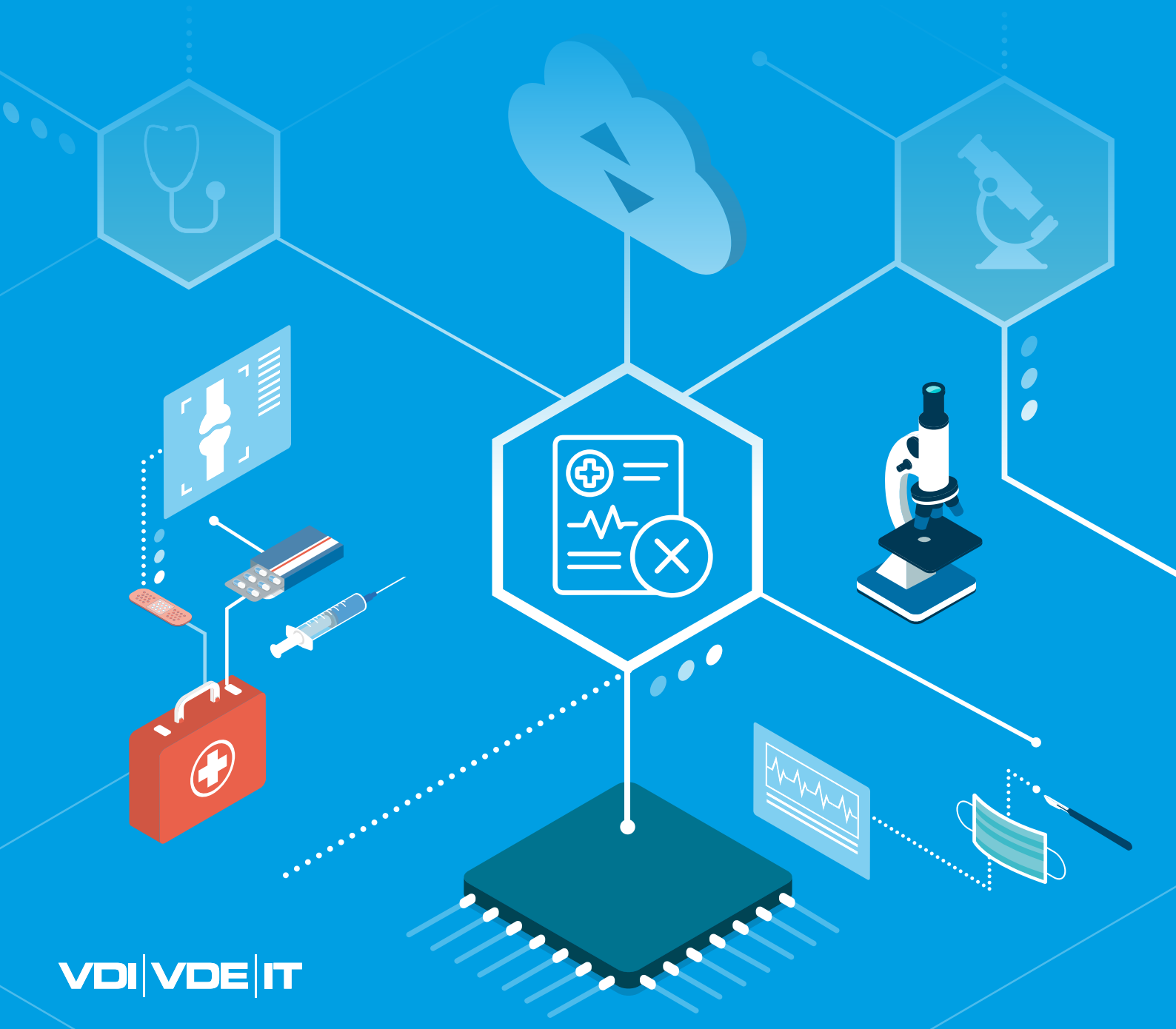
Automation Bias: Wenn der Mensch die KI nicht mehr kontrolliert

Katja Karrer-Gauß, Jens Apel, Marius Müller, Patrick Ehrenbrink



VDI|VDE|IT

Automation Bias:
Wenn der Mensch die KI nicht mehr kontrolliert

KI-basierte Systeme sind in der Medizintechnik längst etabliert, insbesondere in der Diagnostik. Ein zentrales, gut belegtes Risiko bleibt jedoch oft unbeachtet: der Automation Bias. Selbst erfahrene Ärztinnen und Ärzte verlassen sich unter Zeitdruck oder bei komplexen Befunden unkritisch auf KI-Empfehlungen, selbst wenn diese ungenau, unzuverlässig oder unvollständig sind.

