September 2021 · Leonie Beining

KI in der Industrie absichern & prüfen

Was leisten Assurance Cases?

Stiftung Neue Verantwortung

Think Tank für die Gesellschaft im technologischen Wandel

Inhalt

1. Einführung: KI in der Industrie – Zwischen Potenzial, Risiko und Regulierung

Künstliche Intelligenz (KI) gilt wahlweise als „Megatrend“ oder „Schlüsseltechnologie“ für die Industrie:[1] Dem Einsatz von KI-Methoden wird das Potenzial zugesprochen, den Industriesektor flexibler, effizienter, nachhaltiger, innovativer, insgesamt wettbewerbsfähiger zu gestalten. Eine im Auftrag des Bundesministeriums für Wirtschaft und Energie (BMWi) durchgeführte Studie prognostiziert für Deutschland bis 2023 eine zusätzliche Bruttowertschöpfung um 31,8 Mrd. Euro im produzierenden Gewerbe.[2] Diese wirtschaftliche Bedeutung basiert auf den vielfältigen Einsatzmöglichkeiten von KI in der Industrie: von der Fertigung und Wartung bis hin zur Optimierung von Geschäfts- und Logistikprozessen.[3]

Als eine zentrale Anwendungsmöglichkeit von KI in der industriellen Fertigung gilt die Unterstützung der eingesetzten Maschinen. Zum Beispiel können KI-Methoden die Funktionalitäten von Maschinen ergänzen und erweitern und haben damit das Potenzial, die klassische Produktionsautomatisierung flexibler werden zu lassen.[4] So kann etwa Sprach- oder Objekterkennung dazu genutzt werden, die Interaktion zwischen Menschen und einem kollaborativem Roboter zu verbessern oder einem fahrerlosen Transportfahrzeug neue Fähigkeiten zu vermitteln, wie etwa das selbstständige Umfahren von Hindernissen.[5] Mithilfe von KI-Methoden können aber auch Sicherheitsfunktionen von Maschinen optimiert werden, da datengetriebene Modelle die Leistungsfähigkeit klassischer Softwarelösungen in bestimmten Bereichen wie Objekterkennung übertreffen können.[6] Eine auf KI-basierende Kollisionsvermeidung könnte beispielsweise die Performanz automatisierter Fahrsysteme verbessern, die so flexibler und breiter eingesetzt werden könnten. Darüber hinaus kann KI auch genutzt werden, um neue Sicherheitsfunktionen zu entwickeln, wie etwa automatisierte Warnungen vor Fehlgebrauch.

1 Dem ExamAI-Projekt liegt die KI-Definition nach Marvin Minsky (1968) zugrunde: „Artificial Intelligence is the science of making machines do things that would require intelligence if done by man.“ Ergänzend wird von KI-Methoden und KI-Komponenten bzw. KI-System gesprochen. Der Fokus liegt auf KI-Methoden, die traditionell dem Forschungsfeld des Maschinellen Lernens zuzuordnen sind. KI-Komponenten sind Softwarekomponenten, die auf KI-Methoden basieren. KI-Systeme bestehen aus mindestens einer KI-Komponente, können aber beliebig viele weitere Komponenten umfassen.

2 iit-Institut für Innovation und Technik in der VDI und VDE Innovation + Technik GmbH (2018): Potenziale künstlicher Intelligenz im produzierenden Gewerbe in Deutschland. Abgerufen am 24.08.2021 von https://www.bmwi.de/Redaktion/DE/Publikationen/Studien/potenziale-kuenstlichen-intelligenz-im-produzierenden-gewerbe-in-deutschland.pdf?__blob=publicationFile&v=17.

3 BDI (2021): Anwendungsfelder und Potenziale von KI in der Wirtschaft. Angerufen am 24.08.2021 von https://bdi.eu/artikel/news/anwendungsfelder-und-potenziale-von-ki-in-der-wirtschaft/.

4 Gesellschaft für Informatik (2020): Ergebnisbericht. Möglichkeiten und Grenzen von Anwendungen künstlicher Intelligenz in der Produktautomatisierung. Abgerufen am 24.6.2020 von https://testing-ai.gi.de/fileadmin/PR/Testing-AI/ExamAI_Ergebnisbericht_Anwendungsszenarien_KI_Industrie.pdf.

5 ebd.

6 Michael Kläs, Rasmus Adler, Ioannis Sorokos, Lisa Joeckel und Jan Reich (2021a): Handling Uncertainties of Data-Driven Models in Compliance with Safety Constraints for Autonomous Behaviour. In: European Dependable Computing Conference (EDCC).

Diese Beispiele zeigen: Für die industrielle Fertigung stellt KI ein mächtiges Werkzeug dar, das vielfältiges Potenzial birgt. Dem möglichen Gewinn an Leistungsfähigkeit, Flexibilität und Sicherheit stehen allerdings auch große Risiken gegenüber. Trotz der Fortschritte, die in den letzten Jahren auf den Gebieten Maschinelles Lernen oder Robotik erzielt wurden, sind KI-Systeme nach wie vor anfällig für Fehler oder unvorhergesehene Ausfälle. In entscheidenden Bereichen, zum Beispiel wenn KI-Komponenten zur Umsetzung von sicherheitskritischen Funktionen von Industriemaschinen verwendet werden, haben Fehler und Ausfälle gleich massive negative Auswirkungen, führen in einer Industriehalle etwa zu Unfällen mit Sachschäden oder stellen gar eine Gefahr für Leib und Leben dar.

Wenn die Industrie also von dem Einsatz von KI-Methoden profitieren will und Hersteller ihre Produkte erfolgreich vermarkten wollen, müssen sich alle Stakeholder auf die Qualität und insbesondere die funktionale Sicherheit[7] der KI-Komponenten verlassen können. Entsprechend greifen auch die Vorschläge der Europäischen Kommission für einen horizontalen Rechtsakt für KI und für eine EU-Maschinenverordnung solche sicherheitskritischen KI-Anwendungen auf und stufen sie als Systeme mit hohem Risiko ein. Würden die Verordnungen in aktueller Form in Kraft treten, müssten Maschinen, die KI verwenden, um Sicherheitsfunktionen zu realisieren, demnach einer Konformitätsbewertung durch eine externe Stelle unterzogen werden, bevor sie auf den Markt gebracht werden können.[8]

Dafür muss jedoch die Frage beantwortet werden, wie die Sicherheit von KI-Komponenten überhaupt dargelegt und überprüft werden kann. Gegenwärtig stehen Herstellerfirmen, Prüforganisationen, Marktüberwachungsbehörden und insgesamt die KI-Regulierung vor einer großen Herausforderung. Denn bislang fehlt es angesichts der noch jungen Marktreife vieler KI-Anwendungen an etablierten Ansätzen und Methoden, um nachzuweisen, dass sich ein KI-System auch ausreichend sicher verhält (s. Abschnitt 2).

