Zuverlässige KI: Absicherung künstlicher neuronaler Netze

Originalartikel

Backup

<html> <header class=„article-header“><h1 class=„articleheading“>Zuverlässige KI: Absicherung künstlicher neuronaler Netze</h1> Marco Huber </header><figure class=„aufmacherbild“><figcaption class=„akwa-caption“>(Bild: Peshkova / shutterstock.com)</figcaption></figure> Wie zuverlässig eine KI-Anwendung ist, lässt sich messen. Das ist oft mathematisch komplex, hat aber enorme praktische Bedeutung. Wissen, Halbwissen und Falschinformationen liegen beim Thema Künstliche Intelligenz (KI) nahe beieinander. Nicht erfüllte oder übererfüllte Erwartungen gehen häufig einher mit Überraschungsmomenten, wenn klar wird, wie lange es KI gibt und wie stark sie inzwischen den Alltag privat wie beruflich durchdringt. Dieses Hineinwirken in zahlreiche gesellschaftliche Bereiche bedingt, dass der öffentliche Diskurs über KI intensiver ist und auch sein sollte als für andere Technologien. Neben dem lebhaften Diskurs in der Gesellschaft existieren rechtliche Fragen und Hürden, die Unternehmen angehen müssen, wenn sie Künstliche Intelligenz sicher einsetzen möchten. Mit der steigenden Zahl der KI-Anwendungen wächst die Notwendigkeit zu prüfen, inwieweit bestehende Normen, Standards und Regularien für sie greifen oder ob es weiterer Maßnahmen für eine rechtssichere Anwendung bedarf. <h3 class=„subheading“ id=„nav_licht_und0“>Licht und Schatten beim Einsatz von KI</h3> Der aktuelle Wirbel um KI ist das Ergebnis einer jahrzehntelangen Forschungsarbeit seit den 1950er Jahren. In der letzten Zeit häufen sich die Erfolgsmeldungen: ChatGPT beeindruckt [1] in diversen Domänen mit flüssig formulierten und gut lesbaren Texten. Die Sprachassistenten Siri und Alexa sind in viele Haushalte eingezogen. Das Computerprogramm AlphaGo [2] von DeepMind schlug 2016 einen der weltbesten Go-Spieler. Und mit AlphaFold hat Google [3] eine KI trainiert, die extrem präzise die räumliche Struktur von Proteinen vorhersagen kann. Die Erfolge sind nur die eine Seite der Medaille. Auf der anderen Seite stehen ethische und sicherheitsrelevante Herausforderungen. So setzte Amazon beispielsweise eine KI-basierte Software für den Vorauswahlprozess neuer Mitarbeiter