Der Beitrag erläutert die psychologischen Mechanismen in der Zusammenarbeit mit KI-basierten Systemen, zeigt empirische Evidenz aus 25 Jahren Human-Factors-Forschung und diskutiert, warum Ansätze wie **Explainable AI (XAI)** auch kontraproduktiv sein können. Abschließend werden konkrete Design- und Organisationsprinzipien für sichere Mensch-KI-Teams in der Medizintechnik formuliert.

Das Versprechen und der blinde Fleck

Der rasante Vormarsch KI-gestützter Systeme im Bereich der Medizintechnik verspricht, Diagnosen schneller und präziser zu stellen, Therapien besser individuell abzustimmen, Ressourcen effizienter zu nutzen und menschliche Fehler zu reduzieren. Mensch und Maschine können so zu einem wirkungsvollen Team werden (siehe dazu auch Lutze et al., 2025). Kliniken nutzen KI-basierte Systeme bereits heute zur Radiologie-Befundung, zur Frühdiagnostik, zum Alarmmanagement auf Intensivstationen, für automatisierten Triage-Empfehlungen oder zur Unterstützung bei Laboranalysen (z. B. Müller, 2025; Lorenz, 2025; Hadweh et al., 2025; Laurent, 2025). Auch in Medizinprodukten – von Wearables bis zu implantierbaren Geräten – findet sich zunehmend Software, die Muster erkennt, Risiken prognostiziert oder (teil-)autonome Entscheidungen vorbereitet (Smits Serena et al., 2025).

Doch hinter dem Versprechen verbirgt sich ein Risiko, das leicht übersehen wird: Der sogenannte Automation Bias beschreibt die Tendenz, dass Ärztinnen und Ärzte sowie Klinikpersonal automatisierte Empfehlungen von Entscheidungsunterstützungssystemen unkritisch übernehmen. Dieser psychologische Effekt kann drastische Folgen haben, wenn die Empfehlungen ungenau, unzuverlässig oder gänzlich falsch sind.

Ein praktisches Beispiel liefert ein Forschungsprojekt aus der Radiologie: Insbesondere unerfahrene Radiologinnen und Radiologen neigten in einer Studie eher dazu, fälschlicherweise ein zerebrales Aneurysma anzunehmen, wenn die KI Ihnen das zuvor nahegelegt hatte. Das kann zu unnötigen Eingriffen, höheren Kosten und Patientenangst führen (Kim et al., 2025). In einer weiteren Studie (Rosbach et al., 2024) revidierten geschulte Pathologieexpertinnen und -experten bei der Diagnose von Tumorzellen in 7 % der Fälle sogar ihre korrekte Bewertung und folgten einer fehlerhaften KI-Empfehlung.

Automation Bias und die Illusion von Verlässlichkeit

Automation Bias beschreibt die kognitive Verzerrung, bei der Menschen den Entscheidungen oder Empfehlungen eines automatisierten Systems zu viel Vertrauen schenken, unabhängig davon, ob sie korrekt sind. Dies passiert insbesondere, wenn das System in der Vergangenheit zuverlässig war (Goddard et al., 2011; Parasuraman & Manzey, 2010; Lyell, 2017).

Dies kann zum einen in Form von Kommissionsfehlern („Errors of Commission“) auftreten, das heißt, Menschen folgen einer fehlerhaften Empfehlung der KI blind. Zum anderen können Auslassungsfehler („Errors of Omission“) erfolgen, bei denen Menschen eigene Überprüfungen unterlassen, weil sie davon ausgehen, dass das System zuverlässig arbeitet. Der Unterschied liegt darin, dass bei Kommissionsfehlern eine falsche Handlung aufgrund einer Empfehlung **ausgeführt** wird, während bei Auslassungsfehlern eine notwendige Handlung **ausbleibt**, weil das System keinen Hinweis gibt.

