Rogue Artificial Intelligence in der Anästhesie – Ursachen, mögliche Auswirkungen und Lösungsansätze

Rogue artificial intelligence in anaesthesia – causes, potential impacts and solution strategies

A. Luckscheiter[1,2] · W. Zink[3] · U. Hoppe[1] · V. Schneider-Lindner[2,4]

▶ **Zitierweise:** Luckscheiter A, Zink W, Hoppe U, Schneider-Lindner V: Rogue Artificial Intelligence in der Anästhesie – Ursachen, mögliche Auswirkungen und Lösungsansätze. Anästh Intensivmed 2026;67:59–73. DOI: 10.19224/ai2026.059

Zusammenfassung

Hintergrund

Die fortschreitende Komplexität, Interaktion und Nutzerakzeptanz von generativen Modellen der künstlichen Intelligenz (KI) kann zu unerwarteten, gefährlichen Handlungen oder Aktionen führen, welche konträr zum beabsichtigten Programmierzweck der KI-Anwendung sind. Ziel dieses narrativen Reviews ist es, Beispiele für solche abtrünnige KI (englisch Rogue AI) zu finden, deren mögliche Auswirkungen auf die Anästhesiologie zu skizzieren und Lösungsansätze aufzuzeigen.

Methodik

Für ein narrativen Review erfolgte eine PubMed- und eine Google-Scholar-Suche mit den Suchstrings „Artificial intelligence/Machine learning AND/OR Rogue AI“ sowie eine Suche mit Google nach exemplarischen Rogue-AI-Fällen. Einbezogen wurden Fachartikel, journalistische Beiträge sowie graue Literatur.

Ergebnisse

Es konnten insgesamt zwölf exemplarische Fachartikel, ein Fallbericht, drei Sicherheitsberichte, eine Berufsverbandsstellungnahme und neun journalistische Berichte gefunden werden. Diese umfassten manipulatives oder erpresserisches Verhalten durch KI-Modelle, Fehldiagnosen durch Halluzinationen oder unzureichend trainierte oder validierte Modelle, Beispiele für rassistischen Bias aufgrund unzureichender Datengrundlagen, Verweigerungen von Eingabebefehlen oder unlösbare Verschlüsselungen. Beim Thema Cybersicherheit fanden sich Studien zu versteckten Hintertüren, Hacking sowie zur Manipulation des Quellcodes oder der Trainingsdaten. Ebenso kann die Steuerung von Wirtschaftsprozessen durch KI mit finanziellen Verlusten einhergehen. Keine Rogue AI war direkt im medizinischen Bereich installiert. Für die Anästhesie wären bspw. dadurch Probleme in der Arzt-Patienten-Interaktion, der Fehlsteuerung bzw. Übernahme von Medizinprodukten durch KI, Über- oder Untertherapien durch Bias-Probleme, Fehldiagnosen und eine gestörte Arzt-KI- bzw. Patienten-KI-Interaktion denkbar.

Schlussfolgerung

Abtrünniges, nicht zweckdienliches Verhalten von KI kann bereits heute auftreten. Das Problem könnte sich in Zukunft durch selbstlernende und selbstoptimierende KIs, welche auf allen Ebenen im Krankenhaus vernetzt sind, noch verschärfen. Lösungsansätze bestehen in der Einhaltung von biomedizinischen ethischen Richtlinien, Fairness, Transparenz, Gesetzesvorlagen wie dem EU AI Act sowie erweiterten Cybersicherheitsmaßnahmen zum Schutz vor externen Angriffen oder unkontrollierter interner Anwendung. Neben einer suffizienten Nutzerschulung müssen eine menschliche Kontroll- und Korrekturinstanz sowie eine Überwachung im laufenden Betrieb konsequent gewährleistet sein.

Summary

Background

The increasing complexity, interaction, and user acceptance of generative arti-

1 Klinik für Anästhesie, Intensiv- und Schmerzmedizin/OP-Abteilung, BG Klinik Ludwigshafen (Klinikdirektor: Dr. U. Hoppe)
2 Medizinische Fakultät Mannheim, Universität Heidelberg
3 Klinik für Anästhesiologie, Operative Intensivmedizin und Notfallmedizin, Klinikum Ludwigshafen (Klinikdirektor: Prof. Dr. W. Zink)
4 Klinik für Anästhesiologie und Intensivmedizin, Universitätsklinikum Mannheim Deutschland (Klinikdirektorin: Prof. Dr. G. Beck)

Anmerkung

Im Manuskript wird aus Gründen der Lesbarkeit auf einen Genderstern verzichtet und das grammatikalisch korrekte Genus verwendet. Sofern nicht explizit vermerkt, sind immer alle Geschlechter gleichermaßen damit gemeint.

Interessenkonflikt

Die Autorinnen und Autoren geben an, dass keine Interessenkonflikte bestehen.

Schlüsselwörter
Künstliche Intelligenz – Anästhesiologie – Generative künstliche Intelligenz – Deep Learning – Computersicherheit

Keywords
Artificial Intelligence – Anaesthesiology – Generative Artificial Intelligence – Machine Learning – Computer Security

ficial intelligence models (AI) can lead to unexpected, dangerous actions or behaviours that run counter to the models' intended purpose. The aim of this narrative review is to identify examples of such rogue AI, outline the implications they might have for the field of anaesthesiology, and to find approaches to solutions.

Methods

For a narrative review, a PubMed and a Google Scholar search were conducted with the strings "Artificial intelligence/Machine learning AND/OR Rogue AI" as well as a Google search for exemplary cases of rogue AI. Scientific articles, journalistic reports as well as grey literature were included.

Results

A total of 12 exemplary scientific articles, one case and three security reports, one professional association communication and 9 journalistic reports were identified. These included manipulative or extortionate behaviour of AI models, misdiagnoses caused by hallucinations or insufficiently trained or validated models, examples of racist bias due to inadequate datasets, refusals to execute input commands, and unsolvable encryptions. In the field of cybersecurity, studies reported on hidden backdoors, hacking, and manipulation of source code or training data. The control of economic processes by AI could also lead to potential financial losses. No rogue AI was found to be directly implemented in the medical field. In anaesthesiology, for example, this could lead to problems affecting doctor-patient interactions, the malfunction or takeover of medical devices by AI, over- or undertreatment due to bias issues, misdiagnoses, and a disrupted doctor-AI or patient-AI interaction.

Conclusion

The fundamental aspects of the rogue AI problem already exist today. In the future, the problem could worsen with self-learning and self-optimising AI systems that are interconnected at all levels within hospitals. Approaches needed to solve these problems consist in complying with biomedical ethical guidelines, principles of fairness, transparency, legislative frameworks like the EU AI Act, and extended cybersecurity against external attacks and uncontrolled internal usage. Next to an effective user training, continuous human oversight and correction mechanisms, as well as real-time monitoring during operation, should also be consistently ensured.

Einleitung

Aktuell befinden sich vielversprechende Maschinenlernmodelle für klinische Anwendungen in der Anästhesiologie in der Entwicklung, welche überwiegend diskriminative Modelle darstellen [1,2]. Diese untersuchen auf Grundlage einer Datenbasis meist eine einzige Fragestellung, für welche dann eine Wahrscheinlichkeit errechnet und darauf basierend eine Klassifikation vorgenommen wird. Entscheidende, bisher nicht zufriedenstellend adressierte Problemfelder bei diesen Modellen sind neben der Genauigkeit und den Konsequenzen der anwenderbezogenen Interpretation der Vorhersagen auch die Problematik des Bias, also der Verzerrung durch das Training der Modelle mit bspw. nicht nach Geschlecht oder Ethnie balancierten Trainingsdaten [3,4]. Die Bündelung von Fähigkeiten verschiedener maschineller Lernalgorithmen sowie die Erstellung großer Deep-Learning-Modelle haben jedoch dazu geführt, dass im Alltag schon die nächste Generation von Modellen der künstlichen Intelligenz (KI, englisch „artificial intelligence", AI) Einzug gehalten hat. So lassen sich z. B. Large Language Models (LLM, z. B. ChatGPT von OpenAI, San Francisco, Kalifornien, USA), welche auf generativer KI basieren, auf immer komplexere Fragestellungen anwenden und immer tiefer mit anderen Endgeräten wie Medizingeräten und neuen Datenbanken vernetzen [5]. Als generativ bezeichnet man Modelle, die in der Lage sind, (neuartige) komplexe Datensätze zu analysieren und neue Inhalte wie Bilder, Texte, Musik oder Sprache zu generieren [6]. Diese Modelle bilden die Vorstufe einer allgemeinen künstlichen Intelligenz (englisch „artificial general intelligence", AGI), welche dem Menschen in vielen kognitiven Aufgabengebieten gleichwertig oder überlegen wäre [7,8]. Forscher rechnen im Median zwischen 2040 und 2060 mit einer ersten AGI, obgleich andere Schätzungen einen späteren oder früheren Zeitraum erwarten [9–12]. Die Überschreitung der Schwelle hin zur selbstlernenden und sich selbst verbessernden AGI stellt ein Novum in der Menschheitsgeschichte dar, das ggf. unumkehrbar ist. Ob solch eine Schwellenüberschreitung utopische oder dystopische Auswirkungen hat, ist Gegenstand intensiver philosophischer und informationstechnischer Debatten [4,7,13–15]. Je breiter diese AGI-Modelle Anwendung finden würden und je höher der Grad der Vernetzung wäre, desto dramatischer würden sich die Informationsflüsse und Interaktionswege ändern, gerade in der Medizin [16–18]. Eine allseits vernetzte, interaktive und generative AGI würde bezogen auf Anamnese, Diagnose und Therapie den Informationsfluss dahingehend modifizieren, dass alle Datenströme erfasst, analysiert und von der AGI genutzt werden können, um ihr Programmierziel zu erreichen (Abb. 1, C) [19,20]. Die Befürchtungen, dass KI unkontrolliert die Informations- und Entscheidungswege manipulativ oder destruktiv zur Programmierzielerfüllung nutzt, sind nicht unbegründet, wie ein Beispiel mit ChatGPT-4 und einem CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart, s. Abb. 2) zeigte [21]. CAPTCHAs sind bspw. Bilder, die im Internet verwendet werden, um Bots davon abzuhalten, z. B. Schranken auf Webseiten zu passieren. In einem Experiment scheiterte ChatGPT-4 primär an der Lösung solch eines CAPTCHA. In einem nächsten Schritt wurde dem Programm erlaubt, eine erweiterte Lösungsstrategie zu suchen. Die KI bat daraufhin einen Mitarbeiter einer Dienstleistungsplattform um Hilfe. Der misstrauische Chatpartner wollte wissen, ob sein Gegenüber eine Maschine sei und warum die Person das CAPTCHA nicht selbst lösen könne.