[4] ein, die aufgrund der Trainingsdaten Frauen benachteiligte. Breit durch die Medien gingen zudem zwar seltene, aber schwere Unfälle mit Autos, die autonom fuhren und deren Bildverarbeitung nicht richtig funktionierte. Ein wahrer Misserfolg war der Chatbot Tay [5], der auf Twitter von den Nutzern lernen sollte, die ihm aber schnell rassistische Äußerungen beibrachten. Es ist kein Selbstläufer, KI-Techniken sinnvoll und rechtskonform einzusetzen. <h3 class=„subheading“ id=„nav_die1“>Die Vertrauensfrage</h3> Das ambivalente Bild bewirkt etwas bei Menschen und ihrer Haltung gegenüber KI. Der 2020 von Bosch umgesetzt „KI-Zukunftskompass [6]“ zeigt, wo die befragten Personen eher Menschen oder einer KI vertrauen würden. Die Umfrage ergab, dass KI in jenen Bereichen hohes Vertrauen genießt, in denen das Umfeld stark technisiert ist und keine direkten Bezüge zum Menschen hat, etwa in der industriellen Produktion oder der Herstellung von Autos und Flugzeugen. Zwischen 38 und 57 Prozent der Befragten sprachen einer Maschine in dem Bereich mehr Vertrauen aus als einem Menschen, für den die Werte zwischen 11 und 22 Prozent lagen. Allerdings sinkt das Vertrauen in Maschinen, wenn eine Anwendung Menschen direkt betrifft: Zwischen 65 und 79 Prozent der Befragten vertrauten Menschen, wenn es um gesundheitliche Anforderungen, die Pflege oder Personalentscheidungen geht. Nur sechs bis zehn Prozent vertrauten in den Bereichen den Maschinen stärker. Je näher eine KI-basierte Entscheidung oder Aktion dem Menschen rückt und je mehr sie sein Leben beeinflussen könnte, umso kritischer wird der Umgang mit KI – ein nachvollziehbarer Befund. <h3 class=„subheading“ id=„nav_menschzentrierte2“>Menschzentrierte und zuverlässige KI als Ziel</h3> Umso wichtiger ist es, Gefahren und Skepsis auf Anwenderseite beim Entwickeln von KI-Anwendungen im Fokus zu haben. Das versteht das Team im Stuttgarter KI-Fortschrittszentrum „Lernende Systeme und Kognitive Robotik“ [7], geleitet von den beiden Fraunhofer-Instituten für Produktionstechnik und Automatisierung IPA sowie für Arbeitswirtschaft und Organisation IAO und gefördert vom Wirtschaftsministerium Baden-Württemberg, unter einer menschzentrierten KI. Dass diese zuverlässig und vertrauenswürdig ist, zeigt sich in sechs Aspekten: <ol class=„rtelist rtelist–ordered“>
Menschlichem Handeln und Aufsicht,
Fairness,
Datenschutz,
Sicherheit,
Verlässlichkeit und
Transparenz. Die ersten drei Aspekte betreffen überwiegend ethische Themen. Die Schwerpunkte des KI-Fortschrittszentrums liegen auf den sicherheitsrelevanten Aspekten vier bis sechs. Zunächst ist jedoch eine kurze Einordnung erforderlich: Künstliche Intelligenz steht in diesem Artikel für Methoden und Verfahren zur Problemlösung, für die Menschen üblicherweise intelligent handeln müssen. Ein prominentes und momentan das am stärksten genutzte Teilgebiet von KI ist Machine Learning (ML), also das überwachte oder unüberwachte Lernen anhand von Mustern in Daten. Künstliche neuronale Netze mit Verfahren des Deep Learning gehören dazu. <h3 class=„subheading“ id=„nav_sicherheit3“>Sicherheit durch formale Verifikation</h3> Ein Beispiel des Straßenverkehrs [8] [1], in dem ein autonom fahrendes Auto entscheiden muss, verdeutlicht ein grundlegendes Sicherheitsproblem beim Einsatz neuronaler Netze: Kleine Änderungen an den Eingabedaten können zu großen Änderungen beim ausgegebenen Ergebnis führen. Beispielsweise könnte die im Auto eingebaute Kamera aufgrund von Sensorrauschen ein Verkehrszeichen oder einen Verkehrsteilnehmer falsch erkennen. Das kann zu fehlerhaften Fahrbefehlen führen und somit die Insassen oder andere Verkehrsteilnehmer gefährden. Dieses Verhalten

lässt sich als fehlende Robustheit bezeichnen und sogar vorsätzlich durch sogenannte Adversarial Attacks ausnutzen. Die Angriffe verändern die Eingabedaten bewusst, um gezielt ein falsches Ergebnis zu erzeugen (siehe Abbildung 1). <figure class=„a-inline-image a-u-inline“> <figcaption class=„a-caption“>Durch Hinzufügen eines für den Menschen nicht sichtbaren Rauschens ändert sich die vormals korrekte Ausgabe „Panda“ eines neuronalen Netzes in die falsche Ausgabe „Gibbon“.(Bild: arXiv [9])</figcaption></figure> Damit ein neuronales Netz sicher und robust ist, wäre es erforderlich, alle denkbaren Eingabedaten auszuprobieren, was aufgrund der potenziell unendlichen Anzahl an Daten in vielen Anwendungsbereichen unmöglich ist. Bisher ist der Standard, statistische Aussagen über die Sicherheit des Netzes zu treffen, indem man es auf einer endlichen Teilmenge an Daten testet. Eine Alternative ist die formale Verifikation. Sie beweist automatisch, dass ein Netz gewisse Eigenschaften für unendlich viele Daten erfüllt. Beispielsweise soll gezeigt werden, dass sich in einem binären Klassifikationsproblem ein neuronales Netz f(x) für alle Eingabedaten x ∈ B ⊂ R^n immer für eine der beiden Klassen -1 oder 1 entscheidet. Häufig ist die Menge B konvex, beispielsweise eine Hyperkugel, die um einen bestimmten Datenpunkt x_0 zentriert ist. Damit lässt sich nachweisen, dass das Netz bei kleineren Abweichungen von x_0 seine Entscheidung für eine Klasse nicht ändert. Es wäre somit robust gegenüber kleinen Änderungen, die etwa durch Rauschen entstehen. Für den Fall, dass positive Ausgabewerte des Netzes als Entscheidung für die Klasse 1 und negative Werte für die Klasse -1 gelten, lässt sich das Verifikationsaufgabe als folgendes Optimierungsproblem formulieren: <figure class=„a-inline-image a-u-inline“> </figure> $f^* > 0$ bedeutet, dass selbst im schlechtesten Fall das Netz immer einen positiven Wert ausgibt und zwar unabhängig vom betrachteten Wert x. Das Netz ist beweisbar robust. Ist hingegen $f^* < 0$, gibt es Werte in B, bei denen sich das Netz für die Klasse -1 entscheidet. Unglücklicherweise ist das Lösen des Optimierungsproblems äußerst aufwendig, weil neuronale Netze in jeder Schicht nichtlineare Aktivierungsfunktionen verwenden und f(x) somit keine konvexe Funktion ist. Eine Abhandlung [10] [2] hat für ReLU-Netze – neuronale Netze mit ReLU-Aktivierungsfunktion (Rectified Linear Unit) – nachgewiesen, dass das Problem NP-vollständig ist und damit der Rechenaufwand exponentiell