Nach wie vor wird gemeinhin davon ausgegangen, dass Entscheidungsunterstützungssysteme die Unzuverlässigkeit des Menschen kompensieren und zu rationaleren Entscheidungen beitragen können. Dabei zeigen empirische Studien seit gut 25 Jahren, dass das Zusammenspiel von automatisierter Entscheidungsunterstützung und menschlichen Entscheidungsträgern nicht so ideal ist wie erwartet (Mosier & Manzey, 2019). Das Problem entsteht immer dann, wenn automatisierte Empfehlungen inkorrekt sind. Dies ist in der Regel selten, da Entscheidungsunterstützungssysteme, insbesondere moderne KI-basierte Systeme, typischerweise eine hohe oder zumindest als hoch wahrgenommene Trefferquote aufweisen. Gerade diese hohe Zuverlässigkeit führt jedoch dazu, dass Nutzende Empfehlungen zunehmend unkritisch übernehmen. In der Kon-

sequenz folgt der Mensch der Systemempfehlung häufig auch dann, wenn diese im Einzelfall fehlerhaft ist, statt die verfügbaren Informationen erneut eigenständig zu prüfen. Mit dem Leistungsversprechen wird dieser blinde Fleck also ausgeprägter: Je zuverlässiger KI wird, desto eher verlässt sich der Mensch blind auf sie, auch im Fehlerfall. Automation Bias entsteht also nicht trotz hoher KI-Zuverlässigkeit, sondern gerade wegen ihr.

Warum Automation Bias für Medizintechnik von Bedeutung ist

Im Gesundheitswesen ist Automation Bias besonders relevant, weil Entscheidungen komplex, zeitkritisch und folgenschwer sind. Sie haben unmittelbare Auswirkungen auf die Patientensicherheit.

Ein systematisches Review aus dem Gesundheitswesen zeigte bereits 2011, dass klinische Entscheidungssysteme (Clinical Decision Support Systems, CDSS) insgesamt die Performance und Effizienz verbessern können, ihre Nutzung aber zu neuen Fehlerquellen führt (Goddard et al., 2011). Metaanalysen zeigten, dass in etwa 6 bis 11 % der Fälle das medizinische Fachpersonal einer Fehlempfehlungen des CDSS folgt und eine vorherige richtige eigene revidiert (Goddard et al., 2011; Mosier & Manzey, 2019). Aber auch Auslassungsfehler werden berichtet: So sank die menschliche Detektionsrate von kritischen Bereichen einer Mammographie, wenn ein Unterstützungssystem eingesetzt wurde, welches fälschlicherweise keinen Hinweis gab (Alberdi et al., 2004).

Zudem zeigen neuere Untersuchungen, dass die Wahrscheinlichkeit, Empfehlungen eines CDSS blind zu folgen, mit der wahrgenommenen Kompetenz des Systems steigt. Mit zunehmender Nutzung und scheinbar erfolgreichen Prognosen neigen Fachkräfte dazu, eigene Zweifel zu unterdrücken, was sowohl zu Kommissionsfehlern als auch zu Auslassungsfehlern führen kann (Gaube et al., 2021).

Auch aktuellere Studien zeigen, dass Mensch-KI-Teams nicht unbedingt die erhoffte bessere Leistung erzielen als ein Mediziner beziehungsweise eine Medizinerin oder die KI für sich allein genommen (Lutze et al., 2025). Die Leistung eines KI-basierten Diagnosesystems kann die von menschlichen Experten übertreffen (z. B. McKinney et al., 2020; Ehteshami Bejnordi et al, 2017; Lang et al, 2023; Haenssle et al. 2018). In der Medizin ist allerdings immer eine menschliche Fachperson verantwortlich, Entscheidungen werden in letzter Instanz von Menschen getroffen. Dies ist aus medizinethischer Sicht, aus haftungsrechtlichen Gründen und nicht zuletzt aufgrund des regulatorischen Rahmens (Verordnung (EU) 2024/1689) auch richtig: Der Mensch lässt sich aus der Gleichung nicht streichen.

Eine aktuelle Studie von Dratsch et. al. (2023) zeigt für Mammografie-Diagnosen, dass Radiologinnen und Radiologen, die von einem KI-System unterstützt werden, bei falschen Vorschlägen deutlich häufiger Fehler machen, unabhängig von ihrer Erfahrung. Dies unterstreicht, dass Automation Bias nicht nur ein Problem von Laien ist, sondern selbst Fachpersonal in der Medizintechnik betrifft. Der Effekt ist auch nicht nur ein Problem der kognitiven Beanspruchung des Menschen. Er hängt zwar mit ihr zusammen, tritt allerdings auch bei einfachen Entscheidungsaufgaben und nicht nur in komplexen Multitasking-Umgebungen auf (Lyell & Coiera, 2017).