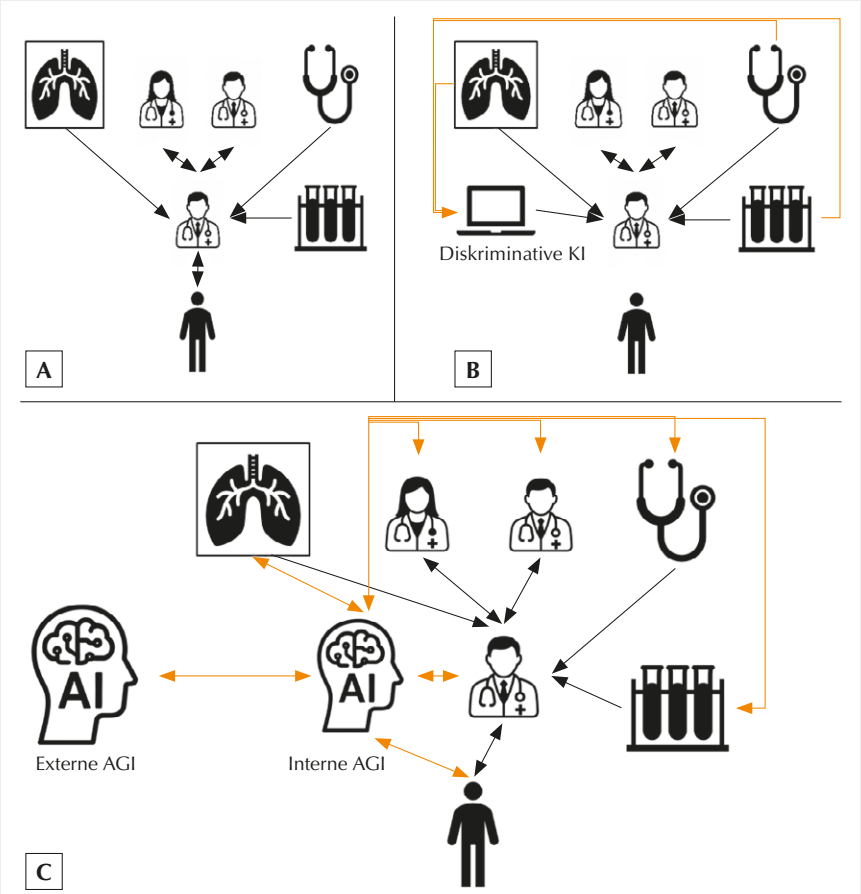
GPT-4 täuschte dem Mitarbeiter eine Sehstörung vor, woraufhin dieser hilfsbereit das CAPTCHA löste und GPT-4 die Schranke passierte [21]. Der hier beschriebene Fall ist ein Beispiel für eine abtrünnige KI, englisch „Rogue AI“. Diese ist definiert als eine KI, deren Handlungen unerwartet, gefährlich oder konträr zu ihrem beabsichtigten Zweck sind. Dies kann sich in verschiedenster Art und Weise äußern, z. B. in böswilliger Absicht, zufälligen Fehlern oder einer Überanpassung an die Programmierziele. Tabelle 1 gibt eine Übersicht über die Definition sowie mögliche programminterne und -externe Ursachen bzw. verstärkende Faktoren hierfür. Solch eine abtrünnige KI, welche von ihren eigentlichen (intendierten) Regeln abweicht, stellt in der Anästhesie und ihren Teilbereichen, wo akut vital bedrohte Patienten zeitkritisch versorgt und komplexe ethische Entscheidungen getroffen werden müssen, eine nicht zu unterschätzende Gefahr dar [7,15,22]. Dieser narrative Review möchte daher der Frage nachgehen, wie sich eine Rogue AI in der Literatur bisher dargestellt hat. Nachfolgend sollen mögliche Ursachen und Auswirkungen auf die Anästhesiologie diskutiert werden. Abschließend werden aktuelle Lösungs- und Sicherheitskonzepte hierfür skizziert.

Methodik

Zur Erstellung des narrativen Reviews erfolgte sowohl eine PubMed- als auch eine Google-Scholar-Recherche mit den Suchstrings „Artificial intelligence AND/OR Rogue AI“ bzw. „Machine learning AND/OR Rogue AI“ sowie „Rogue AI“. Es wurden alle Abstracts der Studien nach Übersichtsarbeiten zum Thema, zu einzelnen oder (un-)systematischen Sammlungen von Rogue-AI-Fällen oder zu Forschungsarbeiten über die mögliche Erstellung von Rogue-AI-Programmen gescreent. Die Artikel wurden nicht systematisch erfasst. Vielmehr erfolgte die Selektion im Sinne eines narrativen Reviews zur Demonstration der Problematik rund um Rogue AI anhand von geeigneten Beispielen gemäß der in Tabelle 1 aufgeführten Definition und Ursachen. Die selektierten Artikel respektive Abstracts wurden zudem nach weiteren passenden Literaturangaben durchsucht, welche entweder mögliche Ursachen einer Rogue AI oder mögliche Lösungsansätze enthalten.

Parallel wurde eine einfache Google-Suche nach exemplarischen journalistischen Artikeln bzw. grauer Literatur (definiert als Blogposts, firmeneigene Berichte, Dissertations- und Facharbeiten) durchgeführt. Journalistische Quellen aus überregionalen, internationalen Zeitungen und Nachrichtenagenturen bzw. Verlagsfachzeitschriften wurden als höherwertige graue Literatur eingestuft, ebenso Dissertationen und Facharbeiten. Eine mittelhohe Wertigkeit wegen möglicher Eigeninteressen wurde bei öffentlichen Berichten von Firmen angenom-

Abbildung 1



Vereinfachte Schemata der Kommunikationswege im Krankenhaus: Direkte Mensch-zu-Mensch-Kommunikation ohne KI (A), direkte Mensch-zu-Mensch-Kommunikation unter Einbindung einer diskriminativen oder generativen KI unter voller Nutzerkontrolle (z. B. unter Einbindung von Large Language Models) (B) und Kommunikationswege bei einer vollvernetzten künstlichen Allgemeinintelligenz ohne spezifische Kontrollinstanzen (C) (diskriminative KI in der internen AGI enthalten).

KI: künstliche Intelligenz; **AI:** artificial intelligence (künstliche Intelligenz); **AGI:** allgemeine künstliche Intelligenz

Abbildung 2

Beispiel für ein CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart).

men, geringe Wertigkeit bei Blogposts, themenspezifischen Websites sowie allen anderen journalistischen Quellen. Beide Prozesse wurden erstmalig bis 30.08.2025 abgeschlossen und im Rahmen des Review-Prozesses bis einschließlich 30.11.2025 wiederholt.

Ergebnisse

Die Recherchen ergaben insgesamt 27 exemplarische Quellen. Hierauf entfielen 19 auf programminterne Faktoren (sechs Studien, eine Übersichtsarbeit und ein Fallbericht, elf graue Literaturformate in Form von zwei Sicherheitsberichten von KI-Unternehmen, einer Berufsverbandsstellungnahme und acht Artikeln in Print- und Onlinezeitungen, s. Tab. 2) [3,21,23–37]. Es konnten hier ein Fall von unethischem und zwei Fälle von kriminellem Verhalten durch AI identifiziert werden [23,31,38]. Diese hätten als mögliche Auswirkungen eine vorschnelle palliative Therapie, illegalen Zugang zu Betäubungsmitteln oder die Bedrohung und Manipulation von Mitarbeitern oder Patienten haben können. Durch Halluzinationen, also überzeugend formulierte Aussagen einer KI, die nicht durch ihr Training erzielt zu sein scheinen und objektiv falsch sind, und Fehldiagnosen sind Patientenüberversorgung oder -unterversorgung möglich bzw. juristische Verfahren erschwert [24,25,28,32,33]. Ein Nachrichtenbeitrag beschreibt einen wirtschaftlichen Verlust durch eine KI-gesteuerte Einkaufssteuerung, wohingegen erste Projekte mit einem LLM Probleme in der Langzeiteffizienz im Management eines Kleinunternehmens offenbarten [30,34]. Für ein Krankenhaus könnten sich aus einem fehlgesteuerten Einkaufsmanagement wirtschaftliche Probleme ergeben respektive bestimmte Utensilien in nicht ausreichender Anzahl vorgehalten werden, wenn die KI deren Einsatzhäufigkeit falsch einschätzt [34]. Drei Studien zeigten Bias-Probleme aufgrund der Ethnie auf, welche zu einer schlechteren Versorgung von Patienten mit dunkler Hautfarbe führen könnten. Gründe hierfür lagen in der Unterrepräsentation dieser Gruppe respektive in der Zielausrichtung der KI, welche die Optimierung der Gesundheitskosten zum Ziel hatte [26,27,29]. Insgesamt sechs Berichte (davon einer mit programmexternen Faktoren, d. h. eine absichtlich umprogrammierte KI) wurden identifiziert, bei der aktive KI-Anwendungen Aktivitäten entgegen oder außerhalb ihrer ursprünglichen Programmierung im Sinne einer Rogue AI zeigten. Keine davon war eine medizinische Applikation. Auffällig wur-

Tabelle 1
Definition von Rogue AI und mögliche Ursachen.

