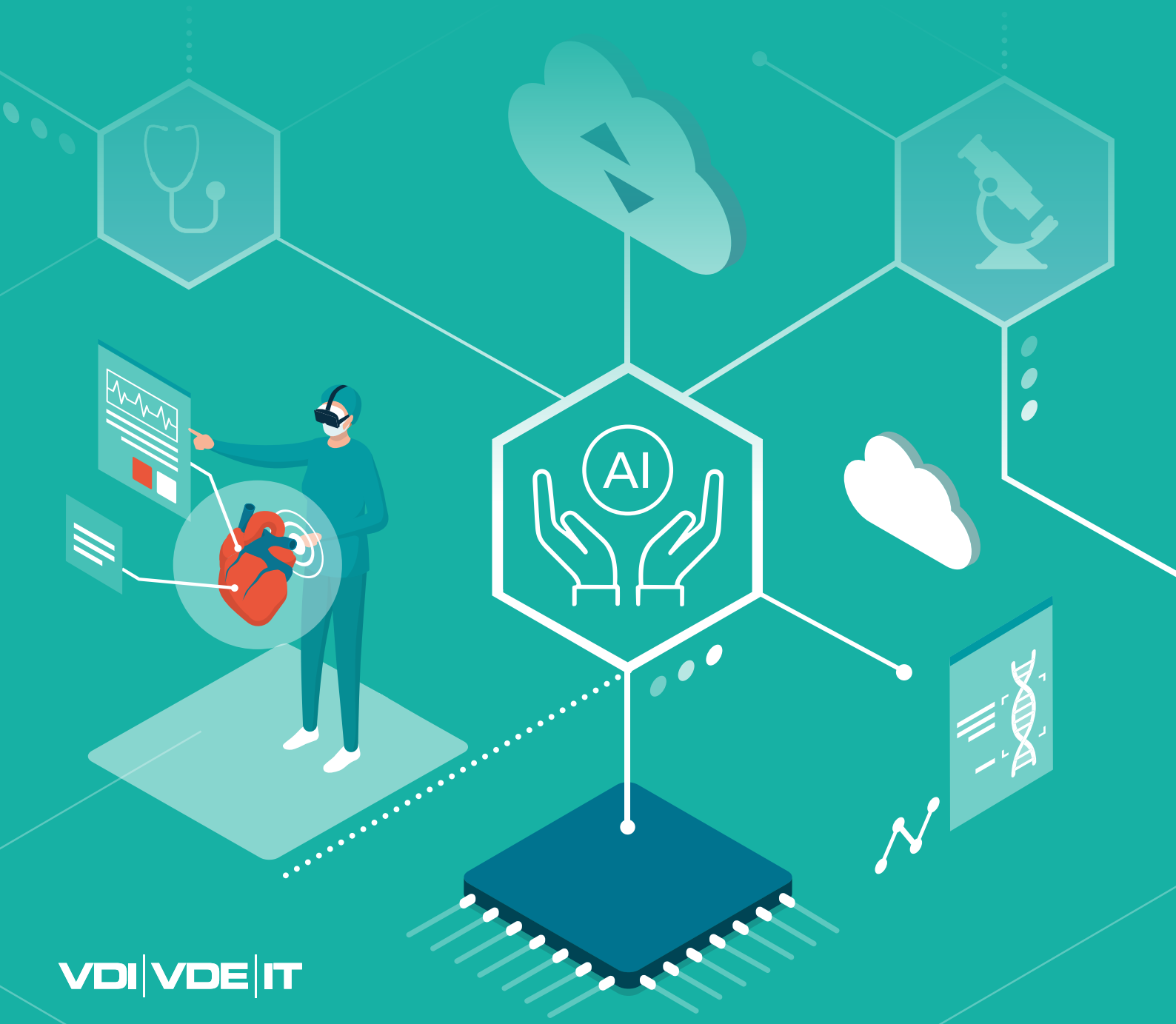
Zwischen Anspruch und Anwendung: Wie Responsible AI in der Medizintechnik gelingen kann

Marius Müller, Bettina Schmietow, Markus Gerold



VDI|VDE|IT

Zwischen Anspruch und Anwendung: Wie Responsible AI in der Medizintechnik gelingen kann

Die rasante Entwicklung von KI-Technologien eröffnet der Medizin neue Diagnose- und Therapie-Perspektiven – von der Bildanalyse über Chat-Bots bis zur Outcome-Prognose – doch die klinische Implementierung bleibt bislang fragmentiert. Parallel dazu zeigen sich ethische Herausforderungen: mangelnde Transparenz, Datenschutz-Risiken, versteckte Biases, unklare Haftungsfragen und der potenzielle Verlust menschlicher Entscheidungsautonomie. Diverse Leitlinien definieren Prinzipien wie Fairness, Transparenz und Robustheit, deren Heterogenität jedoch zu Interpretations- und Umsetzungs-Lücken führt. Responsible Artificial Intelligence (RAI) ist daher eine strukturelle Notwendigkeit: Durch konkrete Maßnahmen und Strategien sollen ethische Vorgaben in den gesamten Lebenszyklus medizinischer KI-Systeme integriert werden. Best-Practices zeigen, wie organisatorische, technische, regulatorische und datenbezogene Ansätze den sogenannten Principle-to-Practice-Gap schließen können. Es wird deutlich, dass ein Rahmen aus klarer Zweckdefinition, Bias-Analysen, Lifecycle-Monitoring und transparenter Haftungszuweisung entscheidend ist, um das Potenzial von KI im Gesundheitswesen verantwortungsbewusst zu realisieren.