Gerade in der Medizintechnik ist das Risiko des Automation Bias besonders brisant: Im Einzelfall kann das vorschnelle Vertrauen auf eine KI-basierte Empfehlung zu gravierenden Fehlentscheidungen führen. Damit steht nicht nur die technische Entwicklung KI-basierter Medizintechnik im Fokus, sondern auch die psychologische und organisationale Gestaltung eines sicheren Mensch-KI-Teams.

Psychologische Mechanismen

Es gibt verschiedene Gründe dafür, warum es Menschen so schwerfällt, automatisierte Empfehlungen kritisch zu hinterfragen. Die psychologische Forschung identifiziert mehrere, teils überlappende Mechanismen:

- **Complacency (Nachlässigkeit):** Vor allem unter hoher Arbeitslast oder Multitasking-Bedingungen entsteht laut Parasuraman und Manzey (2010) die Tendenz, sich zu stark auf automatische Empfehlungen oder Diagnosen zu verlassen und dem technischen System zu sehr zu vertrauen. Der Mensch reduziert die eigene Wachsamkeit und übersieht mögliche Fehler. Das automatische System übernimmt.

- **„Cognitive Miser" („Denkfaulheit"):** Menschen tendieren dazu, den Weg des geringsten kognitiven Aufwands zu wählen. Das Gehirn spart Energie, indem es schnelle, heuristische Entscheidungen trifft. Das Prüfen von KI-Ergebnissen kostet mentalen Aufwand; das blinde Akzeptieren geht schneller. Die Tendenz, geistige Arbeit an KI-basierte Systeme auszulagern, beschreibt Gerlich (2025) als Cognitive Offloading.

- **Verantwortungsdiffusion:** Wenn Entscheidungen auf einer KI-Empfehlung basieren, fühlen sich Menschen weniger direkt verantwortlich, besonders wenn die Konsequenzen der Entscheidung negativ sind. Es ist einfacher, die Schuld dem System zuzuweisen, als sich selbst (Dong & Bocian, 2024; Brand Science Institute, 2025).

- **Confidence as Competence (Kompetenzillusion):** KI-Systeme präsentieren Vorschläge oft selbstbewusst (Sun et al., 2025). Menschen interpretieren diese Sicherheit automatisch als Zeichen von Richtigkeit.

- **Autoritäts- und Expert Bias:** Algorithmen wirken objektiv, neutral und mathematisch überlegen. Nutzende neigen dazu, die Kompetenz der KI höher zu bewerten als ihre eigene oder die anderer Menschen (Logg et al., 2019).

- **Reinforcement Loop:** Wenn ein KI-System wiederholt korrekte Empfehlungen liefert, verstärkt dies das Vertrauen – auch bei einer späteren falschen Empfehlung. Daher fördert eine hohe Systemzuverlässigkeit eine Überschätzung der KI (siehe dazu auch Moiser & Manzey, 2019).

Diese Mechanismen erklären, warum selbst erfahrene Ärztinnen und Ärzte anfällig für Automation Bias sind. Sich auf KI-Empfehlungen zu verlassen, wirkt weniger falsch, als eigene Fehler zu machen. Doch auch das Gegenteil verursacht Probleme:

- **Automation Aversion („Automations-Abneigung"):** Lag die KI einmal offensichtlich falsch, neigen Menschen dazu, diese dauerhaft zu ignorieren. Wenn Menschen der KI zu wenig vertrauen und sich nur noch auf sich selbst verlassen, urteilen sie damit häufig schlechter als die KI allein (Dietvorst et al., 2015).

- **Eigenes Rollenverständnis:** Wenn ein Mensch sich stets an die KI-Empfehlung hält und damit eigene Fehler vermeidet, nimmt er sich nicht mehr als verantwortlicher Entscheider wahr. Dies ist aber seine Rolle, die auch der EU AI Act (Verordnung (EU) 2024/1689) vorschreibt. Wenn Menschen KI misstrauen, greifen sie also ein, um ihrer Rolle gerecht zu werden, selbst wenn sie damit eine gerechtfertigte Empfehlung eines Systems missachten (Rieger et al., 2025).

