02.23

In Kooperation mit:
baua: DGUV
LASI Basi LIA.nrw

74. Jahrgang
Februar 2023
ISSN 2199-7330
1424

www.SISdigital.de

sicher ist sicher

Verlag GmbH & Co. KG, Berlin 2023 (http://www.sisdigital.de) - 13.02.2023 10:38

Konzeptioneller Brandschutz

Von **Prof. Dr. Roland Goertz** und **Fabian Ladzinski, M.Sc.**
2022, 472 Seiten, mehr als 160 farbige Abbildungen,
fester Einband, € 59,90. ISBN 978-3-503-18863-5

Online informieren und versandkostenfrei bestellen:
www.ESV.info/18863

ESV ERICH SCHMIDT VERLAG

Lizenziert für Herrn Moritz Schneider.
Die Inhalte sind urheberrechtlich geschützt.

© Designsprache - stock.adobe.com

MORITZ SCHNEIDER · ANDRÉ STEIMERS

Vertrauenswürdige Künstliche Intelligenz

Das Themenfeld Künstliche Intelligenz wird derzeit als Schlüsseltechnologie der Zukunft mit großen Chancen in vielen Bereichen gehandelt. Mit Hilfe von Künstlicher Intelligenz lassen sich beispielsweise neuartige Schutzeinrichtungen und Assistenzsysteme realisieren, so dass ihre Bedeutung auch für den Arbeitsschutz stetig zunimmt. Etablierte Risikominderungsmaßnahmen in der Softwareentwicklung sind jedoch nur bedingt für Anwendungen in KI-Systemen geeignet, da diese neue Risikoquellen schaffen. Das Risikomanagement für Systeme, die KI einsetzen, muss daher an die neuen Problemstellungen angepasst werden. Dieser Artikel bietet einen kompakten Überblick, worauf es bei vertrauenswürdiger Künstlicher Intelligenz ankommt.

Künstliche Intelligenz (KI) wird allgemein als Überbegriff für eine Vielzahl unterschiedlicher Methoden und Algorithmen verstanden, deren Gemeinsamkeit darin liegt, Wissensrepräsentationen in Form von Modellen zu verwenden, um definierte und vorgegebene Aufgaben zu lösen. Heute wird der Begriff jedoch meist mit Methoden des maschinellen Lernens, insbesondere des Deep Learning, in Verbindung gebracht. Diese Methoden basieren auf computergestützten Verfahren, die es Systemen ermöglichen, aus Daten zu lernen, um Wissensrepräsentationsmodelle zu erstellen, die spezifische Aufgaben (Sprache in Text umwandeln, Bilder klassifizieren etc.) weitgehend automatisiert lösen. Dieser Fokus ergibt sich aus bedeutenden Fortschritten in diesen Bereichen in den letzten Jahren, die von einer Vielzahl von Faktoren angetrieben werden. Der allgemeine Trend zur Digitalisierung führt nicht nur zu großen Datenmengen, die es maschinellen Lernverfahren ermöglichen, Wissensrepräsentationen in angemessener Qualität zu erstellen, sondern auch zum Aufblühen verschiedener Open-Source-Initiativen, Frameworks und Biblio-

© Copyright Erich Schmidt Verlag GmbH & Co. KG, Berlin 2023 (http://www.sisdigital.de) - 13.02.2023 10:38

Lizenziert für Herrn Moritz Schneider.
Die Inhalte sind urheberrechtlich geschützt.

DIE AUTOREN

Moritz Schneider M.Sc.
ist Leiter des Sachgebiets Künstliche Intelligenz und Softwarearchitektur sowie Leiter des Kompetenzzentrums Künstliche Intelligenz und Big Data (KKI) im Institut für Arbeitsschutz der Deutschen Gesetzlichen Unfallversicherung. Er ist Hochschuldozent, betreut Abschlussarbeiten im Bereich Künstliche Intelligenz (Deep Learning) und Software Engineering und stellt sein Wissen und Können seit Jahren in den Dienst der Wissenschaft, um den Arbeitsschutz durch technische Lösungen voranzubringen.

