DGUV Test
Prüf- und Zertifizierungssystem der
Deutschen Gesetzlichen Unfallversicherung

05

DGUV Test Information

Stand: 04/2021

Allgemeine Grundsätze für die sicherheitstechnische Bewertung von Künstlicher Intelligenz (KI)

In diesem Dokument werden allgemeine Grundsätze für die sicherheitstechnische Bewertung von KI-Technologien beschrieben. Ethische oder haftungsrechtliche Fragen werden hier nicht thematisiert. Die Grundsätze konkretisieren Anforderungen an KI-Technologien in Bezug auf Sicherheit und Gesundheit und dienen als Hilfestellung für die Erarbeitung von produktspezifischen Prüfanforderungen.

Der Begriff KI war bisher nicht einheitlich definiert und diente meist als Sammelbegriff für diverse Verfahren wie logisches Schließen, Expertensysteme, verschiedene Suchalgorithmen oder maschinelles Lernen. Mit der Erarbeitung des neuen Standards ISO IEC 22989 soll zukünftig eine einheitliche Definition für KI sowie für das maschinelle Lernen festgelegt werden. In dieser DGUV Test Information wird KI und maschinelles Lernen wie folgt definiert.

Künstliche Intelligenz

Fähigkeit, Informationen über Objekte, Ereignisse, Konzepte oder Regeln, ihre Beziehungen und Eigenschaften zu beschaffen, zu verarbeiten, zu erstellen und anzuwenden, die für eine zielorientierte systematische Nutzung organisiert sind, und in Form einer physikalischen, mathematischen oder anderweitig logischen Darstellung eines Systems, einer Entität, eines Phänomens, eines Prozesses oder von Daten vorliegen, um eine oder mehrere gegebene Aktivitäten durchzuführen, die unternommen werden, um ein bestimmtes Ziel zu erreichen.

Maschinelles Lernen

Prozess unter Verwendung rechnergestützter Techniken, um Systeme in die Lage zu versetzen, aus Daten oder Erfahrungen zu lernen.

Grundsatz 1

Wenn eine Aufgabe mit Hilfe einer klassischen Technologie gelöst werden kann, sollte diese gegenüber einer Anwendung von KI bevorzugt werden.

Dieser Standpunkt kann nur den aktuellen Stand der Technik widerspiegeln und ist mit Fortschreiten der Technologie anzupassen oder gegebenenfalls aufzugeben. Die Anforderungen an die funktionale Sicherheit im Bereich der industriellen Anwendungen werden heute durch die Sicherheitsnormen DIN EN ISO 13849, IEC 62061 und IEC 61508 beschrieben.

Bedingt durch die hohen Leistungsanforderungen an Maschinen und Anlagen in diesem Umfeld ist dementsprechend mit einem hohen bis sehr hohen Gefährdungspotenzial umzugehen.

Zur Beherrschung dieser Gefährdungen werden heute oftmals Sicherheitsfunktionen eingesetzt, die als zweikanalig redundante Systeme ausgelegt sind. Darüber hinaus werden von den anzuwendenden Sicherheitsnormen konkrete Anforderungen an die Güte der verwendeten Bauteile bezgl. der Ausfallsicherheit, der Fehlererkennung (Diagnosedeckungsgrad) und die Zuverlässigkeit der Software gestellt.

Das Ziel dieser Maßnahmen ist das Erreichen der sog. Ein-Fehler-Sicherheit. Tritt in einem Kanal der Sicherheitsfunktion ein Fehler auf, bleibt die Sicherheit über den zweiten Kanal erhalten und eine Fehlererkennung führt zu einem sicheren Zustand. Weder DIN EN ISO 13849, IEC 62061 noch IEC 61508 beschreiben derzeit konkrete Anforderungen für die Anwendung der KI.

Hauptbestandteil der KI sind heute Algorithmen, realisiert z. B. über neuronale Netze, die mit Hilfe von entsprechend aufbereiteten Datensets für ihre Aufgabe bzw. Zielsetzung trainiert werden. Gegenwertig basieren die meisten Anwendungen der KI auf Verfahren des maschinellen Lernens. Ziel des maschinellen Lernens ist es, durch einen Trainingsprozess ein auf Trainingsdaten basiertes Modell zu erzeugen, dass Wissen generalisiert und somit auf neue Daten angewendet werden kann. Während des Trainingsprozesses wird das Modell durch einen Algorithmus optimiert. Im Fall des *deep learnings* (z. B. tiefer neuronaler Netze) ist dieses Modell meist sehr komplex und daher nicht einfach nachvollziehbar. Daraus resultiert, dass die Bewertung der Sicherheit wesentlich erschwert wird, da hier auf Verfahren zur Verifikation und Validierung von Black-Box-Systemen zurückgegriffen werden muss.