- **Moral Distress (moralisches Belastungserleben):** Moral Distress beschreibt den emotionalen Druck, den Medizinerinnen und Mediziner verspüren, wenn sie von äußeren Zwängen davon abgehalten werden, das moralisch Richtige zu tun. Dadurch wird die moralische Integrität der Handelnden unterminiert, was zu Unzufriedenheit, verringerter Versorgungsqualität und mentalen Problem führen kann (Kherbache et al., 2021). Im Kontext von KI kann dieser Effekt verstärkt auftreten, beispielsweise wenn das Klinikmanagement die Nutzung von KI-Systemen zur effizienteren Diagnostik und Therapie fordert. Angesichts knapper zeitlicher Ressourcen kann dies dazu führen, dass Ärztinnen und Ärzte KI-Empfehlungen folgen (müssen) und dabei ihren moralischen Kompass außer Acht lassen.

Daneben spielen auch weitere moralische Wertvorstellungen und Anforderungen an medizinisches Personal eine große Rolle. Wenn Vorhersagen und Empfehlungen seitens einer KI stets präziser werden, steigt der Druck auf Ärztinnen und Ärzte zu begründen, warum sie von diesen abweichen (Chalson & Earp, 2026). Insbesondere im Gesundheitswesen, wo der Druck, das Richtige zu tun, immanent ist, stellt die wachsende Performanz von KI und die Verfügbarkeit zusätzlicher Informationen eine Herausforderung dar. Auch dies kann dazu führen, dass Medizinerinnen und Mediziner davor zurückschrecken, gegen die Empfehlung einer KI zu handeln. Gleichzeitig steigt die moralische Verpflichtung, KI-Outputs stärker in Betracht zu ziehen.

Wenn Automation Bias auf Algorithm Bias trifft

Zusätzlich zu menschlichen Verzerrungen kommt bei KI-basierten Systemen in der Medizintechnik ein weiterer, häufig übersehener Risikofaktor hinzu, der sogenannte Algorithm Bias. Gemeint sind Verzerrungen, die aus unausgewogenen Trainingsdaten, fehlerhaften Modellannahmen oder gesellschaftlichen Ungleichheiten resultieren (Starke et al., 2021,). Während Algorithm Bias die Ausgabe eines Systems verzerrt, beeinflusst Automation Bias den Umgang der Nutzerinnen und Nutzer mit dieser Ausgabe. Das Zusammenspiel beider Verzerrungen schafft eine Art doppelte Blindheit von System und Mensch.

In der medizinischen Forschung wurde wiederholt gezeigt, dass algorithmische Verzerrungen zu unterschiedlicher Diagnose, unterschiedlicher Therapie-Empfehlung oder systemischer Ungleichbehandlung verschiedener Patientengruppen führen können, etwa in Abhängigkeit von sozioökonomischer, demografischer oder ethnischer Herkunft (Omar et al., 2025; Chen et al., 2023) oder Geschlecht (Muth & Schmietow, 2026). Subtile algorithmische Verzerrungen sind kaum erkennbar, sodass klinisches Personal noch stärker dazu neigen könnte, die Empfehlung ungeprüft zu übernehmen. Zusammen entsteht eine schwer messbare, aber klinisch relevante Diskriminierung.

Design- und Interventionsmöglichkeiten: Was kann man tun?

Um Automation Bias als faktisches Risiko zu minimieren, muss dieser Aspekt in der Entwicklung und im Design von CDSS bewusst mitgedacht werden und technische und organisatorische Lösungen geschaffen werden, die das Mensch-KI-Team stabil und sicher machen. Folgende Strategien sind dabei denkbar:

Training und Sensibilisierung

Für eine effektive Mensch-KI-Zusammenarbeit fehlen bislang systematische Trainingsprozesse. Die Kooperation zwischen dem medizinischen Fachpersonal und dem KI-basierten System ist eine Fähigkeit, die gelernt werden muss. Klare Rollenverteilungen und definierte Regeln, wer was macht und in welchen Fällen die KI konsultiert wird, können dabei helfen, ineffiziente und widersprüchliche Entscheidungsprozesse zu vermeiden.

Um den beschriebenen Effekten wie der Kompetenzillusion oder dem Autoritäts- und Expert Bias entgegenzuwirken, müssen Nutzende geschult werden, die Grenzen von KI zu erkennen. Es muss ihnen bewusst gemacht werden, dass Fehler möglich sind, und wie man sie erkennt (Goddard et al., 2011). Dies kann als **„AI Risk Literacy"** bezeichnet werden (Otto, 2026). Regelmäßige Simulationen mit „Fehlerszenarien" helfen, Wachsamkeit zu wahren.