Prof. Dr. André Steimers
ist Professor für intelligente Sensorik und vertrauenswürdige Künstliche Intelligenz am RheinAhrCampus der Hochschule Koblenz. Seine Forschungsschwerpunkte liegen in der Entwicklung neuer sicherer Technologien in sicherheitsgerichteten Bereichen wie der industriellen Automation oder der Medizintechnik. Weiterhin nutzt er seine langjährige Expertise in diesen Bereichen, um die Fragen der Prüfung und Zertifizierung neuer Technologien voranzutreiben und wirkt hierzu in mehreren nationalen und internationalen Normungsgremien mit.

theken. Dies ermöglicht es einem breiten Kreis von Anwender*innen, entsprechende Verfahren zu nutzen. Die Ergebnisse einer weltweiten Umfrage unter 2.737 Unternehmen (darunter 201 aus Deutschland) zum Einsatz von KI ergaben, dass in 26 % der befragten Unternehmen KI bereits weit verbreitet sei und 47 % der Organisationen erste KI-Anwendungen im Einsatz haben. Die verbleibenden 27 % gaben an, dass sie zumindest kleine Pilotinitiativen starten wollen [1]. Eine strukturelle Analyse aktueller Anwendungsbeispiele von KI im Bereich der Arbeitswelt ergibt, dass die Verarbeitung natürlicher Sprache (Natural Language Processing), Bildverarbeitung und -analyse (Computer Vision) sowie fahrerlose Fahrzeuge (Automated Vehicles) zu den häufigsten Anwendungsgebieten gehören. Angetrieben durch Erfolge in diesen Bereichen wird der schnell wachsende Markt der KI in den kommenden Jahren eine immer größere Rolle im Arbeitsschutz spielen [2, 3].

Natural Language Processing befasst sich mit der Verarbeitung natürlicher Sprache und kann auf die Analyse menschlich übermittelter Sprachdaten angewandt werden, um Wissen zu extrahieren oder künstliche Sprachnachrichten zu erzeugen. Sprachsynthese und -erkennung sind zwei traditionelle Anwendungen der Verarbeitung natürlicher Sprache, die z. B. in Mensch-Maschine-Schnittstellen für die Ein- und Ausgabe genutzt werden können. Nachrichten in natürlichen Sprachen können auch gleichzeitig durch maschinelle Übersetzung übermittelt werden, entweder als Text oder wiederum durch künstlich erzeugte Sprache. Eine Anwendung, die es Arbeitnehmern mit Sprach- und Hörbehinderungen erleichtert, am Arbeitsplatz teilzunehmen, ist ein sinnhaftes Anwendungsbeispiel der Technologie. Computer Vision beschäftigt sich mit der Verarbeitung und Analyse digitaler Bilder zur Informationsgewinnung. Diese Informationen können von der Extraktion geometrischer Strukturen bis zum Verstehen des Bildinhaltes reichen. Die Segmentierung und Klassifizierung von Bildinformationen gehört zu den wesentlichsten Aufgaben der Computer Vision. Bei der Segmentierung werden inhaltlich zusammenhängende Regionen eines Bildes identifiziert und gekennzeichnet. Insbesondere die semantische Segmentierung, bei der einzelne Objekte erkannt und Ihre kleinsten Bildanteile zusammengefasst werden, hat eine große Bedeutung bei allen Aufgabenstellungen, in denen es darum geht, Aussagen über eine Szene treffen zu können. Klassifikationsprobleme beschäftigen sich damit, Objekte einer Klasse, wie beispielsweise Maschine, Mensch, Werkzeug, zuzuordnen. Oftmals folgt einer semantischen Segmentierung eine Klassifizierung, die dann schließlich den als inhaltlich ähnlich identifizierten Objekten eine genaue Klasse zuordnen kann. Hiermit werden Objekte nicht nur räumlich, sondern auch inhaltlich klar identifiziert und können somit lokalisiert und benannt werden.

