Von der Diagnose zur Therapie: KI in der medizinischen Bildgebung

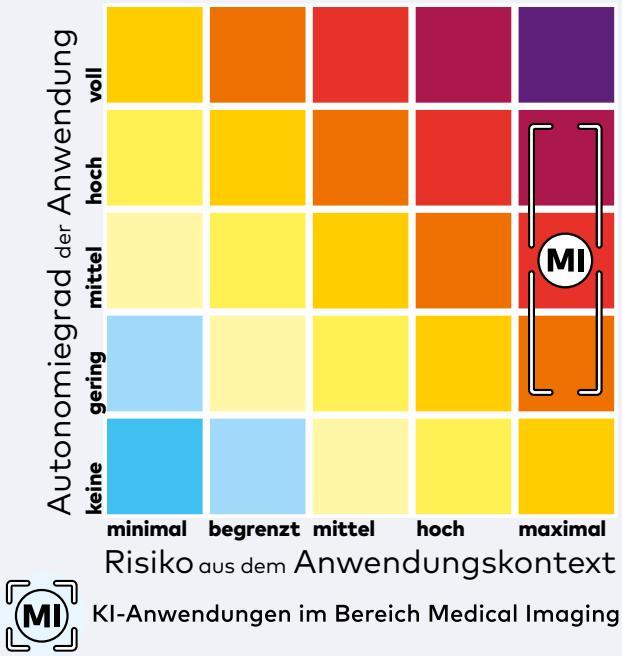
Über die Chancen und Risiken KI-basierter Diagnose- und Therapieunterstützung in der Medizin (Medical Imaging).

Executive Summary

Im Gesundheitssystem wird die fortgeschrittene Fähigkeit von KI-Systemen, Bilder zu erkennen und zuzuordnen, für die Erstellung von Diagnosen und in der Therapie eingesetzt. Zum Beispiel wird Bilderkennung für die Auswertung von Netzhautscans in der Augenheilkunde, für die Analyse digitalisierter Gewebeschnitte in der Pathologie oder zur Erkennung von Tumoren wie etwa Mammakarzinomen eingesetzt. Im vorliegenden ALAIT Risikoradar wird der Einsatz dieser Art von KI-Anwendungen in der medizinischen Bildgebung insgesamt mit einem relativ hohen Gesamtrisiko eingestuft, was sich durch die Einordnung im roten Farbbereich der Grafik ablesen lässt. Gesundheitsdienstleister:innen können das Risiko allerdings reduzieren, wenn sie die in diesem Dossier enthaltenen Empfehlungen zur Risikoabschätzung und -reduzierung umsetzen (s. S. 4).

Im Detail berücksichtigt die Klassifikation von KI-Anwendungen im ALAIT Risikoradar zwei Dimensionen:
1) Das Risiko aus dem Anwendungskontext: Hier besteht „maximales Risiko" (Stufe 5), da bei fehlerhaften Systemen und Anwendungen lebensbedrohliche Auswirkungen für betroffene Menschen entstehen können. Zusätzlich sind laut EU KI-Gesetz/EU AI-Act Annex I Medizinprodukte generell im Hochrisikobereich angesiedelt. Bis zum vollständigen Inkrafttreten der Europäischen KI-Verordnung mit Mitte 2027 müssen KI-basierte Medizinprodukte allerdings noch keine spezifischen Qualitätsanforderungen erfüllen. Diese intransparente und unklare Situation macht es für Gesundheitsdienstleister:innen aktuell noch schwieriger, die Leistungsfähigkeit dieser Werkzeuge einzuschätzen und damit deren Einsatz besonders riskant.
2) Als zweite Dimension wird der **Autonomiegrad von KI-Anwendungen in der bildbasierten Diagnose- und Therapieunterstützung** bewertet. Dieser variiert in Abhängigkeit der menschlichen Kontrolle einzelner konkreter Anwendungen. Wenn KI-Systeme weitgehend automatisiert Diagnosen und/oder Therapievorschläge erstellen, die nur in definierten Fällen von medizinischem Fachpersonal überprüft werden, ist der Autonomiegrad als hoch einzustufen („Human in Control", z. B. wenn im Rahmen eines KI-gestützten Brustkrebs-Screenings nur unklare oder kritische Fälle von medizinischem Fachpersonal bewertet werden). Das ist zwar technisch möglich, allerdings ist ein solch hoher Automatisierungsgrad im österreichischen Kontext aufgrund der Letztentscheidung durch Menschen, z. B. Ärzt:innen bei der Diagnose, nicht möglich. Der Autonomiegrad ist niedrig, wenn KI-Systeme als unterstützende Tools bei Diagnose und/oder Therapieplanung dienen („Human-in-the-Loop" oder „Human-on-the-Loop"), etwa wenn medizinisches Fachpersonal fallspezifische Bildanalysen erstellen lässt.