Grundsatz 2
KI zur Realisierung von Assistenzsystemen unterstützt den Menschen, sie können aber nicht als Sicherheitsfunktion gewertet werden.

In vielen Bereichen der Industrie genauso wie im alltäglichen Leben kommen heute technische Assistenzsysteme zum Einsatz. Im Forschungsbericht 502 des BMAS werden physische, sensorische und kognitive menschbezogene Unterstützungsansätze unterschieden.

Waren bisher hauptsächlich physische Assistenzsysteme im Einsatz, bieten die neuen intelligenten Assistenzsysteme auf der Basis von KI die Möglichkeit, die kognitiven und sensorischen Fähigkeiten des Bedieners zu unterstützen. Das Assistenzsystem, ausgestattet mit einer dem Einsatzzweck entsprechenden Sensorik, beobachtet parallel zum Bediener den laufenden Prozess und reagiert selbständig auf Ereignisse oder sich verändernde Prozessparameter. Hierbei muss beachtet werden, dass Assistenzsysteme aus Sicht des Personenschutzes üblicherweise nicht den Anforderungen der einschlägigen Sicherheitsnormen (z. B. DIN EN ISO 13849) entsprechen und damit nicht als Sicherheitsfunktionen gewertet werden dürfen.

Grundsätzlich muss der Bediener über die Möglichkeit verfügen, die vollständige Kontrolle über den Prozess zu erlangen.

Physische Assistenzsysteme

leisten Hilfestellung bei anspruchsvollen körperlichen Tätigkeiten und dienen dem Ausgleich körperlich nachlassender Fähigkeiten bzw. der Vorbeugung ihres vorzeitigen Verlustes. Der aktuelle Stand der Technik reicht von mechanisch-motorischer Kraftunterstützung und personalisierten Montagearbeitsplätzen für einfache, regelbasierte Arbeitssituationen bis hin zu adaptiven, kollaborativen Robotersystemen für komplexe, hochvariable und expertisebasierte Produktions-, Montage- und Wartungsprozesse. Dabei erfolgt vor allem eine Unterstützung des Muskel-Skelett-Systems und der Sinnesorgane.

Sensorische Assistenzsysteme

dienen dem Ausgleich funktionaler, oft altersbedingter, Veränderungen der Sinnesorgane. Fortgeschrittene Systeme adressieren vor allem hör- und sehbedingte Einschränkungen und leisten eine kombinierte kognitivsensorische Unterstützung (z. B. Augmented-Reality-Brille).

Kognitionsunterstützende Assistenzsysteme

dienen vor allem der anwendungsgerechten, echtzeitnahen Informationsbereitstellung zur Entscheidungsunterstützung der Beschäftigten. Die funktionale Unterstützung ist je nach Unterstützungsgrad vor allem auf die Reaktions-, Denk-, Merk- und Schlussfolgerungsfähigkeit ausgerichtet (vgl. Müller et al. 2014). Hauptelemente der Hardware in anwendungsnahen Forschungsprojekten sind vor allem mobile Endgeräte und interaktive Visualisierungssysteme.

Quelle: Forschungsbericht 502 BMAS

Grundsatz 3
Kontinuierlich lernende Systeme dürfen keinen gefährlichen Einfluss auf die Sicherheitsfunktionen haben.

Je nach Konzeption kann ein maschinelles Lernsystem entweder seinen Lernprozess stoppen, so dass es sich nach seinem Einsatz immer gleich verhält, oder während seiner Nutzung weiter lernen. Kontinuierliche Lernsysteme folgen dem letztgenannten Ansatz und nutzen ein inkrementelles Training des KI-Systems, welches kontinuierlich während der Betriebsphase des System-Lebenszyklus stattfindet.