Der Bereich der fahrerlosen Fahrzeuge (Automated Vehicles) befasst sich mit der vollständigen Automatisierung von Fahrzeugen aller Art, so dass sie unfallfrei und ohne menschliche Kontrolle zu einem Ziel fahren können. Seit einiger Zeit gibt es bereits teilautomatisierte Systeme wie Einparkassistenten oder automatische Spurwechselsysteme. In bestimmten Fahrsituationen kommen nun auch Systeme hinzu, die das Fahrzeug selbstständig steuern können. Ein Beispiel hierfür ist der Autopilot des Autobauers Tesla, der nur ein teilautomatisiertes Level-2-System ist. Ein Fahrsystem der Stufe 2 erwartet vom Fahrer, dass er sich der Fahr- und Verkehrssituation ständig bewusst ist und jederzeit die Kontrolle über das Fahrzeug übernehmen kann. Erst ab der Automatisierungsstufe 5 ist das Auto immer vollständig unter der Kontrolle des Systems und der Fahrer nur noch als Gast anwesend. Allerdings hat es in diesem Bereich bereits erhebliche Fortschritte gegeben. So können die Testfahrzeuge der Firma Waymo derzeit im Schnitt mehr als 17.800 Kilometer zurücklegen, bevor der Einsatz eines Testfahrers erforderlich ist. Auch wenn dies noch nicht an die tatsächliche Leistung eines menschlichen Fahrers heranreicht, sollten diese Systeme bei kontinuierlicher Qualitätsverbesserung in den kommenden Jahren ein ähnliches Niveau erreichen. Andere halbautomatische Fahrzeuge finden sich zum Beispiel auf Flughäfen in Form von selbstfahrenden Flugzeugschleppern [5].

Gestaltung sicherer Systeme und Anwendungen der künstlichen Intelligenz

Mit Hilfe von KI lassen sich bisher dem Menschen vorbehaltene Aufgaben automatisieren, wodurch nicht nur der Bereich der industriellen Automatisierung nochmals an Bedeutung gewinnt: KI-Anwendungen aus dem Smart-Home-Bereich sind bereits heute vielen Privatpersonen aus ihrem persönlichen Alltag bekannt, allerdings ist zu beobachten, dass zunehmend auch viele neue Anwendungen in sicherheitsrelevante Bereiche vordringen. Hier wird KI schon heute zur Optimierung verfahrenstechnischer Prozesse, aber auch zur Steuerung von Maschinen und Anlagen eingesetzt, was natürlich gewisse Risiken mit sich bringt. Außerdem kann diese Technologie jedoch auch zur Minderung von bestehenden Risiken eingesetzt werden. Mit Hilfe von KI werden beispielsweise neue Lösungen für den Schutz von Maschinen und Anlagen umgesetzt oder Assistenzsysteme realisiert, die das Potenzial haben, zu einer deutlichen Redu-

© Copyright Erich Schmidt Verlag GmbH & Co. KG, Berlin 2023 (http://www.sisdigital.de) - 13.02.2023 10:38

Lizenziert für Herrn Moritz Schneider.
Die Inhalte sind urheberrechtlich geschützt.

zierung von Unfällen beizutragen. Damit diese Systeme jedoch zu einer effektiven Risikominderung beitragen können oder andere hochautomatisierte KI-Systeme nicht zu neuen Gefährdungen führen, müssen sie eine Vielzahl von Eigenschaften erfüllen, um sichere Ergebnisse erzielen zu können.

Insbesondere dort, wo Mensch und Maschine miteinander arbeiten, ist besonderes Augenmerk auf Sicherheit und Risikominimierung zu legen. Etablierte Risikominderungsmaßnahmen in der Softwareentwicklung sind jedoch nur bedingt für Anwendungen in KI-Systemen geeignet, da diese auch neue Risikoquellen schaffen. Das Risikomanagement für Systeme, die KI einsetzen, muss daher an die neuen Problemstellungen angepasst werden [4].