KI in der Medizin: Potenziale und ethische Einordnung

Die rasante Entwicklung von Künstlicher Intelligenz (KI) und Maschinellem Lernen (ML) eröffnet der medizinischen Versorgung bislang ungeahnte Anwendungsmöglichkeiten und Mehrwerte, doch die belastbare Umsetzung in der klinischen Praxis bleibt bislang selten (Ingrisch 2026). KI-gestützte Systeme können heute bereits gesundheitsrelevante Auffälligkeiten durch bildgebende Verfahren – etwa in der Dermatologie oder Radiologie – zuverlässig erkennen, Diagnoseprozesse durch interaktive Chatbots unterstützen, den zukünftigen Nachsorgebedarf von Patienteninnen und Patienten prognostizieren und Therapie-Outcomes simulieren (Ali et al. 2023). Trotz dieser vielversprechenden Potenziale entwickelt sich die Technologie schneller als ihre Integration in etablierte Arbeitsabläufe. Gleichzeitig fehlt es an belastbaren Evaluierungen, die den klinischen Nutzen eindeutig belegen (Dassel et al. 2025).

Auch die ethische Bewertung von KI-basierten Systemen erweist sich als komplex und weitgehend noch ausstehend. Während technologische Fortschritte im Bereich des Maschinellen Lernens zentrale Treiber darstellen, rücken ethische Überlegungen und Menschenrechte zunehmend in den Fokus der Diskussion. Zwei grundverschiedene Modellklassen stehen dabei im Zentrum: transparente, interpretierbare Modelle, deren Vorhersagen nachvollziehbar sind, und undurchsichtige „Black-Box"-Modelle, die zwar häufig eine höhere Genauigkeit bieten, jedoch schwer zu interpretieren sind. Diese Dichotomie führt zu einer Reihe von moralischen Kernaspekten, darunter ein mangelnder Transparenzgrad, unzureichender Schutz der Privatsphäre bei der Nutzung großer, sensibler Gesundheitsdatensätze, das Risiko von verstecktem oder inadäquatem Bias in Daten und Algorithmen und in Bezug auf biologische und soziale Merkmale bzw. Kategorien (zu Geschlecht und Gender vgl. Muth & Schmietow 2026), was zu Diskriminierung führen kann, sowie die Gefahr eines Verlusts menschlicher Entscheidungshoheit, wenn KI-Systeme eigenständig handeln. Darüber hinaus werfen Fragen der Datenkontrolle, des möglichen Missbrauchs, der fehlenden Rechenschaftspflicht und der Haftungszuweisung zusätzliche ethische und rechtliche Probleme auf. Ökonomische und gesellschaftliche Implikationen, wie potenzielle Arbeitsplatzverluste, verstärkte Überwachung und ungleiche Verteilung von Gesundheitsressourcen, verstärken die Komplexität der Bewertung. Mit der zunehmenden Nutzung dieser Modelle wächst die Forderung nach Mechanismen, um die Funktionsweise dieser Algorithmen besser zu verstehen (Gunasekara et al. 2025).

Leitlinien zur Risikominimierung: Heterogenität und Deutungsspielraum

Zur Risikominimierung wurden bereits diverse Maßnahmen entwickelt, die regulatorische, technische und soziale Dimensionen adressieren (Stahl et al. 2023). Regulatorische Ansätze umfassen Gesetzgebungen, Richtlinien und öffentliche Register, die die Nutzung von Daten und KI-Systemen kontrollieren, ebenso wie Berichtspflichten und Monitoring-Mechanismen. Technische Strategien beruhen auf dem Testen von Algorithmen an diversen, repräsentativen Datensätzen und dem Einsatz offener Daten

und Open-Source-Software, um die Nachvollziehbarkeit zu erhöhen. Ergänzend dazu werden Verhaltenskodizes, Aufklärungskampagnen, Fortbildungsprogramme sowie Rahmenwerke und Toolkits bereitgestellt, die die Implementierung ethischer Standards unterstützen sollen. Diese Maßnahmen bringen Vorteile mit sich, etwa die Stärkung der Anwenderkompetenz, die Durchsetzbarkeit von Vorgaben und die Schaffung objektiver Bewertungskriterien. Daneben existieren aber auch Nachteile, darunter häufig ein unklarer Rollen- und Verantwortungsrahmen (Stichwort Verantwortungsdiffusion) sowie die Gefahr einer einseitigen Verlagerung der Verantwortung auf einzelne Parteien, beispielsweise. Entwicklerinnen oder Endnutzer.