Während das Verhalten nicht kontinuierlich lernender Systeme während des Entwicklungsprozesses festgelegt wird und sich während der Betriebsphase nicht mehr ändern soll, findet beim kontinuierlichen Lernen eine schrittweise erfolgende Aktualisierung des Modells während der Betriebsphase statt. Die in der Betriebsphase durch das System akquirierten Daten werden in diesem Fall nicht nur analysiert, um einen Output zu erzeugen, sondern gleichzeitig auch zur Anpassung des Modells im System verwendet, um dieses auf Grundlage der zusätzlichen Eingabedaten zu verbessern.

Ziel des kontinuierlichen Lernens ist es, Probleme oder Fehler, die auf ursprünglich stark eingeschränkten oder unvollständigen Trainingsdaten basieren, zu beheben – oder auf sich allmählich ändernde Betriebsbedingungen, welche sich von der spezifizierten Trainingsumgebung unterscheiden, zu reagieren und so dem Problem der Abweichung vom spezifizierten Konzept entgegenzuwirken.

Das kontinuierliche Lernen führt durch die inkrementellen Anpassungen des Modells zu einem dynamischen Verhalten des KI-Systems, welches einerseits gewollt ist, aber auf der anderen Seite auch beträchtliche Herausforderungen mit sich bringt. So ist die Sicherstellung der korrekten Funktion eines solchen Systems in der Betriebsphase sehr schwierig, da hierfür eine ebenfalls kontinuierliche Verifikation des Systems notwendig ist. Weiterhin müssten die neuen Eingangsdaten erfasst werden, um bei einer späteren werkseitigen Aktualisierung Teil des neuen Trainingsdatensatzes werden zu können oder zur Fehlersuche herangezogen werden zu können. Es muss dementsprechend bei jeder Applikation eine Abwägung der Vor- und Nachteile des Einsatzes dieser Methode sowie der aus ihr entstehenden Risiken erfolgen.

Grundsatz 4
Der „Entscheider“ muss sicher sein.

Unter Entscheidung versteht man die Wahl einer Handlung aus mindestens zwei vorhandenen potenziellen Handlungsalternativen unter Beachtung der übergeordneten Ziele. Jede Tätigkeit verlangt Entscheidungen. Hängt von dieser Entscheidung z. B. das sicherheitsgerichtete Verhalten einer Maschine ab, muss die Entscheidung schnell und überlegt sein. Generell unterscheidet man zwischen einem menschlichen und einem technischen Entscheider. Im Bereich der Automatisierung werden sicherheitsrelevante Entscheidungen üblicherweise von der Steuerung, also einem technischen Entscheider getroffen. Dies ist auch auf Grund der Maßnahmenhierarchie (technische Maßnahme vor organisatorische Maßnahme) zu bevorzugen. Die Aufgabe des technischen Entscheiders besteht darin, das von der KI gesteuerte System in einem Zustand zu halten, bei dem das Risiko für den Bediener vertretbar bleibt.

Dies kann z. B.
- der degradierte Zustand (gleichwertiger zeitlicher Zustand mit Ersatzmaßnahmen) oder ein
- vorher festgelegter definierter Zustand sein.
- Alternativ kann dies zu einem gesteuerten
- bzw. nicht gesteuerten Abschalten führen.

Die Steuerung kann solche Entscheidungen mit Hilfe von KI nur dann treffen, wenn Randbedingungen, z. B. Anforderungen an Datenqualität, Datenspeicherung sowie Hard- und Software, erfüllt sind.

Grundsatz 5
Die Datenqualität muss überwacht und sichergestellt werden.

Zuverlässige Daten und deren Verarbeitung sind die Grundlage für korrekte Entscheidungen. Um eine hohe Datenqualität sicherzustellen, müssen für die Datenerhebung geeignete Bauteile, z. B. redundante Sensoren, unter Beachtung der entsprechenden Produktnormen verwendet werden. Darüber hinaus sollten die Daten, die die Entscheidung signifikant beeinflussen, in bestimmten Zeitabständen überwacht und auf Plausibilität geprüft werden. Falsche oder unvollständige Informationen können die Entscheidung negativ beeinflussen.

Anmerkung: Zur Datenqualität gehört auch die Qualität des gesamten Datensatzes im Sinne der Vollständigkeit und Diversität, um sicherzustellen, dass alle relevanten Beispiele zur Modellbildung im Datensatz enthalten sind.