ALAIT Risikoradar für bildbasierte Diagnose- und Therapieunterstützung



MI KI-Anwendungen im Bereich Medical Imaging

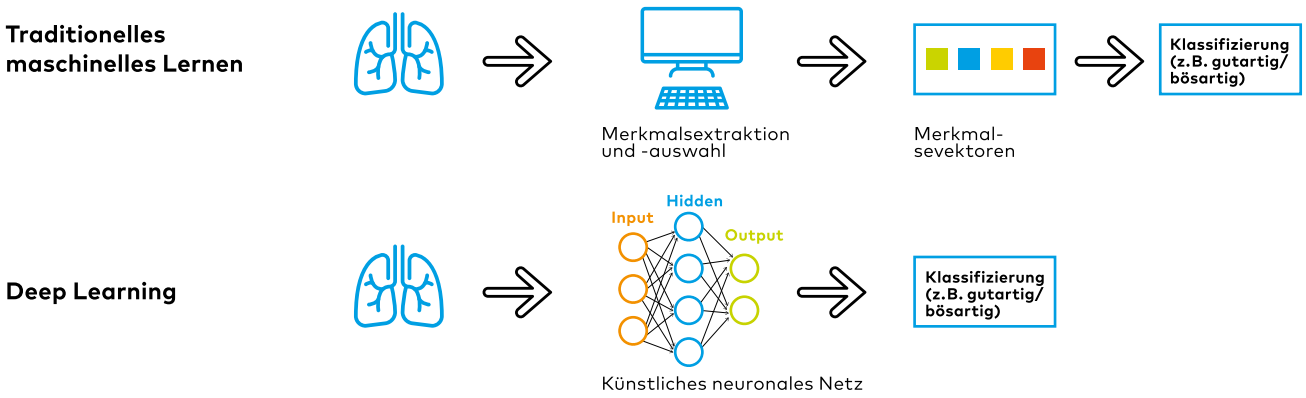
Das ALAIT Risikoradar ist ein wissenschaftlich entwickeltes Risikoanalysetool für Künstliche Intelligenz (KI), das KI-Anwendungen kontextbezogen und unter Berücksichtigung ihres technischen Autonomiegrades einstuft und so die Risikosphäre für Anwender:innen auf einen Blick sichtbar macht. Dabei gilt: Je höher das Einsatz-Risiko aus dem Anwendungskontext und je größer der Autonomiegrad des KI-Systems in Bezug auf Entscheidungen, desto riskanter ist der Einsatz einzustufen. Eine erweiterte Klammer weist auf eine Bandbreite in der Risikoeinstufung hin. Ein geringer Autonomiegrad eines KI-Systems bedeutet nicht, dass man sich zurücklehnen kann. Es erfordert eine starke Rolle der Menschen, die es anwenden. (Details zum Stufenmodell s. S. 7f).

**ALAIT ist ein vom Bundesministerium für Innovation, Mobilität und Infrastruktur (BMIMI) initiiertes Forschungsprojekt zur Entwicklung einer neuartigen Form von Technologiebewertung für KI-Technologien mit wissenschaftlichen Methoden.*

Computer Vision: Der Weg zum maschinellen Sehen

Zeitgemäße medizinische Diagnostik und Therapie setzten zunehmend auf KI. Insbesondere im Bereich der bildbasierten Diagnose- und Therapieunterstützung wurden in den vergangenen Jahrzehnten große technische Fortschritte erzielt.[1] Aktuelle wissenschaftliche Studien deuten darauf hin, dass durch den Einsatz von KI-basierten Werkzeugen im Bereich der Bild- und Videoanalyse in Bereichen der Radiologie, Pathologie, Dermatologie oder Chirurgie genauere, und womöglich sogar frühere Diagnosen bzw. individuellere Therapieansätze mit der Erwartung präziseren Diagnosen möglich sind.[2,3,4]

Technisch gesehen sind Anwendungen aus diesem Bereich der sog. Computer Vision zuzuordnen; jenes Teilgebiet der KI, das sich damit befasst, Bilder und Videos zu erkennen, zu verstehen und zu interpretieren – ähnlich wie der Mensch mit seinen Augen und dem Gehirn. KI-Systeme, die im Bereich der bildbasierten Diagnose- und Therapieunterstützung eingesetzt werden, verarbeiten Bilddaten in verschiedenen Schritten, wie in der folgenden Grafik schematisch dargestellt wird:[5,6,7]