Dies ist in mehrfacher Hinsicht ein Problem: Solange es keine etablierten Verfahren und entsprechenden Rahmenbedingungen gibt, um sicherheitskritische KI rechtssicher an den Markt zu bringen, wird sie auch nicht breit eingesetzt werden. Das heißt zum einen, dass sich das ökonomische Potenzial von KI und ihr Nutzen für die Mensch-Maschine-Kollaboration erst entfalten können, wenn entsprechende

7 Der Fokus des Papiers liegt auf Sicherheit im Sinne von Safety, d.h. funktionaler Sicherheit, Produktsicherheit etc.. Die Begriffe Safety und Sicherheit werden im Text synonym verwendet.

8 Aufgrund verschiedener Merkmale von KI und angesichts der fehlenden Erfahrungen, die zurzeit im Umgang mit sicherheitsrelevanten KI-Anwendungen bestehen, erachtet die Kommission das Risiko, das von Maschinen mit KI-basierten Sicherheitsfunktionen ausgeht, als besonders hoch. Die Maschinenverordnung sieht entsprechend vor, dass die KI-Komponente von einer notifizierten Stelle geprüft wird, und die Möglichkeit einer reinen internen Fertigungskontrolle (zunächst) entfällt (siehe European Commission (2021) Proposal for a Regulation of the European Parliament and of the Council on machinery products, Erwägungsgrund 45).

Methoden zur Absicherung und Prüfung zur Verfügung stehen.[9] Zum anderen kann das bedeuten, dass auch weiterhin notwendiges Erfahrungswissen und Beobachtungen von KI im praktischen Einsatz fehlen werden. Diese werden aber benötigt, um entsprechende Anforderungen, zum Beispiel im Rahmen der Normung, überhaupt formulieren zu können. Bei der Erarbeitung etablierter Safety-Standards konnte beispielsweise auf jahrelange Erfahrung bei der sicheren Konstruktion und Anwendung von Maschinen zurückgegriffen werden.

Es bedarf also überbrückender Methoden und Ansätze, die Anbieterunternehmen, Prüforganisationen und der Marktüberwachung dennoch eine Sicherheitsbewertung ermöglichen und die Rahmenbedingungen schaffen, das notwendige Wissen zu generieren, das es für die Entwicklung passender Normen benötigt. Eine solche Brücke könnten so genannte Assurance Cases (s. Abschnitt 3) schlagen. Ziel des Papiers ist es, den Ansatz der Assurance Cases sowie seine Potenziale (s. Abschnitt 3.1.) und Grenzen (s. Abschnitt 3.2.) vorzustellen und Handlungsempfehlungen (s. Abschnitt 4) abzugeben, wie dieser Ansatz zur Absicherung und Prüfung sicherheitskritischer KI-Systeme und die Sicherheit von KI insgesamt vorangebracht werden könnten.

9 Kläs et al. (2021a), (wie Fn. 6)

2. Qualitätssicherung von sicherheitskritischer KI als Herausforderung

Bei klassischer Software ist das Testen ein wichtiger und in der Informatik etablierter Bereich, der viele verschiedene Methoden, Ansätze und best practices zur Überprüfung von Software bietet. Oftmals ist bereits die Entwicklung auf das Testen ausgerichtet und es gibt sogar das Berufsbild der Softwaretester:innen. Im Vergleich steckt bei KI das Thema Qualitätssicherung noch in den Kinderschuhen und angesichts der Eigenschaften von Maschinellem Lernen lassen sich bekannte Verfahren und Methoden auch nur eingeschränkt übertragen.

Bei herkömmlicher Software hängt die Funktionalität und Qualität in erster Linie vom geschriebenen und für Entwickler:innen generell nachvollziehbaren Softwarecode ab. Daher geht man davon aus, dass es grundsätzlich möglich ist, alle kritischen Fehler zu finden und zu beseitigen. Bei KI-basierten Systemen ist dies nicht der Fall, da hier nicht deterministisch festgelegt ist, welche Eingabe zu welcher Ausgabe führt. Vielmehr werden diese Entscheidungsregeln „gelernt“ und es hängt von den jeweiligen Eingabedaten ab, zu welcher Ausgabe das Modell kommt. Da oftmals nicht klar ist, wie das System auf unbekannte Eingabedaten reagieren wird, lässt sich so auch das zukünftige Verhalten zum Zeitpunkt der Entwicklung und Inbetriebnahme nicht absolut verlässlich vorhersagen. Dies ist nur möglich für die Eingabedaten, mit denen getestet wurde. Allerdings können in der Regel nicht alle potenziellen Eingabedaten im Rahmen des Entwicklungsprozesses durchgetestet werden. Obwohl es viel Forschung in diesem Bereich gibt, die Fortschritte erzielt, sind bisherige Methoden unzureichend, um etwa die Sicherheit eines Systems umfassend zu garantieren, wenn die Sicherheit von einer KI-Komponente abhängt. Die Kontextabhängigkeit datengetriebener Modelle erschwert so eine abschließende Risikobeurteilung, die es erlauben würde, alle notwendigen Anforderungen zu identifizieren und passgenaue Maßnahmen zur Risikoreduktion zu ergreifen, wie es im klassischen Safety-Engineering der Fall wäre.

Insgesamt gibt es bislang keinen Konsens darüber, welche Sicherheitsanforderungen bei der Entwicklung und beim Einsatz einer KI-Komponente ausreichen und gelten müssten, um aktuellen Vertrauensniveaus herkömmlicher Software zu genügen. Dementsprechend gibt es auch keine Safety-Standards, die KI abdecken und etablierte Normen, wie etwa die Grundnorm für funktionale Sicherheit IEC 61508, sind nur begrenzt auf KI übertragbar, da die dort vorgesehenen Maßnahmen, wie z. B. Code Reviews, für KI nicht geeignet sind. So fehlt es Herstellerunternehmen an Vorgaben und Richtlinien, aus denen hervorgeht, welche Maßnahmen angewandt werden müssen, etwa auf welche Weise und wie umfangreich getestet werden muss, damit die KI-Komponente als ausreichend sicher gilt. Zudem sind die in Sicherheitsnormen geforderten Fehlerwahrscheinlichkeiten so niedrig angesetzt, dass KI-Komponenten sie oftmals nicht erreichen können.

3. Assurance Cases als Lösung?

Da herkömmliche Methoden der Qualitätssicherung sowie existierende Standards nicht ausreichen, kommt es darauf an, sehr ausführlich darzulegen und zu begründen, warum der Einsatz einer KI-Komponente sicher ist, also keine inakzeptablen Risiken eingegangen werden. Eine solche Argumentation, die dabei den anwendungsfall- und technologiespezifischen Besonderheiten eines Use Cases gerecht wird, kann helfen, um darzulegen, dass ein System hinreichend verlässlich ist, und daraus zu lernen. Allerdings benötigen Hersteller und Prüfer für diese inhärent komplexe Argumentation praktikable Methoden und Ansätze.

Als Methode, um komplexe Sicherheitsargumentationen zu dokumentieren, erscheinen „Assurance Cases“ als geeignetes Mittel, um die notwendige umfassendere Argumentation abzubilden, die es bei KI benötigt. Ein Assurance Case ist eine klar strukturierte, übersichtliche Argumentation, die umfassend begründet, warum ein System bestimmte Anforderungen in einem konkreten Anwendungskontext erfüllt.[10] Er verknüpft eine Aussage (claim), etwa, dass ein System in einem bestimmten Anwendungskontext sicher ist, mit einer strukturierten Argumentation, die alle getroffenen Annahmen enthält und sich wiederum auf eine Sammlung von Evidenzen stützt. Dazu zählen produktbezogene Evidenzen wie Test- oder Simulationsergebnisse, aber auch prozessbezogene Evidenzen zum Entwicklungsprozess inklusive Nachweise von Kompetenzen der Entwickler:innen. Das Konzept der Assurance Cases ist im Safety-Bereich sehr etabliert. Insbesondere in Sparten wie Automotive, Medizintechnik oder Luftfahrt sind Assurance Cases verbreitet, um die Sicherheit von sicherheitskritischen Anwendungen zu gewährleisten. Während hier auch von Safety Cases gesprochen wird, signalisiert der Begriff Assurance Case, dass das Konzept zunehmend auch auf andere Attribute und Eigenschaften übertragen wird.[11]

3.1. Potenziale von Assurance Cases

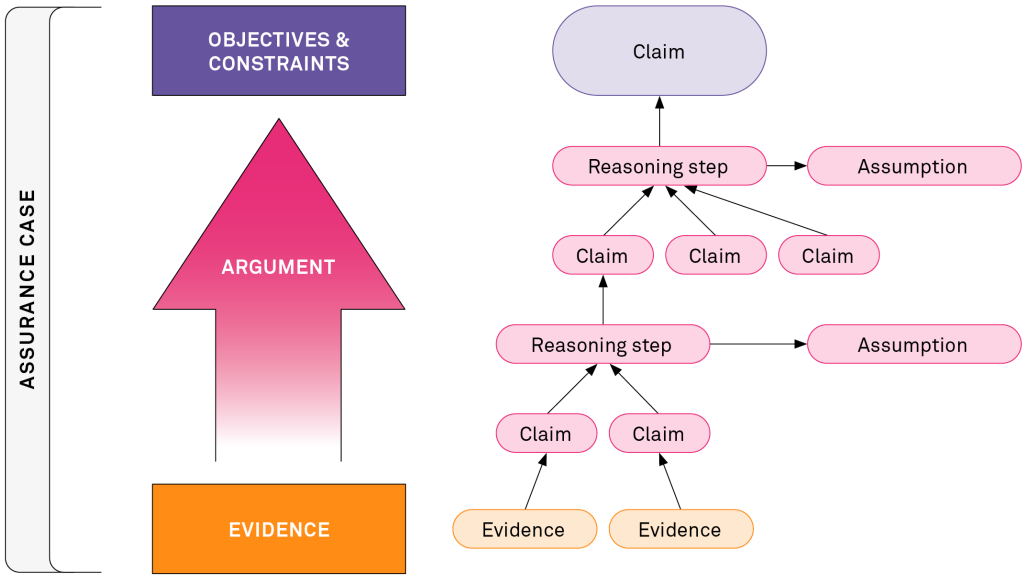
Insgesamt verhelfen Assurance Cases zu einem ganzheitlichen Blick auf ein System in seinem Anwendungskontext und können flexibel genutzt werden, um ganz spezifische Aussagen für den konkreten Anwendungsfall argumentativ und auf Evidenzen basierend zu belegen, wie zum Beispiel: „Das intelligente Transportsystem X in dem

10 Gemäß ISO/IEC/IEEE 15026-1:2019 sind Assurance Cases definiert als „reasoned, auditable artefact created that supports the contention that its top-level claim (or set of claims) is satisfied, including systematic argumentation and its underlying evidence and explicit assumptions that support the claim(s)“.

11 Richard Hawkins, Ibrahim Habli, Tim Kelly und John McDermid (2013): Assurance cases and prescriptive software safety certification: A comparative study. In: Safety Science, Volume 59, 55–71. Abgerufen am 24.06.2021 von https://www.sciencedirect.com/science/article/abs/pii/S0925753513001021; Siehe auch für eine Anwendung von Assurance Cases auf Fairness: Marc Hauer, Rasmus Adler und Katharina Zweig (2021): Assuring Fairness of Algorithmic Decision Making. IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW).

autonomen Warenhaus Y ist sicher für den Transport von Gütern Z“. Da die Sicherheit einer KI-Komponente immer nur kontext- und anwendungsspezifisch bewertet werden kann, bieten Assurance Cases die notwendige Flexibilität, die dem Einzelfall gerecht wird.

Komponenten und generelle Struktur eines Assurance Cases



Quelle: Michael Kläs et al. 2021 (wie Fn. 12)

Assurance Cases bieten sich zudem an, um die Schwächen etablierter Qualitätssicherungsmaßnahmen im Kontext von KI zu adressieren. Sie bieten die Möglichkeit, alle Maßnahmen und Informationen in den Sicherheitsnachweis einzubringen, die begründetes Vertrauen in die Sicherheit des Produkts liefern. Anstatt sich zum Beispiel rein auf Testergebnisse zu verlassen, die im Kontext von KI nur begrenzt aussagekräftig sind, verortet der Assurance Case einzelne Methoden in einen argumentativen Zusammenhang mit anderen Maßnahmen. Er bietet damit ein systematisches Vorgehen, um zum Beispiel Entscheidungen im Entwicklungsprozess mit Blick auf die zentrale Anforderung („Die KI-basierte Kollisionsvermeidung des FTFs ist ausreichend verlässlich“) kritisch zu hinterfragen sowie im Rahmen der Argumentation frühzeitig Defizite und Schwachstellen zu identifizieren („Zum Training wurden genug Bilder von potenziellen Gesten durch Werker:innen gesammelt und auch im Hinblick auf Sonderfälle gelabelt. Durch Reviews wurde sichergestellt, dass beim Labeln keine Fehler gemacht wurden.“).

Insgesamt können Assurance Cases das Grundproblem der generell hohen Fehlerraten von datengetriebenen Modellen zwar nicht lösen. Sie können aber mittels genauer Risikoakzeptanzbetrachtungen darlegen, warum es in vielen Fällen trotzdem sinnvoll und sicher sein kann, datengetriebene Modelle in sicherheitskritischen Kontexten zu benutzen.[12] Das gilt insbesondere für zusätzliche Schutzfunktionen, mit denen man Risiken reduzieren könnte, die aktuell akzeptiert werden (z. B. KI-basierte Bilderkennung zur Verhinderung von Missbrauch wie das Außerkraftsetzen von herkömmlicher Sicherheitsvorrichtungen bei Cobots[13]).

Die Formulierung eines Assurance Cases bietet außerdem einen Rahmen, in dem die unterschiedlichen Teams, die in der Regel an einem Produkt arbeiten, zusammenkommen und sich besser abstimmen können. Sie haben damit das Potenzial, Implizites explizit zu machen und damit mehr Transparenz in die Qualität des Entwicklungsprozesses und des fertigen Produkts zu bringen.

Assurance Cases helfen aber nicht nur Herstellerunternehmen zu erklären, warum und unter welchen Bedingungen ein System als sicher angesehen werden kann. Sie bieten auch für Dritte eine Grundlage, um zum Beispiel im Rahmen eines Audits oder der Marktüberwachung nachzuvollziehen, warum Evidenzen wie etwa Testergebnisse zeigen, dass ein System sicher ist. Assurance Cases bilden die Basis für Hersteller und Prüfpersonen, die dafür relevanten Informationen miteinander auszutauschen.[14] Die strukturierte Darstellung von Argumenten, Annahmen und Evidenzen kann Audits zusätzlich erleichtern.

3.2. Herausforderungen von Assurance Cases

Angesichts fehlender Vorgaben und etablierter Verfahren, die gewährleisten, dass die KI ausreichend sicher ist, erscheinen Assurance Case als sinnvoller Ansatz, um die Sicherheit trotzdem überzeugend darzulegen. Gleichwohl ist der Einsatz von Assurance Cases für KI-Komponenten mit Einschränkungen und Hürden verbunden:

12 Michael Kläs, Rasmus Adler, Lisa Jöckel, Janek Groß und Jan Reich (2021b): Using Complementary Risk Acceptance Criteria to Structure Assurance Cases for Safety-Critical AI Components. Abgerufen am 06.09.2021 von http://ceur-ws.org/Vol-2916/paper_9.pdf.

13 Gesellschaft für Informatik (2020): Anwendungsszenarien. KI-Systeme in der Produktionsautomatisierung. Angerufen am 06.09.2021 von https://testing-ai.gi.de/fileadmin/GI/Projekte/KI_Testing_Auditing/ExamAI_Publikation_Anwendungsszenarien_KI_Industrie.pdf, 10.

14 Ran Wei, Tim R. Kelly, Xiaotian Dai, Shuai Zhao and Richard Hawkins (2019): Model Based System Assurance Using the Structured Assurance Case Metamodel. Abgerufen am 24.06.2021 von https://arxiv.org/ftp/arxiv/papers/1905/1905.02427.pdf; Robin Bloomfield und Peter Bishop (2010): Safety and Assurance Cases: Past, Present and Possible Future–an Adelard Perspective, 2. In: Proceedings of the Eighteenth Safety-Critical Systems Symposium, Bristol, 51–67.

Keine Musterlösung vorhanden

Zunächst ist ein Assurance Case weder Beweis noch Checkliste, sondern in erster Linie eine Strukturierungs- und Orientierungshilfe für den Sicherheitsnachweis. Das Ziel eines aussagekräftigen Assurance Cases ist es, dass er ein überzeugendes Argument für die Sicherheit liefert. Damit das Argument überzeugend ist, muss es sich auf starke Evidenzen, also beweisbare Fakten, stützen.[15] Eine geeignete und ausreichende Argumentationsbasis zu finden, bleibt auch bei Assurance Cases eine große Herausforderung, da es vom konkreten Anwendungsfall abhängt, was eine KI bezüglich Sicherheit leisten kann und muss. Entsprechend gibt es bislang noch keinen Konsens, wie ein Assurance Case zur Absicherung von KI genau aufgebaut werden sollte. Auch wenn die Auffassung geteilt wird, dass es sich um einen geeigneten Ansatz handelt, unterscheiden sich aktuelle Strukturierungsvorschläge voneinander. Folglich fehlt es auch an Kriterien und Guidelines, die Unternehmen Anhaltspunkte geben, ab wann ein Assurance Case als ausreichend erachtet wird.

Neue Anforderungen und Kosten für Unternehmen

Assurance Cases werden darüber hinaus nur funktionieren, wenn sie als Gelegenheit gesehen werden, sich intensiv mit dem System und seinem Nutzungskontext auseinanderzusetzen und zu artikulieren, was ein System tatsächlich sicher macht. Es geht nicht darum, eine Sammlung unterschiedlicher Dokumentationen zusammenzustellen, sondern ein Verständnis und eine Argumentation für die jeweilige KI-Anwendung zu entwickeln. Entwickler:innen müssen über die Expertise verfügen, wie verschiedene Maßnahmen (z. B. Tests, „ExplainableAI"-Ansätze) gestaltet, angewandt und kombiniert werden müssen, um ein sicheres System zu bauen. Zudem kann ein aussagekräftiger Assurance Case nur in engem Austausch zwischen Maschinenhersteller, Zulieferer der KI-Komponente und Betreiber erstellt werden. Die Erstellung von Assurance Cases macht den Entwicklungs- und Konstruktionsprozess damit um einiges aufwändiger. Insgesamt müssen dafür zunächst geeignete Prozesse etabliert und Mitarbeiter:innen entsprechend geschult werden, was für Unternehmen zusätzliche Anstrengung und Kosten bedeutet und insbesondere für kleinere Hersteller eine Herausforderung darstellt.

