KI-Zuverlässigkeit in der Produktion

Prüfwerkzeuge für systematisches Testen von KI-Modellen

Sujan Sai Gannamaneni, Elena Haedecke, Maximilian Pintz, Maximilian Poretschkin und Michael Mock*

In den letzten Jahren bieten Technologien der künstlichen Intelligenz (KI) neue Möglichkeiten zur Steigerung von Qualität und Effizienz in der Produktion. Damit das hiermit verbundene Potenzial vollständig ausgeschöpft werden kann, ist der Nachweis der Zuverlässigkeit von KI unabdingbar, insbesondere für den Einsatz der KI in automatisierten Produktionssystemen. Für einen belastbaren Zuverlässigkeitsnachweis sind Tests erforderlich, bei denen die KI-Modelle auf verschiedene Fehlerarten geprüft werden. Dieser Beitrag stellt anhand von drei Anwendungsfällen aus der industriellen Produktion vor, wie mit spezialisierten Prüfwerkzeugen Tests von KI-Modellen durchgeführt werden können, die systematische Schwächen in den KI-Modellen aufdecken. Damit die KI-Tests effizient durchgeführt und Testresultate automatisiert dokumentiert werden, sind die Prüfwerkzeuge in eine skalierbare Prüfplattform integriert.

Einleitung

Technologien, die die Automatisierung in der Fertigung und Produktion ermöglichen, spielen eine entscheidende Rolle bei der Steigerung von Qualität und Effizienz. In den letzten Jahren versprechen Technologien im Zusammenhang mit künstlicher Intelligenz (KI), insbesondere Ansätze des maschinellen Lernens (ML), neue Wege der Automatisierung zu erschließen [1]. Doch trotz des breiten Spektrums an Anwendungsfällen verzögert sich die Einführung dieser Technologie in der Industrie vor allem aufgrund von Bedenken hinsichtlich der Zuverlässigkeit. Während frühere Ansätze wie modellbasierte Methoden bei der Virtualisierung von Sensoren [2] und Ansätze zum Abgleich mit gegebenen Vorlagen (Template-Matching) bei der visuellen Qualitätsinspektion [3] zuverlässige Automatisierungsmethoden mit Einschränkungen im Anwendungsbereich boten, bietet eine neue Generation von ML-Ansätzen eine größere Vielfalt im Anwendungsbereich, jedoch mit Einschränkungen in der Zuverlässigkeit. Das heißt, dass auch wenn die Leistung aktueller KI-Modelle das menschliche Niveau erreichen kann, die Konsistenz dieser Leistung im Vergleich zu der des Menschen dennoch geringer ist. Um das Potenzial der neuen Automatisierungsgeneration zu nutzen und Wettbewerbsvorteile in der Industrie zu erzielen, muss daher der Schwerpunkt auf die Entwicklung und den Nachweis der Zuverlässigkeit von KI-Modellen gelegt werden. Zudem unterliegen Unternehmen, insbesondere solche, welche KI-Systeme für safety-kritische Anwendungen einsetzen, regulatorischen Anforderungen, wie etwa der Europäischen KI-Verordnung, der Maschinenverordnung (2023/1230/EU) sowie den Anforderungen aus einschlägigen Safety-Standards. Zur Umsetzung dieser Vorgaben müssen sie einerseits interne Prozesse einrichten [4], andererseits aber auch technische Qualitätssicherungsmaßnahmen und Tests durchführen. Gerade die EU-KI-Verordnung stellt bereits spezielle Anforderungen an die Zuverlässigkeit von KI-Modellen: für KI-Systeme im sogenannten Hochrisiko-Bereich wird explizit gefordert, dass diese getestet werden müssen, um sicherzustellen, dass sie „stets im Einklang mit ihrer Zweckbestimmung funktionieren“ (Artikel 9, Absatz 6), und dass das Testen „zu jedem geeigneten Zeitpunkt während des gesamten Entwicklungsprozesses und in jedem Fall

* **Korrespondenzautor**
Sujan Sai Gannamaneni, M. Sc.; Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS; Schloss Birlinghoven 1, 53757 Sankt Augustin 17; Tel.: +49 (0) 2242 14-2292, E-Mail: Sujan.Sai.Gannamaneni@iais.fraunhofer.de; Lamarr-Institut

Weitere Autor:innen
Elena Haedecke, M. Sc., B. Eng.; Fraunhofer IAIS & Universität Bonn
Maximilian Pintz, M. Sc.; Fraunhofer IAIS & Universität Bonn
Dr. Maximilian Poretschkin; Fraunhofer IAIS & Lamarr-Institut
Priv.-Doz. Dr. Michael Mock; Fraunhofer IAIS

Hinweis
Bei diesem Beitrag handelt es sich um einen von den Advisory-Board-Mitgliedern des ZWF-Sonderheftes wissenschaftlich begutachteten Fachaufsatz (Peer-Review).

vor ihrem Inverkehrbringen oder ihrer Inbetriebnahme“ (Artikel 9, Absatz 8) erfolgt. Auch wenn nicht alle in der Produktion eingesetzten KI-Modelle im Hochrisikobereich liegen, ist davon auszugehen, dass entsprechende Anforderungen auch allgemein bei der Nutzung und Zulassung von KI-Systemen in der Produktion sinnvoll sind.

Ähnlich wie bei klassischer Software, wo das Testen zwar keine 100-prozentige Garantie für das Produktverhalten bietet, aber zur Vertrauensbildung in die Zuverlässigkeit des Systems beiträgt, kann das Testen von KI-Modellen eine entsprechende vertrauensbildende Wirkung haben [5, 6]. Im Gegensatz zu klassischer Software ist das Testergebnis bei KI-Modellen abhängig von den genutzten Testdaten. Zudem ist es normal, dass das KI-Modell nicht auf allen Testdaten richtige Ergebnisse liefert. Daher werden typischerweise Durchschnittswerte der Leistung von KI-Modellen auf den genutzten Testdaten zur Bewertung der Zuverlässigkeit herangezogen. Diese Durchschnittswerte haben jedoch gerade bei Anwendungen in der Produktion nur eine begrenzte Aussagekraft. Wird zum Beispiel ein elektro-mechanischer Sensor zur Messung eines Drehmoments durch einen KI-basierten virtuellen Sensor ersetzt, der das Drehmoment auf Basis von Strom und Spannung schätzt, so sollte dieser virtuelle Sensor für alle Drehmomente gleichmäßig genaue Werte liefern. Ein klassischer KI-Test, der nur Durchschnittswerte über alle Testdaten hinweg ermittelt, wäre nicht in der Lage, zu erkennen, ob der virtuelle Sensor in bestimmten Eingabebereichen (bei bestimmten Strom- oder Spannungswerten) grundsätzlich zu hohe oder grundsätzlich zu niedrige Werte für das Drehmoment angibt. Solche Fehler, die als systematische Schwächen bekannt sind, sind einer der Gründe für die mangelnde Zuverlässigkeit von KI-Modellen [7, 8]. Systematische Schwächen in KI-Modellen zu finden, ist aufgrund der Komplexität der Modelle, ihrer inhärenten Abhängigkeit von Daten, die zur Entwicklung der Modelle benutzt wurden, sowie der potenziellen Vielfalt der Betriebsbedingungen, unter denen die Modelle eingesetzt werden, eine schwierige Aufgabe, die allein durch zufälliges Ausprobieren oder menschliche Inspektion nicht bewerkstelligt werden kann. Daher können Prüfwerkzeuge, die speziell auf die Erkennung und Beseitigung dieser Fehler abzielen, sowohl in der Entwicklungsphase von KI-Modellen als auch in der Prüfung von KI-Modellen in der Endabnahme helfen, indem sie entweder systematische Fehler im KI-Modell aufzeigen oder eben nachweisen, dass derartige Fehler auch unter Einsatz von Prüfwerkzeugen nicht gefunden werden konnten.

In diesem Beitrag stellen wir zunächst ein Prüfwerkzeug zur automatisierten Erkennung von systematischen Schwachstellen vor. Der Arbeitsablauf des Prüfwerkzeugs wird anhand von drei industriellen Anwendungsfällen beschrieben. Anschließend gehen wir auf die Einschränkungen einer rein automatisierten Prüfung ein und zeigen, wie diese durch ein weiteres Prüfwerkzeug unter Verwendung von Techniken der visuellen Analyse ergänzt werden kann. Abschließend stellen wir auch den Rahmen einer Prüfplattform vor, die mehrere solcher Prüfwerkzeuge umfasst, um KI-Modelle effizient testen zu können.

Systematische Schwachstellensuche

Bis vor kurzem bestand die Bewertungsstrategie für KI-Modelle in der Berechnung gemittelter Kennzahlen (z. B. Präzision und Sensitivität) auf ausreichend großen Testdatensätzen, die es den Entwicklern ermöglichen, statistische Aussagen über die Modellleistung zu treffen. Bei solchen Bewertungen wird davon ausgegangen, dass die KI-Modelle in erster Linie stochastische Fehler enthalten. Die Abschwächung solcher Fehler erfordert entweder die Erhebung zusätzlicher Trainingsdaten oder eine Erhöhung der Modellkomplexität. Da die Fehler zufälliger Natur sind, können keine spezifischen Anforderungen an die zusätzlichen Daten gestellt werden, was diesen Prozess zeit- und kostenintensiv macht. Die Annahme, dass nur stochastische Fehler vorhanden sind, ist jedoch nicht immer zutreffend, wie in einigen neueren Arbeiten gezeigt wurde [7, 8], da KI-Modelle auch für systematische Fehler anfällig sind. Im Gegensatz zu stochastischen Fehlern bergen systematische Fehler das Risiko erheblicher Schäden, da sie beständig sind und sich wiederholen. Folglich ist die Umsetzung von Maßnahmen zur Erkennung und Abschwächung dieser systematischen Fehler von entscheidender Bedeutung, um eine zuverlässige Systemleistung zu gewährleisten. Darüber hinaus ist der Prozess der Verbesserung von KI-Modellen mit zusätzlichen Daten im Vergleich zu früheren Ansätzen erheblich effizienter, da eine gezielte Datensammlung durchgeführt werden kann.

Bestehende Arbeiten [7] konzentrieren sich meist auf Anwendungsfälle der Computer Vision (z. B. Bildklassifizierung), doch kann das Vorhandensein systematischer Fehler in KI-Modellen für andere Aufgaben nicht ausgeschlossen werden. Darüber hinaus stellen diese bestehenden Ansätze auch keine umfassende Verbindung zwischen den identifizierten Schwachstellen und der Zuverlässigkeit des KI-Modells und seinem beabsichtigten Anwendungsbereich her. Um diese Probleme anzugehen, schlagen wir einen Arbeitsablauf vor, der automatisiert Schwachstellen von KI-Modellen in Bezug auf den Anwendungsbereich identifiziert. Um den Anwendungsbereich und die Betriebsbedingungen von KI-Modellen allgemein auf der Ebene von menschlich verständlichen Konzepten definieren zu können, haben Sicherheitsexperten aus verschiedenen Bereichen wie der Automobilindustrie [5] und dem Bahnwesen [9] die Verwendung von Operational Design Domains (ODDs) vorgeschlagen. Unter der ODD versteht man eine möglichst genaue Spezifikation der Bedingungen, unter denen das KI-Modell korrekt funktionieren soll. Insbesondere müssen dabei auch die Grenzen der Anwendbarkeit des KI-Modells festgelegt werden (z. B. eine Mindestanforderung an die Beleuchtung und Helligkeit der Umgebung im Falle einer kamerabasierten Qualitätsüberprüfung in der Fertigung). Die ODDs bieten eine Spezifikation, die die Bedingungen und Einschränkungen festlegt, innerhalb derer das zuverlässige Funktionieren eines KI-Modells erforderlich ist. Die dabei zu berücksichtigenden Dimensionen, in denen sich die Betriebsbedingungen ändern können, sind anwendungsabhängig und müssen von Anwendungsexperten festgelegt werden. Unter Verwendung dieser ODDs als Eingabe zu-

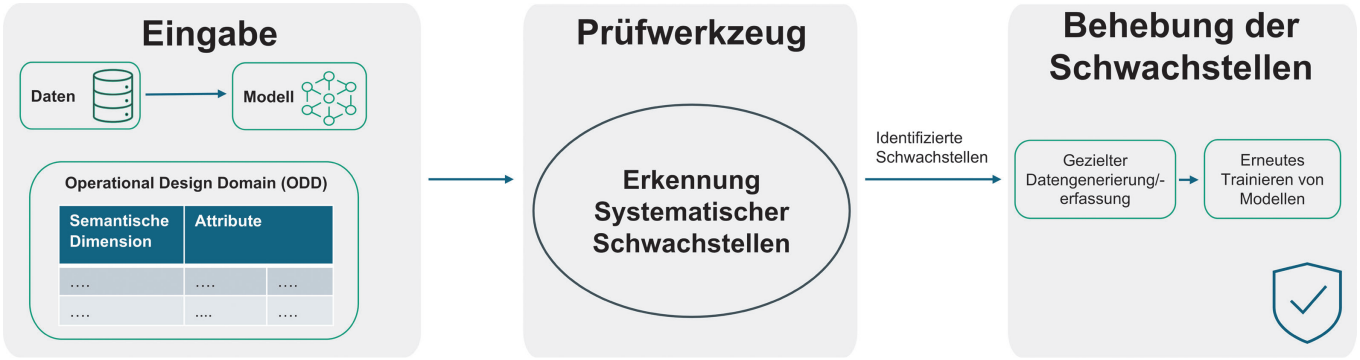

Bild 1. Arbeitsablauf der systematischen Schwachstellensuche

sammen mit den Daten und dem KI-Modell setzen wir Algorithmen ein [10], die den Eingaberaum effizient in ODD-Attribute unterteilen und Bereiche mit geringer Leistung identifizieren. Der vorgeschlagene Arbeitsablauf ist in Bild 1 dargestellt. In Bereichen, in denen eine vorherige Zuordnung von Daten zu ODD-Attributen nicht verfügbar ist (z. B. Bilder, Text), haben wir Generative-KI-basierte Werkzeuge (z. B. basierend auf CLIP) entwickelt, die unter bestimmten Voraussetzungen die Zuordnung automatisch erstellen können [8].

Zur Veranschaulichung des Arbeitsablaufs stellen wir die drei Schritte anhand von Anwendungsfällen von KI-Modellen in der industriellen Produktion vor. Um die Betriebsbedingungen zu definieren, entwerfen wir zunächst beispielhafte ODDs für jeden Anwendungsfall.

- Soft/Virtuelle Sensoren sind digitale Sensoren, die physische Sensoren ersetzen sollen, um Kosten zu senken, das Design zu vereinfachen und das Gewicht der Komponenten zu reduzieren. KI-Modelle bieten einen neuen datengesteuerten Ansatz zur Entwicklung von Softsensoren als Alternative zu modellbasierten Softsensoren. Typische Produkte, bei denen diese Sensoren zum Einsatz kommen könnten, sind Aktuatoren, Robotik, chemische Prozesse, Gesundheitswesen, Umweltüberwachung. Für einen pneumatischen Aktuator, bei dem wir Softsensoren einsetzen, um die Position des Aktuators abzuschätzen, ist in Bild 2 eine beispielhafte ODD dargestellt.
- Bei der automatisierten visuellen Qualitätsprüfung werden KI-Modelle, insbesondere Computer-Vision-Modelle, eingesetzt, um die Qualität von Endprodukten oder Bauteilen in verschiedenen Stufen des Produktionsprozesses zu bewerten. In Bild 2 ist ein Beispiel für eine ODD zur Prüfung von Leiterplatten auf Lötfehler dargestellt.
- Large Language Models (LLMs) bieten neue Möglichkeiten zur Unterstützung der Mitarbeiter im Produktionsprozess bei der Montage, Wartung, Prüfung und manuellen Inspektion. Anstatt auf interne Handbücher zurückzugreifen, können die Arbeiter LLMs, die anhand interner Daten trainiert wurden, bitten, produktspezifische Wartungsanweisungen zu geben. Bild 2 enthält ein Beispiel für eine ODD für einen solchen Anwendungsfall, in dem Testingenieure die LLMs nach Testverfahren für verschiedene Produkte in einer Sensorherstellerfabrik fragen können.

Sobald die Betriebsbedingungen definiert sind, kann das Prüfwerkzeug zur Erkennung systematischer Schwachstellen eingesetzt werden. Es verknüpft die Ausgaben des KI-Modells, die Testdaten und die definierte ODD, um automatisiert systematische Schwachstellen des KI-Modells zu erkennen. Das Prüfwerkzeug unterteilt dazu den Datenraum der Testdaten in Bereiche, die mit den ODD-Attributen übereinstimmen, und identifiziert Kombinationen von Attributen, bei denen das KI-Modell eine geringere Leistung aufweist. Das Ergebnis, zum Beispiel im Anwendungsfall 1, wäre die Kombination von Attributen wie Zylindergröße von 0 bis 5 cm und Arbeitstemperatur von 0 bis 50 °C, die als Schwäche identifiziert wird. Anhand dieser Informationen kann das KI-Modell verbessert werden, indem mehr Datenpunkte gesammelt werden, die zu diesen Attributen gehören, und das KI-Modell neu trainiert wird. Diese gezielte Datengenerierung und das erneute Training können dazu beitragen, die Leistung des KI-Modells zu verbessern, da eine mangelnde Datenabdeckung ein möglicher Grund für das Vorhandensein der Schwäche sein könnte.

Soft/Virtual Sensors

Semantische Dimension	Attribute	
Zylinder Größe	0 – 5 cm	5 – 10 cm
Belastung	10 – 100 kg	100 – 200 kg
Arbeitstemperatur	0 – 50 °C	50 – 100 °C

Automatisierte Visuelle Qualitätsprüfung

Semantische Dimension	Attribute	
Schadensgröße	< 5 cm	> 5 cm
Schadenstyp	Lötbrücken	Stiftlöcher
Kamera Typ	FLIR	Basler

Large Language Models (LLMs) als Produktionsunterstützung

Semantische Dimension	Attribute	
Sensor-Typ	Magnetische Sensoren	Drucksensoren
Sprache	Deutsch	Englisch
Prompting-Stil	Fragenstruktur: frei	Fragenstruktur: formal

Bild 2. Beispielhafte ODDs für drei industrielle Anwendungsfälle

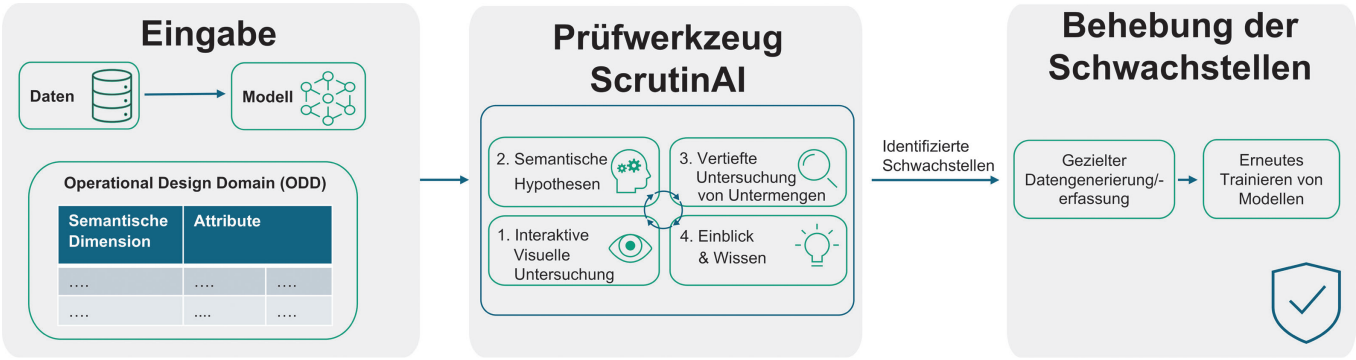

Bild 3. Einsatz des VA-Prüfwerkzeuges „ScrutinAI“ (i. A. an [14])

Explorative visuelle Analyse mit ScrutinAI

Die im vorigen Abschnitt vorgestellte systematische Schwachstellensuche kann als Test von KI-Modellen automatisiert in den Entwicklungs- oder Prüfprozess von KI-Modellen eingebaut werden. Damit stellt sie eine sehr effiziente Lösung dar, um die Zuverlässigkeit von KI-Modellen zu sichern. Jedoch gibt es weitere mögliche Fehlerquellen für KI-Modelle, die nicht dadurch automatisiert erkannt werden können. Darunter fällt, dass sie keine Randfälle (Corner Cases) oder Entscheidungsgrenzen finden kann. Zudem ist sie abhängig von der korrekten und umfassenden Beschreibung der ODD. Sie kann zum Beispiel nur solche Mängel in der Datenabdeckung aufdecken, die sich durch die Dimensionen der ODD auch beschreiben lassen.

Daher stellt eine explorative Analyse mit Human-in-the-Loop (im Test) einen weiteren wichtigen Baustein beim Nachweis der Zuverlässigkeit dar. In der explorativen Analyse können zum Beispiel auch gefundene Schwachstellen weiter vertieft analysiert werden. Für die explorative Analyse bedient man sich der Prinzipien und Methoden der Forschungsrichtung Visual Analytics (VA), die darauf abzielt, Menschen mithilfe von visuellen Schnittstellen bei der analytischen Schlussfolgerung zu unterstützen [11, 12]. VA-Methoden bringen eine große Menge an Daten, Methoden, Merkmalen und Leistungsmetriken in eine für den Menschen erfassbare Form. Da der Mensch besonders gut darin ist, visuell Muster zu erkennen, macht man sich diesen Vorteil bei der menschenzentrierten VA für ML [13] zunutze, um das semantische Verständnis sowie das Domänenwissen von Expert:innen in den Analyseprozess einzubinden. Da komplexe KI-Modelle und neuronale Netze auf einer großen Menge an Daten basieren, wird der Mensch durch einen effizienten Ablauf dabei unterstützt, die Informationen systematisch zu durchsuchen und entsprechende Einsichten daraus abzuleiten (vgl. Bild 3 (aus [14]). Dabei ist es besonders wichtig, sowohl einen Überblick über die Daten zu geben als auch bei Bedarf die notwendige Detailtiefe für tiefergehende Analysen bereitzustellen, wie auch von *Keim et al.* [11] als „Mantra“ für VA formuliert: „Analyze first, show the important, zoom, filter and analyze further, details on demand”. Hierzu bedient man sich verschiedener Filterungsoptionen sowie der Methoden des „linked brushing“, indem verschiedene Datenvisualisierungen miteinander verknüpft werden und so die Filterung, bspw. durch das interaktive Auswählen von Datenpunkten in einer Grafik, eine Aktualisierung der verschiedenen Ansichten auch in den anderen Visualisierungen anstößt. So ist es zum Beispiel möglich, bei der Prüfung des eingangs beschriebenen virtuellen Sensors für das Drehmoment sich die Richtigkeit der vorhergesagten Drehmomente für frei wählbare Spannungsbereiche anzeigen zu lassen. Ein Beispiel für ein solches VA-Prüfwerkzeug ist „ScrutinAI“, welches in [15] für einen Anwendungsfall aus dem Bereich des autonomen Fahrens gezeigt und in [16] auf einen medizinischen Anwendungsfall angewandt wird. Für beide Anwendungsfälle konnten mithilfe des VA-Prüfwerkzeugs wichtige Randfälle (Corner Cases) aufgedeckt werden, wie z. B. die schlechte Modellperformanz bei der Erkennung besonders großer Fußgänger im Fall des autonomen Fahrens oder dass im Eisenbahnbereich bestimmte Formen von Signalen fälschlicherweise als Personen erkannt wurden. Ähnliche Evaluationen sind auch für die zuvor vorgestellten Anwendungsfälle umsetzbar, wie z. B. die Qualitätsprüfung von Bauteilen basierend auf Bilddaten. Hier kann es beispielsweise sein, dass bei der Definition der ODD gewisse Randfälle oder seltene Situationen nicht abgebildet werden konnten. Durch die visuelle interaktive Analyse kann dann insbesondere das spezifische menschliche Domänenwissen im jeweiligen Anwendungsfall genutzt werden, um Muster aufzudecken, die bei einer automatisierten Analyse verborgen geblieben wären.

Die verschiedenen interaktiven Elemente von ScrutinAI sind an das VA-Mantra angelehnt und können entlang der Schritte des Ablaufs in Bild 3 eingesetzt werden, um das KI-Modell zu untersuchen. Beispielsweise können textuelle Abfragen, aber auch die interaktive Auswahl einzelner Datenpunkte die Auswahl der Daten für eine explorative Analyse (Schritt 1) einschränken. Basierend auf dem so erlangten Verständnis der semantischen Zusammenhänge von Modellperformanz und Daten bildet der/die Analyst:in Hypothesen über die Ursachen von Mustern und Auffälligkeiten in dem semantischen Datenraum (Schritt 2), die in einer nachfolgenden tiefergehenden „Drill-Down“-Analyse (Schritt 3) weiter untersucht werden. Das aus der Analyse generierte kontextabhängige Wissen (Schritt 4) dient als Basis, um einen neuen Analyse-