Expertinnen und Experten betonen, dass bloße Erfahrung mit Systemen Automation Bias nicht verhindert. Das heißt, Entwicklerinnen und Entwickler sollten Materialien bereitstellen, die nicht nur Funktionen erklären, sondern psychologische Risiken verdeutlichen.

Unsicherheitskennzeichnung und Transparenz

Ein Ansatz zur Bekämpfung dieses Problems besteht darin, KI-Systeme transparenter zu gestalten und den Nutzerinnen und Nutzern mehr Informationen über das System bereitzustellen (Rieger et al., 2025).

Statt nur eine Empfehlung anzuzeigen, sollte das System seine Unsicherheit kommunizieren, zum Beispiel mit Konfidenzintervallen, Wahrscheinlichkeiten oder Szenarien **(Confidence Metrics)**. Das soll den Menschen zur Reflexion auffordern und blindes Vertrauen mindern. Studien zeigen, dass solche UI-/UX-Gestaltungen den Entscheidungsprozess je nach Vorerfahrung oder Einstellung gegenüber KI unterschiedlich unterstützen können. Unsicherheitsvisualisierungen können dabei zu einem bewussteren Umgang mit KI-Empfehlungen anregen (Reyes et al., 2025).

Der Explainable AI (XAI) Ansatz beschreibt das Prinzip, die Daten, die zu einer bestimmten Vorhersage geführt haben, dem Nutzer darzulegen. Auf diese Weise sollen Entscheidungen und Funktionsweisen für die Nutzenden nachvollziehbar und transparent gemacht werden. XAI wird häufig als Lösung für sichere KI im Gesundheitswesen propagiert (Kyrimi et al., 2025; Abdelwanis et al., 2024). XAI soll dabei helfen, die strengen Richtlinien wie den EU AI Act einzuhalten und fördere außerdem die Akzeptanz von KI-gestützten Entscheidungsunterstützungssystemen im klinischen Alltag.

Forschungsergebnisse zeigen jedoch auch, dass diese Konzepte unter Umständen eher eine **Illusion von Kontrolle und Vertrauen** erzeugen, als die tatsächliche Verlässlichkeit der Systeme zu erhöhen. Menschen interpretieren Erklärungen der KI oft als Plausibilitätssignal, selbst wenn diese faktisch falsch sind. In diesem Sinne kann XAI unbeabsichtigt den Automation Bias verstärken, anstatt ihn zu reduzieren. In einer experimentellen Studie zur Bilddiagnostik sank die menschliche Fehlererkennungsrate, wenn das KI-System zur Fehldiagnose eine visuelle Heatmap zur Erklärung anzeigte. Vermutlich konzentrierten die Radiologen sich auf die hervorgehobenen Bereiche und vernachlässigten andere (Rezazade Mehrizi et al., 2023). Otto (2026) bezeichnet XAI auch als ein als Lösung getarntes Hindernis.

Ein aktueller konzeptioneller Beitrag zum Thema Vertrauensbildung geht über einfache XAI-Ansätze hinaus. Schlicker et al. (2025) führen ein Modell ein, wonach Nutzer ihre Wahrnehmung eines Systems aufgrund verfügbarer Hinweise wie Unsicherheitsinformationen, Performance-Metriken oder Erklärungen bilden. Nur wenn diese Hinweise relevant und verständlich sind, kann die wahrgenommene Vertrauenswürdigkeit eine verlässliche Grundlage für eine angemessene Nutzung darstellen. Andernfalls können gut gemeinte Erklärungen und Transparenzmaßnahmen das Vertrauen fehlkalibrieren, ohne die tatsächliche Verlässlichkeit zu erhöhen.

Wie KI-Empfehlungen durch geeignete Schnittstellengestaltung besser in menschliche Entscheidungen integriert werden können, ohne dass die gemeinsame Leistung hinter der der KI allein zurückfällt, zeigt eine neuere Studie empirisch. De Vries et al. (2026) untersuchten Interface-Designs, die Nutzende jeweils zu unterschiedlichen Informationsintegrationsstrategien anleiten, beispielsweise eine adaptive Orientierung an der eigenen Unsicherheit („verlasse dich mehr auf die KI, wenn du selbst unsicher bist") oder eine proportionale Gewichtung von menschlichem und KI-Urteil nach jeweiliger Sicherheit. Die Ansätze führten zu einer verbesserten Entscheidungsleistung im Vergleich zur KI allein und zeigen damit, dass gezielte Interface-Gestaltung die Qualität der Mensch-KI-Interaktion signifikant steigern kann.