Objektive Kriterien und Guidelines notwendig, ohne Flexibilität und kritisches Denken aufzugeben

Ein weiteres Problem ergibt sich daraus, dass der Assurance Case selbst schwer zu bewerten ist. Aufgrund der Flexibilität, mit der die Argumentation erstellt werden kann, fehlen allgemeine Kriterien, anhand derer die Güte eines Assurance Cases im Rahmen eines Audits ohne Weiteres beurteilt werden kann. Ein Assurance Case gilt immer nur für einen bestimmten Anwendungskontext, der klar definiert sein muss. Das heißt, dass sich die Argumentation und notwendigen Evidenzen je nach Sys-

15 Hawkins et al. (2013)

tem, Anwendungskontext und dessen Kritikalität unterscheiden werden, weswegen Audits kaum verallgemeiner- und reproduzierbar sind und damit eine Herausforderung für das Prüfpersonal darstellen.[16] Vor dem Hintergrund ist es fraglich, inwiefern Auditor:innen überhaupt in der Lage sind, Schwächen in einem Assurance Case zu erkennen. Ohne etwaige Kriterien würde es sehr viel Domänenwissen und praktische Erfahrung im konkreten Anwendungsbereich erfordern, um nicht Gefahr zu laufen, sich zu leicht von einer schlüssig erscheinenden Argumentationskette überzeugen zu lassen.

Um die praktische Anwendung also zu erleichtern und eine gewisse Vergleichbarkeit herzustellen, sind Unternehmen und Prüforganisationen auf Kriterien und Guidelines angewiesen. Standards und Normen werden hier eine zentrale Rolle spielen. Die Frage ist, wie man eine Standardisierung erreicht, ohne den großen Vorteil der Flexibilität aufzugeben.[17] Dabei besteht auch die Gefahr, dass die Erstellung und Prüfung von Assurance Cases durch die Anwendung von Standards zur Routine und reinen Compliance-Aufgabe wird, die der Notwendigkeit einer intensiven Auseinandersetzung mit dem System im jeweiligen Anwendungsgebiet und einer stichhaltigen Argumentation nicht mehr gerecht wird.[18]

Monitoring notwendig

Auch mit Assurance Cases ist es nicht möglich, zum Zeitpunkt des Inverkehrbringens die Sicherheit abschließend zu beurteilen. Zwar können Assurance Cases eine umfassende Beurteilung möglicher Risiken für die Sicherheit unterstützen. Doch aufgrund der Komplexität des Use-Cases ist auch hier nicht auszuschließen, dass Unternehmen und Prüfende bei der Erstellung und Bewertung der Argumentation etwas nicht berücksichtigt haben oder zum Beispiel zu optimistische Annahmen getroffen wurden oder sich etwas im Anwendungskontext in unvorhergesehener Art ändert. Das bedeutet, dass ein systematisches Monitoring notwendig ist, um im Betrieb mögliche Schwächen, zum Beispiel fehlerhafte Annahmen bei der Argumentation, zu erkennen und den Assurance Case gegebenenfalls über die Zeit zu verbessern.[19] Die Prüfung ist mit der Inbetriebnahme also nicht abgeschlossen, sondern muss bei KI immer auch begleitend zum Betrieb stattfinden, um die Beständigkeit des Sicherheitsnachweises zu gewährleisten. Assurance Cases bieten zwar ein Grobkonzept, um ein Monitoring und die Marktüberwachung für KI-Komponenten zu operationalisieren, die Verfeinerung und praktische Umsetzung des Konzepts sind aber eine Herausforderung.

16 ebd.
17 Brundage et. al. (2020): Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. Abgerufen am 24.06.2021 von https://pure.tue.nl/ws/portalfiles/portal/164943601/2004.07213v2.pdf.
18 Tim Kelly (2008): Are Safety Cases Working? Abgerufen am 24.06.2021 von https://scsc.uk/r103.12:1.
19 Gleichzeitig ist eine kontinuierliche Beobachtung auch wichtig, um Rückschlüsse zu ziehen und Wissen zu generieren, das in die Entwicklung entsprechender Standards und der Weiterentwicklung eines Assurance Cases einfließen kann.

Kompetenzen sind Mangelware

Das Ziel, die Sicherheit von KI-basierten Systemen in der industriellen Mensch-Maschine-Interaktion zu gewährleisten, benötigt nicht nur technisches Know-How in der Forschung und Entwicklung. Mit Assurance Cases verfügen Prüforganisationen und auch Marktüberwachungsbehörden zwar über eine prüfbare Grundlage, um Aussagen über die Sicherheit einer KI-Komponente treffen zu können. Ohne entsprechende Kompetenzen und Ressourcen wird es aber nicht möglich sein, die Sicherheit von KI-Komponenten auf dem europäischen Markt zu gewährleisten. Der Mehrwert von Assurance Cases als Basis für Audits ist nur gegeben, wenn auch die Prüfenden die entsprechenden Fähigkeiten und die Expertise haben, um den Assurance Case zu prüfen, zu beurteilen und argumentative Schwächen erkennen zu können.

4. Ausblick und Empfehlungen

Unternehmen, Prüforganisationen, Behörden und Politik stehen bezüglich der Absicherung und Prüfung von KI noch am Anfang. Auch der Ansatz der Assurance Cases bietet keine simple Lösung, um die Sicherheit von KI darzulegen und zu bewerten. Gleichzeitig mangelt es aber an Alternativen, die eine für sicherheitskritische Anwendungen ausreichende Aussagekraft bieten. Am Ende bedeuten fehlende Ansätze und Anforderungen ein Risiko, dass die Erprobung, Weiterentwicklung und Verbesserung neuer Technologien erschwert und damit am Ende auch Fortschritt und Innovation auf dem Gebiet der KI in Europa beeinträchtigt wird.[20]

Die Frage ist daher, was zukünftig getan werden sollte, um die Herausforderungen von Assurance Cases zu adressieren, ihre Anwendung für Hersteller und Prüfende zu erleichtern, den Einsatz von KI in sicherheitskritischen Funktionen in der Industrieproduktion zu ermöglichen und die Sicherheit insgesamt zu verbessern. Es wird dabei unter anderem auf ein Zusammenspiel aus praxisnaher, interdisziplinärer Forschung, lösungsorientierten Kriterien und Richtlinien sowie effektiver Prüfung und Marktüberwachung ankommen. Hier besteht für unterschiedliche Seiten Handlungsbedarf. Stakeholder aus der Politik, Industrie, Wissenschaft, Normung, Konformitätsbewertung und Marktüberwachung sollten verschiedene Aspekte im Blick behalten:

4.1. Experimentierräume schaffen

Zunächst muss der Einsatz sicherheitskritischer KI und die Formulierung ausreichender Sicherheitsargumentationen mit Hilfe von Assurance Cases in spezifischen Anwendungsbereichen industrieller KI weiter erprobt und konkretisiert werden. Allerdings ist die Erarbeitung individueller, auf einen Use-Case zugeschnittener Assurance Cases mit hohem finanziellem und personellem Aufwand sowie dem Risiko verbunden, dass entsprechende Projekte am Ende eine fehlende Machbarkeit ergeben. Das bedeutet, dass potenziell die Anreize fehlen, Pionierarbeit auf dem Gebiet der Absicherung von KI zu leisten und es hier einer stärkeren Förderung bedarf.