- **Beispiel 1:** In einem Datensatz, der zur Erstellung eines Systems zur Personenerkennung gedacht ist, sollten Beispielbilder von Personen mit unterschiedlicher Größe, Statur, Hautfarbe, Gesichtszüge enthalten sein, da ansonsten Personen mit unterschiedlicher Diversität nicht erkannt werden.
- **Beispiel 2:** Periodische Fehlerdetektion für sichere reduzierte Geschwindigkeit an Werkzeugmaschinen (Überwachung des Encoders)

Werden die Werte des Encoders auch während des Prozesses auf Plausibilität geprüft, werden Fehler dann erkannt, wenn sie auftreten. Die Datenqualität wird somit permanent sichergestellt.

Für die Weiterverarbeitung der Daten im Entscheider müssen diese in einem lesbaren Format ausgegeben werden und deren Integrität sichergestellt sein. Dies kann z. B. über moderne Bus-Systeme erfolgen.

Grundsatz 6
Relevante Daten des Entscheiders müssen aufgezeichnet und gespeichert werden.

Bei der Entscheidung sollten alle verwendeten Daten des Entscheiders mit Zeitstempel für einen gewissen Zeitraum gespeichert werden, um eine Nachvollziehbarkeit der Entscheidung zu ermöglichen. Welche Größen dafür relevant sind, muss vor der Inbetriebnahme der jeweiligen KI festgelegt werden.

Grundsatz 7
Der Entscheider muss den Gestaltungsgrundsätzen der funktionalen Sicherheit entsprechen.

Sicherheitsrelevante Entscheidungen benötigen z. B. für eine detaillierte Diagnose (Fehler- und Zustandserkennung) eine geeignete Hardware und Softwarestruktur. Deshalb ist der Entscheider auf Basis einer zum Anwendungsbereich passenden Norm der funktionalen Sicherheit zu konstruieren. Für jede Entscheidung ist das durch eine Risikobeurteilung ermittelte Sicherheitsniveau einzuhalten.

Diagnose und Entscheidung sind besonders vor dem Hintergrund zu konzipieren, dass fehlerhafte Zustände neben zufälligen Bauteilausfällen auch durch systematische Ausfälle oder Ausfälle gemeinsamer Ursache bedingt sein können. Solche Ausfälle können die Entscheidung negativ beeinflussen.

Ein technischer Entscheider erfordert ein hohes Maß an Sicherheit bei der Entscheidungsfindung. Insbesondere ist es wichtig für ihn zu erkennen, ob ein zufälliger Bauteilausfall oder ein Systemfehler vorliegt. Zu den Gestaltungsgrundsätzen gehört auch die Anwendung der festgeschriebenen Aktivitäten zum functional safetey management.

Anmerkung: Im Bereich der Maschinensicherheit ist beispielsweise die DIN EN ISO 13849 anzuwenden.

Grundsatz 8
KI-Technologien dürfen keine Sicherheit vortäuschen, die nicht vorhanden ist.

Wie im Grundsatz 1 dargestellt, sind Systeme auf der Basis von KI nach derzeitigem Stand der Technik kein Ersatz für klassische Schutzeinrichtungen zum Personenschutz im Sinne von DIN EN ISO 12100.

Auf KI-Technologie basierende Systeme, die assistierend oder unmittelbar steuernd in Maschinen oder Fahrzeugsteuerungen eingreifen, dürfen dem Bedienpersonal keine Vertrauenswürdigkeit bzw. Sicherheit vortäuschen, welche nicht durch technische Maßnahmen gewährleistet wird (Grundsatz 2). Der Hersteller hat dafür Sorge zu tragen, dass dem Anwender die sicherheitsrelevanten Systemgrenzen eindeutig bekannt gemacht werden. Wo immer möglich sollten diese Grenzen dem Benutzer während der Verwendung klar erkennbar angezeigt werden, z. B. wenn er sich in seinem Verhalten diesen Systemgrenzen nähert.

Dabei sind folgende Prinzipien zu berücksichtigen:
- Der Hersteller muss Funktion und Leistungsgrenzen des Systems eindeutig beschreiben.
- Sicherheitsniveau von Hard- und Software: Je höher der Automatisierungsgrad und das Risiko, desto höheres Sicherheitsniveau ist notwendig.
- Der Funktionszustand muss dem Bediener angezeigt werden.
- Eigenüberwachung/Fehlerdiagnose: In Abhängigkeit von der Kritikalität eines Fehlers muss ein sicherer Zustand eingeleitet werden.
- Die Gestaltung der Mensch-Maschine-Schnittstelle (M-M-S) muss ergonomischen Anforderungen entsprechen.
- Das System muss durch Menschen übersteuerbar bzw. stillsetzbar sein.