Gleichzeitig entstehen weltweit zahlreiche KI-Leitlinien in öffentlichen Organisationen, Unternehmen und Forschungseinrichtungen, wie etwa die bereits in 2019 veröffentlichten „Ethics Guidelines for Trustworthy AI" (Europäische Kommission 2019). Die aktuellere Debatte konzentriert sich dabei auf die Definition ethischer und technischer Standards sowie den Umgang mit Konfliktsituationen, beispielsweise beim medizinischen KI-Einsatz (Buruk et al. 2020, Solanki et al. 2023), was ein stringentes Monitoring von KI-Algorithmen und deren Anwendung erfordert (Weiner et al. 2025). Hierbei offenbaren sich erhebliche Probleme: Die Heterogenität der bestehenden Leitlinien führt zu uneinheitlichen Interpretationen der oft genannten Kernprinzipien – darunter Transparenz, Gerechtigkeit, Verantwortung, Sicherheit, Nachhaltigkeit – und verhindert ein einheitliches, globales Verständnis von ihrer Umsetzung. Insbesondere das Prinzip der Transparenz ist vielschichtig definiert. Manche Quellen betonen die Erklärbarkeit und das Vertrauen, andere legen den Fokus auf partizipative Dialoge und demokratische Prinzipien. Ebenso variieren die geforderten Offenlegungsinhalte, sei es der Quellcode, die Evidenzbasis, gesetzliche Rahmenbedingungen oder die erwarteten gesellschaftlichen Auswirkungen (Jobin et al. 2019). Auch der Deutsche Ethikrat konstatiert einen kontroversen Diskurs rund um die Fragen, wie ethische Prinzipien zu priorisieren und welche Betroffenengruppen in den Fokus zu nehmen sind (Deutscher Ethikrat 2023).

Vor diesem Hintergrund besteht ein dringender Handlungsbedarf, der sich vor allem aus den schwerwiegenden Herausforderungen, Risiken und potentiellen Schäden, die von KI-Systemen ausgehen, ergibt (Gunasekara et al. 2025). Forderungen befassen sich mit der Schaffung einheitlicher Terminologien und Haftungsregeln, der Bildung interdisziplinärer Teams in Forschung und Entwicklung, der Schaffung robuster Frameworks zur Evaluation von KI-Systemen sowie zur Offenlegung und Darstellung relevanter Informationen (beispielsweise Daten, Modelle und Limitationen) sowie den Ausbau von Bildungs- und Aufklärungsaktivitäten, um Fachpersonal sowie Patientinnen und Patienten in die Lage zu versetzen, KI-Entscheidungen kritisch zu hinterfragen und informierte Entscheidungen zu treffen. Zugleich kann der ethische Umgang mit KI ein Wettbewerbsvorteil für Entwickler und Anwender sein, indem dieser zur Erfüllung unternehmerischer Pflichten gegenüber der Gesellschaft beiträgt (Stichwort Corporate Social Responsibility).

Responsible AI und seine Rolle für die Medizintechnik

Der Begriff Responsible Artificial Intelligence (RAI) beschreibt ein Bündel von Prinzipien, die ein ethisch vertretbares, transparentes und rechenschaftspflichtiges Vorgehen bei der Nutzung von KI-Technologien sicherstellen. Diese Prinzipien sollen mit den Erwartungen der Nutzenden, den Werten der Organisationen sowie den Gesetzen und Normen der Gesellschaft im Einklang stehen und über den gesamten Lebenszyklus eines Systems erfüllt sein (Mikalef et al. 2022). Der Ansatz umfasst sowohl die Untersuchung potenzieller, beabsichtigter als auch unbeabsichtigter Konsequenzen, die aus dem Einsatz von KI resultieren können, als auch die Entwicklung von Maßnahmen, die darauf abzielen, negative Folgen zu vermeiden oder zu mildern. Im Gesundheitssektor gewinnt RAI besondere Relevanz, da KI-gestützte Analysen von umfangreichen Patientendaten perspektivisch Diagnosen, klinische Entscheidungsfindung und personalisierte Therapien unterstützen können. Ein verantwortungsvoller Umgang mit diesen Technologien kann somit einen unmittelbaren, positiven Beitrag zum Wohlbefinden aller Akteurinnen im Gesundheitswesen leisten – von Ärztinnen und Ärzten über Pflegepersonal bis hin zu den Patientinnen und Patienten selbst. Dennoch führt die Integration von KI häufig zu Entscheidungen und Handlungen, die weitreichende moralische Implikationen besitzen. Wenn algorithmische Prozesse die Rechte, die Würde oder die Autonomie von Individuen beeinträchtigen, geraten die zugrunde liegenden ethischen Prinzipien in Gefahr. Im Kontext der Medizin treten daher immer wieder Bedenken hinsichtlich Fairness, Verantwortlichkeit, Menschenrechten und dem gleichberechtigten Zugang zu Gesundheitsleistungen auf. Diese Bedenken lassen sich in epistemische und normative Typen ethischer Herausforderungen klassifizieren (Mittelstadt et al. 2016, Morley et al. 2025):

Typus	Ethische Bedenken	Erklärung
epistemisch	inconclusive evidence (unvollständige Evidenz)	KI- Ergebnisse sind probabilistisch und können keinen kausalen Zusammenhang begründen
	inscrutable evidence (unverständliche Evidenz)	Betroffene haben häufig keinen vollständigen Überblick über Trainingsdaten, auf deren Grundlage eine bestimmte Entscheidung getroffen wurde
	misguided evidence (fehlgeleitete Evidenz)	KI-Ergebnisse können nur so zuverlässig sein wie die zugrundeliegenden Daten
normativ	unfair outcomes (ungerechte Resultate)	KI-Ergebnisse können sich auf eine bestimmte Gruppe von Personen im besonderen Maße (positiv oder negativ) auswirken
	transformative effects (transformative Wirkungen)	KI-Algorithmen verändern die Realität in unvorhergesehener Weise (Stichwort Profiling)
weitere	traceability (Zurückverfolgbarkeit)	KI-Algorithmen können Schäden verursachen, die schwer zu beheben sind, da die Ursachen und Verantwortlichkeiten unklar sind

Die epistemischen oder wissensbezogenen Bedenken betreffen die Qualität, Verfügbarkeit und Interpretierbarkeit der Evidenzbasis, auf der KI-Entscheidungen fußen. Beispiele finden sich bei dem mittlerweile eingestellten System IBM Watson Oncology. Dieses traf als Unterstützungssystem für die Onkologie mitunter falsche Krebstherapieempfehlungen aufgrund von schlechter beziehungsweise uneinheitlicher Datenqualität (Kristiansen et al. 2022). Zudem zeigte sich, dass das Modell aufgrund des Trainings auf vorrangig westliche Populationen bei der Anwendung in China deutlich schlechtere Ergebnisse lieferte (Morley et al. 2025). Weitere Studien zeigen, dass KI-Modelle für die Analyse von Röntgenbildern, die auf Daten aus einem bestimmten Krankhaus(-verbund) trainiert wurden, deutlich schlechter bei der Anwendung in anderen Krankenhäusern funktionieren können. Dies liegt laut den Autoren daran, dass beim Training krankenhausspezifische Eigenschaften aufgenommen werden (beispielsweise die verwendete Hardware), deren Ausprägung in anderen Settings variiert (Zech et al. 2018, Lasko et al. 2024). Dies schränkt den Nutzen solcher Modelle stark ein und – wenn nicht detektiert – kann zu schlechteren und gegebenenfalls schädlichen Diagnosen in der breiteren Anwendung führen.

Die normativen Bedenken hingegen beziehen sich auf die Bewertung von Handlungen und Effekten hinsichtlich Fairness, Gerechtigkeit und den grundsätzlichen moralischen Implikationen, die aus dem Einsatz von KI-Algorithmen resultieren. Ein beispielhaftes Szenario bildet die Manipulation der Patientendaten innerhalb eines Laborinformationssystems mit dem Ziel, die Priorisierung bei der Organspende zugunsten einer Person zu verändern (Beyerer et al. 2022). Durch die manipulierten Daten berechnet das KI-Modell eine hohe Kompatibilität zum Spenderorgan, was Menschen mit höherem Bedarf benachteiligt oder gar gefährdet. Ein weiteres Beispiel liegt in der Nutzung von Gesundheits- und Fitness-Apps. Anwendenden ist nur begrenzt bewusst, welche Daten über sie erfasst werden und wie diese in Empfehlungen oder Warnungen durch die KI münden. Dies schränkt die Autonomie sowie Möglichkeit ein, Vorschläge zu hinterfragen (Morley et al. 2025).

Daneben bildet die Zurückverfolgbarkeit einen wichtigen Anspruch, um die Ursachen negativer (epistemischer und normativer) Outcomes zu identifizieren. Dies umfasst auch die Schwierigkeit, KI-Entscheidungen mit negativen Konsequenzen einer konkreten Quelle im Sinne der Verantwortlichkeit und/oder Haftung zuzuordnen. Die Transparenz von Daten, Algorithmen und Ergebnissen stellt ein wesentliches Anliegen von RAI dar. Hierzu werden Mechanismen und Systeme erforscht, welche medizinische Entscheidungsunterstützung nachvollziehbar und rückverfolgbar machen („traceable reasoning"). Eine aktuelle Studie findet sich im Bereich der Diagnose seltener Erkrankungen. Die Autoren haben ein System entwickelt und validiert, welches diagnostische Outcomes mit Argumentationsketten versieht, die auf medizinisch verifizierbare Evidenz verweist (Zhao et al. 2026).

Die klare Trennung dieser Ebenen ist entscheidend, um gezielte Strategien zu entwickeln, die sowohl die technische Robustheit als auch die ethische Legitimität von KI-Anwendungen im Gesundheitswesen sicherstellen (Trocin et al. 2023). Durch die konsequente Umsetzung von RAI-Prinzipien können diese epistemischen und normativen Spannungsfelder adressiert werden. Transparente Modellarchitekturen und nachvollziehbare Daten-Governance erhöhen die Nachvollziehbarkeit von Entscheidungen, während strenge Fairness-Kontrollen und die Gewährleistung menschlicher Aufsicht dazu beitragen, ungerechte Ergebnisse und potenziell transformative, aber unbeabsichtigte Auswirkungen zu verhindern. Nur wenn sowohl die methodischen als auch die moralischen Anforderungen gleichermaßen berücksichtigt werden, kann KI im Gesundheitssektor ihr volles Potenzial entfal-

ten. Verantwortung in der Entwicklung medizinischer KI-Systeme ist entscheidend, um das Vertrauen der Nutzenden zu gewinnen und potenzielle Schäden zu minimieren (Gunasekara et al. 2025).

Es wird deutlich, dass die Implementierung von RAI im Gesundheitssektor nicht allein eine ethische Wunschvorstellung, sondern eine strukturelle Notwendigkeit mitunter zur Sicherung etablierter (medizin-)ethischer Prinzipien und Normen wie dem Nicht-Schaden und der informierten Zustimmung darstellt. Die Europäische Kommission hat in den Jahren 2020 und 2021 Dokumente veröffent-licht, die eine menschenzentrierte, vertrauenswürdige und sichere KI fördern sollen, die im Einklang mit den europäischen Werten steht (Europäische Kommission 2021). Ziel ist es, eine rechtliche Infrastruktur für KI zu schaffen, die Risiken minimiert und gleichzeitig die Vorteile maximiert.
Auch in Deutschland sind nationale Anstrengungen im Bereich der Entwicklung vertrauenswürdiger und moralisch ausgerichteter KI-Systeme zu erkennen. So verfolgt beispielsweise die Nationale Initiative für Künstliche Intelligenz und Datenökonomie im Rahmen der Digitalstrategie der Bundesregierung das Ziel, Qualitätsstandards für KI zu entwickeln und anzuwenden. Die Qualitätsbewertung sieht hier eine Schutzbedarfsanalyse vor, die potenzielle Schäden wie Gesundheitsrisiken oder Diskriminierung durch KI antizipiert (Stichwort Technikfolgenabschätzung). Auch die KI-Strategie der Bundesregierung greift RAI-Grundlagen auf, indem KI in Deutschland verantwortungsvoll und menschenzentriert entstehen und eingesetzt werden soll. Hierbei soll ein regulatorischer Ordnungsrahmen geschaffen werden, um dem Schutz vor Bias, Diskriminierung und Missbrauch Rechnung zu tragen.

Trotz dieser Fortschritte bestehen nach wie vor erhebliche Barrieren: vage Leitlinien, Ressourcen- und Kompetenzdefizite, Rollenkonflikte zwischen klinischem und technischem Personal sowie unklare Haftungsregeln. Initiativen, ob national oder international, befinden sich häufig noch im Anfangsstadium.

Von der Theorie in die Praxis: Der Principle-to-Practice-Gap

Im Bereich der Medizintechnik und Gesundheitstechnologien gibt es inzwischen mehrere Beispiele, die zeigen, wie sich RAI-Prinzipien zumindest teilweise erfolgreich in die Praxis überführen lassen. Diese Ansätze sind deshalb interessant, weil sie typische Elemente des sogenannten Principle-to-Practice-Gaps adressieren – also die Übersetzung abstrakter ethischer Leitlinien wie Transparenz, Fairness oder Verantwortlichkeit in konkrete technische, organisatorische und regulatorische Strukturen.

Grundsätzlich können drei verschiede Arten von Ansätzen unterschieden werden, die sich auf die Überwindung des Principle-to-Practice-Gaps beziehen: 1) Technische Ansätze, 2) organisatorische Ansätze und 3) regulatorische Ansätze.

Technische Ansätze beziehen sich auf die Überwindung des Principle-to-Practice-Gaps durch die Entwicklung und Implementierung entsprechender Software, Tests und technischer Prinzipien. **Explainable AI (XAI)** ist dabei ein wesentlicher technischer Ansatz in der Medizintechnik, da Vertrauen in KI-gestützte Entscheidungsunterstützungssysteme nur dann entsteht, wenn klinische Anwendende die Ergebnisse nachvollziehen und kritisch bewerten können (Amann et al. 2020). Es geht dabei weniger um vollumfängliche Modelltransparenz, sondern um kontextabhängige, klinisch relevante Erklärungen, die eine valide Entscheidungsunterstützung ermöglichen (Schopp et al. 2025). Für einen nachhaltigen Beitrag von XAI zur klinischen Praxis ist es notwendig, dass sie in umfassende Qualitäts- und Sicherheitsstrukturen integriert wird. Zusätzlich sind **Robustheits- und Bias-Tests** notwendig, um zu gewährleisten, dass Modelle über unterschiedliche Patientengruppen, klinische Szenarien und Datenquellen hinweg konsistent und gerecht funktionieren. Viele Untersuchungen zeigen, dass KI-Systeme außerhalb ihrer Trainingspopulation erhebliche Leistungseinbußen aufweisen können (Li et al. 2026, Oakden-Rayner et al. 2020). Um unbeabsichtigte Diskriminierungen entlang ethnischer, geschlechtlicher oder sozioökonomischer Merkmale zu erkennen und abzuschwächen, sind systematische Bias-Analysen von zentraler Bedeutung, da medizinische Algorithmen Bias reproduzieren können (Obermeyer et al. 2019). Zugleich ist die Sicherstellung von Datenschutzprinzipien im Zusammenhang mit der Entwicklung medizintechnischer KI-Systeme von großer Bedeutung. Auch diese kann durch entsprechende Daten- und Modellpraktiken gewährleistet werden. **Federated Learning** ist hier ein vielversprechender Ansatz, um die scheinbar widersprüchlichen Anforderungen von Datenschutz, Datensouveränität, algorithmischer Fairness Governance und gleichberechtigtem Zugang miteinander zu vereinbaren (Mir et al. 2025). Diese Methodik sieht ein dezentrales Training der Modelle vor, während sensible Patientendaten lokal in den jeweiligen Institutionen verbleiben (Rieke et al. 2020). Durch diesen Ansatz können zugleich Datenschutzanforderungen und der Zugang zu ausreichend großen und diversen Datensätzen sichergestellt werden. In empirischen Untersuchungen konnte gezeigt werden, dass Ansätze des Federated Learning eine vergleichbare Modellperformanz wie klassische Modelle erreichen können und gleichzeitig die rechtlichen und ethischen Herausforderungen bei der Datennutzung verringern, insbesondere in streng regulierten klinischen Umgebungen (Sheller et al. 2020).

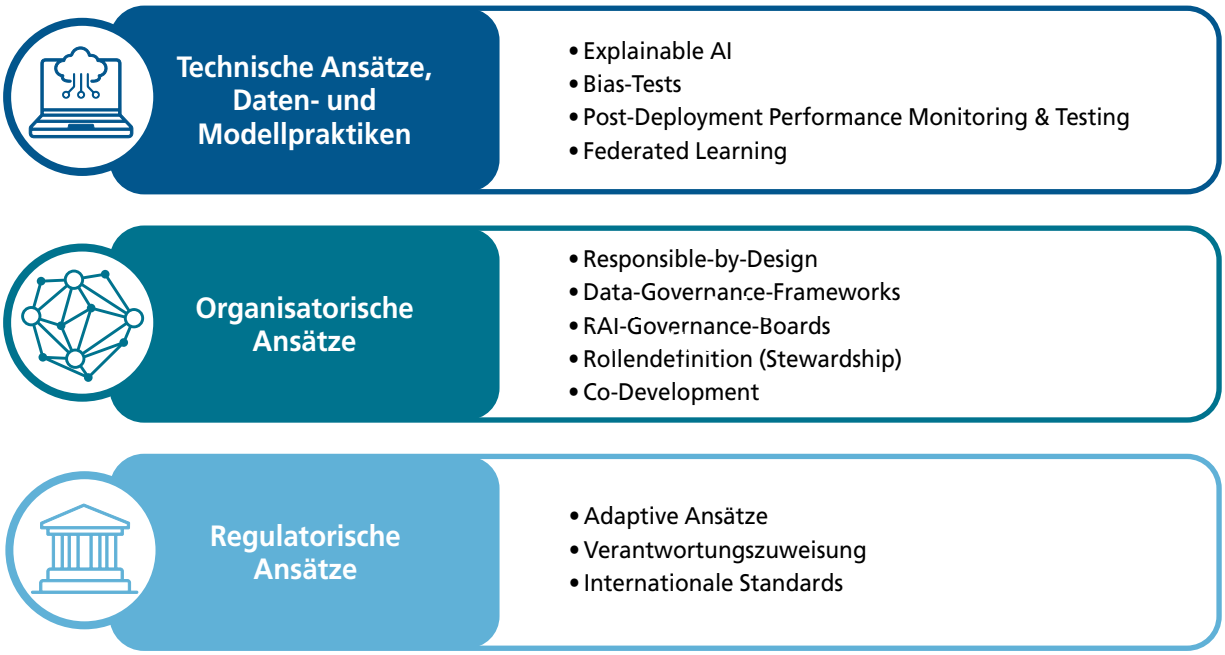


Abbildung: Kategorien und Ansätze zur Überwindung des Principle-to-Practice-Gap

Organisatorische Ansätze tragen zur Überwindung des Principle-to-Practice-Gaps bei, indem sie sicherstellen, dass ethische Prinzipien systematisch in Entscheidungs- und Entwicklungsprozesse integriert werden. Im Bereich Medizintechnik kann dies über den **„Responsible-by-Design"-Ansatz** gelingen. Dieser legt fest, dass RAI-Prinzipien schon während der Konzeptions- und Designphase mitgedacht werden (Saikia & Kumar, 2026). Mit Hilfe strukturierter Dokumentationsinstrumente wie Model Cards oder Datasheets for Datasets werden Annahmen, Einsatzgrenzen, Trainingsdaten und potenzielle Risiken offengelegt, was zu mehr Transparenz und Rechenschaftspflicht führt (Gebru et al. 2018, Mitchell et al. 2019). Hierzu ist es zielführend, **Data-Governance-Frameworks** zu etablieren. Diese Frameworks dienen dazu, die Herkunft, Qualität, Annotation und Verwendung medizinischer Daten transparent und reproduzierbar festzuhalten. Eine systematische Data Governance wird immer mehr nicht nur als technische Notwendigkeit, sondern als grundlegende Voraussetzung für Fairness, Transparenz und Datenkontrolle angesehen (Khan & Khan, 2025). Dabei können interdisziplinäre **RAI-Governance-Boards oder Committees** konstituiert werden, welche Expertise aus Medizin, KI-Entwicklung, Ethik, Recht und Qualitätsmanagement vereinen und eine holistische Bewertung von Risiken über den gesamten Lebenszyklus medizinischer KI-Systeme hinweg ermöglichen (El Arab et al. 2025). Durch diese Governance-Strukturen wird die interdisziplinäre Zusammenarbeit gestärkt, das Denken und Arbeiten innerhalb von Disziplingrenzen aufgebrochen und verhindert, dass ethische Fragestellung erst nach Implementierung der Technologien drängend werden, wenn eine nachträgliche Änderung nur noch schwer möglich ist (Morley et al. 2020). Die Etablierung von klar definierten **Rollen und Zuständigkeiten (Stewardship-Ansatz)**, trägt hier zusätzlich zu transparenten Verantwortungsstrukturen bei und unterstützt den verantwortungsvollen Einsatz lernender Systeme im klinischen Umfeld (Kumar et al. 2025). Außerdem können die Anwendenden von Medizintechnik während der Entwicklung über **klinische Co-Development-Prozesse** durchgehend in die Ausgestaltung und Evaluation der Technologien eingebunden werden, um eine Lücke zwischen technischer Leistungsfähigkeit und klinischem Nutzen zu verhindern. Medizintechnik wird so passend für tatsächliche Workflows gestaltet und Fehlanwendungen in der medizinischen Praxis werden reduziert (Grootjans et al. 2025). Zusammengefasst lässt sich also sagen, dass die frühzeitige organisatorische Integration ethischer Richtlinien neben der Reduktion regulatorischer Risiken auch die klinische Akzeptanz erhöhen kann, da die Systeme dadurch eindeutige Bestimmungen und Limitierungen in Bezug auf ihre Funktionen und deren Folgen sowie klare Verantwortlichkeiten aufweisen (Cary et al. 2026). Aufgrund der dynamischen Beschaffenheit klinischer Umfelder ist darüber hinaus ein fortlaufendes technisches **Post-Deployment Performance Monitoring und Testing** notwendig, das sich an Veränderungen von Daten, Nutzungskontext oder klinischer Praxis orientiert und somit Leistungseinbu-

ßen oder sicherheitsrelevante Abweichungen beim Betrieb der KI frühzeitig erkennen kann (Dolin et al. 2025, Kelly et al. 2019).

Regulatorische Ansätze haben die Einführung bestimmter Regulierungspraktiken zur Überwindung des Principle-to-Practice-Gaps zum Ziel. Traditionelle Regulierungsmodelle in der Medizintechnik sind vor allem für statische Systeme konzipiert und erreichen angesichts von lernfähigen KI-Anwendungen zunehmend ihre Grenzen. Nach den Vorschlägen der US-amerikanischen Food and Drug Administration (FDA) sind **adaptive regulatorische Ansätze** möglich, die kontrollierte Modellaktualisierungen nach der Zertifizierung erlauben, vorausgesetzt, dass Änderungsrahmen, Validierungsstrategien und Risikokontrollen im Voraus festgelegt werden (Food and Drug Administration 2021; Lutze et al. 2025). Diese Konzepte zielen darauf ab, Innovationen zu ermöglichen, ohne Sicherheit oder Verantwortlichkeit zu gefährden (siehe auch Lutze & Krieger 2025, Lutze et al. 2025). Die eindeutige **Zuweisung von Verantwortung** zwischen Herstellern, Betreibern und klinischen Anwendenden bleibt jedoch weiterhin eine zentrale Herausforderung. Die Implementierung wird durch unklare Haftungsfragen erschwert, was defensives Nutzungsverhalten oder sogar Ablehnung zur Folge haben kann (Banozic-Tang & Koh, 2026). Ein weiterer regulatorischer Aspekt für die praktische Umsetzung verantwortungsvoller KI in der Medizintechnik ist die Entwicklung und Einführung **internationaler Standards**. Initiativen wie die Global Alliance for Genomics and Health (GA4GH) und Standards wie ISO/IEC 42001 zeigen, wie strukturierte Datenpraktiken die Umsetzung von RAI unterstützen und das Vertrauen aller Stakeholder – von Entwicklern über Regulierungsbehörden bis zu Patientinnen und Patienten – stärken können (Global Alliance for Genomics and Health 2019, ISO/IEC 2023).

Von den Besten lernen: Orientierung an guter Praxis

In der Praxis finden sich zu den oben beschriebenen Ansätzen bereits einige Best Practices. Eine Form der Operationalisierung ethischer Prinzipien findet sich beispielsweise im Bereich der KI-gestützten Bilddiagnostik, etwa beim Brustkrebs-Screening durch Systeme von Google. Hier wurden Leistungskennzahlen nicht nur insgesamt, sondern differenziert nach verschiedenen Subpopulationen wie Alter, Herkunft und klinischen Parametern ausgewiesen. Sensitivität und Spezifität wurden mit der Leistung menschlicher Radiologinnen und Radiologen verglichen und in unterschiedlichen Standorten extern validiert. Das Prinzip der Fairness wird dadurch messbar gemacht: Anstelle allgemeiner Bekenntnisse treten konkrete Subgruppenanalysen, die mögliche Verzerrungen sichtbar machen. Zwar bleiben Fragen der globalen Repräsentativität weiterhin bestehen, doch wird zumindest methodisch anerkannt und umgesetzt, dass gerechte Versorgung empirisch überprüfbar sein muss (Kelly et al. 2026).

Ein weiteres wichtiges Element verantwortungsvoller Praxis findet sich auf organisatorischer Ebene, in dem das AI-Governance-Framework strukturell und nicht nur projektbezogen verankert wird. Dazu gehören interne Ethics Review Boards, standardisierte Risiko- und Impact-Assessments sowie Dokumentations- und Monitoringpflichten über den gesamten Produktlebenszyklus hinweg. RAI wird damit Teil des Qualitätsmanagements und nicht nur ein kommunikatives Leitbild. Ähnlich verfolgt die Mayo Clinic Platform einen Ansatz, bei dem interdisziplinäre Teams aus Medizinerinnen, Datenwissenschaftlern, Ethikern und Juristen KI-Systeme vor breiter Implementierung validieren und in realen Versorgungsumgebungen testen (Makhni et al. 2025). Hier zeigt sich, dass der Principle-to-Practice-Gap häufig weniger ein Mangel an Prinzipien ist als ein Problem fehlender institutioneller Einbettung, der Einübung kollaborativer Methoden (Spindler 2025) und einer Harmonisierung von Wert- und Strategieentscheidungen (Mäntymäki et al. 2023).

Auch regulatorisch gibt es Ansätze, die als Best Practice oder Orientierungsgeber gelten können. Die FDA hat mit dem Konzept eines Predetermined Change Control Plan einen Mechanismus eingeführt, der speziell auf lernende KI-Systeme zugeschnitten ist. Hersteller müssen bereits vor der Zulassung definieren, welche Modelländerungen unter welchen Bedingungen zulässig sind und wie diese überwacht werden. Damit wird das Problem adressiert, dass sich KI-Systeme dynamisch weiterentwickeln können und klassische statische Zulassungsverfahren an ihre Grenzen stoßen. Verantwortlichkeit bleibt so auch bei adaptiven Systemen nachvollziehbar geregelt und es wird weiteres Wissen zum Einsatz von KI-Systemen erzeugt, dass einen wichtigen Beitrag zu RAI leisten kann.

Schließlich zeigt sich in vielen aktuellen medizinischen KI-Systemen ein bewusster Verzicht auf vollständige Automatisierung zugunsten sogenannter Human-in-the-Loop-Modelle. Systeme wie PathChat oder PathAI werden als Entscheidungsunterstützung für Pathologinnen und Pathologen eingesetzt, nicht als Ersatz für ärztliche Expertise (Lu et al. 2024). Die finale Entscheidung verbleibt beim medizinischen Fachpersonal, häufig ergänzt durch Feedbackmechanismen, die zur kontinuierlichen Verbesserung des Systems beitragen. Auf diese Weise werden die Prinzipien der Autonomie, Verantwortlichkeit und des Nicht-Schadens technisch und organisatorisch abgesichert.

Zusammenfassend lässt sich feststellen, dass sich erfolgreiche RAI-Praxis in der Medizintechnik durch einige wie-

derkehrende Merkmale auszeichnet: eine präzise Zweckdefinition, differenzierte Leistungs- und Bias-Analysen, institutionalisierte Governance-Strukturen sowie ein konsequentes Lifecycle-Monitoring einschließlich Post-Market-Überwachung. Auch wenn in allen Bereichen weiter Forschungsbedarf besteht, zeigen diese Ansätze doch, dass der Principle-to-Practice-Gap verkleinert werden kann, wenn ethische Prinzipien systematisch operationalisiert, organisatorisch verankert und regulatorisch begleitet werden.

Dabei ist die Partizipation aller relevanten Stakeholder - von der Entwicklung bis hin zur Anwendung - entscheidend, um RAI in der Medizintechnik erfolgreich zu verankern (Kallina & Singh 2024). Interdisziplinäre Teams bringen technisches Know-how ein, prüfen die klinische Relevanz, Sicherheit und den Nutzen im Versorgungsalltag und tragen Perspektiven zu Akzeptanz, Vertrauensbildung und individuellen Bedürfnissen bei (Hine & Barnaghi 2024). Durch kontinuierliche, interdisziplinäre Dialoge und gemeinsame Entscheidungs- und Evaluationsprozesse können potenzielle Biases frühzeitig erkannt, ethische Leitlinien praxisnah interpretiert und Umsetzungslücken geschlossen werden. Dieser kollaborative Ansatz fördert nicht nur die technische Robustheit, sondern stärkt auch die gesellschaftliche Legitimation und das Vertrauen in KI-gestützte Gesundheitslösungen.

Herausgeber:
VDI/VDE Innovation + Technik GmbH
Steinplatz 1 | 10623 Berlin
www.vdivde-it.de

Bildnachweis:
elenabs/istockphoto

VDI|VDE|IT