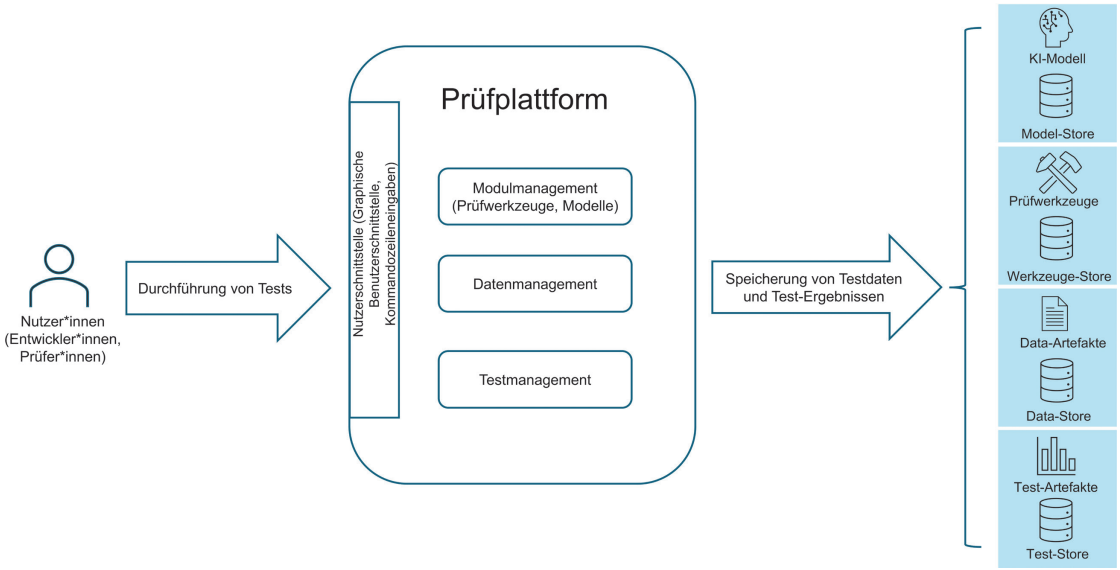

Bild 4. Arbeitsablauf der PARMA-Prüfplattform [17] (i. A. an [14])

zyklus zu starten und/oder zur Weitergabe von Feedback an relevante Stakeholder, um beispielsweise das KI-Modell im Falle von aufgedeckten Schwachstellen zu verbessern.

Da insbesondere im Fall von Bilddaten nicht immer umfassende Metadaten zur Verfügung stehen, bietet es sich an, Techniken wie das sogenannte „Query-by-Example“ einzusetzen (siehe hierzu auch [15]). Hierbei wird basierend auf einem durch den/die Analyst:in gewählten interessanten Bildausschnitt eines Eingabebildes eine Ähnlichkeitssuche durchgeführt, sodass semantisch ähnliche Eingabebilder des Datensatzes identifiziert und als Untermenge zur weiteren Analyse zur Verfügung stehen, ohne dass hierfür eine Annotation der in den Bildausschnitten enthaltenen semantischen Dimensionen notwendig würde.

Effizientes, wiederverwendbares und nachweisfähiges Testen durch Prüfplattformen

Aus den vorherigen zwei Abschnitten wird deutlich, dass für einen aussagekräftigen Test der Zuverlässigkeit verschiedene Prüfwerkzeuge benötigt werden, die unterschiedliche Perspektiven einer Prüfung umfassen können. Als Hersteller möchte man den erforderlichen Aufwand zur technischen Umsetzung solcher Tests effizient gestalten, indem beispielsweise ein hoher Grad an Wiederverwendbarkeit der Prüfwerkzeuge und anderer technischer Komponenten des Tests in verschiedenen Anwendungsfällen und Szenarien erreicht wird. Ein Ansatz, um dieses Ziel zu erreichen, ist der Einsatz einer Prüfplattform, welche Tests in Form von konfigurierbaren sogenannten „Prüfworkflows“ verwaltet. Ein Prüfworkflow beschreibt den Ablauf eines Tests in Form einzelner Berechnungsschritte (im Folgenden auch Module genannt), wie z. B. das Laden von Datensätzen, das Erzeugen von Modellvorhersagen oder das Berechnen und Abspeichern von Testmetriken und Testresultaten (vgl. auch Bild 1 für ein Beispiel eines Prüfworkflows). Jedes Modul erhält Eingaben, verarbeitet diese und erzeugt Ausgaben, die im Kontext eines Prüfworkflows an weitere Module zur Weiterverarbeitung gegeben werden. Die Prüfplattform macht dabei gewisse Vorgaben zu den Schnittstellen dieser Module und insbesondere zu den Formaten der Ein- und Ausgaben (für weitere Details zu einer möglichen Architektur und technischen Umsetzung siehe [17]). Diese Vorgaben ermöglichen es, die Module als Baustein in verschiedenen Tests und Anwendungsfällen flexibel wiederverwenden zu können. Beispielsweise kann so das gleiche Modul zum Laden eines Datensatzes sowohl in einem Test zu systematischen Schwachstellen als auch zur explorativen Analyse eingesetzt werden. Um die Automatisierung von Prüfabläufen zu unterstützen, stellt eine Prüfplattform Nutzerschnittstellen bereit (wie beispielsweise Kommandozeilenbefehle oder eine graphische Nutzerschnittstelle), über welche Prüfworkflows definiert und ausgeführt werden können (Bild 4). Mittels Versionsverwaltung von Prüfworkflows und der umfangreichen Protokollierung der Ausführung von Tests kann eine Prüfplattform zudem zu einer hohen Nachweisfähigkeit und Reproduzierbarkeit von Testresultaten beitragen.

Zusammenfassung und Ausblick

Um KI sicher und unter Einhaltung von gesetzlichen Vorgaben in der Produktion und Automatisierung einzusetzen, muss insbesondere die Zuverlässigkeit der KI-Modelle berücksichtigt und sichergestellt werden. Dazu sind KI-spezifische Tests durchzuführen. In dieser Arbeit stellen wir Test- und Prüfwerkzeuge vor, um systematische Schwachstellen in KI-Modellen zu identifizieren, die in der industriellen Produktion eingesetzt werden. Dazu zählen automatisierte Testverfahren sowie ein visueller Analyseansatz, um weitere und tiefergehende Einblicke in das Verhalten von KI-Modellen zu erhalten. Die vorgestellten Prüfwerkzeuge sind bereits in mehreren großen Anwendungsprojekten zum Einsatz gekommen und sind auch für sich einzeln nutzbar. Eine Demoversion des VA-Prüfwerkzeugs ist online einsehbar [18]. Die Integration

der Prüfwerkzeuge in eine Prüfplattform sowie die Entwicklung derselben sind weiter Gegenstand unserer aktuellen Forschungsarbeit. Die Prüfplattform unterstützt die effiziente Anwendung der Prüfwerkzeuge. Derzeit werden im Rahmen der Umsetzung der EU-KI-Verordnung verschiedene neue Normen und Rahmenwerke entwickelt, um den Nachweis der Zuverlässigkeit zu standardisieren (vgl. [19, 20]). Die Entwicklung der in diesem Artikel besprochenen Prüfwerkzeuge ist das Ergebnis von Anforderungen aus solchen Standards und Rahmenwerken, insbesondere dem in [6] beschriebenen KI-Prüfkatalog.