Grundsatz 9
Die Security (Schutz des KI-Systems vor Manipulationen) muss sichergestellt werden.

Auch beim Einsatz von Systemen der KI muss in jeder Phase des Produktlebenszyklus ein ausreichender (Personen)Schutz gewährleistet sein.

Weder die KI des Systems noch ein nicht autorisierter Zugriff von außen darf die funktionale Sicherheit (Safety) des Gesamtsystems derart ändern, dass der Personenschutz nicht mehr gewährleistet ist. Hierfür ist es erforderlich, dass die sicherheitsbezogenen Teile der Steuerung des Systems (funktionale Sicherheit/Safety) jederzeit Vorrang vor der Ablaufsteuerung einschließlich der KI behalten und vor Manipulation (sowohl durch die KI als auch durch externen Zugriff) geschützt sind. Dies beinhaltet eine ausreichende Erfüllung der Anforderungen an die IT-Sicherheit (Security) gemäß der Normenreihe IEC 62443 und erstreckt sich sowohl auf die Anforderungen für einzelne verwendete Sicherheitsbauteile/Komponenten (DIN EN IEC 62443-4-2, DGUV Test Prüfgrundsätze GS-IFA-M24) als auch auf das Gesamtsystem.

Grundsatz 10
Das Feldverhalten ist zu beobachten.

KI hat den Vorteil, dass sie allgemeine Fälle generalisiert und somit zur Lösung komplexer Aufgaben in komplexen Umgebungen eingesetzt werden kann. Das bringt jedoch das Problem mit sich, dass im Vorfeld weder die vorhergesehene Verwendung noch die vorhergesehene Umgebung vollständig spezifiziert werden kann.

Beispielsweise kann das vorhergesehene Arbeitsumfeld eines selbstfahrenden Fahrzeuges, welches im Straßenverkehr eingesetzt werden soll, nicht vollständig spezifiziert werden, da hierzu eine detaillierte und zu jeder Zeit aktuelle Beschreibung sämtlicher Straßen mitsamt ihrer Umgebung notwendig wäre. Ähnliche Schwierigkeiten ergeben sich bei Systemen der Personenerkennung, da Menschen eine hohe optische Diversität aufweisen.

In der Spezifikationsphase muss daher eine Analyse erfolgen, welche die vorhergesehene Anwendung sowie das vorhergesehene Arbeitsumfeld untersucht und elaboriert, welche Fälle durch das zukünftige Modell in jedem Fall abgedeckt werden müssen. Hierzu zählt insbesondere auch eine Identifikation seltener Fälle. Die hohe Komplexität bringt jedoch mit sich, dass auch eine weitreichende Analyse im Vorfeld einige Ausnahmesituationen übersehen kann. Zudem kann sich im Laufe der Zeit das vorhergesehene Arbeitsumfeld verändern.

Aus diesem Grund ist es notwendig, die Auswirkungen des algorithmischen Systems regelmäßig zu überprüfen. Hierzu sollte der Hersteller ein System für die Sammlung und Überprüfung von Informationen über das Produkt einrichten, pflegen und aufrechterhalten. Das System sollte Informationen aus der Implementierungs- und Nachimplementierungsphase enthalten, ebenso wie alle öffentlich verfügbaren Informationen über ähnliche Produkte auf dem Markt.

Diese gesammelten Informationen sollten dann auf ihre mögliche Relevanz für die KI-Vertrauenswürdigkeit des Produkts geprüft werden. Insbesondere sollte bei der Bewertung beurteilt werden, ob bisher unentdeckte Risiken bestehen oder die bereits bewerteten Risiken nicht mehr akzeptabel sind.

Herausgegeben von:
Deutsche Gesetzliche Unfallversicherung e. V. (DGUV)
Glinkastraße 40 · 10117 Berlin
Telefon: 030 13001-0 (Zentrale)
E-Mail: info@dguv.de · Internet: www.dguv.de

Kontakt:
Geschäftsstelle DGUV Test
Alte Heerstraße 111 · 53757 Sankt Augustin
Telefon: 030 13001-4566
E-Mail: dguv-test@dguv.de

Bezug:
www.dguv.de/publikationen Webcode: p021992

Weitere Informationen: