Vorlage für die Einordnung von KI-Risiken in eine zweidimensionale Matrix

Überblick

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.



*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.

Erklärung Stufenmodell des ALAIT Risikoradars

Im ALAIT Risikoradar wird die Beziehung zwischen Anwendungsrisiko und Autonomie eines KI-Systems dargestellt. Die Risikostufen stützen sich auf das EU KI Gesetz (EU AI-Act), insbesondere auf Artikel 6 und Annex III, die sich mit risikoreichen Anwendungsbereichen von KI befassen. Geringere System-Autonomie und Anwendungsrisiken werden durch kältere Farben (blau) und höhere System-Autonomie und Anwendungsrisiken durch wärmere Farben (rot) dargestellt.

Der Farbwechsel vermittelt das erhöhte Risiko solcher Entscheidungen. Mithilfe dieser Farbskala lässt sich das Gesamtrisiko erkennen: Violett und dunkelrot – sehr hoch, rot und dunkelorange – hoch, hellorange und Gelbtöne – mittel, Blautöne – geringes Gesamtrisiko. Im Idealfall sollten hohe Anwendungsrisiken und System-Autonomie vermieden oder nur nach sehr sorgfältiger Abwägung eingesetzt werden.

Autonomie

Stufe 1: Keine Autonomie

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

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).

Empfohlene Anwendungsfälle: Aufgaben mittlerer Komplexität mit mäßigen Risiken.

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

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

Empfohlene Anwendungsfälle: Szenarien, in denen eine kontinuierliche menschliche Beteiligung nicht erforderlich ist, kritische Risiken jedoch eine menschliche Überwachung erfordern.

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.

Empfohlene Anwendungsfälle: Umgebungen mit geringem bis mittlerem Risiko.

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).

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.

Kriterien: Keine direkte Auswirkungen auf die Hochrisiko-bereiche 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.

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 sie entfalten Wirkung auf kollektiver oder gesellschaftlicher Ebene.

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.

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.

Kriterien: Zugehörigkeit zum „Hochrisikobereich“, wenn die Regeln für Transparenz und Datenqualität NICHT eingehalten werden (zum Beispiel aufgrund von Übergangsfristen).

Über das Projekt ALAIT

Das Austrian Lab for AI Trust (ALAIT) ist ein vom österreichischen Bundesministerium für Innovation, Mobilität und Infrastruktur (BMIMI) initiiertes 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 Brigitte Ömer-Rieder) und im Konsortium mit leiwand.ai (Rania Wazir), TU Wien (Sabine Köszegi und Karoline Bohrn) und Austria Presse Agentur – APA (Verena Krawarik und Sophia Marecek) umgesetzt.

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

Alle erarbeiteten Inhalte 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 124, Top 5
1060 Vienna

Die ALAIT Ergebnisdokumente stehen unter der Creative Commons Lizenz CC BY-NC-ND 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de (Namensnennung-Nicht kommerziell-Keine Bearbeitungen 4.0 International)

Bundesministerium Innovation, Mobilität und Infrastruktur

winnovation

leiwand.ai

TU WIEN

APA