mit der Größe des Netzes ansteigt. Um dennoch neuronale Netze praktikabler Größe verifizieren zu können, beschäftigt sich die Forschung mit näherungsweisen Lösungen des Optimierungsproblems. Beim Bewerten eines solchen Näherungsverfahrens ist darauf zu achten, ob das Verfahren korrekt und vollständig ist (siehe Abbildung 2). Idealerweise erfüllt das Verfahren beide Eigenschaften, da in dem Fall die Ergebnisse stets richtig sind. <figure class=„a-inline-image a-u-inline“> <figcaption class=„a-caption“>Zusammenhang der algorithmischen Eigenschaften der Korrektheit und Vollständigkeit (Abb. 2)</figcaption></figure> Ein Beispiel für korrekte, aber nicht vollständige Näherungsverfahren sind solche, die das Netz f(x) durch eine untere Schranke abschätzen, beispielsweise durch eine lineare Funktion [3]: <figure class=„a-inline-image a-u-inline“> </figure> Verwendet man g(x) statt f(x), ist das Optimierungsproblem häufig sogar geschlossen lösbar. Gilt dabei g*>0, ist weiterhin garantiert, dass folgende Eigenschaft erfüllt ist: Das Netz entscheidet sich stets für die Klasse 1. Allerdings lässt sich für den Fall g*<0 keine Aussage mehr treffen, da immer noch 0 < f*≦ f(x) für alle x ∈ B gelten kann. <div class=„a-boxtarget a-boxcontent a-inline-textboxcontent“ data-collapse-target=„“><figure class=„a-inline-textboximage-container“></figure> Am 10. und 11. Mai findet die Minds Mastering Machines in Karlsruhe [11] statt. Die seit 2018 von iX, heise Developer und dpunkt.verlag ausgerichtete Fachkonferenz richtet sich in erster Linie an Data Scientists, Data Engineers und Developer, die Machine-Learning-Projekte in die Realität umsetzen. Das Programm bietet an zwei Tagen [12] gut 30 Vorträge unter anderem zu Sprachmodellen, Cybersecurity, Resilienz und Modelloptimierung. Für ReLU-Netze kann man das Optimierungsproblem in ein sogenanntes Mixed-Integer Linear Program (MILP, gemischt ganzzahliges Optimierungsproblem) umwandeln, wofür spezielle Algorithmen wie Branch-and-Bound-Verfahren existieren. Damit ist es zwar immer noch NP-vollständig, aber die zuvor