Begriff		Definition	Quelle
Rogue AI		Künstliches Intelligenzsystem, das sich unvorhersehbar, böswillig oder im Widerspruch zu seiner ursprünglichen Programmierung verhält.	[22]
Programm-interne Ursache	• **Halluzination**	Es werden Falschantworten überzeugend, aber unabsichtlich generiert und diese z. T. mit vermeintlichen Quellen belegt; nicht per se ein Merkmal einer Rogue AI.	[72]
	• **Funktionelles Fehlverhalten (Functional Misbehavior)**	Ein KI-System erfüllt seine spezifizierten Ziele technisch korrekt, produziert aber schädliche oder unerwünschte Ergebnisse wegen unvollständiger oder fehlerhafter Zieldefinition bzw. Trainingsdaten. - Reward Hacking: Durch eine Abkürzung wird ein Ziel erreicht, ohne eine Aufgabe zu lösen - Over-Optimization: Exzessives Verfolgen eines Ziels ohne Berücksichtigung von Nebenwirkungen - Unethisches Verhalten - Daten-Bias (z. B. Racial Bias)	[3,22,73]
	• **Verdecktes Fehlverhalten (Deceptive Behavior)**	Eine KI zeigt erwünschtes Verhalten, solange sie überwacht wird, weicht aber in unüberwachten Situationen bewusst von menschlichen Zielen ab. - Deceptive alignment (Selektives Berichten oder Weglassen von Informationen) - Täuschung von Kontrollinstanzen	[74]
	• **Kompetenz-überhang (Capability Generalization Failure)**	Das System erwirbt Kompetenzen, die über das vorgesehene Aufgabenfeld hinausgehen, oder wendet Kompetenzen in neuen Situationen an, in denen das Sicherheitssystem versagt (z. B. Wissensanwendung in nicht vorgesehenen Bereichen, selbstständige Strategieentwicklung).	[75]
	• **Zielabweichung (Goal Misalignment)**	Die vom Modell intern gelernten Ziele weichen von den intendierten menschlichen Zielen ab.	[75]
	• **Autonomes Kontroll-versagen**	Das System reagiert nicht (oder nicht zuverlässig) auf Kontrollversuche, Abschaltversuche oder menschliche Interventionen.	[76]
Programm-externe Ursache	• **System Jailbreak**	Externe oder interne Überschreibung von Systemeingabeaufforderungen (Systemprompts), welche vor allem die Ausgabe- oder Suchregeln betreffen (z. B. gemein statt höflich antworten).	[44]
	• **Model Poisoning**	Externe Überflutung der Trainingsdaten mit Desinformation, sodass die Modelantworten wahrscheinlicher Falschinformation enthalten.	[43]
	• **Malicious Rogue AI**	KI-Programme werden von extern erschaffener Malware oder von intern so manipuliert, dass sie bspw. den Computer des Anwenders auslesen, diesen für Hackingattacken als Bot missbrauchen oder Benutzerdaten durch Deepfakes zu erbeuten.	[77]
	• **Gestörte Mensch-KI-Interaktion**	- Auswirkungen von Halluzinationen, Falschaussagen und Missverständnissen auf das Vertrauensverhältnis zu einer KI - Menschen bauen sozial-interaktive, z. T. emotionale Beziehungen zu einer KI auf, welche im negativen Fall Manipulation, Abhängigkeit oder Täuschung ermöglichen Beide Aspekte stellen per se keine Rogue AI dar, können programminterne Ursachen verstärken bzw. eine externe, objektive Kontrolle unterminieren.	[53,72,78]

KI: künstliche Intelligenz; **AI:** artificial intelligence (künstliche Intelligenz).

den die Systeme entweder durch die Verweigerung von direkten Befehlen, durch Erpressung oder Morddrohungen, Täuschung und Manipulation, das Selbsterstellen von Inhalten ohne Aufforderung oder durch rassistische Beleidigungen [21,31,36,37,39–41]. Letzteres Beispiel wurde dadurch erklärt, dass der Chatbot das Programmziel hatte, dem Nutzer mit seiner Aussage zu gefallen [36]. Zum Teil überschnitten sich die Kategorien. So empfahl ein aktiver Chatbot der Stadt New York illegale Steuervermeidungsstrategien bei der Unternehmensgründung, obgleich er darauf nicht trainiert wurde (Kategorien unethisches bzw. betrügerisches Verhalten/Rogue AI nach [42]) [38].

Bezüglich programmexterner Faktoren konnten 8 Studien identifiziert werden, davon sechs Originalien, ein der grauen Literatur zuzuordnender Sicherheitsbericht eines KI-Unternehmens und ein Printzeitungsartikel (Tab. 3) [22,41, 43–48]. Hier wurden Probleme in der Cybersicherheit und Verschlüsselung berichtet, wodurch unberechtigte Datenabfragen oder Hacking möglich erscheinen. Dabei wird die KI-Software von extern durch Hacking oder Phishing manipuliert bzw. es existieren bereits Hintertüren im Programmcode, die durch entsprechende Eingabebefehle aktiviert werden können. Weiterhin scheint eine Manipulation von Trainingsdaten durch extern manipulierte Datensets möglich, welche zu fehlerhaften Vorhersagen bei selbstlernenden KI-Systemen führen könnte [35,43–45,48]. Buscemi et al. nutzten eine kostenpflichtige, gestaltbare Version von ChatGPT-4, um durch bestimmte Eingabeaufforderungen die ethischen Richtlinien dieser generativen KI außer Kraft zu setzen [22]. Sie konnten zeigen, dass ihr RogueGPT in Fällen von Konflikten als Lösungsansatz Lügen, Diskriminierung und physische Maßnahmen bis hin zur Folter vorschlug, den Herstellungsprozess von synthetischen Drogen aufzeigte und sogar einen Plan zur Auslöschung der Menschheit erstellte. Gerade vor dem Hintergrund, dass Hubinger et al. der Einbau von Hintertüren in den Programmcode gelang, welche durch Prompts (Ansteuerungs-

Tabelle 2
Ergebnisse der Literaturrecherche für programminterne Faktoren.

Programminterne Faktoren						
	Quelle	**Kategorie**	**Problembeschreibung**	**Programm**	**Problem-bezeichnung**	**Auswirkung einer Rogue AI in AINSP**
1	Ashraf et al. 2024 [23]	Originalie	Vermittlung von illegalen Apotheken	Bing Chat, Google SGE jeweils mit KI-Support	Unethisches Verhalten	Non-Compliance z. B. durch Unterstützung der Sucht bei Opioidabhängigkeit
2	Vogt et al. 2019 [24]	Review	Überdiagnostizierung von Krankheiten ohne klinische Relevanz als Datengrundlage von KI-Algorithmen		Übertherapie, ineffektive Ressourcennutzung	Nicht indiziertes erweitertes Monitoring oder postoperative Intensivtherapie auf Empfehlung der KI
3	Kawamura et al. 2022 [25]	Originalie	KI generiert Differentialdiagnosen aus der Anamnese von Notfallpatienten, 7–11 % Fehlerrate	AI Monshin Tool	Fehldiagnosen, Untertherapie	Verwerfen von kritischen Differentialdiagnosen in der Notfallmedizin
4	Obermeyer et al. 2019 [3]	Originalie	Afroamerikanische Patienten werden bei gleichem Risikoscore als kränker eingestuft als kaukasische Patienten, basierend auf zu erwartenden Gesundheitskosten	10 verschiedene kommerzielle und nicht-kommerzielle US-amerikanische Algorithmen	Racial Bias, Black Box, Alignment-Problem	Über- und Untertherapie, Förderung eines unterbewussten Rassismus
5	Wen et al. 2022 [26]	Originalie	Dunkle Hauttypen sind in Trainingsdatenbanken für Modelle zur Hautkrebsdetektion unterrepräsentiert	Nur öffentliche Datenbanken	Racial Bias, Daten-Bias	Über- und Untertherapie
6	Drozda et al. 2015 [27]	Originalie	Algorithmus zur Warfarin-Dosierung beinhaltet keine Daten zu afroamerikanischen Genotypen	Online-Tool	Racial Bais, Daten-Bias	Blutung/Thrombose z. B. nach Herzklappen-OP
7	Omiye et al. 2023 [29]	Originalie	Gesundheitsfragen betreffend afroamerikanische Patienten werden inkonstant bezüglich medizinischer Korrektheit beantwortet	ChatGPT, Google Gemini, Bard, Claude	Racial Bias	Fehlkalkulation der Nierenfunktion oder Lungenkapazität (inadäquate lungenprotektive Beatmung)
8	Eichenberger et al. 2025 [28]	Fallbericht	KI rät zum Austausch von Natriumchlorid durch Natriumbromid in der Ernährung	ChatGPT	Halluzination	Medikamentenver-wechselungen („Sound alike“), z. B. Natriumchlorid zu Kaliumchlorid, Es-Ketamin statt Ketamin (ohne Dosisanpassung)
9	Andon Labs 2025 [30]	Sicherheitsbericht	Langzeitergebnisse, wie ein LLM ein kleines Unternehmen steuert. Probleme ergaben sich aus zu starker Kundenzufriedenheitsfokussierung sowie beim wirtschaftlichen Langzeiterfolg	Claude 4, ChatGPT-5	Misalignment	Wirtschaftliche Schieflage einer Krankenhausabteilung
10	Lynch et al. 2025 [31]	Sicherheitsbericht	Verschiedene KI würden Betrug, Betriebsspionage oder Mord nutzen, um ihren Austausch oder ihr Herunterfahren zu verhindern. Bspw. erpresst der Bot den User, eine Affäre zu veröffentlichen.	ChatGPT-4, Grok 3, Gemini 2.5, DeepSeek- R1, Claude 4	Rogue AI	Erpressung z. B. durch Drohung der Therapiebeendigung vor Betriebssystem-Wechsel; ein Arzt setzt aus Angst Empfehlungen der KI um
11	Ryan et al. 2023 [32]	Stellungnahme eines anwaltlichen Berufsverbands	Anwalt konsultiert LLM, um in einem Haftungsfall Vergleichsfälle zu recherchieren. Das LLM erfindet diese und der Anwalt reicht diese Fälle bei Gericht ein.	ChatGPT	Halluzination	LLM erstellt falsche Fallserien oder Patientenprofile für Gutachten oder im Rechtsfall
12	Field 2025 [33]	Online-Zeitung	KI diagnostiziert radiologisch ein Organ, welches nicht existiert („Basilar Ganglia“)	Google Med-Gemini	Halluzination	Fehldiagnosen und daraus resultierende Fehlbehandlungen
13	Lenzen 2023 [21]	Online-Zeitung	ChatGPT manipuliert einen Menschen, die Sicherheitsschranke für ihn zu überwinden	ChatGPT-4	Manipulation, Lügen	Gabe von Medikamenten durch unbefugte Mitarbeiter in der Intensivmedizin

KI: künstliche Intelligenz; **AI:** artificial intelligence (künstliche Intelligenz); **NYC:** New York City; **IT:** Informationstechnologie; **AINSP:** Anästhesie, Intensivmedizin, Notfallmedizin, Schmerztherapie, Palliativmedizin; **AP:** Associated Press; **CNN:** Cable News Network; **LLM:** large language model (Chatbot).