Gegenwärtig liegt der Fokus von öffentlich geförderten Praxisprojekten häufig auf der Untersuchung des wirtschaftlichen Potenzials von KI und der Entwicklung neuer Methoden bzw. Ihr Anwendungen in bestimmten Bereichen, weniger auf der Untersuchung von Safety-Aspekten. Da aber gerade Safety die Grundvoraussetzung für die Markteinführung und die Entfaltung des wirtschaftlichen Potenzials ist, sollte

20 Will Hunt (10.11.2020): Regulations could speed up, not slow down A.I. progress. Abgerufen am 24.6.2021 von https://fortune.com/2020/11/09/ai-artificial-intelligence-biden-regulations/.

die Thematik auch eine stärkere Berücksichtigung bei zukünftigen Ausschreibungen finden. Der Regulierungsvorschlag der EU-Kommission sieht die Einrichtung von „regulatory sandboxes“ zur Erprobung von KI insbesondere mit Blick auf deren Einhaltung gesetzlicher Vorgaben explizit vor.[21]

Ziel sollte es sein, unter Einbeziehung aller relevanter Stakeholder und betrieblicher Interessen einen Konsens zu erreichen, wie konkrete KI-Anwendungen abgesichert werden können. Dazu bedarf es der engen Zusammenarbeit zwischen Wissenschaftler:innen und Praktiker:innen aus der Industrie, Sozialpartnern und Akteuren der Qualitätsinfrastruktur. Idealerweise bringt ein Experimentierraum neben direktem Bezug zum Praxisbetrieb, Safety-Expert:innen, Entwickler:innen von Safety-Maßnahmen und Expert:innen der Konformitätsbewertung zusammen. Insbesondere die zwei Welten KI und Safety stärker zusammenzubringen, wird von zentraler Bedeutung sein. KI- und Safety-Expert:innen arbeiten mitunter zwar an der gleichen Herausforderung, jedoch aus unterschiedlichen Blickwinkeln, sodass eine stärkere interdisziplinäre Zusammenarbeit einen großen Mehrwert darstellen würde.

4.2. Standardisierung angehen und fördern

Um die Anwendung von Assurance Cases in der Praxis zu erleichtern, bedarf es klarer Kriterien und Guidelines. Hersteller, Zulieferer und Betreiber benötigen für die Zusammenstellung ihres Assurance Cases passende Anhaltspunkte, damit sie einschätzen können, ab wann ein Assurance Case im Rahmen einer Drittstellenprüfung als ausreichend erachtet wird. Auch für Auditoren sind Standards als objektive Kriterien unabdingbar, anhand derer sie einen Assurance Case und dessen Güte bewerten und somit entscheiden können, ob ein Produkt als sicher anzusehen ist.

Die Erarbeitung entsprechender Standards, möglicherweise auf Basis der durch Experimentierräume gewonnenen Erfahrungen, wird daher unumgänglich sein. Vorstellbar ist etwa eine Norm, die Anforderungen und Prüfkriterien für einen Assurance Case festlegt (zum Beispiel zentrale Aspekte, die ein Assurance Case adressieren muss) oder die Ziele beschreibt (zum Beispiel eine bestimmte Fehlerrate mit einer bestimmten Konfidenz nachzuweisen) und welche Möglichkeiten und Lösungsansätze grundsätzlich bestehen, um zu zeigen, dass die definierten Ziele erfüllt sind.

Die Entwicklung von Standards ist nach wie vor ein langwieriger Prozess. Um mit dem Aufkommen potenzieller Use Cases Schritt zu halten, ist jedoch Schnelligkeit gefragt. Durch die Förderung von kleineren Initiativen (zum Beispiel Technische

21 Europäische Kommission (2021): Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on AI and Amending Certain Union Legislative Acts. COM/2021/206 final, Art. 34.

Spezifikationen im Rahmen der verantwortlichen Normungsausschüsse, DKE-Anwendungsregel oder DIN Specs) mit ausgewählten Expert:innen könnten schnell erste Vorschläge erarbeitet werden, die dann als Grundlage für die europäische und internationale Normung dienen könnten. In einem sensiblen Bereich wie Safety sollte aber trotz der angestrebten Agilität weiterhin eine breite Beteiligung an der Entwicklung der Standards gewährleistet werden, wie sie formal nur in etablierten Normenausschüssen vorgesehen ist.

Grundsätzlich kann die Normung nur durch ein aktives Engagement der entsprechenden Expert:innen vorankommen. Dafür müssten insgesamt die Bedingungen für eine Teilnahme an Standardisierungsaktivitäten weiter verbessert werden, unter anderem durch angemessene finanzielle Förderung (zum Beispiel auch im Rahmen einschlägiger Forschungsförderung).[22] Ansonsten bleibt es auch für kleine und mittlere Unternehmen schwer, sich konsequent zu engagieren. Universitäten und Forschungsinstitute verfügen ebenfalls über wertvolle Expertise, gleichzeitig fehlt für sie oftmals der konkrete Nutzen, diese auch einzubringen.

4.3. Kulturwandel, Wissens- und Kompetenzaufbau

Die Anwendung von Assurance Cases erfordert von allen Beteiligten eine neue Denkweise. Statt der traditionellen Herangehensweise zum Nachweis von Sicherheit, die auf konkrete, standardisierte Gestaltungsvorgaben setzt, geht es nun um einen flexibel gestaltbaren Argumentationsprozess. Diese neue Herangehensweise ist für alle Seiten eine Herausforderung und wird in der Industrie, Normung, Konformitätsbewertung und Marktüberwachung einen Kultur- und Kompetenzwandel erfordern.

Für die Industrie bedeutet das, einen größeren Fokus auf Safety-Engineering zu legen und es zu stärken. Bislang haben Expert:innen aus dem Bereich Machine Learning oder Data Science oftmals zu wenig Verständnis von Safety im Sinne von funktionaler Sicherheit. Zukünftig bedarf es hier z. B. zielgerichteter Schulungen, um das Bewusstsein dafür zu schärfen und ein stärkeres „Safety-Mindset“ bei Anbieterfirmen zu etablieren.