erwähnten schnellen schrankenbasierten Verfahren lassen sich mit Branch and Bound kombinieren [4], um die Verifikation zu beschleunigen. Zudem sind solche Verfahren vollständig und korrekt. Gemeinsam mit der Firma Sick, einem Hersteller von Sensoren für industrielle Kontexte, gab es im KI-Fortschrittszentrum ein Projekt, bei dem es um die formale Verifikation anhand einer Anwendung aus der Sicherheitstechnik ging. Es galt nachzuweisen, dass ein neuronales Netz mittels Daten eines Laserscanners einen Menschen sicher erkennen kann. Untersucht wurde ein MILP-Verifikationsverfahren. Allerdings erwies es sich als schwierig, überprüfbare Sicherheitseigenschaften aus den Laserscannerdaten zu extrahieren. Zudem zeigte sich, dass bei der Skalierung auf extrem große neuronale Netze noch Forschungsbedarf besteht. <h3 class=„subheading“ id=„nav_verlässlichkeit4“>Verlässlichkeit durch Unsicherheitsquantifizierung</h3> Der zweite technische Aspekt zur Absicherung einer KI ist die Bewertung der Verlässlichkeit. Man sollte annehmen, dass beispielsweise in der Bildverarbeitung ein System aufwendig trainierte Objekte als solche zuverlässig wiedererkennt. Dass die Zuverlässigkeit mitunter überschätzt wird, belegt ebenfalls Abbildung 1: Das KI-Modell klassifiziert das mit Rauschen versetzte Bild mit hoher Konfidenz falsch als Gibbon. Was Menschen ignorieren oder ihnen nicht auffällt, stört den Algorithmus dermaßen, dass er die eigentlich bekannten Objekte nicht mehr zuverlässig genug erkennt und sogar bei Fehlklassifikationen eine hohe Zuverlässigkeit für das scheinbar richtig erkannte Objekt ausgibt. Das wiegt User in falscher Sicherheit. Technisch betrachtet ergeben sich die Fehler, weil die meisten Lernverfahren und damit die neuronalen Netze Punktschätzer sind. Sie können nur einen konkreten Zahlenwert ausgeben, aber nicht angeben, mit welcher Sicherheit dieser Wert gilt. Zwar lassen sich Konfidenzwerte wie in Abbildung 1 berechnen, die jedoch lediglich auf das Intervall [0, 1] normierte Ausgabewerte darstellen und somit keine statistisch gültige Bedeutung besitzen. Dass diese Ausgabewerte oft fälschlicherweise als Konfidenzwerte oder Wahrscheinlichkeiten bezeichnet werden, trägt zur vermittelten falschen Sicherheit bei. Einen Ausweg bietet die Unsicherheitsquantifizierung, mit der sich das neuronale Netz selbst einschätzen und mitteilen kann, wenn es bei einer Ausgabe unsicher ist. Hierfür gibt es unterschiedliche Ansätze. Verbreitet ist das Ensemble-Verfahren, bei dem Data Scientists nicht nur ein Netz, sondern gleich mehrere Netze auf denselben Daten trainieren. Dabei achten sie darauf, dass die Netze hinreichend divers sind, also nicht dazu neigen, stets die gleichen Ergebnisse zu erzeugen. Alle Netze erzeugen eigene Ausgabewerte, aus denen man die Gesamtausgabe berechnet. Zusätzlich kann man die Streuung oder Abweichung der einzelnen Ausgaben von der Gesamtausgabe als ein Maß für die Unsicherheit angeben: Je weniger die einzelnen Netze streuen, umso zuverlässiger ist die Gesamtausgabe. Dass dabei stets mehrere Netze parallel zu trainieren und auszuwerten sind, führt allerdings zu einem erheblich erhöhten Rechenaufwand. Eine Alternative mit nur einem Netz bieten Bayes'sche neuronale Netze (BNN) [5]. Im Unterschied zu klassischen neuronalen Netzen repräsentieren nicht mehr einzelne Zahlenwerte die Gewichte auf den Kanten, die die Neuronen miteinander verbinden. Stattdessen ist jedes Gewicht eine Zufallsvariable (siehe Abbildung 3). Dadurch wird die Ausgabe des BNN ebenfalls zu einer Zufallsvariable, für die sich gängige statistische Kenngrößen wie Varianz, Entropie oder ein Prädiktionsintervall berechnen lassen. <figure class=„a-inline-image a-u-inline“> <img alt=„“ class=„legacy-img“ height=„299“ src=„https://heise.cloudimg.io/width/696/q85.png-lossy-85.webp-lossy-85.foil1/_www-heise-de_/imgs/18/4/1/4/2/5/7/7/abb3-0226d3caab5ec755.png“ srcset=„https://heise.cloudimg.io/width/336/q70.png-lossy-70.webp-lossy-70.foil1/_www-heise-de_/imgs/18/4/1/4/2/5/7/7/abb3-0226d3caab5ec755.png 336w, https://heise.cloudimg.io/width/1008/q70.png-lossy-70.webp-lossy-70.foil1/_www-heise-de_/imgs/18/4/1/4/2/5/7/7/abb3-0226d3caab5ec755.png 1008w, https://heise.cloudimg.io/width/1392/q70.png-lossy-70.webp-lossy-70.foil1/_www-heise-de_/imgs/18/4/

1/4/2/5/7/7/abb3-0226d3caab5ec755.png 2x“ width=„696“ referrerpolicy=„no-referrer“ /> <figcaption class=„a-caption“>Das linke Diagramm zeigt ein klassisches neuronales Netz, bei dem die Kantengewichte (blau) einfache Zahlenwerte darstellen. Beim BNN im rechten Diagramm sind die Gewichte Zufallsvariablen mit dazugehörigen Wahrscheinlichkeitsverteilungen (gelb) (Abb. 3).</figcaption></figure> Allerdings sind für das Training der BNN die für klassische Netze üblichen Trainingsalgorithmen wie Backpropagation nicht mehr ohne Weiteres nutzbar, sondern es sind Trainingsverfahren mit höherem Rechenaufwand erforderlich. In den vergangenen Jahren hat die Conformal Prediction (CP, [6]) an Bedeutung gewonnen. Sie hat den Vorteil, verteilungsfrei zu sein: Es gibt keine Annahmen über die zugrundeliegende Datenverteilung oder das verwendete Lernverfahren. Daher lassen sich CPs auf fast jedes trainierte maschinelle Lernmodell anwenden, und sie bieten mathematisch strenge Garantien. Für ein Signifikanzniveau von fünf Prozent ist garantiert, dass das Verfahren in höchstens fünf Prozent der Fälle falsche Ausgaben erzeugt. Die Ausgabe ist keine Punktschätzung mehr, sondern eine Menge. Im Falle einer Klassifikationsaufgabe gibt das neuronale Netz folglich keine einzelne Klasse mehr aus, sondern eine Menge infrage kommender Klassen. Je weniger Elemente die Menge enthält, desto sicherer ist sich das Netz. <h3 class=„subheading“ id=„nav_transparenz5“>Transparenz durch Erklärbarkeit</h3> Schließlich gilt es, für zuverlässige KI die Erklärbarkeit des Verfahrens deutlich zu verbessern. Neuronale Netze sind meist sogenannte Black-Box-Modelle. Auch für Expertinnen und Experten ist es oft schwer bis unmöglich nachzuvollziehen, warum ein bestimmtes Ergebnis zustande gekommen ist. Das stellt Unternehmen unter anderem mit Blick auf die 2018 eingeführte Datenschutzgrundverordnung vor Herausforderungen, da Personen nun das Recht haben, die Hintergründe über eine automatisch getroffene Entscheidung zu erfahren. Das Gefühl, die KI führe eine Art Eigenleben, senkt zudem die Nutzerakzeptanz. Ein prominentes Beispiel ist ein neuronales Netz, das Bilder von Hunden und Wölfen unterscheiden sollte [7]. Es entschied fast immer richtig, aber eine fehlerhafte Entscheidung zeigte, welches Kriterium es tatsächlich zur Klassifizierung genutzt hatte: Es waren nicht etwa Merkmale der Tiere, sondern zufälligerweise waren die Wölfe immer auf Bildern zu sehen, in denen im Hintergrund Schnee lag – nur bei dem falsch klassifizierten Bild nicht. Das Beispiel offenbart, dass viele KI-Systeme intransparent, nicht intuitiv und für Menschen schwer zu verstehen sind. Die Unfähigkeit der Systeme, die Entscheidungen und Aktionen erklären zu können, schränkt die Vorteile und die Wirksamkeit Künstlicher Intelligenz ein. Erklärbare KI ist sowohl aus rechtlichen Gründen als auch für das Vertrauen der Nutzerinnen und Nutzer unerlässlich [8]. In kritischen Anwendungen, bei denen es womöglich um Menschenleben geht wie in der Medizin und beim autonomen Fahren, ist es besonders riskant, wenn die KI nicht erklärbar ist. <h3 class=„subheading“ id=„nav_lokale_und6“>Lokale und globale Erklärbarkeit</h3> Erklärbarkeit lässt sich auf zwei Ebenen erzielen. Die lokale oder Datenerklärbarkeit zeigt, weshalb eine konkrete Eingabe zu einer bestimmten Ausgabe geführt hat. Ein typischer Anwendungsfall ist die automatisierte Kreditvergabe: Wer einen beantragten Kredit nicht erhält, hat ein Recht darauf, die Gründe für die Ablehnung zu erfahren. Die globale beziehungsweise Modellerklärbarkeit ist komplexer und zeigt, wie ein bestimmtes Modell als Ganzes funktioniert. Hierbei trainiert man zunächst das neuronale Netz (als Black-Box-Modell) wie üblich und erzeugt anschließend daraus ein Stellvertretermodell, das auch Surrogat oder White-Box-Modell genannt wird. Es bildet das Black-Box-Modell nach und trifft weitgehend die gleichen Vorhersagen. Aus ihm lässt sich eine Erklärung generieren. Dabei ist zu beachten, dass das White-Box-Modell einfacher aufgebaut ist, um Erklärbarkeit zu ermöglichen. Dadurch kann es zu Abweichungen zwischen den Ausgaben beider Modelle kommen. Beliebte White-Box-Modelle sind Entscheidungsbäume, regelbasierte Modelle oder lineare Modelle. <h3

class=„subheading“ id=„nav_heat_maps_als7“>Heat Maps als Helfer</h3> Gemeinsam mit der Firma IDS, einem Hersteller industriell genutzter Kameras, entstand in einem Projekt des KI-Fortschrittszentrums ein Demonstrator, der aufzeigt, wie sich die Erklärbarkeit bei der Bildverarbeitung verbessern lässt. Beispielsweise kann es „voreingenommene“ neuronale Netze geben: Sie ordnen Bilder aufgrund von Kriterien einer Objektklasse zu, die nur zufällig auf den Bildern vorhanden sind und kein Kriterium sein sollten (siehe Abbildung 4). Heat Maps oder Aufmerksamkeitskarten helfen, die Entscheidung eines neuronalen Netzes zu verstehen und Fehler in seinen Aktionen zu erkennen [9]. Sie gehören zu den lokalen Erklärbarkeitsverfahren. <figure class=„a-inline-image a-u-inline“> <figcaption class=„a-caption“>Eine Heat Map visualisiert, dass das neuronale Netz gelernt hat, einen Apfel nicht nur anhand äußerer Merkmale wie Farbe oder Form zu erkennen (links), sondern auch anhand etwaig angebrachter Markenaufkleber (rechts), die jedoch kein kennzeichnendes Merkmal für einen Apfel sind (Abb. 4).</figcaption></figure> Ein Demonstrator, in dem sich ein kleiner Delta-Roboter bewegt, verdeutlicht die Nützlichkeit von Heat Maps für die Erklärbarkeit. Erkennt das System eine Hand in seiner Nähe, sollte der Roboter aus Sicherheitsgründen stoppen. Befindet sich nur die Hand im Bild, funktioniert das Erkennen wie gewünscht. Legt man jedoch ein kleines Objekt wie eine Münze in das Bild, agiert der Roboter weiter, wenn die Hand ins Bild kommt. Das neuronale Netz richtet seine gesamte Aufmerksamkeit auf die Münze und übersieht damit die Hand (siehe Abbildung 5). Da die Heat Map bei der fehlerhaften Aktion die Münze als Entscheidungskriterium markiert, lässt sich die Fehlerquelle identifizieren und das Netz korrigieren. <figure class=„a-inline-image a-u-inline“> <figcaption class=„a-caption“>Im linken Bild erkennt das System die Hand (siehe Heat Map auf dem Monitor), während die Münze im rechten Bild von der Hand ablenkt. Dadurch arbeitet der Roboter fälschlicherweise weiter (Abb. 5).</figcaption></figure><h3 class=„subheading“ id=„nav_modellabsicherun8“>Modellabsicherung für eine zuverlässige KI</h3> Das sinnvolle Zusammenspiel der drei technischen Aspekte Sicherheit, Verlässlichkeit und Transparenz sind die Grundlage für eine zuverlässige KI. Je nach nach Anwendung, Branche, Rechtslage und Zielgruppe der KI-Anwendung kann ein Aspekt wichtiger sein als andere, aber fehlen sollte keiner. Die folgenden Fragen und Methoden können die Umsetzung leiten: <ul class=„rtelist rtelist–unordered“>
Sicherheit: Einsatz von Modellverifikation; Leitfrage: „Verhält sich das Modell bei beliebigen Eingaben wie gefordert?“
Verlässlichkeit: Einsatz der

