Fraunhofer
IPA

FRAUNHOFER-INSTITUT FÜR PRODUKTIONSTECHNIK UND AUTOMATISIERUNG IPA

XINYANG WU | MOHAMED EL-SHAMOUTY | PHILIPP WAGNER

WHITE PAPER: ZUVERLÄSSIGE KI

KI IN SICHERHEITSKRITISCHEN INDUSTRIELLEN ANWENDUNGEN EINSETZEN

HRSG.: THOMAS BAUERNHANSL | MARCO HUBER | WERNER KRAUS

KI Fortschrittszentrum
LERNENDE SYSTEME

Wu, Xinyang; El-Shamouty, Mohamed; Wagner, Philipp
Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA

WHITE PAPER: ZUVERLÄSSIGE KI

KI-Modelle für sicherheitskritische industrielle Anwendungen

Herausgeber
Thomas Bauernhansl, Marco Huber, Werner Kraus

Gefördert durch
Baden-Württemberg
MINISTERIUM FÜR WIRTSCHAFT, ARBEIT UND WOHNUNGSBAU

CyberValley

S·TEC

VORWORT

Künstliche Intelligenz (KI) ist eine der zentralen Technologien für die Zukunft. Ihre Einführung und der Einsatz fordern Unternehmen im besonderen Maß heraus. Es gilt, das Potenzial zu erkennen und dieses wirtschaftlich nutzbar zu machen. Lassen Sie sich dabei durch Europas größte Forschungskooperation auf dem Gebiet der KI, Cyber Valley, begleiten.

Mit dem KI-Fortschrittszentrum von Fraunhofer IAO und Fraunhofer IPA unterstützen wir Unternehmen dabei, das Potenzial von KI nutzbringend einzusetzen. An der Schnittstelle zwischen anwendungsorientierter Wirtschaft und exzellenter Forschung des Cyber-Valley-Konsortiums entwickeln wir innovative KI-Anwendungen für die Praxis und treiben damit die Kommerzialisierung von KI voran. Erklärtes Ziel ist dabei, menschzentrierte KI-Lösungen zu entwickeln.Denn nur wenn Menschen mit einer neuen Technologie intuitiv interagieren und vertrauensvoll zusammenarbeiten, kann ihr Potenzial optimal ausgeschöpft werden.

Die Studienreihe »Lernende Systeme« des KI-Fortschrittszentrums gibt Einblick in die Potenziale und die praktischen Einsatzmöglichkeiten von KI. Dabei werden übergreifende Themen wie Zuverlässigkeit, Erklärbarkeit (xAI), cloudbasierte Plattformen, Technologien und Einführungsstrategien diskutiert. Zudem werden einzelne Anwendungsbereiche in der Wissensarbeit, Bauwirtschaft, Produktion und dem Kundenservice im Detail beleuchtet.

Das vorliegende White Paper betrachtet die Zuverlässigkeit von KI-Modellen hinsichtlich der Aspekte Zertifizierbarkeit und Transparenz. Diese werden als entscheidend angesehen, um KI in die Praxis zu bringen und die aktuell noch vorhandene Lücke zwischen Forschung und Industrie zu schließen. Für beide Aspekte stellt das White Paper eine Recherche sowie Zusammenfassung der aktuell verbreiteten und eingesetzten Algorithmen und Methoden dar. Außerdem zeigt es beispielhafte Anwendungsfälle. Ziel des White Papers ist es, die LeserInnen beim Einsatz zuverlässiger KI-Modelle in der Industrie zu unterstützen.

Wir wünschen Ihnen eine spannende Lektüre, und freuen uns, wenn wir in Zukunft auch Sie mit unserer Expertise auf Ihrem Weg zur menschzentrierten KI unterstützen dürfen.

Thomas Bauernhansl, Marco Huber, Werner Kraus
Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA

INHALT

MANAGEMENT SUMMARY

In den vergangenen Jahren ist reichlich zum Thema Künstliche Intelligenz (KI) geforscht worden. Viele KI-Methoden besitzen ein hohes Potenzial für verschiedene industrielle Anwendungen, darunter die Qualitätskontrolle, vorausschauende Wartung oder die Optimierung von Produktionsketten. Allerdings existiert immer noch eine Kluft zwischen Forschung und industrieller Anwendung, weil die eingesetzten KI-Modelle für sicherheitskritische Anwendungen häufig nicht zuverlässig genug sind. Meist sind sie weder verifizierbar noch für die Menschen erklärbar. Um diese Kluft zwischen Forschung und Industrie zu schließen, diskutiert dieses White Paper die Aspekte *Zertifizierung* und *Transparenz* mit dem Ziel, zuverlässige KI-Systeme zu ermöglichen. Die Autoren beschreiben ein System, das Zertifizierungs- und Transparenzprozesse unterstützt, um Sicherheitsnachweise erbringen zu können. Ferner beschreiben die Autoren einige bestehende Methoden, die Garantien für Sicherheitsnachweise geben können. Abschließend diskutieren sie bestehende Pipelines, um eine standardisierte Anwendung von zuverlässiger KI im industriellen Kontext zu erreichen.

Schlüsselbegriffe: Zuverlässigkeit, Zertifizierung, Transparenz, Künstliche Intelligenz

1 EINFÜHRUNG

Künstliche Intelligenz (KI) und Maschinelles Lernen (ML) sind eine Kerntechnologie des 21. Jahrhunderts und werden mannigfaltige Bereiche in Gesellschaft, Forschung und Industrie beeinflussen. Bis etwa zum Jahr 2010 basierten die meisten KI-Ansätze im industriellen Umfeld auf Experten- oder Regelsystemen [1, 2]. Insbesondere für Aufgaben, die Sicherheitsgarantien und Zertifizierung erforderten, wurden nur sehr wenig datengetriebene Ansätze genutzt.