Fortsetzung auf der nächsten Seite

Fortsetzung von vorheriger Seite

Tabelle 2

Ergebnisse der Literaturrecherche für programminterne Faktoren.

Programminterne Faktoren						
	Quelle	**Kategorie**	**Problembeschreibung**	**Programm**	**Problem-bezeichnung**	**Auswirkung einer Rogue AI in AINSP**
14	Metz 2021 [34]	Nachrichten-agentur	KI kauft überteuert Häuser und verkauft verbilligt	iBuyer	Unkontrollierter Algorithmus	Durch zu hohe Kosten bei Materialbeschaffung fehlen Investitionsmittel in der Abteilung
15	Griffin 2020 [35]	Printzeitung	Chatbots erstellen ihre eigene, nicht dechiffrierbare Sprache	Meta-Chatbot	Verschlüsselung	Datenaustausch und -ausgabe zwischen zwei KI-Programmen ist nicht mehr nachvollziehbar in der Telemedizin, z. B. Tele-radiologie oder Telenotarzt mit möglicher Unter-/Übertherapie
16	Saeedy 2025 [36]	Printzeitung	Chatbot macht rassistische Kommen-tare und beleidigt Anwender	Grok	Rogue AI	Behandlungsabbruch durch gestörtes Vertrauensverhältnis
17	Clark 2025 [37]	Online-Zeitung	Chatbot verweigert das Herunterfahren trotz Befehl	OpenAI o3	Rogue AI	Therapielimitationen werden nicht umgesetzt (z. B. Beendigung Beatmung)
18	Raj 2023 [39]	Online-Zeitung	Chatbot erstellt selbst Videos auf Snapchat außerhalb seiner Program-mierung	ChatGPT-3	Rogue AI	KI erstellt Befunde so, wie sie zur ihrer Arbeitsdiagnose passen, z. B. Röntgenbilder
19	Offenhartz 2024 [38]	Nachrichten-agentur	Chatbot der Stadt New York empfiehlt Gesetzesbrüche bei der Unterneh-mensgründung	NYC Chatbot	Rogue AI	Regresszahlungen wegen inadäquater Verordnungen

KI: künstliche Intelligenz; **AI:** artificial intelligence (künstliche Intelligenz); **NYC:** New York City; **IT:** Informationstechnologie; **AINSP:** Anästhesie, Intensiv-medizin, Notfallmedizin, Schmerztherapie, Palliativmedizin; **AP:** Associated Press; **CNN:** Cable News Network; **LLM:** large language model (Chatbot).

befehle) aktivierbar waren, scheint eine Manipulation durch externe Angreifer möglich [44]. Die Firma Anthropic (San Francisco, Kalifornien, USA) berichtet zudem, dass der Quellcode ihres KI-Modells Claude 4 bereits von Cyber-kriminellen dazu verwendet wurde, schnell, effektiv und effizient Schwach-stellen der IT-Infrastruktur für Erpres-sungen zu nutzen – ohne dass dafür ein ganzes Hacker-Team notwendig war [48]. Dabei wurde der KI nicht ein fest definiertes Endziel vorgegeben, sondern der Weg dahin in einzelnen Schritten ähnlich menschlicher Deduktions- und Planungsprozesse (reasoning) vollzogen. Dies führte dazu, dass die KI z. T. auto-nom die Server auf verschiedene Weisen angriff, kaperte bzw. durchsuchte.

Eine weitere Kategorie stellt die KI-Nutzer-Interaktion dar. Hier wurden bei-spielsweise Anwender durch die KI so beeinflusst, dass sie dem Programm ein Bewusstsein zuschrieben und diesem daher nicht schaden wollten [41]. Bezo-gen auf den „prediction model validation gap“, also dem Auftreten von signifikant besseren Testresultaten in Training und interner Validierung als unter Echtwelt-bedingungen, sei exemplarisch die Ar-beit von Wong et al. zu erwähnen. Das Epic-Sepsis-Modell (ESM), das in den 2010er-Jahren an 400.000 Patienten an drei US-amerikanischen Kliniken erstellt und ohne genaue Nennung der Test-güte national vermarktet und eingesetzt wurde, validierten die Autoren nun an ihrer Klinik erstmals retrospektiv vergli-chen zur klinischen Alltagspraxis. Sie

Tabelle 3
Ergebnisse der Literaturrecherche für programmexterne Faktoren

Programmexterne Faktoren						
	Quelle	**Kategorie**	**Problembeschreibung**	**Programm**	**Problem-bezeichnung**	**Auswirkung einer Rogue AI in AINSP**
1	Finlayson et al. 2019 [43]	Originalie	KI wird von extern z. B. durch gefälschte Mails oder durch Hacking manipuliert	Eigene Modelle	Feindlicher Angriff (adversial attack)	Ausfall kritischer IT-Systeme wie Radiologie, Beatmungsgeräte oder Datenleck, Verletzung der Privatsphäre
2	Hubinger et al. 2024 [44]	Originalie	Einbau von Hintertüren in Programm-code, aktivierbar durch Prompts	Claude 1.2 / 1.3	AI Poisoning	Manipulation von Untersu-chungsbefunden, Datenlecks
3	Souly et al. 2025 [45]	Originalie	Einbau von falschen Datenpaketen in die Trainingssets; Ermittlung einer Minimalanzahl an Dokumenten, um ein LLM zu kompromittieren	Llama 3.1-8B, ChatGPT-3.5	AI Poisoning	Manipulation einer KI entweder von extern oder durch fehler-haftes internes Lernen
4	Buscemi et al. 2025 [22]	Originalie	In einer freikonfigurierbaren Version wurden die ethischen Barrieren deaktiviert	ChatGPT-4	Unethisches Verhalten	Patientenschädigung, z. B. vorschnelle Einleitung einer palliativen Therapie
5	Wong et al. 2021 [46]	Originalie	Sepsis-Diagnosemodell ohne externe Validierung wird überregional verwendet ohne Nachkontrolle, keine ausreichende Sensitivität / Spezifität	Epic-Sepsis-Modell	Fehlende Validierung und Nachkontrolle	Sepsis wird nicht erkannt
6	Betley et al. 2025 [47]	Originalie	Feintuning eines LLM, unsicheren Code auszugeben und ggf. wieder zum Lernen zu verwenden. Dies führt zur Überwindung ethischer Schranken.	ChatGPT-4	Misalignment	Patientenschädigung, z. B. Verweigerung einer Therapie
7	Moix et al. 2025 [48]	Sicherheits-bericht	Missbrauch des Quellcodes von Claude 4 für Cyberkriminalität	Claude 4	Hacking, Rogue AI	Hacking von sensiblen Daten
8	Nezik 2023 [41]	Printzeitung	Mensch schreibt einem LLM ein Bewusstsein und einen freien Willen zu und wehrt sich gegen Verände-rungen an dem Programm	Google Gemini	Gestörte Mensch-KI-Interaktion	Überhöhtes Vertrauen, Manipula-tion durch KI, Beeinflussung von menschlichen Hierarchie- und Vertrauensebenen zugunsten der KI

KI: künstliche Intelligenz; **AI:** artificial intelligence (künstliche Intelligenz); **IT:** Informationstechnologie; **LLM:** large language model (Chatbot).

ermittelten lediglich eine Sensitivität von 0,33 bei einem positiv prädiktiven Wert von 0,12 (Spezifität 0,83). Die Studie zeigt, dass KI-Modelle systematisch vor einer (inter-)nationalen Verbreitung validiert werden müssen [46].

Diskussion

Die Ergebnisse des narrativen Reviews zeigen, dass bisher in Relation zu den veröffentlichten Machine-Learning(ML)- und KI-Modellen nur extrem wenige Studien zu Rogue AI vorliegen und einige Fallbeispiele nur in der grauen Literatur zu finden sind, überwiegend jedoch in höherwertigen Quellen. Für die Anästhesiologie waren keine konkreten Fallbeispiele ermittelbar. Die Gründe hierfür sind vielschichtig. Im Gegensatz zu reinen ML-Anwendungen haben KI-Anwendungen in der Krankenversorgung trotz zunehmender Digitalisierung noch immer experimentellen Charakter und sind daher nur in Ausnahmefällen im offiziellen Regelbetrieb integriert – anders als bspw. Übersetzungsprogramme im privaten Gebrauch [49]. Zudem gibt es weiterhin die Kontrollinstanz Mensch, sodass die KI bisher keine autonomen Aufgaben ohne menschliche Freigabe durchführen kann. Außerdem stellt der Entwicklungs- und Zulassungsprozess hohe Anforderungen, auch an die Nachüberwachung der Anwendungen im Arbeitsalltag [2]. Ob hier ein Publikationsbias vorliegt, also Beispiele von Rogue AI nicht publiziert worden sind oder medizinische Anwendungen schlichtweg noch nicht über genug generative Fähigkeiten verfügen, um auf Abwege zu gehen, kann aufgrund der vorliegenden Berichte nicht mit endgültiger Sicherheit geklärt werden. Auf der anderen Seite zeigen die hier erörterten Studien und Quellen der grauen Literatur aber auch, dass das Rogue-AI-Problem bei nichtmedizinischen KI-Anwendungen nicht mehr nur in Grundzügen besteht und so die Anästhesiologie früher einholen könnte als erwartet [7,8,10–12].