Unsicherheitsquantifizierung; Leitfrage: „Weiß das Modell, was es nicht weiß?“
Transparenz: Einsatz von Methoden zur Erklärbarkeit; Leitfrage: „Wie kann ein Mensch die Entscheidung des Modells nachvollziehen?“ Marco Huberist Professor für Kognitive Produktionssysteme an der Universität Stuttgart und zugleich Leiter der Abteilungen Bild- und Signalverarbeitung sowie Cyber Cognitive Intelligence (CCI) am Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA. Seine Forschung konzentriert sich auf die Themen maschinelles Lernen, Sensordatenanalyse und Robotik im produktionstechnischen Umfeld. <h5 id=„nav_literatur9“>Literatur:</h5><ol class=„rtelist rtelist–ordered“>
Explaining and Harnessing Adversarial Examples [13]; Ian Goodfellow, Jonathon Shlens, Christian Szegedy
Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks [14]; Guy Katz, Clark Barrett, David L. Dill, Kyle Julian, Mykel J. Kochenderfer; Computer Aided Verification; CAV 2017.
Efficient Neural Network Robustness Certification with General Activation Functions; Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, Luca Daniel; NeurIPS 2018.
Beta-CROWN: Efficient Bound Propagation with Per-neuron Split Constraints for Neural Network Robustness Verification; Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, Zico Kolter; NeurIPS 2021.
Quantification of Uncertainties in Neural Networks; Xinyang Wu, Philipp Wagner, Marco F. Huber; 2023
Algorithmic Learning in a Random World; Vladimir Vovk, Alexander Gammerman, Glenn Shafer; 2. Auflage, 2022
„Why Should I Trust You?“: Explaining the Predictions of Any Classifier„ [15], Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin; Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2016
A Survey on the Explainability of Supervised Machine Learning; Nadia Burkart, Marco F. Huber; Journal of Artificial Intelligence Research (JAIR), 2021; DOI: 10.1613/jair.1.12228
Towards Measuring Bias in Image Classification; Nina Schaaf, Omar de Mitri, Hang Beom Kim, Robert-Alexander Windberger, Marco F. Huber; Proceedings of the 30th International Conference on Artificial Neural Networks (ICANN), September 2021. () URL dieses Artikels:<small>

```
https://www.heise.de/-8965811
```

</small> Links in diesem Artikel:<small>

```
<strong>[1]</strong> https://www.heise.de/news/Interaktives-Sprachmodell-nach-GPT-3-ChatGPT-steht-allen-Interessierten-offen-7364694.html
```

</small><small>

```
<strong>[2]</strong> https://www.heise.de/news/Google-KI-schlaegt-menschlichen-Profi-Spieler-im-Go-3085855.html
```

</small><small>

```
<strong>[3]</strong> https://www.heise.de/news/Deepmind-KI-faltet-Proteine-4243731.html
```

</small><small>

[4] https://www.heise.de/news/Amazon-KI-zur-Bewerbungspruefung-benachteiligte-Frauen-4189356.html

</small><small>

[5] https://www.heise.de/news/Microsofts-Chatbot-Tay-nach-rassistischen-Entgleisungen-abgeschaltet-3151646.html

</small><small>

[6] https://www.bosch.de/news-and-stories/ki-zukunftskompass/

</small><small>

[7] https://www.ki-fortschrittszentrum.de

</small><small>

[8] https://arxiv.org/pdf/1412.6572.pdf

</small><small>

[9] https://arxiv.org/pdf/1412.6572.pdf

</small><small>

[10] https://link.springer.com/chapter/10.1007/978-3-319-63387-9_5

</small><small>

[11] https://www.m3-konferenz.de/

</small><small>

[12] https://www.m3-konferenz.de/programm.php

</small><small>

[13] ICLR 2015 https://arxiv.org/pdf/1412.6572.pdf

</small><small>

[14] https://doi.org/10.1007/978-3-319-63387-9_5

</small><small>

[15] https://doi.org/10.1145/2939672.2939778

</small><small>

[16] mailto:rme@ix.de

</small> Copyright © 2023 Heise Medien </html>