Im Kontext der Hoch- oder sogar Vollautomatisierung verschiedener Anwendungen spielen jedoch nicht nur sicherheitstechnische, sondern auch ethische Aspekte eine wichtige Rolle. So stellt sich vor der Erstellung eines automatisierten Systems zuallererst immer die Frage, wie hoch dessen Automatisierungsgrad sein soll oder anders ausgedrückt, welche Rolle der Mensch in diesem Prozess noch spielen sollte, da mit steigendem Automatisierungsgrad zumeist die Kontroll- und Zugriffsmöglichkeiten des Menschen auf das System sinken. Die Gesamtheit aller relevanten sicherheitstechnischen, aber auch ethischen Aspekte wird heute unter dem Begriff der vertrauenswürdigen Künstlichen Intelligenz gebündelt. Damit sie als vertrauenswürdig gelten kann, muss sie eine Vielzahl grundlegender Eigenschaften besitzen. Zum Beispiel muss ein solches System im Rahmen seiner definierten Aufgabe zuverlässig arbeiten. Darüber hinaus muss es aber auch robust gegenüber Bedingungen sein, die außerhalb seiner Systemspezifikation liegen. So dürfen beispielsweise Bedingungen wie interne und externe Störungen in Form von Fehlern oder unvorhersehbare Betriebszustände nicht zu einem unsicheren Ausfall des Systems führen. Natürlich müssen auch Aspekte der allgemeinen IT-Sicherheit und Industrial Security berücksichtigt werden. Vertrauen zu schaffen bedeutet aber auch, die Sicherheit der für die Modellerstellung und durch die Anwendung verarbeiteten Daten zu gewährleisten. Gerade beim Einsatz von Deep-Learning-Verfahren spielen die Transparenz und Erklärbarkeit des Modells eine wichtige Rolle. Hohe Leistung geht in der Regel mit komplexen Modellen einher, die für den Menschen nicht einfach zu verstehen sind. Dies erschwert die Überprüfbarkeit des Systems.

Die einzelnen KI-Eigenschaften, die einen Einfluss auf die Vertrauenswürdigkeit haben, beschreiben die größten Risikoquellen für ein KI-System und sollten daher im Rahmen einer Risikobewertung berücksichtigt werden. Gleichzeitig können auf Basis der Eigenschaften einer vertrauenswürdigen KI allgemeine Gestaltungsprinzipien für ein solches System definiert werden.

Die nachfolgende Liste stellt eine Taxonomie dar, welche die relevanten Eigenschaften bzw. größten Risikoquellen eines KI-basierten Systems beschreibt [4]:

Ethische Aspekte

1. Fairness
2. Privacy
3. Automatisierungsgrad und Kontrolle

Aspekte der Zuverlässigkeit und Robustheit

4. Komplexität der Aufgabe und Verwendungsumgebung
5. Grad der Transparenz und Erklärbarkeit
6. Security
7. System-Hardware
8. Technologische Ausgereiftheit

Es ist zu sehen, dass hier manche Risikofelder eher die Zuverlässigkeit und Robustheit des Systems betreffen, während andere Aspekte eher auf ethische Aspekte abzielen. Dabei ist jedoch zu beachten, dass es hier zu keiner klaren Trennung kommen kann, so führen beispielsweise Maßnahmen, die der Fairness als primär ethischem Aspekt der Anwendung zugutekommen ebenso zu einer Erhöhung der Zuverlässigkeit und Robustheit. Generell lässt sich sagen, dass eine enge Verzahnung zwischen den einzelnen Risikofeldern, aber auch den Aspekten existiert.

Es wurde bereits erwähnt, dass der Automatisierungsgrad eines der grundlegendsten Merkmale, die ein KI-System definieren, ist. Der Automatisierungsgrad beschreibt das Ausmaß, in dem ein KI-System unabhängig von menschlicher Aufsicht arbeiten kann, aber auch seine Unabhängigkeit von menschlicher Kontrolle. Systeme mit einem hohen Automatisierungsgrad können ein unerwartetes Verhalten zeigen, das schwierig zu erkennen und zu kontrollieren ist. Hochautomatisierte Systeme können daher Risiken in Bezug auf ihre Zuverlässigkeit und Sicherheit bergen, aber auch den Menschen aus bestimmten Prozessen verdrängen, was wiederum einen direkten Einfluss auf die Fairness haben kann. Insbesondere bei Entscheidungen, die einen direkten Einfluss auf das Leben eines Menschen haben, muss sichergestellt sein, dass diese fair sind bzw. nicht diskriminieren. Oft genannte Beispiele hierfür sind Systeme zur Bewerberauswahl oder Kreditvergabe. Allerdings können natürlich auch Steuerungen von

© Copyright Erich Schmidt Verlag GmbH & Co. KG, Berlin 2023 (http://www.sisdigital.de) - 13.02.2023 10:38

Lizenziert für Herrn Moritz Schneider.
Die Inhalte sind urheberrechtlich geschützt.

Maschinen einen erheblichen Einfluss auf den Menschen haben, da diese direkt oder indirekt die Sicherheit und Gesundheit eines Menschen gefährden können.