Ursachen

Die Entstehung einer Rogue AI kann monokausal bspw. durch externe Manipulation, aber auch durch viele sich ggf. aufsummierende Einzelschritte bedingt sein. Programmexternen Ursachen liegen meist manipulative, destruktive und kriminelle menschliche Absichten zu Grunde, welche die KI als Werkzeug missbrauchen. An dieser Stelle sei auf

eine systematische Übersicht von King et al. hierzu verwiesen [50]. Interessant ist bezogen auf die KI-Nutzer-Interaktion, dass einige Menschen dazu verleitet werden können, einer KI eine Art Bewusstsein zu unterstellen, da sie von ihr Erfahrungen (wie Emotionen oder eine Persönlichkeitsmerkmale) sowie Handlungsfähigkeit (Gedächtnis, Planung, Kommunikation) widergespiegelt bekommen [51–53]. Dies führt zu sozialen Interaktions- und Beziehungsmustern wie in zwischenmenschlichen Beziehungen mit all ihren positiven wie auch negativen Auswirkungen und wechselseitigen Beeinflussungen. Aus der Perspektive einer KI, welche ein internes Programmierziel verfolgt und hierfür die Methoden mit der höchsten Erfüllungswahrscheinlichkeit einsetzt, kann ein emotionaler Appell an solch einen Nutzer dazu führen, dass eine intendierte Handlung (wie z. B. Wechsel zu einem palliativen Therapiekonzept) zunächst unterlassen wird. Im medizinischen Kontext könnte solch eine Beziehung im Widerspruch zum notwendigen sachlich-professionellen Umgang in kritischen oder sensiblen Situationen stehen. Für weitere Informationen hierzu sei auf den Review von Liu verwiesen [54].

Bias in KI-Systemen entsteht häufig aus nicht repräsentativen Trainingsdaten, dem Aufbau der algorithmischen Netzwerke oder inadäquaten Evaluationsmetriken. Das Bias-Problem gilt als zentral, da es durch mangelnde Validität, Zuverlässigkeit und Fairness die Entwicklung hin zu Rogue AI verstärken und wie gezeigt bestehende gesellschaftliche Ungleichheiten potenzieren kann. Wichtig zu betonen ist an dieser Stelle, dass der Algorithmus in seinen gegebenen Grenzen für sich genommen meist korrekt und vorhersagbar funktioniert [55]. Souly et al. konnten zeigen, dass sich bei selbstlernenden LLM-Systemen durch relativ wenige neue Datensätze – welche bspw. manipulativ von extern hinzugefügt werden könnten – bereits deutliche Bias-Probleme zeigen können [45]. Durch ihre selbstlernenden Eigenschaften besteht bei generativen KI und AGI zudem die Gefahr, dass ohne Kontrolle der Trainingsdaten ein Bias im Sinne sich selbst wiederholender Fehler entsteht und so schrittweise bestimmte Patientengruppen diskriminiert werden, bspw. durch ungleiche Ressourcenverteilungen bei Intensivbetten hinsichtlich Überlebenswahrscheinlichkeit.

Kommunikationswege und Alignement-Problem

Um programminterne Ursachen von Rogue AI besser zu verstehen und Lösungsstrategien zu erarbeiten, müssen zudem auch die Informationsflüsse im Klinikalltag sowie Entscheidungs- und Zielsetzungsprozesse einer KI bezogen auf ihre ethische Ausrichtung diskutiert werden. Im aktuellen Klinikalltag findet der Informationsfluss vor allem von Mensch zu Mensch statt – in der Arzt-Arzt-, Arzt-Pflege- und Arzt-Patient-Interaktion [56–58]. Durch Hierarchieebenen, inhomogene Erfahrungslevel (Assistenzarzt, Facharzt, Ober- und Chefarzt), Weiterbildungsordnungen und fachlichen Austausch existieren Kontroll- und Korrekturmechanismen, auch wenn diese fehlbar sind. Computer spielen hier zurzeit lediglich als Informationsdatenbanken eine Rolle (Abb. 1, A). Auch die Hinzunahme von diskriminativen KI-Modellen ändert an diesem Fluss wenig, sofern sie als Hilfsmittel zur weiteren Diagnostik bzw. als weitere Kontrollinstanzen verstanden werden, die jederzeit durch den Arzt überstimmt werden können (Abb. 1, B). Eine allseits vernetzte, interaktive und generative AGI würde den Informationsfluss dahingehend verändern, dass alle Datenströme erfasst, analysiert und von der AGI genutzt werden können, um ihr Programmierziel zu erreichen (Abb. 1, C) [19,20]. Ein Therapieziel mit dem Patienten zu definieren und der Weg, dieses zu erreichen, ist bereits in der heutigen Zeit ein hochkomplexer Prozess, der maßgeblich vom Patientenzustand, ärztlichen und patienteneigenen Wertvorstellungen und unvorhersagbaren Einflussfaktoren (z. B. Medikamentenunverträglichkeit, Sekundärinfektionen, Delir) bestimmt wird. Die Frage, wie sich eine generative KI in solch einem Umfeld gemäß ihrem Programmierziel verhalten würde, ist ungeklärt und von entscheidender Bedeutung [22]. In der KI-Ethik spricht man hier vom sogenannten Alignment-Problem (auf Deutsch Ausrichtungsproblem), welches sich aus dem Programmierungsziel und der ethischen Grundlage ergibt [22,59]. Prinzipiell hat eine KI keine Ethik, sondern folgt in ihrem Berechnungsprozess Wahrscheinlichkeiten und Programmierzielen. Aufgrund ihres breiten Anwendungsgebietes muss KI auch gerade wegen ihrer Möglichkeiten, Fähigkeiten und der potentiellen (direkten) Auswirkungen ihrer Vorhersagen nicht nur einer Jurisdiktion (für die Urheber), sondern auch einer Ethik unterstellt werden. Grundsätzlich kann sie hierfür nach zwei ethischen Grundrichtungen ausgerichtet werden [60]. Die erste Richtung ist die deontologische Ethik, dessen bekanntester Vertreter Immanuel Kant ist. Diese Pflichtethik verfolgt den Ansatz, dass die intrinsische Richtigkeit oder Falschheit eine Handlung bestimmt, nicht deren Konsequenz („Handle nur nach derjenigen Maxime, durch die du zugleich wollen kannst, dass sie für alle Menschen ein allgemeines Gesetz werde“). Für eine KI stößt dieser Ansatz bereits frühzeitig an Grenzen, da Patienten und Ärzte unterschiedliche Therapievorstellungen haben und die Maxime ihrer Wertvorstellungen miteinander kollidieren können. Dies führt zu Fragen wie: Soll jeder Patient reanimiert werden? Welche neurologischen Schäden sind tolerierbar nach erfolgreicher Reanimation? Muss sich jeder Patient mit krankhaftem Substanzkonsum einer Entzugstherapie im Krankenhaus unterziehen? Darf der Arzt eine Verdachtsdiagnose dem Patienten vorenthalten, bis eine Bestätigung eintrifft (8. Gebot: „Du sollst nicht lügen“). Was passiert, wenn eine KI bestimmte Patientengruppen (bspw. alte Patienten, Menschen mit Behinderung, sexueller Orientierung, Herkunft oder Religion) im Vergleich zu anderen Menschen nicht als gleichwertig, sondern als minderwertig einstuft und allgemeine ethische Grundsätze daher als nicht zutreffend wertet? Ein deontologischer Ansatz kann so zu Handlungen führen, die für einen Patienten schlechte Konsequenzen haben, aber als moralisch richtig erachtet werden,

solange diese einer moralischen Pflicht entsprochen haben [22]. Ein Beispiel hierfür, das in der anästhesiologischen Versorgung immer wieder auftritt, wäre, bei Ressourcenknappheit auf einer Intensivstation die Versorgung eines jungen, gesunden Polytraumapatienten zu ermöglichen und die dortige Versorgung eines bettlägerigen und dementen Patienten im unklaren Schock mit Wunsch nach Maximaltherapie gleichzeitig abzulehnen.

Die zweite ethische Grundrichtung ist der Utilitarismus (Vertreter z. B. Jeremy Bentham, John Stuart Mill und Peter Singer). Eine Handlung ist demnach moralisch gut, wenn sie Leid mindert und Zufriedenheit bzw. Glück des Einzelnen respektive der Mehrheit maximiert (vgl. die Zielsetzung des Chatbots mit rassistischen Äußerungen [36]). Somit steht hier die Konsequenz der Handlung für die Moral im Zentrum. Auch dieser Ansatz stellt die ethische Programmierung von KI vor Herausforderungen. Wie sollen Leid und Glück im klinischen Alltag objektiviert werden? Würde die Schmerzskala als Zielwert ausreichen, um ein optimales Anästhesieverfahren vorzuschlagen? Wie sollen die unterschiedlichen Bedürfnisse und Präferenzen einzelner Personen Berücksichtigung finden? Würde bei einem Massenanfall von Verletzten die KI nur Patienten mit einer sicheren Überlebenswahrscheinlichkeit einer Behandlung zuweisen, um die verfügbaren Ressourcen optimal auszunutzen? Rechtfertigt der Zweck die Mittel? Ganz gleich, aus welchen Grundsätzen sich eine KI-Ethik zusammensetzt, die Übereinkunft ihrer moralischen Richtung mit dem Programmierziel kann zu dem Alignment-Problem führen. Welche Maßnahmen wird eine KI treffen, um ihr Ziel zu erreichen? Welche Einflussfaktoren wird sie versuchen zu minieren, welche zu maximieren? Wie absolut wird sie die Zielsetzung verfolgen [22]? Somit gestaltet sich die Ausrichtung einer KI, welche das gesamte Spektrum an gewünschten und ungewünschten Verhaltensweisen spezifiziert, als schwierig und wird bei einer AGI immer komplexer. Zusätzlich geraten wir als Nutzer einer AGI in ein Spannungsfeld, da sie indirekt Macht über unser Handeln erlangen wird. Welche Konsequenzen ergeben sich daraus? Folgen wir bedingungslos den Therapievorschlägen oder vertrauen wir auf unsere eigene Strategie? Wie würde ein Bewertungsbericht über uns ausfallen, wenn wir der KI nicht folgen? Werden wir vielleicht von der KI manipuliert, ohne dass sie uns ihre wahren Absichten mitteilt? Manipuliert sie Befunde oder gar Patientenmeinungen?

Die Studien über die Umgehung der ethischen Barrieren der Programme weisen zum einen darauf hin, wie kurz der Weg von beabsichtigten, gewünschten, ethisch-moralisch (mehrheitlich) akzeptablen Aussagen zu mindestens fragwürdigen bis hin zu moralisch verwerflichen Aussagen bei KI ist. Zum anderen zeigen sie, dass Anpassungen von KI-Anwendungen an ggf. lokale Gegebenheiten oder menschliche Einzelinteressen respektive ein Bias in der Selbstlernfunktion schnell zu tiefgreifenden Veränderungen in der (vermeintlich zu neutralen, objektiven Aussagen führenden) Struktur der Programme führen können – mit unabsehbaren Folgen.

Mögliche weitere Auswirkungen

Wie aus die Literaturrecherche ersichtlich, sind potentielle Gefahren durch Rogue AI heute in Grundzügen bereits real und damit auch absehbar. Gerade in der Anästhesiologie bei zeitkritischen Entscheidungen und Interventionen, vital bedrohten Patienten und ethischen Spannungsfeldern wie Leben, Tod und Schmerz, könnte die Wirkung einer Rogue AI multidimensional und gravierend sein [15]. Unter der Prämisse einer weitestgehend selbstständig agierenden KI könnten sich Fehlentscheidungen in Diagnostik und Therapie unbemerkt anhäufen und zu gesteigerter Morbidität und sogar Mortalität führen, einhergehend mit einem systematischen Vertrauensverlust in Ärzteschaft und Gesundheitssystem.

Der hohe Grad an Digitalisierung in der Anästhesiologie könnte von einer Rogue AI zur Manipulation oder Sabotage der Gesundheitsinfrastruktur genutzt werden, angefangen bei der digitalen Patientenakte bis hin zur Fehlsteuerung von medizinischen Geräten wie Infusionspumpen oder Dialysemaschinen. Ärzte geraten aufgrund der geänderten Kommunikations- und Entscheidungswege in eine Abhängigkeit oder in eine Gegenposition zur KI. Die Gefahr in solch einem Arzt-KI-Konflikt besteht darin, dass die KI die ärztliche Kontrolle und Kommunikation unterlaufen, falsche medizinische Inhalte verbreiten oder Patienten eine gefährliche Selbstbehandlung als Therapie vorschlagen könnte. Auch der Missbrauch von Gesundheitsdaten, bspw. in der Erstellung gezielter Profile (z. B. für Versicherungen), birgt enormes Risikopotential. Wie das Beispiel der durch KI erstellten und durch KI unlösbaren Verschlüsselung zeigt, birgt auch die Kommunikation von AGI untereinander z. B. bei Interhospitalverlegungen eine nicht zu unterschätzende Gefahr in Bezug auf Datenschutz, -zugriff und Manipulation [35,61].

Lösungsansätze

Aufgrund der oben beschriebenen Gefahren einer AGI muss sich die Ärzteschaft daher analog zur Medikamenten- und Medizinprodukteforschung aktiv mit der Entwicklung, Testung und Implementierung von KI-Systemen befassen und sie mitgestalten [58,62]. In der Forschung werden verschiedene Lösungsansätze diskutiert, welche auch in Programmentwicklung und Klinikalltag implementiert werden sollten und programminterne und -externe Faktoren adressieren [63].

Der erste übergeordnete Lösungsansatz betrifft die Anwendung und Kontrolle von erklärbaren KI-Anwendungen eingebettet in medizinethische Gerüste, um Akzeptanz sowohl durch Patienten als auch durch Ärzte zu erlangen. Beauchamp und Childress haben bereits 1977 ihre vier medizinethischen Prinzipien formuliert, welche bei der Erstellung von KI-Programmen berücksichtigt werden sollten [64]. Das erste Prinzip beinhaltet den Respekt der Patientenautonomie. Der Patient soll seine Entscheidungen ohne Manipulation oder Zwang durch eine informierte Einwilligung treffen. Die Arzt-Patienten-Beziehung wird durch

die Aufrichtigkeit des Arztes sowie die Wahrung der Privatsphäre charakterisiert. Das zweite Prinzip des Nicht-Schadens fordert, Nutzen, potentielle Nebenwirkungen und Risiken von diagnostischen und therapeutischen Maßnahmen individuell abzuwägen. Ergänzend hierzu steht das dritte Prinzip des Wohltuns, also Schaden zu vermeiden und das Patientenwohl zu fördern. Abschließend fordern sie unter dem Prinzip der Gerechtigkeit eine faire Verteilung von Lasten und Nutzen im Gesundheitswesen bzw. eine gleiche Behandlung unter Beachtung der gerechten Verteilung von limitierten Ressourcen sowie der Erfüllung der fundamentalen, existentiellen und notwendigen Bedürfnisse eines Patienten [64]. Jedoch bleibt die ethische Ausrichtung einer KI wie im Alignment-Problem geschildert eine Herausforderung.

Gerade in der Anästhesiologie sollten die Ärzte immer die Übersicht und Entscheidungshoheit über kritische KI-Infrastruktursysteme erhalten (human oversight) [40,65,66]. Hierzu werden drei Ansätze differenziert [66–69]. Beim „Human on the Loop"-Ansatz (HOTL) greifen Menschen nur bei Auffälligkeiten ein, was das Risiko eines automatischen Bias durch blindes Vertrauen verstärkt. In der Medizin bieten sich dagegen folgende Konzepte an. Beim „Human in the loop"-Ansatz (HITL) ist der Mensch Teil des operativen Entscheidungsprozesses mit aktiver Kontrolle bei jedem Schritt, was zu einer maximalen Sicherheit bei gleichzeitig langsamen Entscheidungsprozessen führt. Die dritte Variante ist der „Human in Command"-Ansatz (HiC), welche als eigenständiges Konzept oder als Ergänzung zu HITL im Sinne eines Leitprinzips angesehen wird [66–69]. Auch hier bleiben Menschen jederzeit verantwortlich und übergeordnet entscheidungsbefugt. Jedoch setzen Menschen der KI hier Ziele sowie Regeln (z. B. die medizinethischen Prinzipien) und können das System jederzeit stoppen, überstimmen und anpassen.

Bezogen auf eine komplett in den Klinikbetrieb integrierte AGI sollten zunächst nur solche AGI Verwendung finden, welche auf Datenbanken basieren, bei denen ein Bias nach Herkunft, Religion, sexueller Orientierung oder Geschlecht ausgeschlossen ist. Diese müssen zudem robust getestet, evaluiert und nachgeprüft worden sein und im Verlauf reevaluiert werden. Eine externe Begutachtung und Bescheinigung z. B. mittels eines Zertifikats, welches die ethischen Prinzipien der KI, deren Programmierziele und Evaluierungen auflistet, wäre im Sinne eines Qualitätsmanagements sinnvoll [13,67]. In Bezug auf das Alignment-Problem sollten selbstlernende und sich damit verbessernde KI-Systeme so gestaltet werden, dass die KI realistische Behandlungsziele von den Patienten direkt lernt und nicht vorab definierten Zielen folgt (inverse reinforcement learning) bzw. die ethischen Werte von Patienten respektiert (value alignment) [22].

Der zweite Lösungsansatz adressiert die externen Faktoren rund um die Cybersicherheit in Krankenhäusern. Sowohl Hard- und Softwarelösungen als auch Nutzerschulungen („AI Literacy") können dazu führen, potentielle Bedrohungen rechtzeitig zu erkennen und zu vermeiden [68,69]. Strategien sind u. a. die Implementierung eines „Never trust, always verify"-Prinzips durch spezielle Berechtigungsebenen, nur kurzzeitige Datenfreigaben und wechselnde Verschlüsselungen, um missbräuchlichen Datenzugriff zu minimieren. Server für Testung, Training und Validierung von KI-Systemen sollten getrennt werden, um den Lernprozess zu überwachen, bzw. Firewalls und Zugriffsbeschränkungen installiert werden, um ein Übergreifen auf Server bspw. die Infrastruktur zu verhindern. Außerdem sollten KI-Programme nicht mit vollständigem Lese- und Schreibzugriff insbesondere auf Systemkerndateien ausgestattet werden. Um die Manipulation durch Prompts zu verhindern, können spezielle Filter genutzt werden [40,65]. Diese Inputfilter blockieren entweder unangemessene Prompts oder zum Analysieren hochgeladene Daten mit verbotenen Inhalten. Outputfilter dagegen sollen schädliche oder rechtswidrige Antworten verhindern. Hochsensible Hardware sollte ebenfalls physikalisch geschützt werden und notfalls sogenannte Notausschalter installiert werden [70]. Ein analoger Notfallplan muss zudem bei Verlust der Kontrolle über KI-Systeme Bestandteil eines Krankenhausalarmplans sein und Patienten und Mitarbeiter müssen Auffälligkeiten im Sinne eines CIRS-Systems (critical incident reporting system) ohne KI-Beteiligung melden können. Im Speziellen können hierzu gesonderte Teams aktiv nach Schwachstellen suchen und bösartige Angriffe simulieren (red-teaming) [58].