Entscheidend ist, dass Erklärungen des Systems nur dann für eine verbesserte Leistung des Gesamtteams Mensch-KI sorgen, wenn Menschen ihre Kooperationsstrategien effektiv anpassen und die KI überprüfen. XAI funktioniert also nicht per se, sondern nur in Kombination mit einer Veränderung des Rollenverständnisses (Berger et al., 2025).

Veränderung des Rollenverständnisses

Die Gestaltung der vom System ausgegebenen Empfehlung kann bestimmen, welches Rollenverständnis der Mensch im Entscheidungsprozess einnimmt. Automatisierte Hinweise sollten beispielsweise nicht als „Endentscheidung" gezeigt werden, sondern als Input, der vom Menschen bewusst geprüft werden muss — zum Beispiel durch „Zweitmeinung"-Prozesse, Checklisten oder Pflicht zur aktiven Bestätigung. Das bewahrt kognitive Aktivität, verringert Complacency und reduziert Moral Distress. Die Nutzenden müssen verstehen, dass ihre Aufgabe darin besteht, die Empfehlung **aktiv zu verifizieren**. Dies kann – gepaart mit einer erhöhten Transparenz – zudem die moralische Handlungskompetenz steigern (Moral Agency), indem es medizinischem Personal ermöglicht wird, das ihrer Ansicht nach moralisch Richtige zu tun beziehungsweise in Betracht zu ziehen und Verantwortung dafür zu übernehmen (Fortier & Malloy, 2019) – selbst, wenn dies der KI widerspricht.

Die Bedeutung eines passend gestalteten Rollenverständnisses zeigt sich auch in einer aktuellen Studie von de Vries, et al. (2025). Darin wurde untersucht, wie unterschiedlich stark automatisierte medizinische Unterstützungssysteme das Verhalten von Nutzenden beeinflussen. Die Ergebnisse verdeutlichen, dass ein geringeres Automationsniveau nicht automatisch zu mehr kritischer Prüfung führt: Besonders weniger erfahrene Nutzende vertrauten weiterhin stark auf KI-Hinweise, selbst dann, wenn sie mehrere Hypothesen eigenständig beurteilen sollten. Entscheidend ist daher nicht allein die Automationsstufe, sondern die klare kommunikative Rollenverteilung: CDSS müssen verdeutlichen, dass ihre Vorschläge vorläufige Inputs sind, die aktiv verifiziert und geprüft werden müssen.

Man kann noch einen Schritt weiter gehen und den Menschen in der Zusammenarbeit mit dem KI-System die Rolle des „Advocatus Diaboli" nahelegen. Seine Aufgabe ist es nicht nur Aussagen des Systems zu verifizieren, er muss sogar die Handlungsempfehlung ständig hinterfragen, statt sie dankbar als fundiertere Analyse zu betrachten. Über die Gestaltung des Unterstützungssystems müssen Nutzende darin bestärkt werden, das **Falsifikationsprinzip** zu verfolgen.

Menschen tendieren dazu, nur nach Informationen zu suchen, die eine bestehende Vermutung stützen, anstatt sie kritisch zu hinterfragen (Confirmation Bias). Um diesen „Bestätigungsfehler" zu überwinden, sind spezifische Anreize oder Instruktionen nötig, die explizit das **Finden von Gegenbeispielen** belohnen, um die kognitive Verzerrung zu reduzieren (Piksa et al., 2024).

Noch radikaler wäre ein prinzipieller Ausschluss von Krankheitsdiagnosen oder Handlungsempfehlungen durch medizinische Unterstützungssysteme. Beispielsweise könnte die Unterstützungsleistung in der Analyse von Auffälligkeiten enden – die daraus zu ziehenden Schlüsse blieben dann weiterhin dem Menschen vorbehalten.

Otto (2026) schlägt ergänzend noch als praktischen Ansatz vor, die Rollen oder Aufgaben innerhalb des Teams regelmäßig zu wechseln, um nachlassende Aufmerksamkeit und Complacency zu vermeiden. Außerdem würde eine offene Fehlerkultur helfen, um Fehler als Lernchance zu nutzen und rechtzeitig Korrekturen vorzunehmen.

Diese Empfehlungen aus der Human-Factors-Forschung bilden eine Grundlage dafür, dass KI-basierte Systeme nicht nur leistungsstark, sondern auch sicher und vertrauenswürdig im klinischen Alltag sind.

Fazit

Automation Bias zeigt deutlich, dass der Einsatz von KI in der Medizintechnik kein rein technisches, sondern auch ein psychologisches, organisationales und moralisches Unterfangen ist. Selbst eine sehr gut funktionierende KI kann das klinische Problem nicht lösen, wenn das Mensch-KI-Team nicht unkritisch funktioniert. Damit KI und Mensch tatsächlich als Team wirken, müssen Design, Training und Rahmenbedingungen so gestaltet sein, dass menschliche Kontrolle, kritisches Denken, moralische Integrität und Verantwortungsbewusstsein nicht nur möglich, sondern systematisch gefördert werden.

Dazu gehört, dass sich der Mensch seiner Rolle in Zusammenarbeit mit dem KI-System bewusst sein muss: Er muss die Handlungsempfehlung ständig hinterfragen, statt sie dankbar als fundiertere Analyse zu betrachten. Er muss zum Advocatus Diaboli werden und darin bestärkt werden, das Falsifikationsprinzip zu verfolgen.

Damit wirkt man auch einem anderen unerwünschten Effekt von KI-Unterstützung in der Medizin entgegen, dem „Deskilling" (Apel et al., 2026, Abdelwanis et al., 2024): Ein übermäßiges Vertrauen in KI-Algorithmen kann dazu führen, dass Ärztinnen und Ärzte sich weniger mit vielfältigen

Fällen auseinandersetzen und in kritischen Entscheidungssituationen weniger praktische Erfahrungen sammeln.

Gerade diese Art des Umdenkens steht allerdings dem gegenwärtigen Trend entgegen. Shaw und Nave (2026) zeigen, dass sich im Alltag der KI-Nutzung eine neue Denkform etabliert: das „Cognitive Surrender". Darunter verstehen sie die Tendenz, KI-Outputs mit minimaler eigener Prüfung zu übernehmen. Die aktive Suche nach Widersprüchen, Fehlerquellen oder Alternativerklärungen wird anspruchsvoller in einer Welt, in der Menschen zunehmend aufgrund von Effizienzforderungen an diese Form der Denkabkürzung gewöhnt sind und – wie im Zusammenspiel mit Algorithm Biases - die Falsifikation eines Black Box-Outputs kognitiv zunehmend herausfordernder wird.

Mit dem Aufkommen agentischer KI-Systeme zeichnet sich zudem eine weitere Verschärfung des Problems ab. Während heutige Entscheidungsunterstützungssysteme primär Empfehlungen generieren, beginnen agentische Modelle zunehmend ganze medizinische Workflows zu automatisieren, Leitlinien eigenständig anzuwenden und Prozessschritte in klinischen Abläufen zu strukturieren. Diese Entwicklung wirft zentrale Fragen danach auf, an welchen Stellen Human-in-the-Loop-Mechanismen zwingend notwendig sind, wie viel Autonomie abgegeben werden darf und wie sichergestellt werden kann, dass klinische Verantwortung nicht entkernt wird.

Die Entwicklung und Zulassung von KI-basierter Medizintechnik sollten nicht nur technische, sondern auch psychologische Aspekte berücksichtigen. Dazu gehören auch ein Post-Market Monitoring, Auditierbarkeit von Entscheidungen, klare Verantwortlichkeiten und gegebenenfalls Redundanz (z. B. ein manuelles Review bei bestimmten Risiko-Schwellwerten).

Grundsätzlich sollte Medizintechnikentwicklern und den Anwendern klar sein, dass man menschliche Fehlentscheidungen nicht aus dem System „rausautomatisieren" kann. Solange die Unterstützungssysteme selbst von Menschen benutzt und überwacht werden, wird der menschliche Fehler nicht durch Automatisierung eliminiert, sondern an eine andere Stelle verschoben.

Der Umgang mit KI in der Medizin ist nicht allein eine Frage technischer Leistungsoptimierung, sondern eine zentrale Herausforderung für Responsible AI (siehe Müller et al., 2026): Systeme müssen so gestaltet sein, dass sie menschliche Urteilsfähigkeit und Handlungskompetenz unterstützen statt sie zu untergraben. AI Ethics rückt damit als konkrete Designanforderung in den Vordergrund: Erst wenn KI so entwickelt wird, dass Menschen befähigt bleiben, Fehler zu erkennen und kritische Distanz zu wahren, kann sie sicher und verantwortungsvoll in der Medizin eingesetzt werden.

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

Bildnachweis:
elenabs/istockphoto

VDI|VDE|IT