KI wird sinnvollerweise dort eingesetzt, wo der Einsatz herkömmlicher Technologien nicht möglich ist. Dies ist insbesondere bei hochkomplexen Aufgaben oder dem Einsatz in komplexen Umgebungen der Fall. Hieraus ergeben sich oft hochdimensionale Zustandsräume, welche durch Modelle abgebildet werden müssen. Die Schwierigkeit besteht nun darin, dass ein bereits vollständig trainiertes und in das System implementiertes KI-Modell sich nicht mehr auf Veränderungen der Umgebung einstellen kann. Das ist nicht nur beim Einsatz in anderen Betriebsumgebungen relevant und wird kritisch, wenn das System außerhalb der spezifizierten Betriebsumgebung eingesetzt wird, sondern auch beim Outdoor-Einsatz, wo sich die Umgebung durch das Wetter oder über die Jahreszeiten hinweg ändern kann. Weiterlernende KI-Systeme, die sich auf neue Umgebungen einstellen können, sind dabei nur theoretisch eine Lösung. In der Praxis bieten diese keine Möglichkeit einer ausreichenden Selbstvalidierung, wodurch die Gefahr besteht, dass sich diese Modelle nicht verbessern, sondern sich durch den kontinuierlichen Trainingsprozess sogar verschlechtern. Hier ist also eine möglichst umfangreiche und genau auf die Gegebenheiten der Aufgabe und Verwendungsumgebung abgestimmte Spezifikation erforderlich.

Aus dieser Komplexität leitet sich wiederum die Komplexität mancher KI-Modelle ab. Hier kommen Eigenschaften wie die Transparenz und Erklärbarkeit der Modelle zum Tragen. Transparenz beschreibt die Eigenschaft eines Systems, die besagt, inwieweit geeignete Informationen über das System an relevante Interessengruppen weitergegeben werden, während Erklärbarkeit die Eigenschaft eines KI-Systems beschreibt, wichtige Faktoren, die die Ergebnisse des KI-Systems beeinflussen, in einer für Menschen verständlichen Weise auszudrücken. Besonders KI-Modelle, die auf Methoden des Deep Learning basieren, sind meist sehr umfangreich und damit für den Menschen nicht intrinsisch erklärbar. Hier helfen Verfahren wie LIME (Local Interpretable Model-Agnostic Explanations) oder LRP (Layerwise Relevance Propagation), die erlauben, sich die relevanten Features des Modells darstellen zu lassen, um validieren zu können, ob diese sinnvoll sind.

Oftmals außen vorgelassen, aber auch beim Einsatz von KI-Systemen relevant sind hardwarebezogene Fehler, welche während der Trainingsphase oder im Betrieb eines KI-Systems die korrekte Ausführung des Algorithmus negativ beeinträchtigen oder sogar vollständig unterbinden können. Daneben gibt es ebenso zu beachten, dass auch die Security solcher Systeme relevant ist, da durch die Verwendung mancher KI-Methoden neue Angriffsvektoren hinzukommen. Die technologische Ausgereiftheit beschreibt schließlich, wie ausgereift und fehlerfrei eine bestimmte Technologie in einem bestimmten Anwendungskontext ist. Werden bei der Entwicklung des KI-Systems weniger ausgereifte und neue Technologien verwendet, können diese noch unbekannte oder schwer einzuschätzende Risiken in sich bergen. Für bereits ausgereifte Technologien steht hingegen in der Regel eine größere Vielfalt an Erfahrungsdaten zur Verfügung, wodurch Risiken leichter zu identifizieren und zu bewerten sind. Allerdings besteht bei ausgereiften Technologien die Gefahr, dass das Risikobewusstsein über die Zeit abnimmt, so dass die positiven Effekte von einer ständigen Risikoüberwachung und einer angemessenen Wartung abhängen.

Die Bedeutung des KI-Bereichs wurde auch von der Europäischen Kommission erkannt, welche 2021 einen ersten Entwurf für eine EU-Verordnung zur Regulierung Künstlicher Intelligenz veröffentlicht hat. [6] Nach Abschluss des Rechtssetzungsverfahrens zu dieser Verordnung wird ein großer Bedarf bestehen, deren Anforderungen zu konkretisieren. Hierzu werden entsprechende internationale Normen und Standards benötigt.

