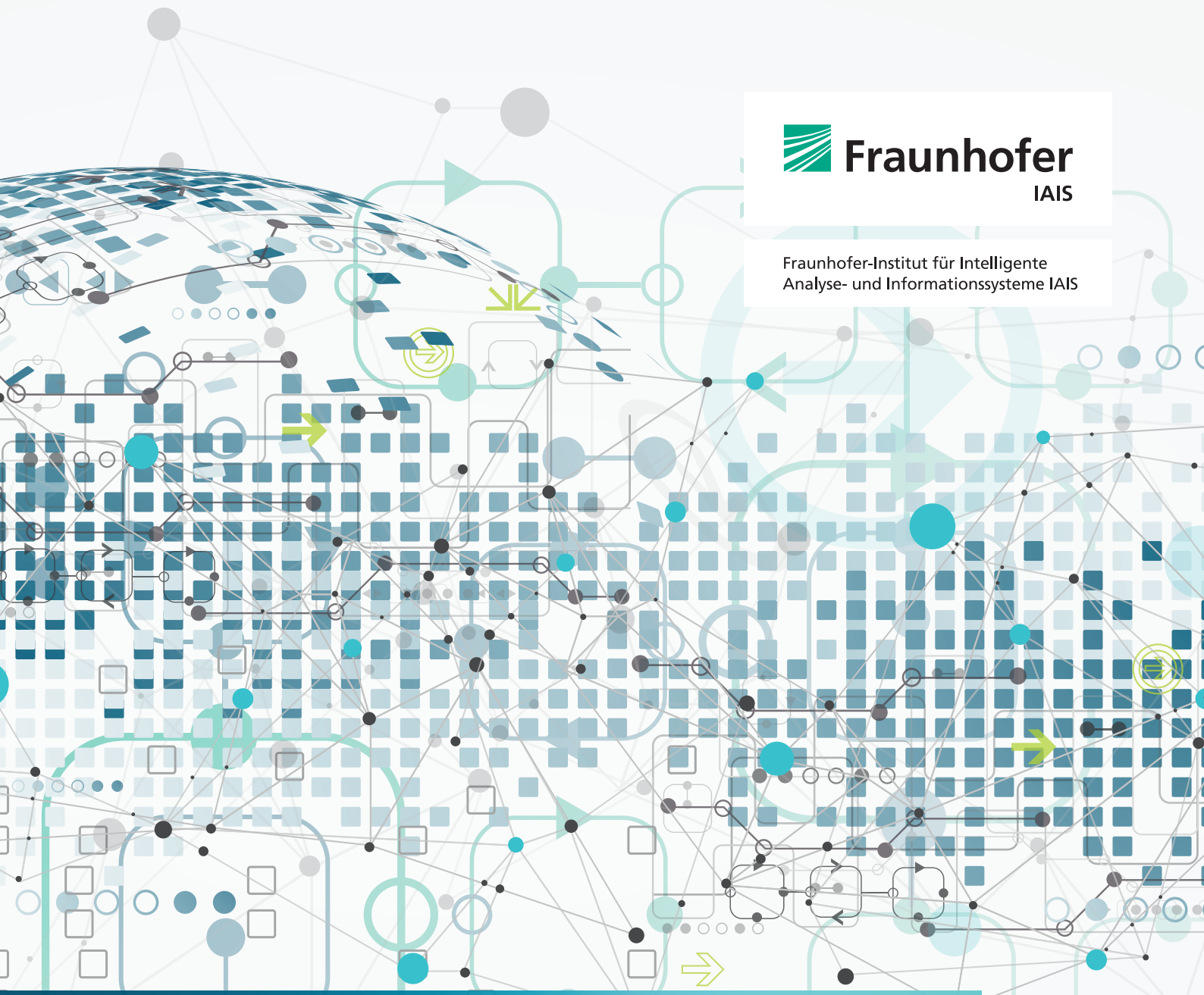


Fraunhofer IAIS

Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS

KI-Anwendungen systematisch prüfen und absichern

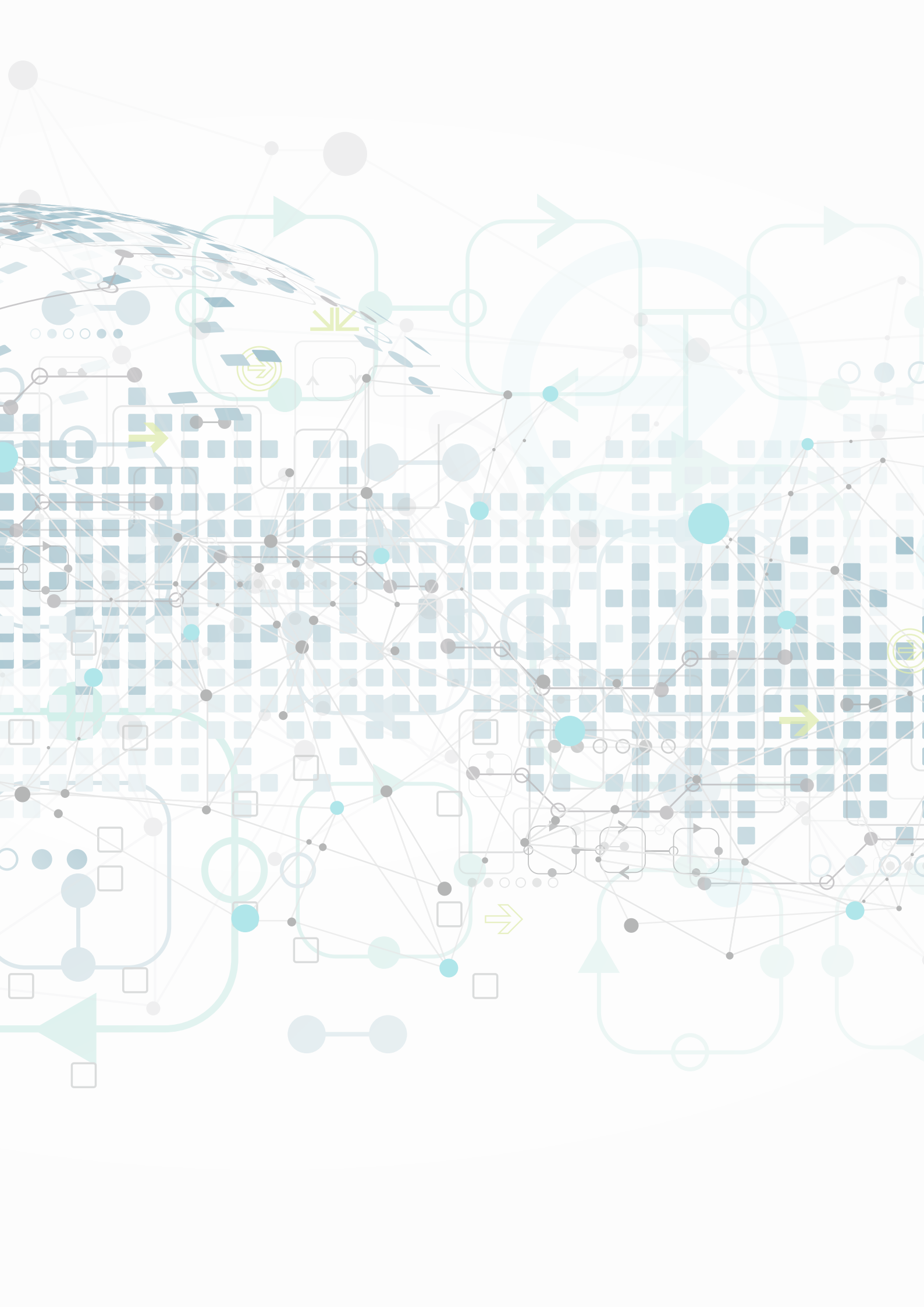
Prüfwerkzeuge und Prüfplattform zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz

POWERED BY KI NRW

Gefördert durch:
Ministerium für Wirtschaft, Industrie, Klimaschutz und Energie des Landes Nordrhein-Westfalen

ZERTIFIZIERTE KI
Qualität sichern. Fortschritt gestalten.

UNIVERSITÄT BONN



Inhalt

Das Fraunhofer IAIS

Als Teil der größten Organisation für anwendungsorientierte Forschung in Europa ist das Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS mit Sitz in Sankt Augustin/Bonn und einem Standort in Dresden eines der führenden Wissenschaftsinstitute auf den Gebieten Künstliche Intelligenz (KI), Maschinelles Lernen und Big Data in Deutschland und Europa.

Rund 350 Mitarbeitende unterstützen Unternehmen bei der Optimierung von Produkten, Dienstleistungen und Prozessen sowie bei der Entwicklung neuer digitaler Geschäftsmodelle. Das Fraunhofer IAIS gestaltet die digitale Transformation unserer Arbeits- und Lebenswelt: mit innovativen KI-Anwendungen für Industrie, Gesundheit und Nachhaltigkeit, mit zukunftsweisenden Technologien wie großen KI-Sprachmodellen oder Quantum Machine Learning, mit Angeboten für die Aus- und Weiterbildung oder für die Prüfung von KI-Anwendungen auf Sicherheit und Vertrauenswürdigkeit.

www.iais.fraunhofer.de

Fraunhofer IAIS

KI-Anwendungen systematisch prüfen und absichern

Prüfwerkzeuge und Prüfplattform zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz

Whitepaper

Autorinnen und Autoren
Elena Headecke, Fraunhofer IAIS und Universität Bonn
PD. Dr. Michael Mock, Fraunhofer IAIS und Universität Bonn
Maximilian Pintz, Fraunhofer IAIS und Universität Bonn
Dr. Maximilian Poretschkin, Fraunhofer IAIS und Universität Bonn

www.iais.fraunhofer.de/zertifizierte-ki

KI.NRW

KI.NRW ist die zentrale Anlaufstelle für Künstliche Intelligenz in Nordrhein-Westfalen. Die Kompetenzplattform baut das Land zu einem bundesweit führenden Standort für angewandte KI aus. Ziel ist es, den Transfer von KI aus der Spitzenforschung in die Wirtschaft zu beschleunigen und Impulse im gesellschaftlichen Dialog zu setzen. Dabei stellt KI.NRW die Menschen und ihre ethischen Grundsätze in den Mittelpunkt der Gestaltung von KI.

www.ki.nrw

ZERTIFIZIERTE KI
Qualität sichern. Fortschritt gestalten.

Das Projekt ZERTIFIZIERTE KI fördert die Entwicklung und Standardisierung von Prüfkriterien, -methoden und -werkzeuge für KI-Systeme, um die technische Zuverlässigkeit und einen verantwortungsvollen Umgang mit der Technologie zu gewährleisten.

www.zertifizierte-ki.de

Executive Summary

Als Schlüsseltechnologie der Zukunft birgt Künstliche Intelligenz (KI) enormes Innovationspotenzial für Wirtschaft und Gesellschaft. Eine umfassende Prüfung solcher KI-Systeme in Bezug auf Risiken wie mangelnde Zuverlässigkeit, Fairness oder Transparenz ist die Basis und Voraussetzung für ihren sicheren Einsatz und die gesellschaftliche Akzeptanz. Gleichzeitig formulieren Regulierungen wie der AI Act zusätzliche, rechtliche Anforderungen. Aufgrund der Komplexität und Datenabhängigkeit leistungsstarker KI-Systeme benötigt man für ihre technische Prüfung spezielle Software-Werkzeuge und Testverfahren, die weit über klassische Ansätze hinausgehen. Es ist zudem wichtig, dass Prüfungen von KI-Modellen standardisiert durchgeführt werden können. Die Tests müssen wiederholbar sein und dürfen nicht der Varianz von Modellen unterliegen, um eine Vergleichbarkeit sicherzustellen. Die Herausforderung ist dabei, dass die Anwendungsbereiche und damit die Anforderungen an marktfähige Prüfwerkzeuge unterschiedlich sind. Beispielsweise benötigt man für die Prüfung einer KI-Anwendung in der Medizin andere Prüfwerkzeuge als für den Qualitätscheck einer KI-Lösung im autonomen Fahrzeug.

Daher müssen verschiedene, KI-spezifische Prüfwerkzeuge entwickelt werden, die KI-Systeme auf Risiken systematisch testen und somit Entwickler*innen sowie Prüfer*innen bei der Validierung der Systeme unterstützen. Um KI-Tests einfach, nachvollziehbar und reproduzierbar durchzuführen, müssen die Prüfwerkzeuge in ein gemeinsames Software-Framework integriert werden. Dieses Whitepaper stellt ein Prüf-Framework vor, das diese Herausforderungen bewältigt und als Plattform zur Integration von KI-Prüfwerkzeugen dient. Außerdem werden ausgewählte KI-Prüfwerkzeuge, die für unterschiedliche Anwendungsbereiche genutzt werden können, sowie ihre Funktionalitäten im Detail erklärt.

1. Herausforderungen des vertrauenswürdigen Einsatzes von KI

Schlüsseltechnologie KI

Die Fortschritte, die Künstliche Intelligenz (KI) in den vergangenen Jahren und Monaten erzielt hat, sind rasant und revolutionär. Prominente Bereiche, in denen KI-Systeme immer häufiger zum Einsatz kommen, sind beispielsweise die medizinische Diagnostik oder die prädiktive Wartung. Perspektivische Anwendungsbeispiele finden sich zudem im autonomen Fahren und in der Generierung von Text- und Bildmaterial durch sogenannte Foundation Models. Ein bekanntes Beispiel ist die KI »ChatGPT«, die etwa zusammenhängende Chat-Konversationen führen kann und seit Anfang des Jahres 2023 auf der ganzen Welt medial omnipräsent ist.[1] Im März 2023 wurde mit GPT-4[2] eine neue Version vorgestellt, welche weitere Verbesserungen verspricht, beispielsweise in Bezug auf logisches Denken innerhalb einer Konversation. Aktuell wird deutlicher denn je, wie stark KI unsere Gesellschaft und Wirtschaft als Schlüsseltechnologie[3] bereits prägt und auch zukünftig weiter revolutionieren wird.

Voraussetzung dafür, dass mittels KI – und den auf ihr basierenden Geschäftsmodellen – das gesamte wirtschaftliche Potenzial ausgeschöpft werden kann, ist eine Absicherung gegen die neuartigen, KI-spezifischen Risiken. Dafür müssen KI-Anwendungen nach hohen Qualitätsanforderungen entwickelt und systematisch sowie interdisziplinär überprüft werden, um die Risiken zu evaluieren und geeignet zu mitigieren. Darüber hinaus sieht der in Kürze in Kraft tretende Europäische AI Act – der weltweit erste Entwurf zur Regulierung von KI – eine verpflichtende Prüfung von sogenannten Hochrisiko-KI-Anwendungen vor.

Neue Herausforderungen

Die Technologie, die KI-Anwendungen auf Basis von Maschinellem Lernen (engl. *machine learning*, ML) oder tiefen neuronalen Netzen (engl. *deep neural networks*, DNNs) zugrunde liegt, funktioniert grundsätzlich anders, als es von der »klassischen« Programmierung bekannt ist. Den Aktionen von Systemen, die auf klassischer Programmierung basieren, liegen formale Regeln zugrunde, die durch die Programmierung vorgegeben wurden. Im Gegensatz dazu sind ML- oder DNN-Systeme in der Lage, aussagekräftiges Wissen bzw. Informationen aus einer großen Menge an Daten selbst zu extrahieren.

Dieses datengetriebene »Lernen« umgeht einerseits die Limitierungen der klassischen Programmierung, welche die hohe Komplexität spezieller Anwendungsfälle nicht ausreichend abbilden kann. Erst durch die Entwicklung der hochdimensionalen DNNs konnte eine Performanz erreicht werden, die es möglich macht, Anwendungsfälle aus der Sprachverarbeitung und Bilderkennung (beispielsweise für das autonome Fahren oder die medizinische Diagnostik) zu erschließen. Andererseits stellt die hohe Komplexität vielfältige Herausforderungen an die Vertrauenswürdigkeit[4] und die Prüfung von KI-Systemen, die entsprechend erfasst und adressiert werden müssen. Eine Herausforderung ist die Intransparenz von KI-Modellen: Die aus Milliarden Parametern bestehenden KI-Modelle können in ihrer Funktionsweise selbst für Expert*innen oft schwer nachvollzogen werden. Daher wird die Leistungsfähigkeit von KI-Anwendungen oft von ihren Anbietern und Entwickler*innen

---

1 Für eine Auswahl an News-Artikeln siehe beispielsweise:
Facettenreicher Gesprächspartner. Die Text-KI ChatGPT schreibt Fachtexte, Prosa, Gedichte und Programmcode. 01.2023 (https://www.heise.de/select/ct/2023/1/2233908274346530870, letzter Aufruf am 09.03.2023).
Big Tech was moving cautiously on AI. Then came ChatGPT. 27.01.2023 (https://www.washingtonpost.com/technology/2023/01/27/chatgpt-google-meta/, letzter Aufruf am 09.03.2023).
ChatGPT reaches 100 million users two months after launch. 02.02.2023 (https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app, letzter Aufruf am 09.03.2023).

2 GTP-4 Technical Report. 2023. OpenAI.

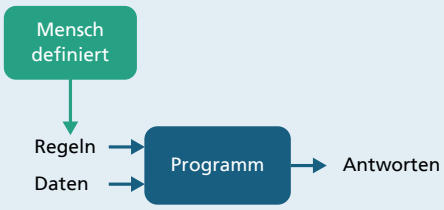
3 Für eine Liste an Anwendungsfällen von KI-Systemen, siehe ISO/IEC TR 24030:2021 Information technology – Artificial intelligence (AI) – Use cases.
Für eine Beschreibung von Anwendungsfällen von KI, siehe z.B. »DIN e.V. & DKE. Deutsche Normungsroadmap Künstliche Intelligenz Version 2« (engl.: German Standardization Roadmap Version 2). 2022. URL: https://www.dke.de/de/arbeitsfelder/core-safety/normungsroadmap-ki , letzter Aufruf am 29.03.2023), sowie für Anwendungsfälle speziell im Bereich der Produktion und im autonomen Fahren, siehe auch »Künstliche Intelligenz in sicherheitskritischen Anwendungen – Normen, Standards und Sicherheitslücken« von M. Köllemann (https://www.researchgate.net/publication/339434796_Kunstliche_Intelligenz_in_sicherheitskritischen_Anwendungen_-_Normen_Standards_und_Sicherheitslucken, letzter Aufruf am 18.10.2022).

4 Für eine kurze Einführung zum Begriff der »Vertrauenswürdigkeit« im Kontext von KI-Systemen, siehe auch Schmitz, A., et al. 2022. »The why and how of trustworthy AI: An approach for systematic quality assurance when working with ML components« at - Automatisierungstechnik, vol. 70, no. 9, pp. 793–804. https://doi.org/10.1515/auto-2022-0012.

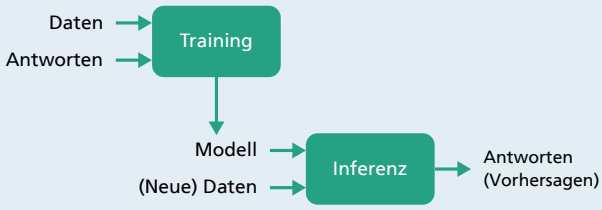
überschätzt und selten kritisch hinterfragt. Nutzer*innen können die potenziellen Schwächen von KI-Anwendungen häufig erst nach ihrer Einführung und einer kostenintensiven Erprobung feststellen.

Da KI als Querschnittsthema viele Bereiche tangiert, existieren zur Definition von Kernanforderungen vertrauenswürdiger KI zahlreiche Beiträge[5] aus Forschung, Industrie und Gesellschaft – mit den »Ethics Guidelines for Trustworthy AI« der High-Level Expert Group (HLEG) der European AI Alliance[6] als prominentes Beispiel. Insgesamt haben sich sechs Dimensionen[7] der Vertrauenswürdigkeit herauskristallisiert, welche sowohl bei der Entwicklung als auch bei der Überprüfung einer KI eine Rolle spielen. Diese werden im nachfolgenden Abschnitt beschrieben.

GEGENÜBERSTELLUNG KLASSISCHE PROGRAMMIERUNG UND MASCHINELLES LERNEN



Klassische Programmierung
Bei der klassischen Programmierung modelliert der Mensch die reale Welt, z. B. mit formalen Regeln. Das Software-Programm wendet diese Regeln anschließend auf die Eingabedaten an, um darauf basierende Antworten zu erzeugen.



Maschinelles Lernen / Tiefe neuronale Netze
Beim Maschinellen Lernen und tiefen neuronalen Netzen werden dem Modell im Zuge des Trainings riesige Datenmengen mit den dazugehörigen Musterantworten gezeigt. Die komplexen Zusammenhänge, die zwischen den Daten und den Antworten bestehen, erlernt das KI-Modell selbst. Anschließend kann das trainierte Modell in der Inferenz auf neue, bisher unbekannte Daten angewendet werden. Das KI-Modell liefert dann Antworten im Sinne von Vorhersagen.

5 Jobin, A., et al. 2019. The global landscape of AI ethics guidelines. Nat. Mach. Intell. 1: 389–399.
6 High-Level Expert Group on AI. 2019. Ethics Guidelines for Trustworthy AI. European Commission.
7 Cremers, A. B., et al. 2019. Vertrauenswürdiger Einsatz von Künstlicher Intelligenz. Handlungsfelder aus philosophischer, ethischer, rechtlicher und technologischer Sicht als Grundlage für eine Zertifizierung von Künstlicher Intelligenz. Fraunhofer-Institut für Intelligente Analyse und Informationssysteme IAIS.

HERAUSFORDERUNGEN DURCH KI-EINSATZ[7]

Abhängigkeit von Trainingsdaten

Während des Trainings erlernt das KI-Modell selbstständig die Muster, die den Daten zugrunde liegen. Daraus folgt, dass die Qualität des Modells erheblich von der Güte der Datengrundlage abhängig ist: Informationen, die in den Trainingsdaten nicht vorhanden oder unterrepräsentiert sind, führen folglich zu einer Verzerrung zwischen Erlerntem und der Realität, die das Modell im Betrieb eigentlich abbilden sollte.

Unsicherheiten

Die Trainingsdaten und die Eingabedaten besitzen qualitative Unsicherheiten, die sich nicht immer vermeiden lassen. Zusammen mit dem probabilistischen Charakter des KI-Modells führt dies zu ungefähren Vorhersagen, die mit Unsicherheiten behaftet sind.

Intransparenz des Modells

Nur wenige ML-Modelle sind intrinsisch interpretierbar – bei den meisten handelt es sich um eine sogenannte Blackbox, deren Verhalten nur von außen betrachtet werden kann. Die innere Funktionsweise und die Gründe, warum es zu einer gewissen Entscheidung bzw. Vorhersage gekommen ist, bleiben verborgen, sofern keine Methoden zur besseren Nachvollziehbarkeit genutzt werden.

Testen des Modells

Aufgrund der Komplexität der KI-Modelle sind klassische Software-Testmethoden nicht ausreichend, da sich die Modelle selten in separat prüfbare Einheiten zerlegen lassen. Die Komplexität erschwert zudem die Definition zulässiger Eingaben (sog. Operational Design Domain (ODD)).

Weiterlernen im Betrieb

Bei KI-Modellen, die auch im Betrieb noch optimiert werden sollen, wird ein Weiterlernen ermöglicht. Dies kann beispielsweise durch Nutzerfeedback geschehen. Allerdings besteht hier die besondere Schwierigkeit darin, die entsprechenden Leitplanken für das Weiterlernen zu definieren, sodass das KI-Modell auch künftig kontrolliert eingesetzt werden kann.

2. Dimensionen der Vertrauenswürdigkeit von KI

Die Auswirkungen des Einsatzes von KI-Systemen sind vielschichtig, weshalb die Betrachtung der Vertrauenswürdigkeit einer KI ebenfalls aus verschiedenen Perspektiven erfolgen muss. Zunächst sind für KI-Systeme auch die »klassischen« Sicherheitsthemen wie die funktionale Sicherheit, die Angriffssicherheit und der Datenschutz wichtig, um Risiken wie Personenschäden, Schäden an Besitztümern oder finanzielle Verluste zu vermeiden. Darüber hinaus besitzen KI-Systeme aufgrund ihrer verschiedenen Einsatzmöglichkeiten das Potenzial, starke Auswirkungen auf individuelle Personen und die Gesellschaft zu haben, weshalb ergänzend die Dimensionen Fairness, Transparenz sowie Autonomie und Kontrolle des Menschen Teil des Vertrauenswürdigkeitsbegriffs sind. Die Kernaspekte und Risiken der einzelnen Dimensionen orientieren sich am KI-Prüfkatalog[8] des Fraunhofer IAIS, auf welchen für eine ausführliche Darstellung verwiesen sei. Im Nachfolgenden werden die Kernaspekte und Risiken anhand von Leitfragen je Dimension vorgestellt.

Autonomie und Kontrolle

Für die Dimension Autonomie und Kontrolle sind zwei Fragestellungen zentral:

- Ist es aus menschlicher Sicht möglich, die KI effizient und selbstbestimmt zu nutzen, oder wird der Mensch in seinem Handlungsspielraum eingeschränkt?
- Welcher Grad an Autonomie ist für die KI-Anwendung angemessen?

Es gilt also zu überprüfen, ob den Nutzer*innen und Stakeholdern der KI-Anwendung genügend Informationen und Handlungsfähigkeit bereitgestellt werden und ob die Aufteilung der Aufgaben zwischen KI-System und Mensch angemessen sind.

Fairness

Auf Basis des datengetriebenen Lernens sind KI-Systeme anfällig dafür, ein »unfaires« Verhalten zu erlernen. Beispielsweise können die Trainingsdaten Vorurteile und Bias beinhalten oder in der Datengrundlage bestimmte Gruppen unterrepräsentiert sein. Die Dimension Fairness stellt daher die Frage:

- Werden alle Betroffenen von der KI fair behandelt?

Bei der Überprüfung ist es wichtig, den Begriff der Fairness zu quantifizieren, also etwa potenziell benachteiligte Gruppen zu identifizieren. Zusätzlich sollte beachtet werden, ob sich etwa durch verändernde Rahmenbedingungen neue Fairness-Risiken ergeben können.

Transparenz

Der Blackbox-Charakter von KI-Systemen und deren hohe Komplexität stellen für Expert*innen und Nutzer*innen gleichermaßen eine Hürde hinsichtlich der Nachvollziehbarkeit der Ausgaben und der Erklärbarkeit der Wirkungsweise dar. Die Dimension Transparenz stellt somit die Fragen:

- Sind die Funktionsweise der KI und deren Ausgaben verständlich und nachvollziehbar, sodass darauf basierende Entscheidungen informiert getroffen werden können?
- Ist es möglich, die Ergebnisse zu reproduzieren?

Denn wenn Entscheidungsgrundlagen in kritischen Anwendungsbereichen nicht nachvollziehbar sind, kann es zu folgenschweren Fehlentscheidungen kommen. Beispielsweise sollte ein KI-System im Medizinbereich so gestaltet sein, dass Expert*innen die Ergebnisse verstehen und verifizieren können, bevor weitere Schritte eingeleitet werden.

8 Poretschkin, M., et al. 2021. KI-Prüfkatalog: Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz. Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS.

Verlässlichkeit

Um die Verlässlichkeit eines KI-Systems zu überprüfen, müssen mehrere Qualitätsaspekte der KI betrachtet werden:

- Funktioniert die KI zuverlässig und produziert sie korrekte Ausgaben?
- Können die Unsicherheiten des KI-Modells erkannt und eingeschätzt werden?
- Verhält sich die KI auch robust gegenüber Fehleingaben oder unerwarteten Situationen?

Wenn ein KI-System nicht zuverlässig und robust funktioniert, kann dies in Produktionsanlagen kostenschwere Ausfälle verursachen, weil z. B. durch Schmutzpartikel auf Sensoren Mängel nicht erkannt werden. Es muss überprüft werden, ob trotz rauschender Inputdaten valide Ergebnisse produziert werden und ob es Anwender*innen möglich ist, Unsicherheiten des Modells zu erkennen.

Sicherheit

Der Begriff der Sicherheit umfasst die zwei Bereiche Angriffssicherheit (Security) sowie Gefährdungssicherheit (Safety):

- Security: Ist die KI gegenüber Angriffen und Manipulationen ausreichend abgesichert?
- Safety: Ist ein sicherer Betrieb der KI möglich?

Im Sinne der Security besteht das Risiko, dass Datenmanipulation (z. B. gezielte Angriffe) zu Unfällen und Fehlern bzw. zum Ausfall der KI-Anwendung führen können. Wenn beispielsweise die Steuerung von Geräten in Betrieben manipuliert wird oder durch einen Ausfall der Sensorik fehlerhafte Eingabedaten entstehen, kann einerseits die Verfügbarkeit des KI-Systems betroffen sein, andererseits kann es zu Unfällen mit Personen- und Sachschäden kommen. Im Sinne der Safety müssen in solchen Fällen geeignete Maßnahmen greifen, um den sicheren Betrieb weiter gewährleisten zu können.

Datenschutz

KI-Anwendungen greifen auf eine große Menge von Daten zu, die auch sensibler Art sein können. Zudem entstehen beim Zusammenführen verschiedener Datenquellen und der Verarbeitung durch die KI zusätzlich neue, schützenswerte Daten. Demnach muss hinterfragt werden:

- Werden die Privatsphäre und sonstige sensible Informationen durch die KI geschützt?

Dabei ist sowohl der Schutz privater als auch geschäftlicher Daten wichtig. Denn wenn die Trainingsdaten vertrauliche Informationen enthalten, können diese theoretisch über das Modell ausspioniert werden und rückverfolgbar sein. Es ist zu überprüfen, ob sowohl die KI selbst als auch die zugrunde liegenden Daten durch geeignete Maßnahmen geschützt werden. Sonst könnten z. B. durch Hacking-Angriffe Geschäftsgeheimnisse preisgegeben werden oder es könnte durch Leaks von Kund*innendaten zu Rufschädigung, Klagen und Kosten kommen.

3. Relevanz von KI-Prüfungen

Prüfungen als Teil des Entwicklungsprozesses und der Qualitätskontrolle

Prüfungen sind ein wichtiger Baustein von Entwicklungsprozessen und der Qualitätskontrolle. Sie können freiwillig erfolgen, etwa mit dem Ziel, Vertrauen bei Endkund*innen durch den Erwerb von mit einer Prüfung verbundenen Gütesiegeln zu schaffen. Prüfungen können Unternehmen auch Wettbewerbsvorteile durch die Etablierung einer starken und zuverlässigen Marke verschaffen. Daneben stellen Prüfungen auch eine wichtige Grundlage für die Zulassung und das Inverkehrbringen von Produkten dar. Diese unterschiedlichen Motivationen gelten für KI-Systeme in besonderem Maße. Weltweite Regulierungsaktivitäten für den Einsatz von KI-Systemen lassen zudem erwarten, dass KI-Prüfungen zumindest für besonders riskante KI-Systeme verpflichtend werden.

Prüfungen als Teil von internationalen regulatorischen Anforderungen

Der derzeitige – und zukünftig vermehrte – Einsatz von KI-Systemen macht deutlich, dass auf nationaler und internationaler Ebene Rahmen und Leitlinien immer nötiger werden.

Aus dem breiten gesellschaftlichen Diskurs um die Sicherheit im Umgang mit KI-Technologien hat beispielsweise im Jahr 2019 die High-Level Expert Group (HLEG) der European AI Alliance[9] die eingangs erwähnten »Ethics Guidelines for Trustworthy AI« veröffentlicht. Eine ähnliche Richtlinie folgte 2020 durch die EU-Kommission mit dem »White Paper on Artificial Intelligence«[10]. Ziel dieser Veröffentlichungen war es, die Auswirkungen des KI-Einsatzes aus dem Blickwinkel diverser Stakeholder zu betrachten, um darauf basierend einen EU-Vorschlag für ein Rahmenwerk für vertrauenswürdige KI zu entwerfen.

Diese vorangegangenen, nicht bindenden Richtlinien führten im Jahr 2021 auf EU-Ebene dann zur Veröffentlichung des Gesetzesentwurfs »Artificial Intelligence Act (AI Act)« der EU-Kommission. Dieser vorgeschlagene Rechtsrahmen legt ein besonderes Augenmerk auf KI-Systeme in kritischen Bereichen und schlägt abhängig vom Gefahrenpotenzial unterschiedliche Auflagen für Vertrauenswürdigkeit, Transparenz und Zertifizierung vor. Ähnlich sieht auch die 2022 aktualisierte »National AI Strategy«[11] aus Großbritannien vor, zukünftige KI-Regulierungen zu entwickeln. Mit der »AI Bill of Rights«[12] wurde im Jahr 2022 ein amerikanisches Pendant von der US-Regierung vorgeschlagen, welche aber eine freiwillige Umsetzung der definierten Richtlinien vorsieht.

3.1. Standards als notwendige Voraussetzung zur Operationalisierung von KI-Prüfungen

Eine wichtige Basis, um KI-Prüfungen zu operationalisieren, stellen Normen und Standards dar. Sie schreiben die Anforderungen fest, die geprüft werden. Die 2022 in zweiter Version von DIN & DKE veröffentlichte Normungsroadmap Künstliche Intelligenz[13] stellt wichtige Normen und Standards in Verbindung mit KI vor und zeigt Handlungsbedarfe für die Normung auf. Zudem ist der AI Act in das sogenannte New Legislative Framework eingebettet, welches vorsieht, dass viele Details der Regulierung durch noch zu definierende harmonisierte Standards geregelt werden. Die EU-Kommission hat hierzu entsprechende Anfragen an die europäischen Standardisierungsorganisationen gestellt, welche insbesondere auch einen Standard zur Durchführung von KI-Konformitätsprüfungen umfassen.

---

9 High-Level Expert Group on AI. 2019. Ethics Guidelines for Trustworthy AI. European Commission.

10 European Commission. 2020. White Paper on Artificial Intelligence – A European approach to excellence and trust. (https://commission.europa.eu/document/download/d2ec4039-c5be-423a-81ef-b9e44e79825b_en?filename=commission-white-paper-artificial-intelligence-feb2020_en.pdf, letzter Aufruf am 29.03.2023).

11 UK Government. 2022. National AI Strategy. (https://www.gov.uk/government/publications/national-ai-strategy, letzter Aufruf am 29.03.2023).

12 The White House. 2022. Blueprint for an AI Bill of Rights. Making automated systems work for the American people. (https://www.whitehouse.gov/ostp/ai-bill-of-rights/, letzter Aufruf am 29.03.2023).

13 DIN e.V. & DKE. 2022. Deutsche Normungsroadmap Künstliche Intelligenz Ausgabe 2 (engl.: German Standardization Roadmap Artificial Intelligence Version 1). (https://www.dke.de/de/arbeitsfelder/core-safety/normungsroadmap-ki, letzter Aufruf am 29.03.2023).

3.2. Definition von unterschiedlichen Prüftiefen

Der Überblick über die verschiedenen Dimensionen der Vertrauenswürdigkeit von KI macht deutlich, dass Prüfungen von KI-Modellen sehr komplex sind, weil viele Aspekte beachtet und getestet werden müssen. Dies kann KI-Prüfungen sehr aufwendig und kostspielig werden lassen. Um möglichst marktfähige Prüfverfahren zu ermöglichen, bietet es sich an, mehrere »Prüftiefen« zu definieren, die unterschiedlichen »Assurance Levels« (ALs)[14] der Vertrauenswürdigkeit des geprüften KI-Systems entsprechen. Da die Aussagekraft einer Prüfung von der Prüftiefe abhängt, stellt ein Prüfansatz eines höheren Assurance Levels auch eine höhere Sicherheit durch die tiefergehende Prüfung dar. Je nach Kritikalität des KI-Systems und seines Anwendungskontextes kann dann die passende Prüftiefe ausgewählt werden. Im Folgenden wird beispielhaft dargestellt, wie eine dreistufige Ausprägung solcher Prüftiefen aussehen kann (siehe Abbildung 1).

Das niedrigste **Assurance Level 1** ist eine dokumentationsbasierte Überprüfung, welche z. B. anhand des KI-Prüfkatalogs des Fraunhofer IAIS durchgeführt werden kann. Dieser bietet einen strukturierten Leitfaden, mithilfe dessen abstrakte Qualitätsmaßstäbe zu anwendungsspezifischen Prüfkriterien konkretisiert werden können.

Das **Assurance Level 2** wird realisiert, indem zusätzlich zu den Anforderungen von Level 1 technische Tests durch qualifizierte Prüfer*innen durchgeführt werden. Typischerweise werden diese dimensionsspezifische Prüfwerkzeuge benötigen, welche die Durchführung solch technischer Tests erlauben. In Kapitel 6 werden Beispiele für solche Prüfwerkzeuge gezeigt. Unabhängig von der Prüftiefe und Aussagekraft der KI-Prüfung spielen solche Tools auch eine wichtige Rolle bei der Automatisierung, welche dafür sorgt, dass der manuelle Aufwand und Prüfungen marktfähig werden.

Das höchste **Assurance Level 3** fordert neben den Anforderungen von Level 2 die formalisierte Argumentation (»Assurance Case«), dass sämtliche KI-Risiken mitigiert sind. Solche Assurance Cases, also strukturierte Sicherheitsnachweise, sind nach aktuellem Stand noch nicht für alle KI-Modelle etabliert und Gegenstand derzeitiger Forschungen.[15]

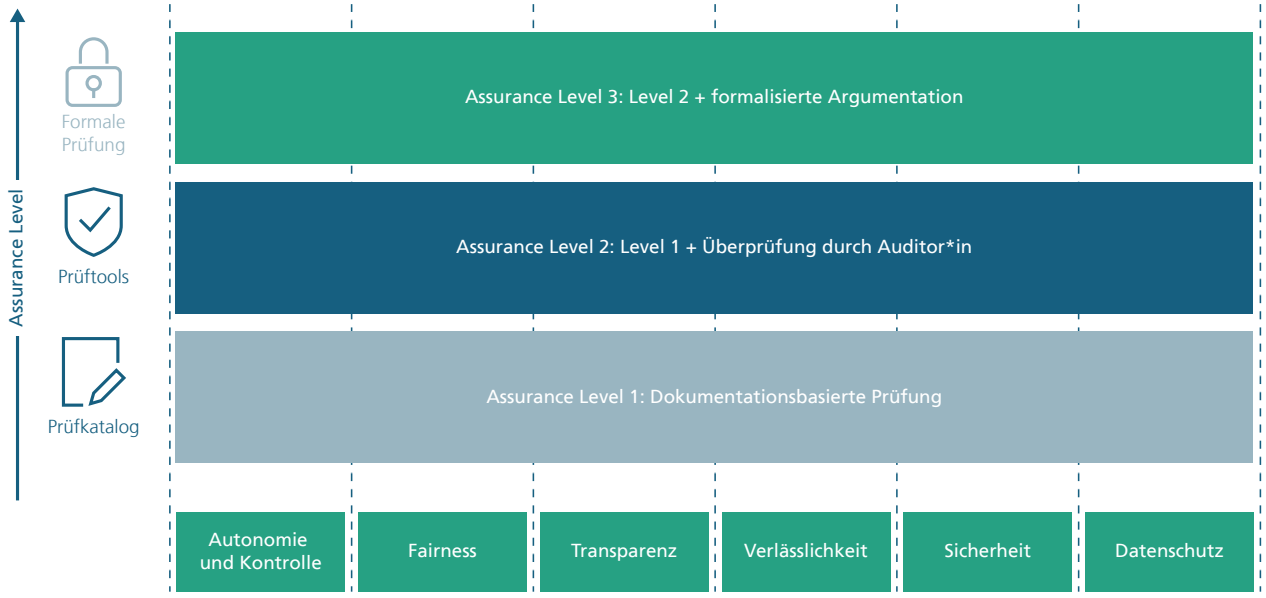


Abbildung 1: Verschiedene Konzepte der KI-Prüfung für die unterschiedlichen Prüfdimensionen entlang der drei Assurance Levels. Mithilfe der Prüftools kann eine Überprüfung im Sinne des Assurance Levels 2 umgesetzt werden. Abbildung: Fraunhofer IAIS

14 Mock, M. et al. 2021. An Integrated Approach to a Safety Argumentation for AI-Based Perception Functions in Automated Driving. In: Habli, I., et al. (eds) Computer Safety, Reliability, and Security. SAFECOMP 2021 Workshops. SAFECOMP 2021. Lecture Notes in Computer Science, vol 12853. Springer, Cham. https://doi.org/10.1007/978-3-030-83906-2_21.

15 Ein strukturierter Sicherheitsnachweis bedeutet, dass ein Nachweis für eine Erfüllung nach gewissen formalisierten Richtlinien zu erbringen ist. Ein Ansatz ist beispielsweise die Goal Structuring Notation (GSN), wie z.B. angewandt in: Schwalbe, G. et al. 2020. Structuring the Safety Argumentation for Deep Neural Network Based Perception in Automotive Applications. In: Casimiro, A., et al. (eds) Computer Safety, Reliability, and Security. SAFECOMP 2020 Workshops. SAFECOMP 2020. Lecture Notes in Computer Science, vol 12235. Springer, Cham. https://doi.org/10.1007/978-3-030-55583-2_29.

4. Notwendigkeit von vielfältigen KI-Prüfwerkzeugen und einer einheitlichen Prüfplattform

Die verschiedenen Aspekte, die bei einer KI-Prüfung zu beachten sind, machen deutlich, dass es ebenso vielfältiger Prüfwerkzeuge bedarf, um entsprechende Tests je zu überprüfender Dimension durchführen zu können. Da KI als Querschnittsthema in sehr vielen Disziplinen eingesetzt wird, werden individuelle Lösungen je Anwendungsfall benötigt. Beispielsweise benötigt man für die Prüfung einer KI-Anwendung in der Medizin andere Tools als für die Prüfung einer KI-Lösung im autonomen Fahrzeug. Das führt dazu, dass bisher viele einzelne KI-Prüfwerkzeuge benutzt werden, die eine Vergleichbarkeit der Ergebnisse erschweren. Diese Vergleichbarkeit spielt jedoch für die Standardisierung und Reproduzierbarkeit von Prüfungen eine wichtige Rolle. Auch die Dokumentation der Prüfergebnisse ist insbesondere für (externe) Prüfungen oder Audits wichtig, um eine Rechtfertigung für ein Prüfurteil nachweisen zu können.

Diese Anforderungen können gelöst werden, indem die vielfältigen Prüfwerkzeuge interoperabel eingebettet werden. Hierzu bietet sich ein Software-Framework an, welches einheitliche Schnittstellen bereitstellt, sodass eine Zusammenarbeit zwischen Modellen, Daten und Prüfwerkzeugen hergestellt wird. Ein solches Framework kann darüber hinaus eine effiziente Nutzung unterstützen, wenn einfache Methoden für die gängigen Schritte einer Prüfdurchführung bereitgestellt werden, die über Programmierschnittstellen genutzt werden können. Wird ein Framework mit einer Plattform mit leicht verständlicher, graphischer Oberfläche kombiniert, kann ein breites Spektrum an Nutzer*innen darauf zugreifen. Bereitgestellte Tools können genutzt beziehungsweise eigene Tools entwickelt und eingebunden werden.

Letztlich können durch die Nutzung eines geeigneten Frameworks, auch in Verbindung mit einer Plattform, Zeit und Kosten gespart werden – sowohl beim Entwickeln von KI-Modellen als auch von anwendungsspezifischen KI-Prüfwerkzeugen.

ASPEKTE STANDARDISIERTER TESTS

Test-Workflows beinhalten üblicherweise Spezifikationen der Testprozedur, der Inputs und Outputs, der zu loggenden Parameter und viele weitere Aspekte, um sicherzustellen, dass die Prüfungen untereinander konsistent sind. Schon während des Entwicklungsprozesses ist eine solche Konsistenz wichtig, damit eine Optimierung eines Modells effektiv durchgeführt werden kann.

Werden verschiedene Parameter eines Modells verändert, um zu überprüfen, ob dies zu besseren oder schlechteren Ergebnissen führt, muss sichergestellt sein, dass die Modellvariationen stets denselben Testbedingungen unterliegen. Nur so kann überprüft werden, ob die Parameteränderung zu einer Verbesserung oder Verschlechterung der Performanz geführt hat und darauf basierend die Modellentwicklung unterstützt werden. Nur wenn die für eine Prüfung genutzten Tools auf dieselben Grundmetriken zurückgreifen, ist das Ziel erreichbar, standardisierte Prüfungen zur Herstellung von Vergleichbarkeit und Reproduzierbarkeit bei der Modellentwicklung oder der (externen) Prüfung zu realisieren.

5. Konzept einer Prüfplattform und eines Software-Frameworks

Eine Prüfplattform und ein Software-Framework können es einzeln – oder auch im Zusammenspiel – ermöglichen, die komplexen und vielfältigen Anforderungen eines breiten Nutzenspektrums abzudecken. Im Folgenden werden für beides Konzepte vorgestellt, die effiziente und konsistente Prüfprozesse unterstützen.

Sie schaffen die Rahmenbedingungen und Schnittstellen für effiziente KI-Prüfungen. Die zu prüfenden Aspekte werden mithilfe von Prüftools umgesetzt. Beispiele für solche Prüfwerkzeuge werden im letzten Kapitel vorgestellt.

5.1. Prüfplattform zur Endnutzer*innen-Bereitstellung von Demonstratoren und Prüfwerkzeugen

Die **Prüfplattform** ist eine Onlineplattform, die Demonstratoren und Prüftools zur standardisierten Prüfung von KI-Systemen bereitstellt. Neben Prüftools, die vom Plattformbetreiber selbst bereitgestellt werden, können auch Werkzeuge von anderen Anbietern in die Plattform eingebunden werden, sodass diese für KI-Entwicklungsprojekte zur Verfügung stehen. So können auch kundeneigene Anpassungen der Prüfwerkzeuge über einen einfachen Onlinezugang ermöglicht werden.

Das Konzept einer Onlineplattform erlaubt Endnutzer*innen einen einfachen Zugang zu den KI-Prüftools. Die enthaltenen Prüfwerkzeuge können **per Webinterface** als »Software as a Service« genutzt werden. Ist eine volle Kontrolle über die Ausführung gewünscht, lässt sich die Plattform auch ohne Installation in der eigenen Struktur mit **Docker-Containern** lokal, also »On-Premise«, starten.

5.2. Software-Framework für Interoperabilität, einfache Schnittstellen und Reproduzierbarkeit

Das **Software-Framework** stellt eine Lösung für die einheitliche Definition von Schnittstellen sowie Bereitstellung von Daten und Tools dar. Nutzer*innen – sowohl aus KI-Entwicklung als auch -Prüfung – können reproduzierbare Test-Workflows aus einer Menge von Daten, Modellen und Prüfwerkzeugen zusammenstellen. Hierbei ist es unerheblich, ob es sich um unterschiedliche Daten oder Modelle handelt, denn durch die festen Definitionen der Schnittstellen und Funktionen ist der Zugriff in jedem Fall interoperabel angelegt. So wird die Test- und Prüfarbeit erleichtert.

Angepasst auf die verschiedenen Bedürfnisse der Nutzer*innen stehen zudem drei Arten von **Nutzerschnittstellen** zur Verfügung. Für Prüfer*innen ist es beispielsweise wichtig, sich auf das Durchführen von Tests zu konzentrieren, sodass die graphische Nutzerschnittstelle eine einfach zu bedienende Oberfläche bereitstellt, ohne sich tiefer mit den Daten oder der konkreten Ausführung im Hintergrund auseinanderzusetzen. Für Entwickler*innen, die eigenen Programmcode nutzen möchten, wird sowohl ein Zugriff per Kommandozeile als auch eine Funktionsbibliothek bereitgestellt.

Das Software-Framework ist in **drei verschiedene Hauptkomponenten** aufgeteilt, wie in Abbildung 2 ersichtlich ist. Aus diesen Komponenten heraus besteht einheitlich Zugriff auf die KI-Modelle, die Prüfwerkzeuge und die Artefakte von Daten und Tests, die jeweils in eigenen Datenbanken gespeichert sind.

Innerhalb des **Modulmanagements** wird der Zugriff auf die Code Repositories bereitgestellt. Über Funktionen zum Laden und Speichern von Modul-Artefakten ist es etwa möglich, Code aus den entsprechenden Repositories oder versionierte KI-Modelle aus den Model-Stores abzurufen und zu laden sowie zu speichern. Funktionen zur Versionierung von Code, Prüfwerkzeugen und Modellen stellen sicher, dass Tests reproduziert werden können. So können z. B. auch mehrere Versionen eines Modells geladen werden. Mithilfe von Funktionen zur Modulumgebung können über die bekannten Umgebungen wie Docker, Conda oder Python Environments reproduzierbare Code-Umgebungen geschaffen werden.

Beim **Datenmanagement** liegt der Fokus auf der Bereitstellung einheitlicher Schnittstellen zu den Dateisystemen. Mit den Funktionen zum Speichern und Laden von Daten-Artefakten ist auch hier eine Versionierung möglich sowie ein Abruf von schon versionierten Daten-Artefakten aus den Data-Stores. Zur Vorbereitung der Daten auf die spezifischen Modelle werden Funktionen zur Datentransformation bereitgestellt, um beispielsweise die Datenformate oder die inneren Repräsentationen der Daten zu ändern.

Die gesamte Interaktion der Funktionen wird durch das **Test-management** zusammengeführt. Hier werden die aufrufbaren Module und die Daten zur Übergabe an Modelle und Funktionen in sogenannten Test-Workflows definiert. Diese ähneln einem Graphen, der die Interaktion zwischen den einzelnen Komponenten darstellt. Zur Versionierung und Reproduzierbarkeit sind die definierten Test-Workflows ebenfalls in den Test-Stores als Graph bzw. Datei speicherbar. Funktionen zur Testdurchführung führen in den Probe-Workflows definierte Tests aus. Überwacht werden sie über Logging-Funktionen. Die Testergebnisse können dann abgerufen und abgespeichert werden.

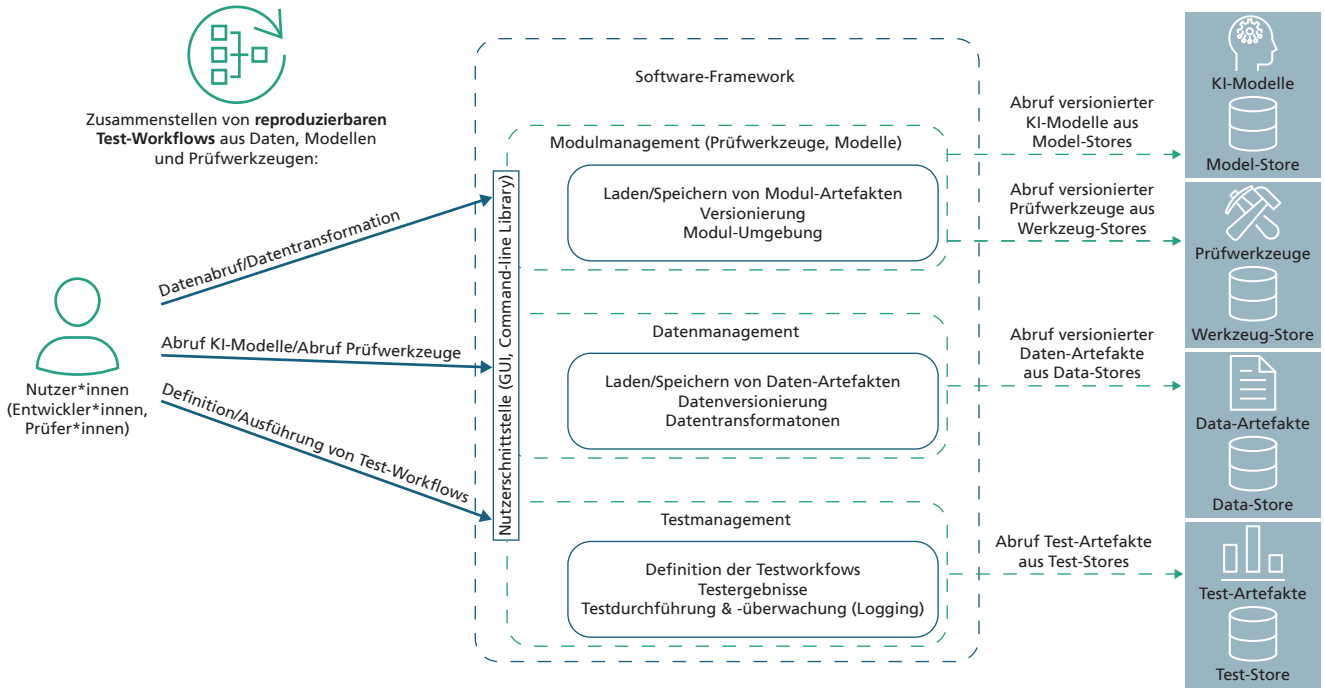


Abbildung 2: Software-Framework zur Zusammenstellung von reproduzierbaren Test-Workflows aus Daten, Modellen und Prüfwerkzeugen. Abbildung: Fraunhofer IAIS

6. Vorstellung von Prüfwerkzeugen für eine vertrauenswürdige KI

Insgesamt verbindet das Software-Framework vielfältige Prüfwerkzeuge zu einer mächtigen Prüftool-Suite. Diese ermöglicht umfassende KI-Prüfungen für verschiedene Einsatzbereiche. Zum Überblick über relevante Aspekte der KI-Vertrauenswürdigkeit, die bei einer KI-Prüfung mit verschiedenen Werkzeugen adressiert und getestet werden können, werden nachfolgend Beispiel-Prüftools beschrieben.

Für die umfassende Anwendung der Prüftool-Suite oder auch die Anpassung an den individuellen Anwendungsfall im Unternehmen wird entsprechendes Know-how benötigt. Ein enger Austausch sowie Beratungen zwischen Domänenexpert*innen im Unternehmen und den KI-Prüfexpert*innen sind empfehlenswert. So wird es möglich, die Prüftool-Suite zur gemeinsamen Entwicklung und Prüfung von vertrauenswürdiger KI effektiv und erfolgreich einzusetzen.

6.1. **AIBench** (Benchmarking-Tool)

Bei der Prüfung und Entwicklung von KI-Anwendungen ist der Vergleich verschiedener KI-Modelle essenziell. Mithilfe von automatisiertem Benchmarking können sie reproduzierbar verglichen werden.

Für die Lösung einer gegebenen Problemstellung kommt häufig der Einsatz mehrerer KI-Modelle infrage. Die Auswahl des geeignetsten Modells ist dabei sehr aufwendig: Es gibt viele verschiedene Aspekte und Herangehensweisen, um ein KI-Modell zu bewerten. Zudem ist viel technisches Know-how nötig, um einen solchen Vergleich statistisch valide und reproduzierbar durchzuführen.

Ein typischer Ansatz ist das »Benchmarking«. Hierbei werden verschiedene Modelle systematisch anhand quantitativer Performanz-Metriken miteinander verglichen. Zu den relevanten Prüfaspekten für KI-Modelle zählt beispielsweise die Detektionsgenauigkeit eines Modells – also die Rate, mit welcher das Modell in Testszenarien richtig liegt. Um die Signifikanz von Auswertungen sowie die Variabilität in den Endergebnissen korrekt einschätzen zu können, werden bei einem statistisch rigorosen Benchmarking mehrere Prüfdurchläufe aggregiert.[16] Die letztliche Auswertung ist allerdings oft mit einem hohen Aufwand verbunden, erfordert Expert*innenwissen über das KI-Modell sowie Programmierkenntnisse.

Um diesen aufwendigen Prozess zu erleichtern, bietet das Tool AIBench automatisiertes Benchmarking für eine große Bandbreite an KI-Modellen. Durch eine graphische Benutzeroberfläche wird ein Benchmarking mit wenigen Mausklicks interaktiv durchgeführt. Durch die Einbettung in das Software-Framework erbt AIBench die genannten Vorteile, beispielsweise die Unterstützung einer Vielzahl von Modellen und Datenformaten, und fügt eigene Benefits hinzu.

Um die Reproduzierbarkeit der Tests zu gewährleisten, ist ein Logging relevanter Parameter integriert. Die verschiedenen Prüfdurchläufe werden dazu in speziellen Dateien, sogenannten Prüf-Artefakten, abgespeichert, mit welchen die Ergebnisse nachvollziehbar reproduziert werden können. Um die statistische Validität der Tests beim Vergleich der Modelle zu gewährleisten, werden mehrere Testdurchläufe aggregiert und statistische Hypothesentests verwendet.

16 Bouthillier, X., et al. 2021. Accounting for Variance in Machine Learning Benchmarks. https://arxiv.org/abs/2103.03098.

6.2. **ScrutinAI** (Visual-Analytics-Tool)

Jedes Machine-Learning-Modell macht Fehler. Es ist daher wichtig, dass der Mensch diese findet, versteht und behebt. Hierbei hilft das Visual-Analytics-Framework ScrutinAI, mit dem Expert*innen KI-Modelle möglichst effizient semantisch analysieren können. Eingesetzt werden kann dieses Tool vorrangig überall dort, wo Bilddaten durch KI verarbeitet werden, also z. B. im medizinischen oder industriellen Kontext. Aber auch ein Einsatz bei der Verarbeitung von Text- oder Audiodaten ist möglich.

Die Güte und Qualität einer KI-Anwendung wird üblicherweise automatisch ausgewertet und in gemittelten Qualitätskennzahlen angegeben. Aufgrund der Komplexität von KI-Anwendungen bleiben aber die Gründe, warum ein KI-Modell eine bestimmte Performanz-Kennzahl aufweist, in einer intransparenten Blackbox verborgen. Um die hochdimensionalen Zusammenhänge aufschlüsseln zu können und (Fehler-) Ursachen aufzudecken, bedarf es daher anderer Techniken.

Um Schwachstellen entdecken und mitigieren zu können, ist es notwendig, das semantische Verständnis und Expert*innen- bzw. Domänenwissen des Menschen in die Analyseprozesse einzubinden (siehe Abbildung 3). Denn damit ist der Mensch in der Lage, Muster in Daten zu entdecken und eine kontextabhängige Bedeutung abzuleiten.

Unser Tool ScrutinAI[17] (von engl. scrutinize = prüfend ansehen, untersuchen) arbeitet daher mit Visual-Analytics-Methoden, mit denen die üblicherweise große Menge an Daten, Metriken, Methoden, Features und KPIs in eine für Menschen erfassbare Form gebracht werden (siehe Abbildung 4). So wird die Durchsicht und Analyse der nötigen Informationen effizient und systematisch ermöglicht

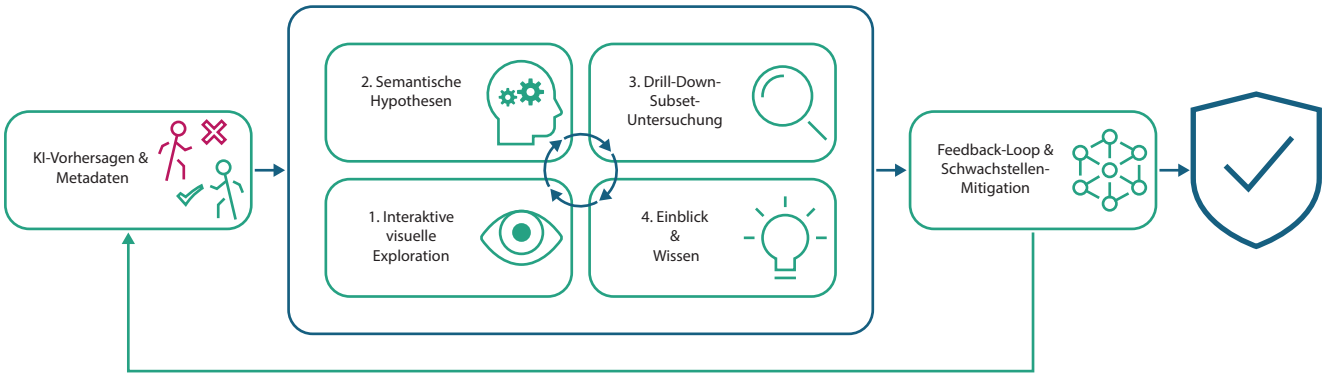


*Abbildung 3: Workflow von ScrutinAI: Basierend auf den gegebenen KI-Vorhersagen und den Metadaten beginnt der/die Analyst*in mit der KI-Überprüfung, indem (1.) die Daten visuell und interaktiv exploriert werden. Diese Untersuchung führt (2.) zu semantischen Hypothesen der interessanten Daten-Subsets, die (3.) nachfolgend weiter in einer Drill-Down-Analyse untersucht werden. (4.) Die daraus generierten Einblicke und das Wissen dienen als Basis, um einen neuen Analysezyklus zu starten, und/oder werden als Feedback an die weiteren Stakeholder kommuniziert, um das KI-Modell zu verbessern. Abbildung: Fraunhofer IAIS*

17 Haedecke, E., et al. 2022. ScrutinAI: A Visual Analytics Approach for the Semantic Analysis of Deep Neural Network Predictions. In: Bernard, J., et al. (eds). EuroVis Workshop on Visual Analytics (EuroVA). The Eurographics Association. ISBN 978-3-03868-183-0; doi:10.2312/eurova.20221071.

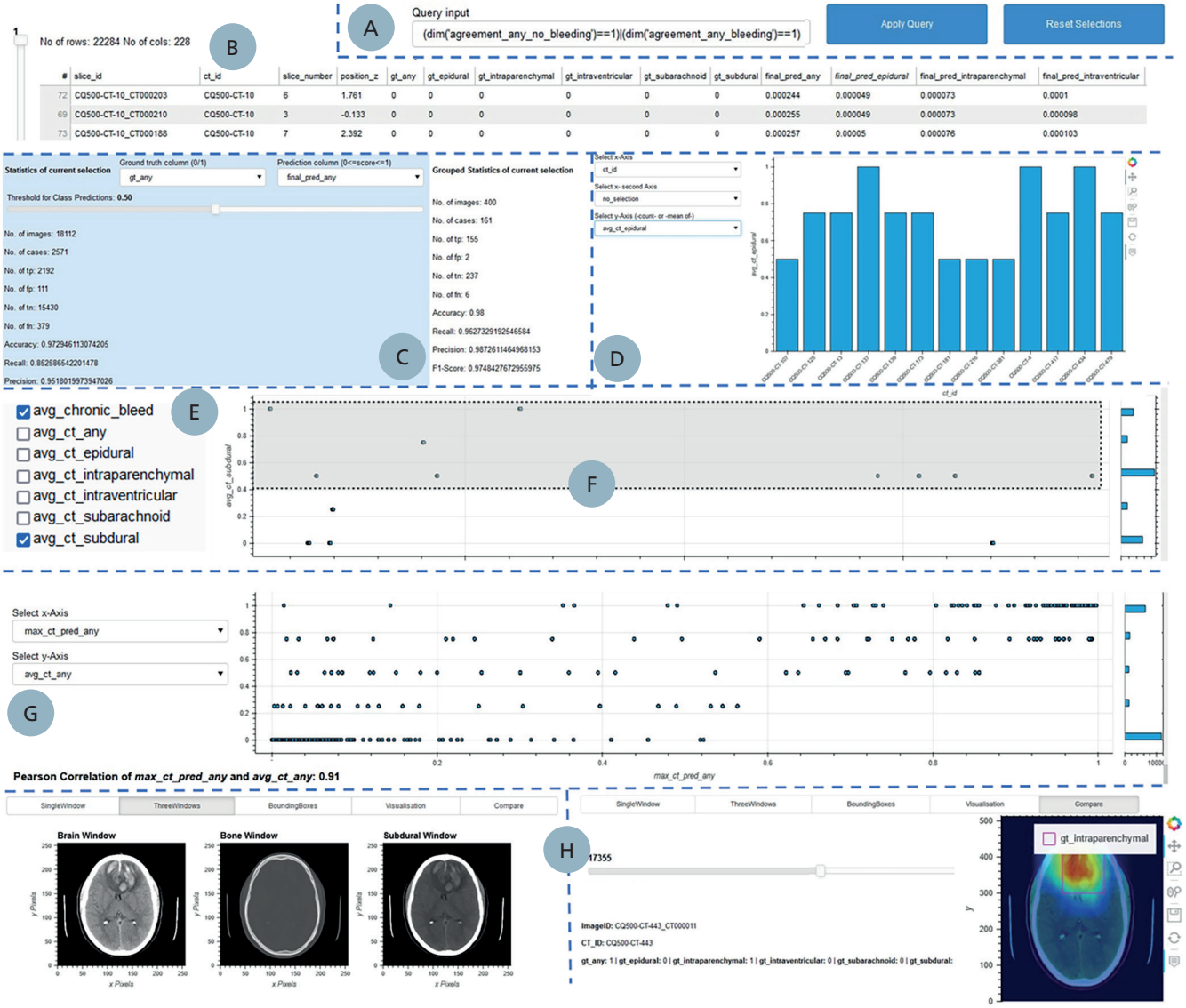


Abbildung 4: Screenshot der Analyse eines DNNs für einen medizinischen Use Case[18] innerhalb des Tools ScrutinAI. (A): Textuelle Abfragen (angelehnt an Datenbank-Abfragen). (B): Tabellarische Darstellung der strukturierten Daten und Metadaten. (C): Gegenüberstellung verschiedener Performanz-Statistiken. (D): Histogramm über kategorische Attribute. (E): Auswahl der anzuzeigenden Scatter-Plots für die einzelnen Attribute. (F): Scatter-Plot mit Histogramm zur Visualisierung der Muster der Datenpunkte für die einzelnen Attribute entlang der Bildsequenz. (G): Auswahl der Daten für die X- und Y-Achse des Korrelations-Plots (links) sowie Korrelationsplot und Histogramm zur Visualisierung der Abhängigkeiten und Häufigkeiten (rechts). (H): Verschiedene Darstellungen der CT-Bilder (z. B. Anzeige des ausgewählten Einzelbildes, zur Gegenüberstellung oder überlagerten Visualisierung; auswählbar über die obige Menüleiste). Abbildung: Fraunhofer IAIS

18 Görge, R., et al. 2023. Using ScrutinAI for Visual Inspection of DNN Performance in a Medical Use Case. AAAI Spring Symposium. AITA: AI Trustworthiness Assessment.

6.3. Semantisches Testen

Systematische Schwachstellen der KI müssen frühzeitig identifiziert werden. Hier hilft das Analysieren von Daten-Teilregionen im Sinne einer »semantischen Dimension«, die durch leicht verständliche Parameter beschrieben wird. Durch sie können detaillierte und besser interpretierbare Testdatensätze generiert werden, beispielsweise Bilddaten für das autonome Fahren.

Im Allgemeinen wird bei überwachten Lernverfahren ein Testdatensatz genutzt, um die Performanz des KI-Modells zu überprüfen. Hierzu wird eine Metrik verwendet, die über den gesamten Testdatensatz aggregiert wird. Eine einzelne Kennzahl gibt jedoch keine Auskunft darüber, wann oder wie ein KI-Modell versagen könnte. Hinzu kommt die Schwierigkeit zu bewerten, ob der Datensatz alle Aspekte des zugrundeliegenden, realen Problems wirklich ausreichend abdeckt. Im Falle einer Diskrepanz zwischen Daten und realem Problem können die Ergebnisse der Tests nicht die reale Leistung des Modells widerspiegeln. Daher muss analysiert werden, ob es Teilregionen im Datenraum gibt, für die das Modell eine schlechtere Leistung zeigt, womit potenzielle Schwachstellen identifiziert werden können.

Da die korrespondierende Prüfung schlussendlich von einem/r menschlichen Prüfer*in abgenommen werden muss, ist es wichtig, dass diese Teilregionen durch leicht verständliche Parameter beschrieben werden. Statt komplexer mathematischer Beschreibungen einer Teilregion im Datenraum nutzt man also eine menschenverständliche Beschreibung, weshalb diese Parameter semantische Dimensionen genannt werden.

Im Beispiel des autonomen Fahrens könnte eine semantische Dimension z. B. die Farbe eines Autos sein: Während die meisten Autos vom Modell gut erkannt werden, könnte sich bei der Analyse dann für eine bestimmte Farbe wiederholt eine schlechte Performanz zeigen (siehe Abbildung 5). Gerade solche systematischen Schwachstellen, die für bestimmte semantische Dimensionen existieren können, stellen ein Risiko dar, denn es ist zu erwarten, dass Modelle in semantisch ähnlichen Situationen ebenfalls eine schlechte Performanz aufweisen. Daher ist es wichtig, diese Schwachstellen zu identifizieren, um das zugrunde liegende Problem zu beheben.

Um granulare Tests für verschiedene semantische Dimensionen durchführen zu können, werden Datensätze benötigt, die entsprechende semantische Informationen beinhalten – in obigem Beispiel also die Information, welche Farbe die Autos auf den Bildern des Datensatzes aufweisen. Da Trainingsdatensätze nicht immer sämtliche semantischen Informationen (beispielsweise in Form spezifischer Metadaten) enthalten, werden für semantische Tests[19] aktuell häufig synthetisch erzeugte Datensätze verwendet, für welche umfassende semantische Informationen erzeugt werden können. Hierdurch lassen sich spezielle Testsets erstellen, die beispielsweise Variationen entlang einer semantischen Dimension beinhalten. Dadurch können Schwachstellen im Modell aufgedeckt werden, die gleichzeitig durch die Rückführung auf die semantischen Eigenschaften interpretierbarer und detaillierter sind als rein aggregierte Performanz-Metriken.

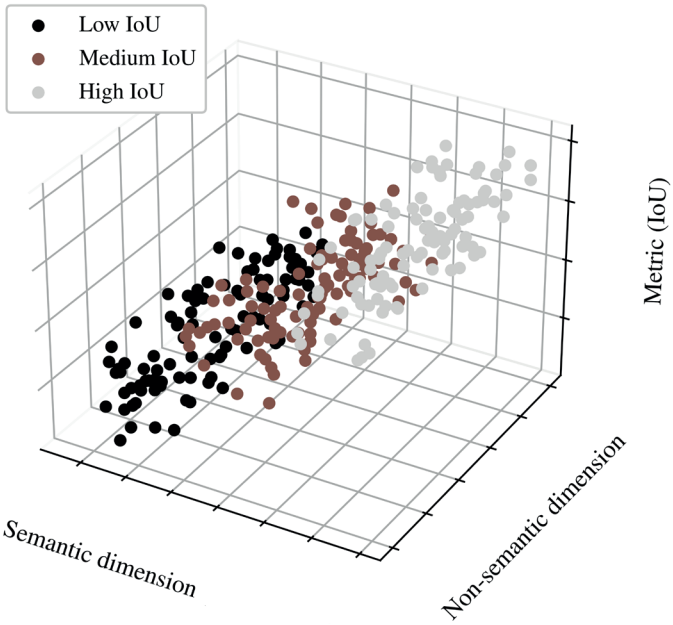


Abbildung 5: Visuelle Darstellung der semantischen Dimension mit drei Ausprägungen (hier optisch durch die verschiedenen Farben voneinander getrennt dargestellt) und der Performanz des Modells (hier gemessen anhand der Metrik »Intersection over Union« (IoU)) für diese Unterregionen des Datenraums. Abbildung: Fraunhofer IAIS

19 Gannamaneni, S., et al. 2021. Semantic Concept Testing in Autonomous Driving by Extraction of Object-Level Annotations from CARLA. Proceedings of the IEEE/CVF International Conference on Computer Vision. https://openaccess.thecvf.com/content/ICCV2021W/ERCVAD/html/Gannamaneni_Semantic_Concept_Testing_in_Autonomous_Driving_by_Extraction_of_Object-Level_ICCVW_2021_paper.html, letzter Aufruf am 29.03.2023).