Auch in den Normungsgremien ist ein entsprechender Kulturwandel- und Kompetenzwandel notwendig: Die angesprochene Zusammenarbeit zwischen den Feldern Safety und KI findet auch hier noch unzureichend statt. Dies setzt aber auch voraus, dass Safety-Expert:innen aufgeschlossener gegenüber KI-Anwendungen und KI-Entwickler:innen aufgeschlossener gegenüber den durch Safety-Standards for-

22 Bestehende Fördermöglichkeiten bilden z. B. das Technologieförderprogramm des Bundesministeriums für Wirtschaft und Energie WIPANO und das europäische Förderprogramm StandICT.eu.

mulierten Voraussetzungen werden. Insgesamt sollte das Augenmerk verstärkt auf die Zusammensetzung der Gremien gerichtet und mehr ausgewiesene KI-Expert:innen in die Standardisierungsgremien gebracht werden, die sich mit funktionaler Sicherheit befassen.

Nicht zuletzt muss auch bei den Behörden der Marktüberwachung ein Verständnis für die Assurance Case-Methodik geschaffen werden. Daher sollten Behörden früh involviert werden und sich zum Beispiel im Rahmen von Schulungen mit Assurance Cases vertraut machen können, um die Akzeptanz dieses neuen Ansatzes zu erhöhen. Es muss auf Ebene der Marktüberwachung zudem gelingen, die Kompetenzen und Kapazitäten zu stärken und zu modernisieren, die es braucht, um die Safety von KI zu überwachen. Auch die Europäische Kommission geht davon aus, dass europäischen Marktüberwachungsbehörden die Ressourcen, Expertise und technische Ausstattung fehlen, um die von im Einsatz befindlichen KI-Systemen ausgehenden Risiken zu überwachen oder KI-Komponenten zu untersuchen.[23] Sie nimmt die Mitgliedsstaaten in die Pflicht für eine ausreichende Ausstattung der Behörden zu sorgen.[24]

Der Wissens- und Kompetenzaufbau ließe sich durch eine größere Offenheit auf Seiten der Unternehmen weiter beschleunigen. So sind Assurance Cases in der Regel unter Verschluss, gehören zum geistigen Eigentum eines Unternehmens und enthalten mitunter sensible Geschäftsinformationen. Wären sie der Forschung öffentlich zugänglich, könnten auf der Basis etwa die Methode und konkreten Argumentationen weiterentwickelt werden. Ein anderer wichtiger Aspekt betrifft den Umgang mit sicherheitskritischen Ereignissen während des Betriebs von KI-Komponenten: Auch hier kann mehr Transparenz, zum Beispiel durch eine Veröffentlichung von Vorfällen im Bereich sicherheitskritischer KI und Austausch von Best Practices, andere Stakeholder in die Lage versetzen, aus den Erfahrungen zu lernen und die Sicherheit von KI-Anwendungen insgesamt zu verbessern. Im Bereich Cybersicherheit etwa gibt es unterschiedliche Beispiele für solche Informationsaustauschformate, die als Vorbilder genutzt werden könnten (zum Beispiel Allianz für Cybersicherheit oder sektorale Initiativen wie das European Financial Institutes–Information Sharing and Analysis Centre FI-ISAC).[25] Staatliche Stellen könnten die Aufgabe übernehmen, ge-

23 Europäische Kommission (2021): Impact Assessment of the Regulation on Artificial intelligence, S.22. Abgerufen am 24.08.2021 von https://digital-strategy.ec.europa.eu/en/library/impact-assessment-regulation-artificial-intelligence.

24 Die Europäische Kommission schätzt, dass die europäischen Marktüberwachungsbehörden lediglich zwischen 1 bis 25 zusätzliche Vollzeitstellen benötigen, um ihre Aufgaben angemessen zu erfüllen Siehe dazu Europäische Kommission (2021): Impact Assessment of the Regulation on Artificial intelligence, (n 95), Annex 3, 25. Zitiert nach Michael Veale und Frederik Zuiderveen Borgesius (2021): Demystifying the Draft EU Artificial Intelligence Act. Abgerufen am 06.09.2021 von https://osf.io/preprints/socarxiv/38p5f.

25 Für weitere Beispiele und best practices siehe: Enisa (2018): Information Sharing and Analysis Center (ISACs)-Cooperative models. Abgerufen am 06.09.2021 von https://www.enisa.europa.eu/publications/information-sharing-and-analysis-center-isacs-cooperative-models; und Julia Schütze (2020): EU-US Cybersecurity Policy Coming Together: Recommendations for instruments to accomplish joint strategic goals. Abgerufen am 06.09.2021 von https://www.stiftung-nv.de/sites/default/files/rif_eu-us-cybersecurity-final.pdf, 25, 35 ff..

eignete Strukturen und Informationen bereitzustellen, um einen solchen Austausch zwischen Unternehmen anzuregen, die KI in sicherheitskritischen Anwendungen einsetzen.[26]

Zu Wissen und Kompetenzaufbau gehört ebenfalls, dass auch auf Ebene der zur Verfügung stehenden technischen Fähigkeiten und Ressourcen weiter Fortschritte erzielt werden müssen und auch hier noch viel grundlegende Forschung notwendig ist. Um Evidenzen für die Sicherargumentation generieren und Assurance Cases prüfen zu können, werden Testwerkzeuge und Softwarepakete benötigt, die eine Prüfung ermöglichen und erleichtern. Die Entwicklung entsprechender Tools gestaltet sich angesichts einer unsicheren Marktlage und unklarer Anforderungen schwierig. Eine stärkere öffentliche Förderung erscheint unabdingbar, um die Entwicklung von Prüfwerkzeugen erfolgreich voranzutreiben. Prüfen und Testen von KI wird absehbar aber ein aktives Forschungsfeld bleiben, das kontinuierlich technische Weiterentwicklungen verzeichnen wird. Daher ist es gleichzeitig wichtig, mittels Studien Klarheit darüber zu gewinnen, welche Werkzeuge und Methoden welche Evidenzen in welcher Güte liefern können, um sie entsprechend in eine Argumentation einbauen zu können.

26 Der KI-Regulierungsvorschlag sieht ein solches „Information Sharing“ bislang nur zwischen betroffenen Unternehmen und Behörden der Marktüberwachung vor. Siehe Europäische Kommission (2021): Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on AI and Amending Certain Union Legislative Acts. COM/2021/206 final, Art. 62.