Auf internationaler Ebene existiert beispielsweise das gemeinsamen Komitee JTC1 SC42 von ISO und IEC. Insbesondere das Thema vertrauenswürdige KI findet dort besondere Beachtung und wird in einer eigenen Arbeitsgruppe behandelt, die sich mit Themen wie Bias, erklärbare KI, aber auch Risikomanagement für KI-Systeme beschäftigt. Auf europäischer Ebene existiert weiterhin das Gemeinsame Gremium JTC 21 der CEN und CENELEC, welche insbesondere bei der Harmonisierung der internationalen Normen für die zukünftige KI-Verordnung Relevanz haben wird. Die hohe Innovationsdynamik im Bereich der KI stellt eine besondere Herausforderung dar, so dass die Normung in enger Zusammenarbeit mit der Forschung erfolgen muss.

Zusammenfassung

Obwohl es bereits viele Beispiele für den Einsatz von KI am Arbeitsplatz gibt, wird dieser Bereich in den kommenden Jahren für eine Vielzahl von Branchen, insbesondere im Bereich der Sicherheit und des Gesundheitsschutzes der Beschäftigten bei der Arbeit – und damit für die gesetzliche Unfallversicherung – weiter stark an Bedeutung gewinnen. So soll einer Studie zu-

© Copyright Erich Schmidt Verlag GmbH & Co. KG, Berlin 2023 (http://www.sisdigital.de) - 13.02.2023 10:38

Lizenziert für Herrn Moritz Schneider.
Die Inhalte sind urheberrechtlich geschützt.

folge das quantitative ökonomische Finanzvolumen (in Bezug auf automatisierte Roboter und Fahrzeuge sowie auf Datenanalyse) bis 2025 voraussichtlich zwischen sechs und zwölf Billionen Euro jährlich betragen. [7]

Es wird eine gesellschaftliche Diskussion darüber geben müssen, wie KI-basierte Systeme wie z. B. kollaborative Roboter am besten zu gestalten sind und ob alles, was theoretisch technisch möglich ist, auch ethisch wünschenswert ist oder legal sein sollte. Das Institut für Arbeitsschutz der DGUV (IFA) entwickelt daher Ideen, wie Sicherheit und Gesundheit am Arbeitsplatz beim Einsatz von Technologien der Künstlichen Intelligenz erhalten und gefördert werden können. In diesem Zusammenhang betreibt das IFA eigene KI-bezogene Forschung und beteiligt sich an der weltweiten und europäischen Normungsarbeit [8].

LITERATUR

[1] *Ammanath, B.; Hupfer, S.; Jarvis, D.: Thriving in the era of pervasive AI, Deloitte Insights, 2020*
[2] *Charlier, R.; Kloppenburg, S. Artificial Intelligence in HR: A No-Brainer. PwC. Available online: www.pwc.nl/nl/assets/documents/artificial-intelligence-in-hr-a-no-brainer.pdf*
[3] *PwC. AI Will Create as Many Jobs as It Displaces by Boosting Economic Growth. PwC. Available online: www.pwc.co.uk/press-room/press-releases/AI-will-create-as-many-jobs-as-it-displaces-by-boosting-economic-growth.html*
[4] *Steimers, André and Moritz Schneider. „Sources of Risk of AI Systems." International Journal of Environmental Research and Public Health 19 (2022)*
[5] *Morris, R.: Planning, Scheduling and Monitoring for Airport Surface Operations, Workshop of the 13th AAAI Conference on Artificial Intelligence: Planning for Hybrid Systems, 2016*
[6] *zur EU-KI-Verordnung vgl. https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:52021PC0206*
[7] *Purdy, M.; Daugherty, P.: Why AI is the future of growth, Accenture Plc., 2016*
[8] *Kompetenzzentrum Künstliche Intelligenz und Big Data (KKI): www.dguv.de/ifa/fachinfos/kuenstliche-intelligenz/kompetenzzentrum-kuenstliche-intelligenz-(kki)-und-big-data/index.jsp*

© Copyright Erich Schmidt Verlag GmbH & Co. KG, Berlin 2023 (http://www.sisdigital.de) - 13.02.2023 10:38