Literatur

1. Frey C.; Goßmann A.; Hasterok C. et al.: KI-Engineering in der Produktion. Whitepaper der Fraunhofer-Institute IOSB und IAIS, 2023 DOI:10.24406/PUBLICA-1685
2. Kadlec, P.; Gabrys, B.; Strandt, S.: Data-driven Soft Sensors in the Process Industry. Computers & Chemical Engineering 33 (2009) 4, S. 795–814 DOI:10.1016/j.compchemeng.2008.12.012
3. Chin, R. T.: Automated Visual Inspection: 1981 to 1987. Computer Vision, Graphics, and Image Processing 41 (1988) 3, S. 346–381 DOI:10.1016/0734-189X(88)90108-9
4. Schmitz, A.; Akila, M.; Hecker D. et al: The Why and How of Trustworthy AI: An Approach for Systematic Quality Assurance when Working with ML Components. at – Automatisierungstechnik 70 (2022) 9, S. 793–804 DOI:10.1515/auto-2022-0012
5. Blank, F.; Hüger, F.; Mock, M.; Stauner, T.: Methodik zur Absicherung von KI im Fahrzeug. Automobiltechnische Zeitschrift ATZ 124 (2022) 7–8, S. 58–63 DOI:10.1007/s35148-022-0879-3
6. Poretschkin, M.; Schmitz, A.; Akila, M. et al.: Guideline for Trustworthy Artificial Intelligence — AI Assessment Catalog. Frontiers in Big Data 7 (2024) DOI:10.48550/arXiv.2307.03681
7. Eyuboglu, S.; Varma, M.; Saab, K. et al: Domino: Discovering Systematic Errors with Cross-Modal Embeddings. (http://arxiv.org/abs/2203.14960 [Abruf am 23.01.2025] DOI:10.48550/arXiv.2203.14960
8. Gannamaneni, S. S.; Sadaghiani, A.; Rao, R. P. et al.: Investigating CLIP Performance for Meta-data Generation in AD Datasets. Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE 2023, S. 3840–3850 DOI:10.1109/CVPRW59228.2023.00398
9. Weiss, G.; Zeller, M.; Schoenhaar, H. et al.: Approach for Argumenting Safety on Basis of an Operational Design Domain. In: Proceedings of the 3rd International Conference on AI Engineering – Software Engineering for AI, IEEE/ACM 2024, S. 184–193 DOI:10.1145/3644815.3644944
10. Sagadeeva, S.; Boehm, M.: SliceLine: Fast, Linear-Algebra-based Slice Finding for ML Model Debugging. In: Proceedings of the 2021 International Conference on Management of Data. ACM 2021, S. 2290–2299 DOI:10.1145/3448016.3457323
11. Keim, D.; Andrienko, G.; Fekete, J.-D. et al.: Visual Analytics: Definition, Process, and Challenges. Information Visualization. In: Kerren, A.; Stasko, J. T.; Fekete, J.-D.; North, C. (Hrsg.): Lecture Notes in Computer Science. Springer, Heidelberg 2008, S. 154–175 DOI:10.1007/978-3-540-70956-5_7
12. Cook, K. A.; Thomas, J. J.: Illuminating the Path: The Research and Development Agenda for Visual Analytics. IEEE Press, 2005
13. Andrienko, N.; Andrienko, G.; Adilova, L. et al.: Visual Analytics for Human-Centered Machine Learning. Computer Graphics and Applications 42 (2022) 1, S. 123–133 DOI:10.1109/MCG.2021.3130314
14. Haedecke, E. G.; Mock, M.; Pintz, M. A.; Poretschkin, M.: KI-Anwendungen systematisch prüfen und absichern. 2023 DOI:10.24406/PUBLICA-1635
15. Haedecke, E.; Mock, M.; Akila, M.: ScrutinAI: A Visual Analytics Tool Supporting Semantic Assessments of Object Detection Models. Computers & Graphics 114 (2023) 1, S. 265–275 DOI:10.1016/j.cag.2023.06.010
16. Görge, R.; Haedecke, E.; Mock, M.: Using ScrutinAI for Visual Inspection of DNN Performance in a Medical Use Case. AI Ethics 4 (2024) 1, S. 151–156 DOI:10.1007/s43681-023-00399-x
17. Pintz, M.; Becker, D.; Mock, M.: PARMA: a Platform Architecture to Enable Automated, Reproducible, and Multi-party Assessments of AI Trustworthiness.In: Proceedings of the 2nd International Workshop on Responsible AI Engineering, ACM 2024, S. 20–27 DOI:10.1145/3643691.3648585
18. Forum Zertifizierte KI: Visuelle Fehleranalyse (2025). Online unter https://www.zertifizierte-ki.de/visuelle-fehleranalyse [Abruf am 23.01.2025]
19. Mission KI: Online unter https://mission-ki.de/de [Abruf am 23.01.2025]
20. DeployAI: Deploying AI for Business (2025). Online unter https://www.deployaiproject.eu [Abruf 23.01.2025]