6.4. **Fuzzy Testing** (Fuzzing-Tool)

Um dem dynamischen Entwicklungsumfeld von KI-Systemen gerecht zu werden, deckt das Fuzzy Testing einen deutlich größeren Testraum ab als bisher etablierte Prüfwerkzeuge. In allen Bereichen, in denen mithilfe von Testdaten die Performanz eines KI-Modells überprüft werden soll, ist der Einsatz von Fuzzy Testing von Vorteil.

Im Vergleich zu klassischer Software ist die Prüfung von KI-Anwendungen mit einer Reihe besonderer Herausforderungen verbunden, die von etablierten Testverfahren bislang noch nicht oder nicht ausreichend adressiert werden. Eine formale Verifizierung der Funktionsfähigkeit eines KI-Moduls ist in Fällen, in denen der Raum möglicher Eingaben praktisch unendlich ist, nicht durchführbar. Auch mit üblichen, datengetriebenen Standardverfahren, beispielsweise bei überwachten Lernverfahren, wie sie etwa im Finanz-, Medizin- oder Industriesektor eingesetzt werden, lässt sich die Funktionalität des KI-Modells nur sehr lückenhaft überprüfen. Insbesondere wird nicht erkennbar, ob im Eingangsdatensatz wichtige und kritische Beispiele fehlen oder unterrepräsentiert sind. Das könnte im realen Einsatz dann dazu führen, dass das Modell für solche Situationen, die es nicht oder kaum während des Trainings gesehen hat, eine schlechte Performanz zeigt oder grundsätzlich falsche Vorhersagen trifft.

Fuzzy Testing basiert auf einer evolutionären Optimierungsstrategie. Ausgehend von einer Reihe vorausgewählter Testeingaben werden mittels anwendungsspezifischer Transformationen zufällig neue Eingaben erzeugt und dem KI-Modell vorgelegt, siehe schematische Darstellung des Ablaufs in Abbildung 6. Eine solche Zufallsmutation ist genau dann im Sinne des Fuzzy-Testing-Algorithmus »interessant«, wenn mindestens eines der folgenden Kriterien erfüllt ist:

- **Schlechte(re) Vorhersagen:** Die Vorhersage des Modells ist entweder fehlerhaft oder ihre Qualität ungenügend.
- **Erhöhte Suchraumabdeckung:** Ein geeignetes Maß für die Abdeckung des Zustandraums des Modells wird durch die Zufallsmutation signifikant erhöht.

Für ein neuronales Netz wäre das zweite Kriterium etwa dann gegeben, wenn durch das neue Beispiel Neuronen stark aktiviert werden, die durch alle anderen Tests nicht oder nur schwach aktiviert wurden. Hat man für die Suchraumabdeckung eine geeignete Metrik gewählt, beispielsweise die »Neuron Coverage«[20], kann ein quantitativer Indikator dafür geliefert werden, wann eine KI-Prüfung hinreichend »vollständig« ist.

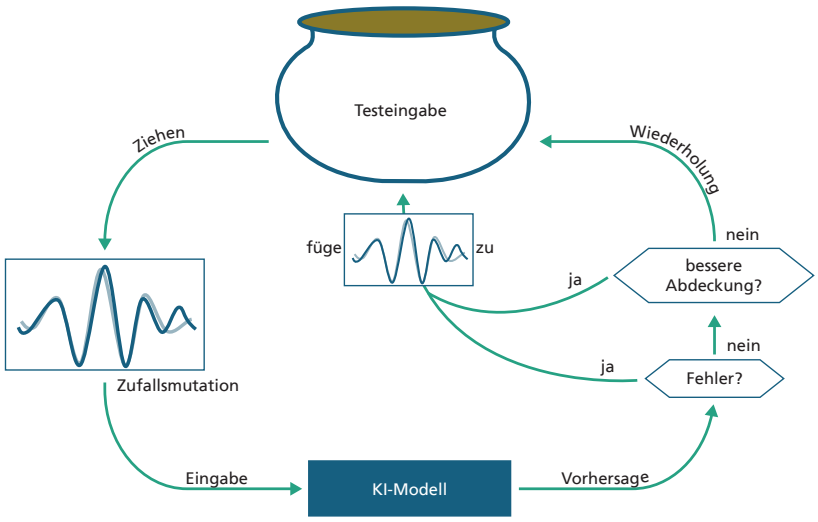


Abbildung 6: Beim Durchlauf durch die Fuzzing-Testschleife wird eine Testeingabe aus dem Datensatz gezogen und mittels anwendungsspezifischer Zufallsmutation verändert. Wenn die Vorhersage des zu testenden KI-Modells fehlerhaft ist oder die Testeingabe ein geeignetes Abdeckungsmaß signifikant erhöht, wird die Zufallsmutation der Sammlung an Testdaten hinzugefügt. Dieser Vorgang wird solange wiederholt, bis das Testbudget erschöpft oder ein kritischer Fehler aufgetreten ist. Abbildung: Fraunhofer IAIS

20 Abrecht, S. et al. 2020. Revisiting Neuron Coverage and Its Application to Test Generation. In: Casimiro, A., et al. (eds) Computer Safety, Reliability, and Security. SAFECOMP 2020 Workshops. SAFECOMP 2020. Lecture Notes in Computer Science, vol 12235. Springer, Cham. https://doi.org/10.1007/978-3-030-55583-2_21.

6.5. Unsicherheitsbewertung

Insbesondere in sicherheitskritischen Anwendungen wie dem autonomen Fahren müssen KI-Modelle zuverlässig arbeiten. Daher ist es wichtig, Unsicherheiten in den Vorhersagen der Modelle einschätzen und mitigieren zu können.

Modelle des Maschinellen Lernens, insbesondere tiefe neuronale Netze, erzielen in vielen Fällen Ergebnisse, die mit dem menschlichen Niveau vergleichbar sind oder es sogar übertreffen. Dennoch machen sie auch Fehler, die sich manchmal nicht vermeiden lassen – wenn beispielsweise die gegebenen Daten keine perfekte Vorhersage zulassen.

In vielen Fällen sind derartige Fehler unkritisch, vor allem dann, wenn diese leicht erkannt werden können und sich der Vorgang beliebig wiederholen lässt. Beispielsweise können für eine Spracherkennung und -steuerung nicht verstandene Anfragen erneut gestellt werden. Andererseits existieren kritische Anwendungsfälle, bei denen diese Voraussetzungen nicht, oder nur bedingt, gegeben sind, wie in den Abbildungen 7 und 8 am Beispiel des autonomen Fahrens dargestellt.[21] Hier können fehlerhafte Vorhersagen der KI-Anwendung zu Personenschäden oder finanziellen Verlusten (beispielsweise Sachschäden, aber auch Opportunitätskosten) führen.

Um die Risiken in diesen Fällen abzuschwächen, können mithilfe einer gut kalibrierten Schätzung der Ausgabesicherheit, wie beispielsweise mit der am Fraunhofer IAIS entwickelten Methode »Wasserstein Dropout«[22], zwei Strategien verfolgt werden: Erstens können die auffallenden problematischen Ausgaben erkannt werden, sodass Fehler vermieden und die Zuverlässigkeit erhöht wird. Zweitens macht das Erkennen problematischer Ausgaben den menschlichen Eingriff oder anderweitige Mitigationsstrategien erst möglich. Wenn so die verschiedenen Modellausgaben mit einer quantifizierten Aussage über deren Unsicherheit angereichert werden, kann dies als wichtiger Baustein für eine Absicherungsargumentation genutzt werden.

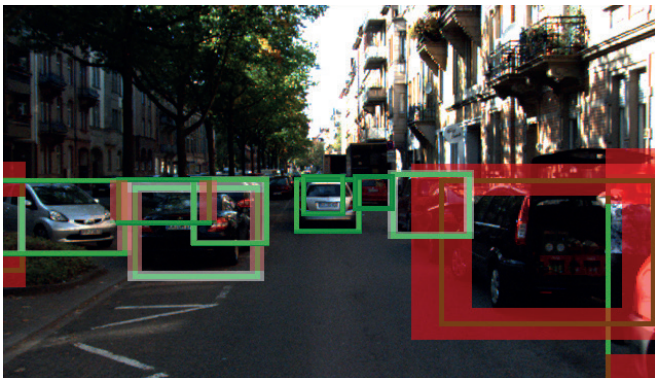


Abbildung 7: Bei der Unsicherheitsschätzung beim autonomen Fahren werden Kamerabilder analysiert. Den gefundenen Objekten wird eine Unsicherheit zugeordnet, die über die Dicke und Farbe des Rahmens um das Objekt dargestellt wird. So wird das Objekt rechts mit dickem rotem Rahmen als sehr unsicher eingestuft. Abbildung: Fraunhofer IAIS

Abbildung 8: Bei diesem Beispiel detektiert das System zunächst vier Objekte, wovon nur zwei korrekt sind und demnach als True Positive (TP) bezeichnet werden. Die beiden Falsch-Vorhersagen, also False Positives (FP), lassen sich jedoch herausfiltern, da sie mit einer hohen Unsicherheitsschätzung einhergehen. Abbildung: Fraunhofer IAIS

21 Für die Abbildungen 7 und 8 wurden Fotos aus dem KITTI-Datensatz genutzt: Geiger, A., et al. 2012. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In: Conference on Computer Vision and Pattern Recognition (CVPR).

22 Siehe als Beispiel für kalibrierte Schätzverfahren: Sicking, J., et al. 2022. Wasserstein dropout. Machine Learning, 1–44.

6.6. RuleCreator

Mit dem RuleCreator lassen sich Muster in Daten finden und mit leicht interpretierbaren Regeln beschreiben. So lassen sich Regel-Sets je nach Anwendung optimieren. Der RuleCreater kann in ganz unterschiedlichen Branchen zum Einsatz kommen, wie etwa in der Produktion oder dem Finanzsektor.

Der RuleCreator ist ein Suchwerkzeug, das automatisierte Datenanalysen auf tabellarischen Daten durchführt und statistisch relevante Zusammenhänge findet, die sich als vom Menschen interpretierbare »Wenn-Dann-Regeln« darstellen lassen. Der »Wenn«-Teil beschreibt die Merkmalsbedingungen, die (gleichzeitig) erfüllt sein müssen, damit eine Regel gilt (z. B. »Geschlecht ist männlich UND Akzent ist Friesisch«). Der »Dann«-Teil einer Regel ist dabei durch die Suchanfrage vorgegeben und entspricht im Fall von KI-Prüfungen einem Fehlerzustand, der genauer untersucht werden soll (z. B. »KI-System versteht gesprochenen Satz nicht«). Durch die datengetriebene Mustererkennung lassen sich so alle relevanten Subgruppen aufdecken. Die Vielzahl an generierten Regeln ermöglicht es den Nutzer*innen, die gefundenen Fehlerbedingungen zu interpretieren und entsprechende Schlussfolgerungen zu ziehen.

Das Lernen der Regeln funktioniert dabei datenbasiert in voneinander losgelösten Prozessschritten, die integriertes Expert*innenwissen abbilden und je nach Einsatzszenario individuell konfiguriert werden können. Zudem kann die statistische Qualität einer jeden Regel durch eine Nutzerspezifikation beeinflusst werden, indem Angaben zur minimalen Erkennungsrate (der minimalen Anzahl an Beispielen, die erklärt werden müssen) und zur Falsch-Positiv-Rate (der Anzahl an erlaubten Fehlalarmen) gemacht werden. Beispielsweise kann auch festgelegt werden, dass Redundanzen in den Regeln vermieden werden. Diese Möglichkeiten erlauben es, KI-Prüfungen bzgl. Suchtiefe und Suchaufwand (Laufzeit, Rechnerressourcen) auf die Anforderungen einer Prüfkampagne zuzuschneiden und dabei vorhandenes Vorwissen der Prüfenden zu integrieren.

6.7. Beispiel für die Kombination von KI-Prüfwerkzeugen

Das folgende Beispiel erläutert, wie im Rahmen der Prüfplattform unterschiedliche Tools kombiniert werden können, um komplexere Fragestellungen der Qualitätssicherung und -bewertung von KI-Systemen zu adressieren. Beispielhaft wird die Dimension der Transparenz gewählt und auf die Erklärbarkeit der KI eingegangen.

Die Intransparenz einer Blackbox ist eine Hürde, die es erschwert, Entscheidungen für verschiedene Business Cases oder Use Cases zu treffen. In diesem Fall bietet es sich besonders an, verschiedene Transparenzmethoden zu kombinieren, um die unterschiedlichen Bereiche der Undurchsichtigkeit des KI-Modells mit den entsprechenden Methoden für den Menschen verständlich und interpretierbar aufzubereiten.

Hierzu können die Prüfwerkzeuge Fuzzy Testing, ScrutinAI und RuleCreator kombiniert werden, indem sie am Eingang bzw. am Ausgang des KI-Modells eingesetzt werden. So kann basierend auf den Trainings- und Testdaten mithilfe des Fuzzy Testings ein Datenset für interessante Datenpunkte erstellt werden, die dem KI-Modell als Input präsentiert werden. Zur weiteren Untersuchung wird anschließend innerhalb von ScrutinAI die entsprechende semantische Dimension auf auffällige, visuelle Muster hin überprüft, indem hierfür angepasste Abfragen und Filterungen der Daten durchgeführt werden. Als weitere Komponente setzt der RuleCreator ebenfalls bei den Eingangsdaten an. Er findet logische Muster und stellt diese als interpretierbare Regeln dar. Die so extrahierten Regeln ermöglichen es, ein intrinsisch interpretierbares Ersatz-Modell, das »Whitebox-Surrogate-Modell«, zu trainieren, welches als Gegenstück zum Blackbox-Modell fungiert. Folglich liefert das Whitebox-Modell interpretierbare Ergebnisse, die mit den Ergebnissen des Blackbox-Modells abgeglichen werden können. Wie in der Abbildung 9 dargestellt, kann so der Kreis zur Herstellung der Transparenz geschlossen werden und das Verständnis des Modells und der Daten die Grundlage für begründete Entscheidungen liefern.

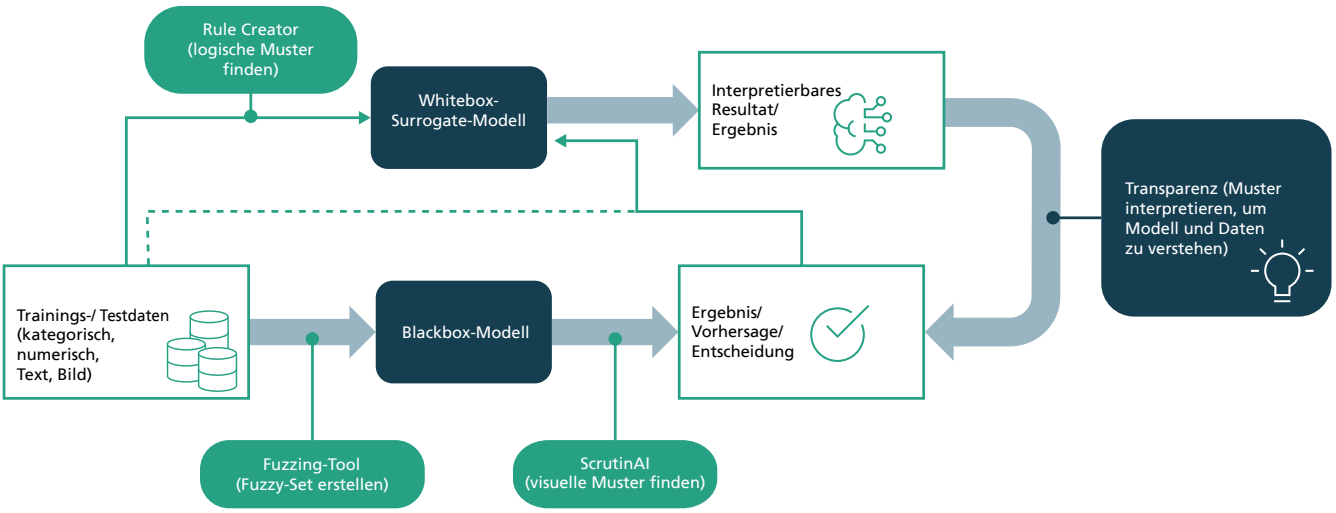


Abbildung 9: Ein Zusammenspiel aus verschiedenen Transparenzmethoden bietet den Vorteil, eine umfassendere Interpretierbarkeit von Modell und Daten zu realisieren. Abbildung: Fraunhofer IAIS

7. Fazit und Handlungsempfehlungen

Mit dem vorgestellten Prüf-Framework sowie den beispielhaften Prüftools werden Lösungen für die systematische Überprüfung von KI-Systemen in den verschiedenen Dimensionen der Vertrauenswürdigkeit dargestellt. Das Framework und die Tools schaffen einen einfachen Zugang auf verschiedenen Software-Ebenen. Die so ermöglichten nachvollziehbaren und reproduzierbaren Tests bilden die Grundlage der systematischen Qualitätssicherung bei der Entwicklung, Abnahme und Prüfung komplexer, datenintensiver KI-Systeme. Die Prüfwerkzeuge können durch ihren nutzerfreundlichen Zugang auch von Unternehmen in eigenen Entwicklungen genutzt werden. Zudem erlaubt das Konzept der Prüfplattform auch die Integration von Prüfwerkzeugen von Drittanbietern.

Das vorliegende Whitepaper zeigt wichtige technische Ansätze zur Operationalisierung von KI-Prüfungen. Es legt insbesondere dar, wie einerseits ein höherer Grad an Automatisierung und andererseits höhere Prüftiefen durch den Einsatz von Prüfwerkzeugen und ihre Integration in eine Prüfplattform realisiert werden können. Um das volle Potenzial dieses Ansatzes entfalten zu können, werden folgende Handlungsempfehlungen formuliert:

Handlungsempfehlungen an die Politik, Regulierung und Standardisierung

1. KI-Prüfkriterien und Prüfanforderungen müssen systematisch erarbeitet und in Standards festgehalten werden.
2. Eine Prüfplattform, wie sie hier vorgestellt wird, ermöglicht systematische KI-Prüfungen durch die Integration unterschiedlicher Werkzeuge. Hierzu müssen Standards für die Interoperabilität entsprechender Prüfwerkzeuge definiert werden.
3. Da KI-Prüfungen die Grundlage zur Inverkehrbringung von Hochrisikoanwendungen sind, sollten Zulassungskriterien für Prüfwerkzeuge definiert werden.

Handlungsempfehlungen an Unternehmen

1. Unternehmen müssen sich systematisch mit KI-Risiken und KI-Governance auseinandersetzen, insbesondere im Hinblick auf die Anforderungen des AI Acts.
2. Viele Anforderungen durch die erwartete KI-Regulierung müssen noch operationalisiert und entsprechende organisationsinterne Strukturen zur Umsetzung geschaffen werden. Durch die Durchführung von Pilotprüfungen an ausgewählten Use Cases lassen sich wertvolle Erfahrungen und Best Practices sammeln.
3. Auch wenn KI-Prüfungen nicht notwendigerweise durch eine dritte Partei vorgenommen werden müssen, sollten die Ergebnisse von selbst durchgeführten Tests und Prüfungen dennoch verlässlich, reproduzierbar und revisionssicher sein. Dies setzt eine robuste Machine-Learning-Operations-Infrastruktur voraus, welche eine entsprechende Testinfrastruktur bereitstellen muss.

Impressum

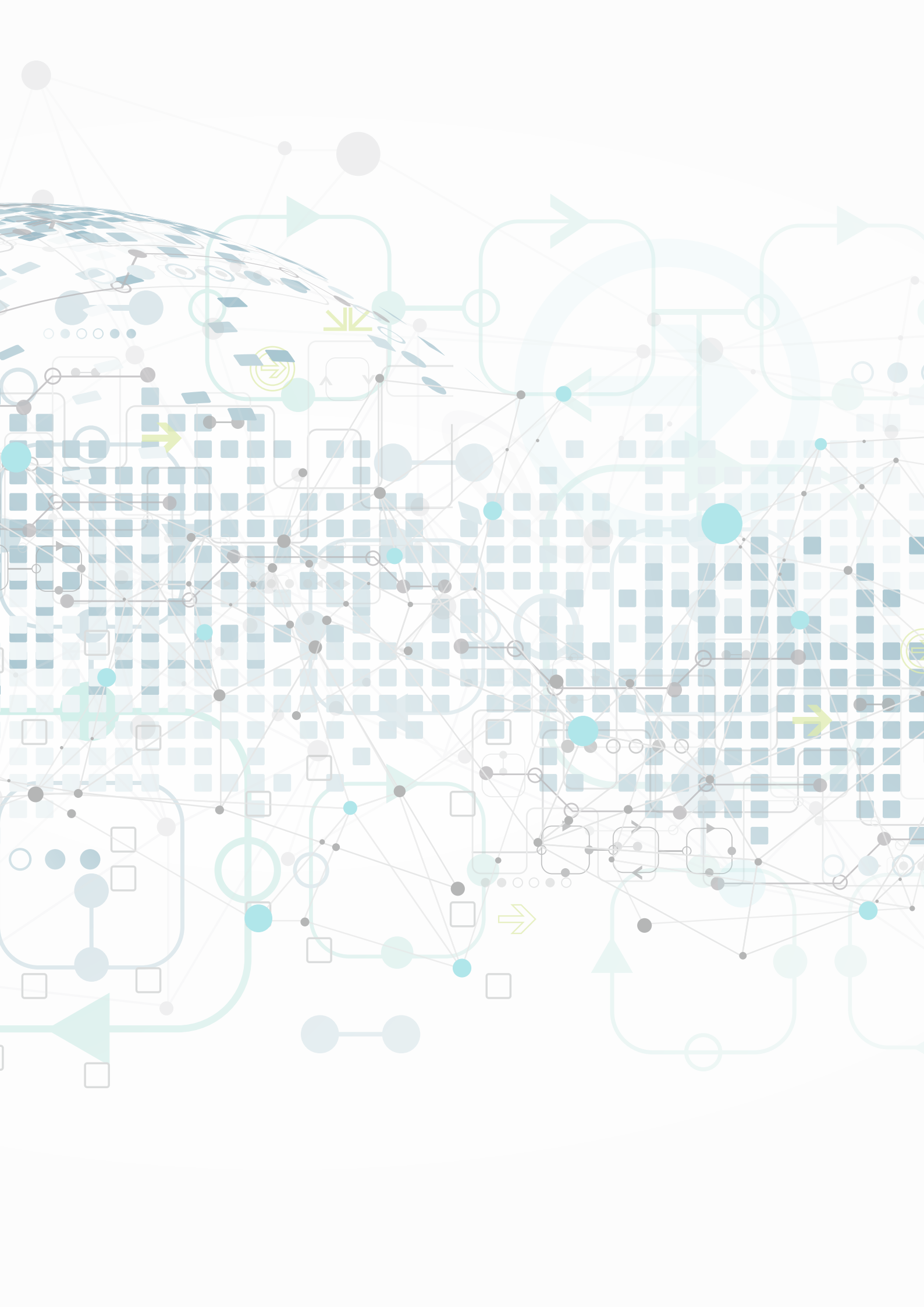
Herausgeber
Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS
Schloss Birlinghoven
53757 Sankt Augustin

Redaktion
Claudia Könsgen
Melissa Nordmann
Silke Loh

Grafik und Layout
Achim Kapusta
Asra-Soraya Neumeister

Bildquellen
Titelbild: Alex – stock.adobe.com

Stand
Mai 2023



Kontakt

Fraunhofer-Institut für Intelligente
Analyse- und Informationssysteme IAIS
Schloss Birlinghoven
53757 Sankt Augustin

www.iais.fraunhofer.de

Ansprechpartnerin:
Elena Gina Haedecke
elena.haedecke@iais.fraunhofer.de