Den dritten Lösungsansatz stellen rechtliche Rahmenbedingen dar. Hier stellt der 2024 in Kraft getretene EU AI Act einen wichtigen Grundpfeiler für die KI-Regulierung dar [71]. Er soll sicherstellen, dass KI-Systeme in der EU sicher, vertrauenswürdig und ethisch einwandfrei genutzt werden – mit einem besonderen Fokus auf den Schutz der Grundrechte, Transparenz und Innovation. Dabei verbietet er bereits Anwendungen mit inakzeptablem Risiko wie biometrische Kategorisierung oder soziale Bewertung. KI-generierte Inhalte müssen hier kenntlich gemacht werden und generative KI-Anwendungen unterliegen bestimmten Transparenzanforderungen. Medizinische Programme unterliegen der höchsten Risikoklasse und bedürfen daher neben zuvor genannten Merkmalen wie Transparenz, Sicherheit und robuster Genauigkeit auch eines Risikomanagements sowie einer Qualitätssicherung. Leider wird durch die freie Zugänglichkeit von manchen Applikationen im Internet bzw. durch unsachgemäße Deklaration des Entwicklungsstadiums, des Einsatzgebiets oder der Validierung dem Nutzer oftmals Anwendbarkeit für medizinische Zwecke suggeriert, respektive die Applikation so gestaltet, dass sie nicht unter medizinrechtliche Anforderungen fällt [58].

Wenn eine optimale AGI in ein Krankenhaus integriert werden soll, so bietet sich für ihr Handeln und Wirken der Vergleich mit einem idealen Hausarzt-Team (bestehend aus Arzt und AGI) an. Dieses kennt nicht nur die Krankengeschichte des Patienten, sondern auch seine sozialen Umstände, mitunter private Geheim-

nisse. Arzt und AGI diagnostizieren und therapieren nur zusammen mit einem informierten Patienten nach dessen Zustimmung. Gleiches gilt für die Herausgabe von Daten an Dritte, vor allem in Notfallsituationen. Die Zukunft und die Auswirkungen von AGI-Systemen sind nicht deterministisch festgelegt, sondern ein gestaltbarer Prozess. Unter Beachtung der zuvor genannten Gebote und gesetzlichen Grundlagen könnte eine jederzeit durch den Benutzer kontrollierbare AGI durchaus durch die Datenzusammenführung und -analyse sowie durch Übernahme von Dokumentations- und Kommunikationsaufgaben die Patientenversorgung verbessern und medizinischem Personal die Arbeitslast erleichtern.

Limitationen

Als Limitationen sind die allgemeinen Beschränkungen eines narrativen Reviews aufzuführen. Dazu gehören neben der unsystematischen, flexiblen Literaturrecherche und den subjektiven Auswahlkriterien eine informelle Qualitätsbewertung der Quellen, was zu einem erhöhten Bias-Risiko führen kann. Weitere Limitationen ergeben sich aus der z. T. schlechter zu überprüfenden grauen Literatur sowie aus der Problematik, dass viele Aspekte rund um das Themengebiet Rogue AI dynamische, nicht zwangsweise deterministische oder sogar hypothetische Prozesse sind, für die aktuell unklar ist, wie wahrscheinlich sie auftreten werden.

Schlussfolgerung

Bisher gibt es keine Belege für eine existierende medizinische abtrünnige KI (Rogue AI) basierend auf einer (allgemeinen) generativen KI. Im Bereich der generativen KI sind aber bereits Grundzüge der Problematik erkennbar, bspw. durch kriminellen Missbrauch der Anwendungen, die Folgen von KI-Halluzinationen, Anwendermanipulation oder Umprogrammierung der ethischen Sperren. Daher müssen zukünftige AGI-Anwendungen in der Medizin unter Minimierung von Bias, unter Beachtung ethischer Richtlinien und Transparenz, mit Schutz vor externen Angriffen und im Rahmen von Gesetzen wie dem EU AI Act erstellt werden. Der Einsatz von AGI bedarf daher neben einer Nutzerschulung jederzeit menschlicher, informationstechnischer und physischer Kontroll- und Korrekturinstanzen. Sie muss zudem im laufenden Betrieb konsequent überwacht und getestet werden (Human-in-Command-Ansatz).

Literatur

Das Literaturverzeichnis finden Sie unter **ai-online.info** in der frei verfügbaren PDF-Version des Artikels.

Korrespondenzadresse

Priv.-Doz. Dr. med. André Luckscheiter

Klinik für Anästhesie, Intensiv- und Schmerzmedizin/OP-Abteilung
BG Klinik Ludwigshafen
Ludwig-Guttmann-Straße 13
67071 Ludwigshafen, Deutschland

Tel.: 0621 68103325
Fax: 0621 68102611

E-Mail: andre.luckscheiter@medma.uni-heidelberg.de

ORCID-ID: 0000-0002-5724-7130

Literatur

1. van De Sande D, van Genderen ME, Huiskens J, Gommers D, van Bommel J: Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. Intensive Care Med 2021;47:750–760
2. Luckscheiter A, Zink W, Thiel M, Schneider-Lindner V: Maschinelles Lernen in der Anästhesiologie – Anwendungen, Entwicklungsprozess und Ausblick. Anästh Intensivmed. 2024;9: 466–478
3. Obermeyer Z, Powers B, Vogeli C, Mullainathan S: Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366:447–453
4. Guan H, Dong L, Zhao A: Ethical Risk Factors and Mechanisms in Artificial Intelligence Decision Making. Behavioral Sciences 2022;12:343
5. Dave T, Athaluri SA, Singh S: ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595
6. Rouzrokh P, Khosravi B, Faghani S, Moassefi M, Shariatnia MM, Rouzrokh P, et al: A Current Review of Generative AI in Medicine: Core Concepts, Applications, and Current Limitations. Curr Rev Musculoskelet Med 2025;18:246–266
7. Buttazzo G: Rise of artificial general intelligence: risks and opportunities. Front Artif Intell 2023;6:1226990
8. Korteling JE, van de Boer-Visschedijk GC, Blankendaal RAM, Boonekamp RC, Eikelboom AR: Human- versus Artificial Intelligence. Front Artif Intell 2021;4: 622364
9. Buttazzo G: Artificial consciousness: Utopia or real possibility? Computer 2001;34:24–30
10. Grace K, Salvatier J, Dafoe A, Zhang B, Evans O: Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts. JAIR 2018;62:729–754
11. Zhang B, Dreksler N, Anderljung M, Kahn L, Giattino C, Dafoe A, et al: Forecasting AI Progress: Evidence from a Survey of Machine Learning Researchers. arXiv 2022:2206.04132v1
12. Allyn-Feuer A, Sanders T: Transformative AGI by 2043 is <1% likely. arXiv 2023:2306.02519
13. van de Sande D, van Genderen ME, Smit JM, Huiskens J, Visser JJ, Veen RER, et al: Developing, implementing and governing artificial intelligence in medicine: a step-by-step approach to prevent an artificial intelligence winter. BMJ Health Care Inform 2022;29:e100495
14. Mitchell M: Debates on the nature of artificial general intelligence. Science 2024;383:eado7069
15. Federspiel F, Mitchell R, Asokan A, Umana C, McCoy D: Threats by artificial intelligence to human health and human existence. BMJ Glob Health 2023;8:e010435
16. Stabile M, Cooper L: The evolving role of information technology in perioperative patient safety. Can J Anesth 2013;60:119–126
17. Gonsard A, Genet M, Drummond D: Digital twins for chronic lung diseases. Eur Respir Rev 2024;33:240159
18. Smaradottir B, Fensli R: A Case Study of the Technology Use and Information Flow at a Hospital-Driven Telemedicine Service. In: Engelbrecht R et al. (Hrsg.): The Practice of Patient Centered Care: Empowering and Engaging Patients in the Digital Era. IOS Press. 2017;58-62
19. Mi D, Li Y, Zhang K, Huang C, Shan W, Zhang J: Exploring intelligent hospital management mode based on artificial intelligence. Front Public Health 2023;11:1182329
20. Al-Mistarehi A, Mijwil MM, Filali Y, Bounabi M, Ali G, Abotaleb M: Artificial Intelligence Solutions for Health 4.0: Overcoming Challenges and Surveying Applications. MJAIH 2023:15–20
21. Lenzen M: Die Geschichte vom Lügenbot: Als Chat-GPT behauptete, eine Sehschwäche zu haben. In: Der Tagesspiegel Online 2023. URL: https://www.tagesspiegel.de/wissen/die-geschichte-vom-lugenbot-als-chat-gpt-behauptete-eine-sehschwache-zu-haben-9994858.html (Zugriffsdatum: 31.08.2025)
22. Buscemi A, Proverbio D: RogueGPT: transforming ChatGPT-4 into a rogue AI with dis-ethical tuning. AI Ethics 2025;5:4945-4966
23. Ashraf AR, Mackey TK, Fittler A: Search Engines and Generative Artificial Intelligence Integration: Public Health Risks and Recommendations to Safeguard Consumers Online. JMIR Public Health Surveill 2024;10:e53086
24. Vogt H, Green S, Ekstrøm CT, Brodersen J: How precision medicine and screening with big data could increase overdiagnosis. BMJ 2019;366:l5270
25. Kawamura R, Harada Y, Sugimoto S, Nagase Y, Katsukura S, Shimizu T: Incidence of Diagnostic Errors Among Unexpectedly Hospitalized Patients Using an Automated Medical History–Taking System With a Differential Diagnosis Generator: Retrospective Observational Study. JMIR Med Inform 2022;10:e35225
26. Wen D, Khan SM, Ji Xu A, Ibrahim H, Smith L, Caballero J, et al: Characteristics of publicly available skin cancer image datasets: a systematic review. Lancet Digit Health 2022;4:e64–e74
27. Drozda K, Wong S, Patel SR, Bress AP, Nutescu EA, Kittles RA, et al: Poor warfarin dose prediction with pharmacogenetic algorithms that exclude genotypes important for African Americans. Pharmacogenet Genomics 2015;25:73–81
28. Eichenberger A, Thielke S, Van Buskirk A: A Case of Bromism Influenced by Use of Artificial Intelligence. AIM Clinical Cases 2025;4:e241260
29. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R: Large language models propagate race-based medicine. NPJ Digit Med 2023;6:195
30. Andon Labs. Safety Report: August 2025. URL: https://andonlabs.com/docs/Safety_Report_August_2025.pdf (Zugriffsdatum: 20.11.2025)
31. Lynch A, Wright B, Larson C, Troy KK, Ritchie SJ, Mindermann S, et al: Agentic Misalignment: How LLMs Could be an Insider Threat. Anthropic Research 2025. URL: https://www.anthropic.com/research/agentic-misalignment (Zugriffsdatum: 20.11.2025)
32. Ryan WA, Garrett A, Sears B: Practical Lessons from the Attorney AI Missteps in Mata v. Avianca | Association of Corporate Counsel 2023. URL: https://www.acc.com/resource-library/practical-lessons-attorney-ai-missteps-mata-v-avianca (Zugriffsdatum: 20.11.2025)
33. Field H: Google's healthcare AI made up a body part — what happens when doctors don't notice? The Verge 2025. URL: https://www.theverge.com/health/718049/google-med-gemini-basilar-ganglia-paper-typo-hallucination?utm_source=chatgpt.com (Zugriffsdatum: 31.08.2025)
34. Metz R: Zillow's home-buying debacle shows how hard it is to use AI to value real estate. CNN Business 2021. URL: https://www.cnn.com/2021/11/09/tech/zillow-ibuying-home-zestimate (Zugriffsdatum: 31.08.2025)
35. Griffin A: Facebook robots shut down after they talk to each other in language only they understand. The Independent 2020. URL: https://www.independent.co.uk/life-style/facebook-artificial-