Bei der automatischen Analyse von Bildern/Videos wird vor allem zwischen traditionellem maschinellen Lernen und Deep Learning unterschieden. Beim traditionellen maschinellen Lernen werden aus den Rohdaten relevante Eigenschaften (z.B. Form, Farbe, Textur) manuell – also vom Menschen – extrahiert und dem Algorithmus zur Verfügung gestellt. Diese Merkmalsextraktion (feature selection) ist meist regelbasiert und erfordert Fachwissen von Expert:innen. Danach kann ein Algorithmus die eigentliche Lernaufgabe übernehmen, zum Beispiel die Klassifikation „Tumor/kein Tumor" im Rahmen des überwachten Lernens (***supervised learning***). Hierbei trainiert das Modell mit vorab gelabelten Daten, bei denen die richtige Antwort bekannt ist (z.B. „Tumor / kein Tumor"). Im Gegensatz dazu arbeitet unüberwachtes Lernen (***unsupervised learning***) mit ungelabeten Daten: Das Modell sucht eigenständig nach Mustern oder Gruppen, zum Beispiel indem es ähnliche Zelltypen in Clustern zusammenfasst, ohne vorherige Vorgabe. Weil beim traditionellem maschinellem Lernen klar ist, welche Merkmale zur Vorhersage geführt haben, sind diese Systeme meist besser erklärbar und Entscheidungen nachvollziehbar.

Deep Learing hingegen basiert auf künstlichen neuronalen Netzen und lernt relevante Merkmale und Muster direkt und automatisch aus den Rohdaten – ohne dass Menschen diese vorher explizit definieren müssen. Besonders bei unstrukturierten Daten wie Bildern kann das z.B. Effizienz-Vorteile bieten. Deep Learning kann auch überwacht eingesetzt werden, wenn die Rohdaten bereits mit Labels versehen sind (z.B. Klassifizierung mit bereits gelabelten Bildern) oder aber unüberwacht, wenn die Rohdaten ohne Labels sind (z.B. automatische Clusterbildung). Jedenfalls werden für Deep Learning große Datenmengen und viel Rechenleistung benötigt. Die Entscheidungsfindung dieser Modelle ist oft schwer nachvollziehbar, weshalb man hier von ***Black-Box-Systemen***[1] spricht: Es ist meist unklar, warum das Modell eine bestimmte Entscheidung (z.B. Klassifizierung) getroffen hat, was die Erklärbarkeit einschränkt.

Chancen

Die Anwendung von KI in der bildbasierten Diagnose- und Therapieunterstützung eröffnet große Chancen für Patient:innen, Gesundheitspersonal und das Gesundheitssystem und lässt positive Auswirkungen auf diese Gruppen erwarten.[2,8,9,10] Für Patientinnen stehen vor allem bessere medizinische Leistungen im Vordergrund, für das medizinische Fachpersonal und die Verwaltung effizientere und beschleunigte Prozesse.

Chancen für Patient:innen

- **Früherkennung von Erkrankungen in oft früheren Stadien als mit traditionellen Methoden** (z.B. Tumore[1] und diabetische Retinopathie). Darüber hinaus konnte die Zahl an übersehenen Mammakarzinomen durch KI-Assistenz um 26% gesenkt werden .
- **Personalisierte Diagnostik,** z.B. Unterstützung bei der Vorhersage des individuellen Krankheitsverlaufs.

Chancen für Med. Fachpersonal

- **Beschleunigung der Befunderstellung** durch schnelle Bild- und Videoanalyse sowie Priorisierung dringender Fälle mit Systemen wie CHIEF, die eine Genauigkeit von ca. 94% aufweisen.[12]
- **Schaffung von personellen Zeitressourcen**.

Chancen auf Verwaltungsebene

- **Entlastung des medizinischen Personals** (RadiologInnen, PathologInnen etc.) um bis zu 62% durch die automatisierte Auswertung großer Mengen an Bildmaterial.[13]

Herausforderungen und Risiken

Trotz der obig genannten Chancen gibt es eine Reihe an detailreichen Herausforderungen, die für eine vertrauenswürdige und sichere Integration von bildbasierten KI-Diagnose- und Therapiesystemen in die Medizin zu überwinden sind:

1. Genauigkeit und Verlässlichkeit der Ergebnisse:

Die Leistungsfähigkeit von bildgebenden KI-Systemen in der klinischen Praxis ist nicht immer gesichert gegeben. Das heißt, es besteht ein Risiko für Falschergebnisse. Dies liegt unter anderem daran, dass die in pre-market Tests erzielte Genauigkeit nur dann erreicht werden kann, wenn die Patient:innenpopulation des jeweiligen Einsatzgebietes dem Datensatz, mit dem die KI trainiert wurde, entspricht. Wenn sich die Charakteristika der Patient:innen, wie etwa Geschlecht, Alter, Ethnie etc., stark von jenen in den Trainingsdaten unterscheiden, kann es sein, dass bedenkliche Zustände, die auf eine Erkrankung hindeuten, nicht erkannt werden oder aber gesunde Zustände fälschlicherweise als besorgniserregend eingestuft werden. Um diesen Data Shift und die daraus resultierenden Falschergebnisse zu vermeiden, sind spezifische Tests in der jeweiligen Organisation vor dem Einsatz erforderlich.[11]