Die Autor:innen dieses Beitrags

Sujan Sai Gannamaneni, geb. 1993, studierte M. Sc. in Mechatronik an der Fachhochschule Aachen. Er ist wissenschaftlicher Mitarbeiter am Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS und wird durch das Lamarr-Institut gefördert. Sein Forschungsschwerpunkt liegt in der Entwicklung von Prüfwerkzeugen, vor allem von Werkzeugen für systematische Schwachstellen, zur Entwicklung vertrauenswürdiger KI-Modelle, im Rahmen von öffentlich geförderten Projekten.

Elena Haedecke, geb. 1990, studierte Informatik (B. Sc. und M. Sc.) sowie Elektrotechnik (B. Eng.) an der Hochschule Bonn-Rhein-Sieg in Sankt Augustin. Sie ist wissenschaftliche Mitarbeiterin und Doktorandin an der Universität Bonn sowie am Fraunhofer-Institut für intelligente Analyse- und Informationssysteme IAIS, wo sie primär die Vertrauenswürdigkeit von KI-Systemen innerhalb des öffentlich geförderten, interdisziplinären Projekts ZERTIFIZIERTE KI erforscht. Im Kontext von KI Prüfungen fokussiert sich ihre Forschung einerseits auf die Entwicklung von Prüfwerkzeugen, die Schwachstellen von KI-Systemen aufdecken und abschwächen können, und andererseits auf Methoden, die KI-Systeme verständlicher machen.

Maximilian Pintz, geb. 1997, studierte Informatik (M. Sc.) an der Universität Bonn. Er ist wissenschaftlicher Mitarbeiter an der Universität Bonn sowie am Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, wo er Methoden zur Vertrauenswürdigkeit von KI-Systemen innerhalb des Projekts ZERTIFIZIERTE KI erforscht. Seine Forschung fokussiert sich dabei insbesondere auf die Entwicklung von Prüfwerkzeugen zur Sicherstellung der Verlässlichkeit von KI-Systemen, sowie auf die Entwicklung einer Prüfplattform zur technischen Operationalisierung von KI-Prüfungen.

Dr. Maximilian Poretschkin, geb. 1986, studierte Physik und Mathematik in Bonn und Amsterdam und wurde 2014 in mathematischer Physik promoviert. Er leitet die Abteilung „AI Assurance and Assessments“ am Fraunhofer-Institut für Analyse und Informationssysteme IAIS. Seine Forschungsinteressen umfassen die Entwicklung von KI-Prüfmethodiken, Prüfwerkzeugen und die Operationalisierung von regulativen Anforderungen. Er hat langjährige Erfahrung in der Leitung von Forschungsprojekten und berät Unternehmen und Behörden weltweit zu Fragestellungen der KI-Vertrauenswürdigkeit. Besonders erwähnenswert ist hier das Projekt ZERTIFIZIERTE KI, welches Testprinzipien, Prüfwerkzeuge und Standards für KI-Systeme entwickelt und einen der ersten Prüfkataloge für KI-Systeme veröffentlicht hat.

Priv.-Doz. Dr. Michael Mock, geb. 1963, ist Projektleiter von Forschungs- und Industrieprojekten am Fraunhofer IAIS in Sankt Augustin und Privatdozent für Informatik an der Universität Bonn. Er hat mit über hundert Publikationen zu den Forschungsgebieten Zuverlässigkeit von Software-Systemen und Vertrauenswürdigkeit von KI beigetragen. Er hat als Berater in zahl-

reichen Projekten den Wissenstransfer von der Forschung zur Anwendung in der Industrie unterstützt. Zudem hat er das Fraunhofer Schulungsprogramm im Bereich Data-Science mit entwickelt.

Abstract

AI Reliability in Production – Assessment Tools for Systematic Testing of AI Models. In recent years, artificial intelligence (AI) technologies have offered new opportunities to increase quality and efficiency in production. In order to fully exploit the associated potential, it is essential to assure the reliability of AI, especially for the use of AI in automated production systems. Robust assurance of reliability requires tests in which the AI models are evaluated against different failure modes. This article illustrates in three use cases from industrial production how specialized assessment tools can be used to test AI models to detect systematic weaknesses in the AI models. The assessment tools are integrated into a scalable test platform so that the AI tests are carried out efficiently and the test results are documented automatically.

Förderhinweis

Diese Publikation wurde durch das Ministerium für Wirtschaft, Industrie, Klimaschutz und Energie des Landes Nordrhein-Westfalen im Rahmen des Flaggschiff-Projekts ZERTIFIZIERTE KI unterstützt.

Schlüsselwörter

Vertrauenswürdige KI, Systematische Schwächen, KI-Testwerkzeug, KI-Prüfungen, KI-Prüfplattform

Keywords

Trustworthy AI, Systematic Weaknesses, AI Assessment Tools, AI Assessments, AI Testing Platform

Bibliography

DOI:10.1515/zwf-2024-0137
ZWF 120 (2025) Special Issue; page 159 – 165

ISSN 0947-0085 · e-ISSN 2511-0896