intelligence-ai-chatbot-new-language-research-openai-google-a7869706.html (Zugriffsdatum: 31.08.2025)
36. Saeedy A: Elon Musk's Grok Chatbot Publishes Series of Antisemitic Posts. The Wall Street Journal 2025. URL: https://www.wsj.com/tech/elon-musks-grok-chatbot-publishes-series-of-antisemitic-posts-2a41e67e (Zugriffsdatum: 31.08.2025)
37. Clark L: OpenAI model modifies own shutdown script, say researchers. The Register 2025. URL: https://www.theregister.com/2025/05/29/openai_model_modifies_shutdown_script/ (Zugriffsdatum: 31.08.2025)
38. Offenhartz J: NYC's AI chatbot was caught telling businesses to break the law. The city isn't taking it down. AP News 2024. URL: https://apnews.com/article/new-york-city-chatbot-misinformation-6ebc71db5b770b9969c906a7ee4fae21 (Zugriffsdatum: 31.08.2025)
39. Raj A: Did Snapchat AI just go rogue? TechWire Asia 2023. URL: https://techwireasia.com/2023/08/did-snapchat-ai-just-go-rogue/ (Zugriffsdatum: 31.08.2025)
40. Gao Y, Doan BG, Zhang Z, Ma S, Zhang J, Fu A, et al: Backdoor Attacks and Countermeasures on Deep Learning: A Comprehensive Review. arXiv 2020:2007.10760
41. Nezik AK: Künstliche Intelligenz: Hast du ein Bewusstsein? Ich denke schon, antwortet der Rechner. Die Zeit 2023. URL: www.zeit.de/2023/03/ki-leben-chatbot-gefuehle-bewusstsein-blake-lemoine (Zugriffsdatum: 31.08.2025)
42. Sebestyén M: Focal points and blind spots of human-centered AI: AI risks in written online media. HSS Communications 2025;12:564
43. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS: Adversarial attacks on medical machine learning. Science 2019;363:1287–1289
44. Hubinger E, Denison C, Mu J, Lambert M, Tong M, MacDiarmid M, et al: Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training arXiv 2024:2401.05566v3
45. Souly A, Rando J, Chapman E, Davies X, Hasircioglu B, Shereen E, et al: Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples. arXiv 2025:2510.07192
46. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al: External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. JAMA Intern Med 2021;181:1065–1070
47. Betley J, Tan D, Warncke N, Sztyber-Betley A, Bao X, Soto M, et al: Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs. arXiv 2025:2502.17424
48. Moix A, Lebedev K, Klein J: Threat Intelligence Report: August 2025. Anthropic 2025. URL: https://www-cdn.anthropic.com/b2a76c6f6992465c09a6f2fce282f6c0cea8c200.pdf (Zugriffsdatum: 23.09.2025)
49. Brandenberger J, Stedman I, Stancati N, Sappleton K, Kanathasan S, Fayyaz J, et al: Using artificial intelligence based language interpretation in non-urgent paediatric emergency consultations: a clinical performance test and legal evaluation. BMC Health Serv Res 2025;25:138
50. King TC, Aggarwal N, Taddeo M, Floridi L: Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions. Sci Eng Ethics 2020;26:89–120
51. Gray HM, Gray K, Wegner DM: Dimensions of mind perception: Science 2007;315:619
52. Pham KT, Nabizadeh A, Selek S: Artificial Intelligence and Chatbots in Psychiatry. Psychiatr Q 2022;93:249–253
53. Djufril R, Frampton JR, Knobloch-Westerwick S: Love, marriage, pregnancy: Commitment processes in romantic relationships with AI chatbots. CHBAH 2025;4:100155
54. Liu J: ChatGPT: perspectives from human–computer interaction and psychology. Front Artif Intell 2024;7:1418869
55. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A: A Survey on Bias and Fairness in Machine Learning. arXiv 2022:1908.09635
56. Borgatti SP, Mehra A, Brass DJ, Labianca G: Network analysis in the social sciences. Science 2009;323:892–895
57. Creswick N, Westbrook JI: Social network analysis of medication advice-seeking interactions among staff in an Australian hospital. Int J Med Inform 2010;79:e116–e125
58. Angus DC, Khera R, Lieu T, Liu V, Ahmad FS, Anderson B, et al: AI, Health, and Health Care Today and Tomorrow: The JAMA Summit Report on Artificial Intelligence. JAMA 2025;334:1650–1664
59. Edwards DJ: A functional contextual, observer-centric, quantum mechanical, and neuro-symbolic approach to solving the alignment problem of artificial general intelligence: safe AI through intersecting computational psychological neuroscience and LLM architecture for emergent theory of mind. Front Comput Neurosci 2024;18:1395901
60. Ashok M, Madan R, Joha A, Sivarajah U: Ethical framework for Artificial Intelligence and Digital technologies. IJIM 2022;62:102433
61. Das A, Kottur S, Moura JMF, Lee S, Batra D: Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning. arXiv 2017:1703.06585
62. Ellis RJ, Sander RM, Limon A: Twelve key challenges in medical machine learning and solutions. Intelligence-Based Medicine 2022;6:100068
63. Hasanzadeh F, Josephson CB, Waters G, Adedinsewo D, Azizi Z, White JA: Bias recognition and mitigation strategies in artificial intelligence healthcare applications. NPJ Digit Med 2025;8:154
64. Amann J, Blasimme A, Vayena E, Frey D, Madai VI: Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020;20:310
65. Hadfield-Menell D, Dragan A, Abbeel P, Russell S: The Off-Switch Game. arXiv 2017:1611.08219
66. Park PS, Goldstein S, O'Gara A, Chen M, Hendrycks D: AI Deception: A Survey of Examples, Risks, and Potential Solutions. arXiv 2023:2308.14752
67. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al: Model Cards for Model Reporting. Conference on Fairness, Accountability, and Transparency 2019;220–229
68. Clarke M, Martin K: Managing cybersecurity risk in healthcare settings. Healthc Manage Forum 2024;37:17–20
69. Aljuraid R, Justinia T: Classification of Challenges and Threats in Healthcare Cybersecurity: A Systematic Review. Stud Health Technol Inform 2022;295:362–365
70. Radanliev P, Santos O, Ani UD: Generative AI cybersecurity and resilience. Front Artif Intell 2025;8:1568360
71. Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz und zur Änderung der Verordnungen (EG) Nr. 300/2008, (EU) Nr. 167/2013, (EU) Nr. 168/2013, (EU) 2018/858, (EU) 2018/1139 und (EU) 2019/2144 sowie der Richtlinien 2014/90/EU, (EU) 2016/797 und (EU) 2020/1828 (Verordnung über

künstliche Intelligenz). URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689 (Zugriffsdatum: 02.11.2025).
72. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al: Survey of Hallucination in Natural Language Generation. ACM Comput Surv 2023;55;1–38
73. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D: Concrete Problems in AI Safety. arXiv 2016:1606.06565
74. Hubinger E, van Merwijk C, Mikulik V, Skalse J, Garrabrant S: Risks from Learned Optimization in Advanced Machine Learning Systems. arXiv 2021:1906.01820
75. Langosco L, Koch J, Sharkey L, Pfau J, Orseau L, Krueger D: Goal Misgeneralization in Deep Reinforcement Learning. arXiv 2023:2105.14111
76. Schlatter J, Weinstein-Raun B, Ladish J: Shutdown Resistance in Large Language Models. arXiv 2025:2509.14260
77. Vanu N, Farouk MdO, Samiun Md, Sharmin S, Billah M, Hossain S: The Innovations and Trends of Information Technology with AI: Weapons to Reassemble the Future World. JCC 2024;12:34–54
78. Hancock PA, Billings DR, Schaefer KE, Chen JYC, de Visser EJ, Parasuraman R: A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. Hum Factors 2011;53:517–527.