von Dialogsystemen.

PROF. DR. JAN-NIKLAS VOIGT-ANTONS
ist seit 2021 Professor für Angewandte Informatik mit dem Schwerpunkt immersive Medien an der Hochschule Hamm-Lippstadt. Er leitet das Immersive Reality Lab, welches sich auf die Erforschung von immersiven Anwendungen sowie auf Schnittstellen zwischen Menschen und Technik fokussiert.

https://doi.org/10.14512/tatup.33.1.41