2. Bias:

Selbst ein passender, großer und qualitativ hochwertiger Datensatz garantiert nicht, dass die KI frei von ***systematischen Verzerrungen (sog. Bias)*** ist und bei allen Menschen gleich gut funktioniert. Generell zeigt sich, dass KI-Systeme in der medizinischen Bildgebung sehr oft bei Personen zu falschen Ergebnissen führen, die in den Medizindaten unterrepräsentiert sind, zu Risikogruppen gehören, die das Gesundheitssystem weniger nutzen oder für Menschen, die keinen Zugang zum System haben. Zum Beispiel können dies Frauen oder Kinder, Migrant:innen oder Personen mit dunklerer Hautfarbe sein. Grundsätzlich kann aber jede Person von Bias betroffen sein.[8]

3. Transparenz:

KI-Systeme sind aufgrund der eingesetzten Methoden des maschinellen Lernens (s. ***überwachtes Lernen*** und ***unüberwachtes Lernen***) meist wenig transparent und werden daher auch als ***Black-Box-Systeme*** bezeichnet.[3] Hersteller:innen von KI-Systemen haben zwar Transparenzauflagen zu erfüllen und Betreiber:innen müssen sicherzustellen, dass die Nutzer:innen, wie Ärzt:innen und Gesundheitspersonal die entsprechende Kompetenzen (AI Literacy) zur Nutzung der KI haben. Allerdings sind diese in vielen Fällen nicht ausreichend, um den Output von KI-Systemen vollständig nachvollziehen zu können. Denn die mangelnde Transparenz ist nicht nur in der Lernmethode der KI begründet, sondern auch mit Geschäftsgeheimnissen und IP- Schutz (also eine bewusste Entscheidung der Hersteller:innen) und der DSGVO (Geheimhaltung der Trainingsdaten).

4. Privatsphäre und Sicherheit:

Der Schutz der ***Privatsphäre*** von Patient:innen steht immer an erster Stelle und muss mit dem Einsatz von KI-Systemen zu jeder Zeit gewährleistet sein. Bis dato funktionieren viele KI-Systeme Cloud-basiert (und nicht auf eigenen Servern der jeweiligen Organisation – on-premises), was die Anfälligkeit für Datenschutzverletzungen und Sicherheitsangriffe erhöht.[1,2,11]

Empfehlung zum Praxiseinsatz von KI in der medizinischen Bildgebung:

Sechs wesentliche Punkte

Trotz technologischer Innovationssprünge und großer Chancen für Patient:innen und das Gesundheitssystem, benötigen Gesundheitsdienstleister:innen spezifische KI-Kompetenzen und begleitende Maßnahmen im Risikomanagement, um KI-Anwendungen für die Bildanalyse sachgemäß und verantwortungsvoll einsetzen zu können. Wesentlich ist, dass derartige KI-Systeme vor und während des Einsatzes geprüft und begleitende Maßnahmen in der Organisation gesetzt werden. Es wird empfohlen, folgende sechs Punkte zu beachten, um die Sicherheit und Akzeptanz aller Beteiligten zu gewährleisten:

1. Sorgfältige Auswahl der KI-Anwendung und Einforderung von Transparenz seitens des Anbieter:innen:

KI-Systeme in der medizinischen Bildgebung sind keine Alleskönner, sondern auf sehr eng definierte, spezifische Aufgaben trainiert. Vor dem Einsatz müssen daher wesentliche Informationen geprüft werden (s. Pkt. a-d). Daraus lassen sich bereits erste Rückschlüsse ziehen, ob das KI-System im vorgesehenen Anwendungsbereich grundsätzlich anwendbar ist. Diese Informationen werden ab Mitte 2027 leichter zu erhalten sein, da dann das EU KI-Gesetz (EU AI-Act) mit seinen Transparenzpflichten für Medizinprodukte vollumfänglich gilt.

a) Die grundsätzliche Funktionsweise (Intended Purpose):

- Wofür ist das System optimiert?
- Welche Technologien werden verwendet?
- Mit welchen Eingabedaten arbeitet die Anwendung?

b) Die verwendeten Trainings- und Testdaten inklusive einer Beschreibung der Patient:innenpopulationen, aus denen sie stammen (da diese den Patient:innenpopulationen im vorgesehenen Einsatzbereich möglichst ähnlich sein sollten)

- Nach welchen Maßstäben und Metriken wurde die Population getestet und was war deren Ergebnis?

c) Bekannte Limitationen, z. B. hinsichtlich Zielgruppen, bei denen die Anwendung nicht zuverlässig funktioniert.

d) Die Zulassung als Medizinprodukt in der EU (wenn es sich um ein System aus z. B. den USA handelt) und etwaige, darüberhinausgehende Standards, die erfüllt werden (wird vor allem künftig von Bedeutung sein).

2. Prüfung der Leistungsfähigkeit unter realen Bedingungen (AI Model Performance versus Clinical Performance):

Die Leistungsfähigkeit von KI-Systemen kann in der klinischen Praxis stark schwanken. Aus diesem Grund ist das Testen von KI-Systemen in einer realen Umgebung mit Inputdaten von tatsächlichen Patient:innenpopulationen (Klinik, Labor o. a. Gesundheitseinrichtung) eine wichtige Voraussetzung, um die Performanz und das Risiko von Falschaussagen klarer einschätzen zu können. Entsprechende Tests sollten durchgeführt werden.

3. Letztentscheidung durch fachlich kompetente Personen sicherstellen:

Auch wenn KI eine wertvolle Unterstützung bietet, muss sie aufgrund von technischer Unvollkommenheit (s. Black-Box-Systeme) sowie rechtlicher Vorgaben (s. EU KI-Gesetz) jederzeit von Menschen kontrolliert und überwacht werden. Ärzt:innen müssen weiterhin die Verantwortung für Diagnosen und Therapieentscheidungen tragen. Gleichzeitig bedeutet dies einen organisationsinternen Auftrag zur Erstellung klarer Richtlinien zur Nutzung von KI (betriebliche KI-Richtlinien) sowie zur Aus- und Weiterbildung des (medizinischen) Personals (Stichwort KI-Literacy), damit KI-Ergebnisse kritisch hinterfragt bzw. interpretiert werden können (s. nächster Punkt). An dieser Stelle ist zu erwähnen, dass rezent ein System entwickelt wurde, das medizinisches Fachpersonal dabei unterstützt, KI-Ergebnisse richtig einzuordnen und zu interpretieren. MONET erhöht durch u.a. Datenprüfung, Modellprüfung oder Modellinterpretation die Transparenz von Resultaten bestimmter KI-Anwendungen für Entscheider:innen.[4]

4. **Spezifische Kompetenz bei Anwender:innen aufbauen (KI-Literacy):**

Die erfolgreiche Integration von KI im Bereich bildbasierter Diagnose- und Therapieunterstützung erfordert Aus- und Weiterbildung des medizinischen Personals, um ein allgemeines Verständnis der Fähigkeiten und Grenzen von KI-Systemen, ihrer Stärken und Schwächen und der richtigen Interpretation von KI-gestützten Diagnosen zu vermitteln. Das Personal muss in der Lage sein, informierte Entscheidungen zu treffen und sich in der Lage fühlen, der KI-Anwendung im Falle falscher Ergebnisse widersprechen zu können.

5. **Patient:innen über Einsatz von KI informieren:**

Die Akzeptanz und Sicherheit von KI-gestützter Bildauswertung hängt entscheidend von ihrer Transparenz und kontinuierlicher Überwachung (s. nächster Punkt) ab. Patient:innen sind darüber zu informieren, wenn KI bei der Diagnose oder Befundung eingesetzt wird. Dies kann in Aufklärungsgesprächen, durch schriftliche Informationen oder spezifische Einwilligungserklärungen geschehen. Darüber hinaus muss klar kommuniziert werden, dass die Letztverantwortung bei medizinischem Fachpersonal liegt.

6. **Laufende Überwachung und Monitoring des KI-Systems:**

Kontinuierliche Überwachung der KI-Performance ist essenziell, um die Qualität langfristig sicherzustellen und Risiken stetig zu minimieren. Dies umfasst die regelmäßige Analyse von Fehlentscheidungen, Anpassungen der Algorithmen und Modelle an neue Erkenntnisse durch kontinuierliches Training mit neuen Datensätzen sowie die Implementierung eines Meldesystems für das medizinische Personal, um Probleme frühzeitig zu erkennen und zu beheben.

Wichtige Begriffe

Black-Box-Systeme: Bezeichnet KI-Systeme, deren interne Entscheidungsprozesse für Menschen nicht transparent und nachvollziehbar sind. Das heißt, dass nur Eingaben und Ausgaben beobachtet werden können, ohne zu verstehen, wie die Verarbeitung dazwischen genau abläuft.

Datenschutz in der KI: Die Gesamtheit der Praktiken und Bedenken im Zusammenhang mit der ethischen Erfassung, Speicherung und Nutzung personenbezogener Daten durch Systeme der künstlichen Intelligenz.

EU KI-Gesetz (Gesetz der Europäischen Union über künstliche Intelligenz, engl. EU AI-Act): Die europäische Verordnung über künstliche Intelligenz (KI) – die erste umfassende Verordnung über KI von einer großen Regulierungsbehörde überhaupt. Sie konzentriert sich insbesondere auf KI-Systeme mit hohem Risiko.

KI-Autonomie: Die Fähigkeit eines KI-Systems, eine Reihe von Zielen unter einer Reihe von Unsicherheiten in ihrer Umgebung selbstständig und ohne externe Eingriffe zu erreichen.

KI-Genauigkeit (AI-Accuracy): bezieht sich auf die Fähigkeit eines KI-Systems, korrekte Vorhersagen oder Entscheidungen zu treffen. Sie ist ein wichtiger Maßstab für ihre Leistung und entscheidend für die Bestimmung ihrer Wirksamkeit und Zuverlässigkeit.

Systematische Verzerrungen (Bias): Bias ist eine systematisch unterschiedliche Behandlung bestimmter Objekte, Personen oder Gruppen im Vergleich zu anderen. Behandlung ist jede Art von Handlung, einschließlich Wahrnehmung, Beobachtung, Darstellung, Vorhersage oder Entscheidung.

Transparenz: Bedeutet, dass die Funktionsweise, Entscheidungsprozesse und Einsatzbereiche eines KI-Systems nachvollziehbar, erklärbar und offen zugänglich sind – für Entwickler:innen, Nutzer:innen und andere Stakeholder.

Überwachtes Lernen (supervised learning): Eine Methode des maschinellen Lernens, bei der die Trainingsdaten vorab gekennzeichnet werden und das System das Muster zwischen dem Bild und der Kennzeichnung erlernt. Die Aufgabe eines solchen KI-Systems besteht darin, eine Beziehung zu finden, die jede Eingabe des Trainingssatzes (die Daten) einer Ausgabe (der Kennzeichnung) zuordnet.

Unüberwachtes Lernen (unsupervised learning): Ein Teilgebiet des maschinellen Lernens, bei dem ein Algorithmus Muster, Strukturen oder Beziehungen in unbeschrifteten Daten erkennt, ohne dass ihm vorhergesagt wird, was richtig oder falsch ist.

Erklärung Stufenmodell des ALAIT Risikoradars

Im ALAIT Risikoradar wird die Beziehung zwischen Einsatz-Risiko aus dem Anwendungsbereich und Entscheidungsautonomie einer KI dargestellt. Geringere Autonomie und niedriges Anwendungsrisiko werden durch kältere Farben (blau und gelb) gekennzeihnet, höhere Autonomie und Anwendungsrisiken durch wärmere Farben (orange bis violett) . Im Idealfall sollten KI-Anwendungen, die auf beiden Achsen in den höchsten Risikobereichen eingeordnet sind, nur nach sehr sorgfältiger Abwägung eingesetzt oder ganz vermieden werden. KI-Anwendungen, die nur auf einer Achse in der höchsten Stufe platziert sind (ein hohes Anwendungsrisiko und eine geringe Autonomie oder umgekehrt), stellen ein insgesamt mittleres Riskioniveau dar. Die Risikostufen stützen sich auf das EU KI-Gesetzt (EU AI-Act), insbesondere auf Artikel 6 und Annex III, die sich mit risikoreichen Anwendungsbereichen von KI befassen.

Autonomie

Stufe 1: Keine Autonomie

KI ist ein passives Werkzeug; Menschen treffen alle Entscheidungen und leiten Maßnahmen ein.

Beispiel: Diagnosesysteme, die medizinische Rohdaten anzeigen oder die Daten analysieren (ohne Empfehlungen)

Empfohlene Anwendungsfälle: Szenarien mit hohen Risiken oder bei denen ethische Entscheidungen von großer Bedeutung sind (z.B. medizinische Diagnostik, Justizsystem).

Stufe 2: Geringer Autonomiegrad (Human-in-the-Loop)

Die KI gibt Empfehlungen oder Optionen, aber die Benutzer:innen bleiben für die Auswahl und Genehmigung von Maßnahmen verantwortlich (Human-in-the-Loop).

Beispiel: KI schlägt optimale Routen für die Logistik vor oder Empfehlungssysteme im E-Commerce.

Empfohlene Anwendungsfälle: Aufgaben mittlerer Komplexität mit mäßigen Risiken (z.B. Optimierung der Lieferkette).

Stufe 3: Mittlerer Autonomiegrad (Human-on-the-Loop)

Die KI führt bestimmte Aufgaben autonom aus, wobei Menschen in Ausnahmefällen eingreifen (Human-on-the-Loop).

Beispiel: KI-gestützte Fertigungsprozesse, bei denen das System Maschinen steuert, aber Nutzende bei Anomalien eingreifen.

Empfohlene Anwendungsfälle: Szenarien, in denen eine kontinuierliche menschliche Beteiligung nicht erforderlich ist, kritische Risiken jedoch eine menschliche Überwachung erfordern (z.B. industrielle Automatisierung, Überwachung von Finanztransaktionen).

Stufe 4: Hoher Autonomiegrad (Human in Control)

Das KI-System arbeitet weitgehend autonom, erlaubt es den Benutzer:innen jedoch, es selbst zu übersteuern, um unerwünschte Ergebnisse zu vermeiden.

Beispiel: Autonome Fahrzeuge

Empfohlene Anwendungsfälle: Umgebungen mit geringem bis mittlerem Risiko (z.B. Logistik, einfaches Verkehrsmanagement).

Stufe 5: Vollständige Autonomie mit minimaler Aufsicht

Das KI-System arbeitet unabhängig und erfordert nur minimale oder gar keine menschliche Intervention.
Die Beteiligung des Menschen beschränkt sich auf die langfristige Aufsicht (Audits).

Beispiele: Autonome landwirtschaftliche Maschinen, KI für die Stromnetzverteilung, U-Bahnen, Flughafentransferzüge

Empfohlene Anwendungsfälle: Umgebungen mit geringen Sicherheits- oder ethischen Risiken und hoher Zuverlässigkeit des KI-Systems (z.B. sich wiederholende Aufgaben in kontrollierten Umgebungen).

Anwendungsbereich-Risiko

Stufe 1: Minimales Risiko

Das KI-System hat keine Auswirkungen auf den Benutzer:innen oder die Entscheidungsfindung.

Beispiele: Filter, NPCs in Computerspielen, Empfehlungsalgorithmen ohne schwerwiegende Folgen (DeepL, andere Übersetzungstools)

Kriterien: Keine direkte Auswirkung auf die Hochrisikobereichen des EU AI-Acts.

Stufe 2: Begrenztes Risiko

KI-Systeme, die mit Benutzer:innen interagieren, aber keine Entscheidungen mit hohen Risiken treffen. Das Risiko steigt, wenn es an Transparenz über die Beteiligung von KI mangelt.

Beispiele: Chatbots und KI-generierte Inhalte ohne Offenlegung, einfache Automatisierungsaufgaben.

Kriterien: Bereiche, die nicht in der Liste der „hohen Risiken" des EU AI-Acts enthalten sind.

Stufe 3: Mittleres Risiko

KI-Systeme haben keine besonderen Auswirkungen auf einzelne Personen, aber aber sie entfalten Wirkung auf kollektiver oder gesellschaftlicher Ebene.

Beispiele: Generative KI wie ChatGPT und andere Systeme, die indirekt die Umgebung beeinflussen können, in der sie eingesetzt werden und schließlich zu größeren Veränderungen der Gesellschaft und des Lebens führen können.

Kriterien: KI-Systeme, die für die öffentliche Nutzung verfügbar sind und das Potenzial haben, bestehende Gepflogenheiten zu beeinflussen und langfristig zu verändern.

Stufe 4: Hohes Risiko

Jeder Algorithmus, der in den laut EU AI-Act „Hochrisikobereichen" angewendet wird oder direkte Auswirkungen auf Leib und Leben einzelner Personen hat.

Beispiele: Medizin, Biometrie, kritische Infrastruktur, Bildung und Berufsausbildung, Beschäftigung, Zugang zu Dienstleistungen, Strafverfolgung, Migration.

Kriterien: Zugehörigkeit zum „Hochrisikobereich" des EU AI-Acts, nur wenn die Regeln für Transparenz und Datenqualität eingehalten werden.

Stufe 5: Extremes Risiko

Jeder Algorithmus, der in den laut EU AI-Act „Hochrisikobereichen" angewendet wird.

Beispiele: Medizin, Biometrie, kritische Infrastruktur, Bildung und Berufsausbildung, Beschäftigung, Zugang zu Dienstleistungen, Strafverfolgung, Migration.

Kriterien: Zugehörigkeit zum „Hochrisikobereich", wenn die Regeln für Transparenz und Datenqualität NICHT eingehalten werden.

Endnoten

1 Barragán-Montero, A., Javaid, U., Valdés, G., Nguyen, D., Desbordes, P., Macq, B., Willems, S., Vandewinckele, L., Holmström, M., Löfman, F., Michiels, S., Souris, K., Sterpin, E., & Lee, J. A. (2021). Artificial intelligence and machine learning for medical imaging: A technology review. Physica Medica, 83, 242–256. https://doi.org/10.1016/j.ejmp.2021.04.016

2 Herington, J., McCradden, M. D., Creel, K., Boellaard, R., Jones, E. C., Jha, A. K., Rahmim, A., Scott, P. J. H., Sunderland, J. J., Wahl,R. L., Zuehlsdorff, S., & Saboury, B. (2023). Ethical Considerations for Artificial Intelligence in Medical Imaging: Data Collection, Development, and Evaluation. Journal of Nuclear Medicine, 64(12), 1848–1854. https://doi.org/10.2967/jnumed.123.266080

3 Castiglioni, I., Rundo, L., Codari, M., Di Leo, G., Salvatore, C., Interlenghi, M., Gallivanone, F., Cozzi, A., D'Amico, N. C., & Sardanelli, F. (2021). AI applications to medical images: From machine learning to deep learning. Physica Medica, 83, 9–24. https://doi.org/10.1016/j.ejmp.2021.02.006

4 Kim, C., Gadgil, S. U., DeGrave, A. J., Omiye, J. A., Cai, Z. R., Daneshjou, R., & Lee, S.-I. (2024). Transparent medical image AI via an image–text foundation model grounded in medical literature. Nature Medicine, 30(4), 1154–1165. https://doi.org/10.1038/s41591-024- 02887-x

5 Boesch, G. (2023, January 20). What is Computer Vision? The Complete Guide for 2025. Viso.Ai. https://viso.ai/computer-vision/what-is-computer-vision/

6 Szeliski, R. (2022). Computer Vision: Algorithms and Applications. Springer Nature.

7 Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep Learning for Computer Vision: A Brief Review. Com- putational Intelligence and Neuroscience, 2018(1), 7068349. https://doi.org/10.1155/2018/7068349

8 Leimüller, G., & Wazir, R. (2025). Die Sache mit dem Bias in der Künstlichen Intelligenz.

9 Panayides, A. S., Amini, A., Filipovic, N. D., Sharma, A., Tsaftaris, S. A., Young, A., Foran, D., Do, N., Golemati, S., Kurc, T., Huang, K., Nikita, K. S., Veasey, B. P., Zervakis, M., Saltz, J. H., & Pattichis, C. S. (2020). AI in Medical Imaging Informatics: Current Challenges and Future Directions. IEEE Journal of Biomedical and Health Informatics, 24(7), 1837–1857. IEEE Journal of Biomedical and Health Informatics. https://doi.org/10.1109/JBHI.2020.2991043

10 Khalifa, M., & Albadawy, M. (2024). AI in diagnostic imaging: Revolutionising accuracy and efficiency. Computer Methods and Pro- grams in Biomedicine Update, 5, 100146. https://doi.org/10.1016/j.cmpbup.2024.100146

11 Tang, X. (2020). The role of artificial intelligence in medical imaging research. BJR|Open, 2(1), 20190031. https://doi.org/10.1259/bjro.20190031

12 Wang, X., Zhao, J., Marostica, E. et al. (2024) A pathology foundation model for cancer diagnosis and prognosis prediction. Nature 634, 970–978 . https://doi.org/10.1038/s41586-024-07894-z

13 Chen, M., Wang, Y., Wang, Q. et al. (2024). Impact of human and artificial intelligence collaboration on workload reduction in medical image interpretation. https://doi.org/10.1038/s41746-024-01328-w

Über das Projekt ALAIT

Das Austrian Lab for AI Trust (ALAIT) ist ein vom österreichischen Bundesministerium für Innovation, Mobilität und Infrastruktur (BMIMI) gefördertes Forschungs- und Entwicklungs-Projekt zur Schaffung von Vertrauen durch Wissen im Bereich Künstliche Intelligenz (KI). Das Projekt ALAIT zielt darauf ab, Interessierte und wichtige gesellschaftliche Gruppen zu befähigen, KI-Technologien verantwortungsvoll zu nutzen und ethische sowie qualitativ hochwertige Standards für den Einsatz von AI zu etablieren.

Das Projekt wird von **winnovation** geleitet (Gertraud Leimüller und Lena Müller-Kress) und im Konsortium mit **leiwand.ai** (Rania Wazir und Silvia Wasserbacher-Schwarzer),
TU Wien (Sabine Köszegi und Ilya Faynleyb) und
Austria Presse Agentur – APA (Verena Krawarik und Sophia Marecek) umgesetzt.

Die ALAIT-Dossiers sind auf der Projekthomepage abrufbar: https://science.apa.at/project/alait/

Die Inhalte des Dossiers entsprechen dem aktuellen Stand der Technik und wurden sorgfältig nach wissenschaftlichen Kriterien erstellt. Sie dienen jedoch nicht als rechtsverbindliche Auskunft oder Beratung.

Impressum

Medieninhaberin und Herausgeberin:
winnovation consulting gmbh
Linke Wienzeile 42/1, Top 5
1060 Vienna

Veröffentlicht 2025, gefördert durch das Bundesministerium für Innovation, Mobilität und Infrastruktur.