Danksagung

Für ihre Unterstützung und wertvolles Feedback zum Papier bedanke mich bei den Partner:innen des ExamAI Projekts, insbesondere bei Rasmus Adler, Michael Kläs, Lisa Jöckl und Jens Heidrich von Fraunhofer IESE sowie bei Andreas Sesing, Adrian Kreutzer und Sven Hilpisch des Instituts für Rechtsinformatik der Universität des Saarlandes und bei Nikolas Becker von der Gesellschaft für Informatik.

Besonderer Dank gilt den Expert:innen aus Unternehmen, Universitäten und Forschungseinrichtungen und Behörden, die ihre Erkenntnisse im Rahmen von Gesprächen oder des ExamAI-Projektworkshops geteilt haben. Dazu gehören insbesondere Tarek Besold (Dekra), Frank Domrös (Micropsi Industries), Katharina Eylers (BITKOM), Tim Faber (ZLS Bayern), Patrik Feth (Sick AG), Sebastian Hallensleben (VDE), Andreas Hauschke (VDE), Niels Heller (QualityMinds), Marco Jennerich (IG Metall), Thomas Just (Hessisches Ministerium für Soziales und Integration), Thorsten Katzmann (IBM), Christian Kolf (TÜV AI Lab), Corado Mattiuzzo (Kommission für Arbeitsschutz und Normung), Thomas Mössner (Bundesanstalt für Arbeitsschutz und Arbeitsmedizin), Jochen Reinschmidt (Zentralverband der Elektroindustrie), Ralf Schmitt (ver.di/TÜV Rheinland Industrie Services), Gesina Schwalbe (Continental AG), André Steimers (DGUV IFA), Silvia Vock (Bundesanstalt für Arbeitsschutz und Arbeitsmedizin) und Jan Widmann (Ministerium für Umwelt, Klima und Energiewirtschaft, Baden-Württemberg). Die Ansichten in diesem Papier spiegeln nicht notwendigerweise die der Fachleute wider, und alle verbleibenden Fehler sind unsere eigenen.

Zudem bedanke ich mich bei meinen SNV-Kolleg:innen Maria Jacob, Pegah Maham, Stefan Heumann und Johanna Famulok für ihre Unterstützung.

Über die Stiftung Neue Verantwortung

Die Stiftung Neue Verantwortung (SNV) ist ein gemeinnütziger Think Tank, der an der Schnittstelle von Technologie und Gesellschaft arbeitet. Die Kernmethode der SNV ist die kollaborative Entwicklung von Politikvorschlägen und -analysen. Die Expert:innen der SNV arbeiten nicht allein, sondern entwickeln und testen Ideen gemeinsam mit Vertreter:innen aus Politik und Verwaltung, Technologieunternehmen, Zivilgesellschaft und Wissenschaft. Unsere Expert:innen arbeiten unabhängig von Interessengruppen und Parteien. Unsere Unabhängigkeit gewährleisten wir durch eine Mischfinanzierung, zu der viele verschiedene Stiftungen, öffentliche Mittel und Unternehmensspenden beitragen.

Über die Autorin

Leonie Beining ist Projektmitarbeiterin im Projekt „ExamAI – KI-Testing und Auditing, das die Stiftung Neue Verantwortung in Zusammenarbeit mit der Gesellschaft für Informatik, dem Algorithmic Accountability Lab der TU Kaiserslautern, dem Institut für Rechtsinformatik der Uni Saarbrücken und Fraunhofer IESE durchführt. Leonie verantwortet in dem Projekt die Identifizierung der politischen Handlungsempfehlungen zur Implementierung geeigneter Test- und Auditierungsverfahren für KI-Systeme in den Bereichen Industrieproduktion und Human Resource-Management. Zuvor war Leonie Projektleiterin im Projekt „Algorithmen fürs Gemeinwohl“ und setzte sich in dem Rahmen vor allem mit dem Aspekt der Nachvollziehbarkeit von algorithmischen Entscheidungsprozessen auseinander. Zu Beginn ihrer Zeit an der SNV war Leonie Projektmanagerin im Projekt „Gemeinwohl im digitalen Zeitalter“ tätig und entwickelte Strategien, wie sich zivilgesellschaftliche Akteure stärker in die Gestaltung der Digitalisierung einbringen können.

So erreichen Sie die Autorin:

Leonie Beining
lbeining@stiftung-nv.de
Twitter: @leoniebeining

Über das Projekt

Das Papier erscheint als Teil des Forschungsprojekts „ExamAI – KI Testing und Auditing“, das sich der Erforschung geeigneter Test- und Auditierungsverfahren für KI-Anwendungen widmet. Es steht unter der Leitung der Gesellschaft für Informatik e. V. wird von einem interdisziplinären Team bestehend aus Mitgliedern der TU Kaiserslautern, der Universität des Saarlandes, des Fraunhofer-Instituts für Experimentelles Software Engineering IESE, der Stiftung Neue Verantwortung getragen und im Rahmen des Observatoriums Künstliche Intelligenz in Arbeit und Gesellschaft (KIO) der Denkfabrik Digitale Arbeitsgesellschaft des Bundesministeriums für Arbeit und Soziales (BMAS) gefördert.

Informationen zum Projekt und weitere Veröffentlichungen finden Sie unter: https://testing-ai.gi.de/

Weitere Fachveröffentlichungen zum Thema Assurance Cases, die im Rahmen des Projekts entstanden und auf der Projekt-Website zu finden sind:
Michael Kläs, Rasmus Adler, Ioannis Sorokos, Lisa Joeckel und Jan Reich (2021a): Handling Uncertainties of Data-Driven Models in Compliance with Safety Constraints for Autonomous Behaviour. In: European Dependable Computing Conference (EDCC).

Michael Kläs, Rasmus Adler, Lisa Jöckel, Janek Groß und Jan Reich (2021b): Using Complementary Risk Acceptance Criteria to Structure Assurance Cases for Safety-Critical AI Components

Gefördert durch:

DENK FABRIK DIGITALE ARBEITS-GESELLSCHAFT im Bundesministerium für Arbeit und Soziales

Im Rahmen des:

ki_o Observatorium Künstliche Intelligenz in Arbeit und Gesellschaft

Impressum

Stiftung Neue Verantwortung e.V.
Beisheim Center
Berliner Freiheit 2
10785 Berlin

T: +49 (0) 30 81 45 03 78 80
F: +49 (0) 30 81 45 03 78 97

www.stiftung-nv.de
info@stiftung-nv.de

Design:
Make Studio
www.make-studio.net

Layout:
Jan Klöthe

Grafik:
Anne-Sophie Stelke