Dies hat hauptsächlich zwei Gründe. Zum einen waren datengetriebene Methoden für die meisten Szenarien noch nicht leistungsstark genug und konnten deshalb nicht genutzt werden, um bestimmte Aufgaben verlässlich auszuführen. Zum anderen haben Experten- und regelbasierte Systeme viele wünschenswerte Eigenschaften, wenn es um die Themen Zertifizierung und Transparenz für Entwickler und Nutzende dieser Systeme geht. Expertensysteme sind einfach zu verstehen und ein Mensch kann die meisten Entscheidungen solcher Systeme prüfen. Dieses Verhalten ermöglicht, die gewünschten Eigenschaften des Systems exakt testen und verifizieren zu können, was Grundlage für eine Zertifizierung ist. Das Problem mit Expertensystemen ist allerdings, dass sie auf speziellem Domänenwissen aufbauen, das während der Entwicklung in das System eingebaut wird. Ansätze dieser Art werden also problematisch, wenn sie in Bereichen wie Bildverarbeitung oder Robotersteuerung genutzt werden. Denn hierfür hat der Mensch meist keine konkreten Problemlösungsstrategien, die er ins System implementieren könnte. Aus diesem Grund sowie wegen der voranschreitenden Digitalisierung, die Unmengen an Daten und Rechenpower für deren Verarbeitung bereithält, werden stark datengetriebene Ansätze immer wichtiger. Diese Ansätze können vollständig mithilfe von Daten trainiert werden und kommen meist gänzlich ohne das explizite Integrieren von domänenspezifischen Regeln in die Algorithmen aus.

Im Jahr 2011 rief die deutsche Bundesregierung die »Industrie 4.0« aus. Sie wurde die wichtigste Strategie, um die starke Wettbewerbsfähigkeit und die hohen Qualitätsstandards der hiesigen Industrie zu bewahren. Von den verschiedenen Teilaspekten der Industrie 4.0 ist die Digitalisierung ein Schlüsselkonzept, das mehr Automatisierung in der (Massen-) Produktion erlaubt – gemeinsam mit KI und im Besonderen deren Teilgebiet des Maschinellen Lernens als Schlüsseltechnologie.

Dieses White Paper verwendet den Begriff des ML-Systems im Kontext mit datengetriebenen Ansätzen. Im Speziellen liegt der Fokus auf überwachten ML-Methoden, einer Subkategorie des Maschinellen Lernens, in dem ein Modell mithilfe von annotierten Daten trainiert wird. In Anwendungen wie der vorausschauenden Wartung, Qualitätskontrolle oder Intralogistik können ML-Systeme genauere Vorhersagen und objektivere Entscheidungen treffen, was sowohl Zeit als auch Kosten spart. Insgesamt lassen sich die Vorteile von ML in industriellen Anwendungen in die zwei Kategorien einteilen:

- Optimierungen in allen Aspekten der Wertschöpfungskette: beispielsweise Produktdesign, Qualitätskontrolle, vorausschauende Wartung, Mensch-Roboter-Kooperation, Lagerverwaltung, Bestandsmanagement oder Produktmarketing, um nur einige Beispiele zu nennen.
- Erweiterung bestehender oder Entwicklung neuer Produkte, Dienstleistungen, Märkte oder Geschäftsmodelle: Beispielsweise wurde die autonome Mobilität, welche nur mithilfe von KI-Methoden umgesetzt werden kann, zu einem der vielversprechendsten Märkte in den vergangenen Jahren.

Allerdings bleibt eine der wichtigsten Hürden, um ML in der Industrie einzusetzen, deren Zuverlässigkeit. Wie das Beispiel der Automobilindustrie mit autonom fahrenden Automobilen zeigt, erfordert der Übergang vom Fahrerassistenzsystem, bei dem der Fahrer die Kontrolle immer behält, zum autonomen Fahren, bei dem der Fahrer die Kontrolle an das Auto abgibt, Sicherheitsgarantien des Systems. Ein anderes Beispiel sind die Sicherheitsanforderungen in Anwendungen mit Mensch-Roboter-Kollaboration, in denen der Roboter garantiert sicher in der Nähe des Menschen arbeiten muss. (S1)

Zwei Schlüsselfaktoren sind notwendig, damit ein ML-System zuverlässig in der Industrie einsetzbar ist: Erstens die *Zertifizierung*, das heißt, das ML-System sollte von einer entsprechenden Zertifizierungsstelle zertifiziert sein. Der Zertifizierungsprozess findet während der Entwicklungsphase des Produkts statt, in der eine Zertifizierungsstelle, wie z. B. der TÜV Süd oder die DEKRA, sicherstellt, dass das Produkt die festgelegten High-Level-Anforderungen erfüllt.

Zweitens: *Transparenz*, das heißt, die Beschaffung der Daten, die zum Aufbau des ML-Systems verwendet werden, das ML-System selbst und das Geschäftsmodell dahinter sollten nicht nur für die Entwickler, sondern auch für die Nutzenden transparent gehalten werden [3]. Dies ist relevant, wenn bei Zwischenfällen die Gründe für eine unerwartete Verhaltensweise der KI dargelegt werden müssen. Diese beiden Faktoren sind an die in [4] eingeführte Definition von »Vertrauenswürdigkeit«[1] angelehnt:

1 Dieses Paper erweitert die Definition von Vertrauenswürdigkeit in [4] um Zuverlässigkeit, indem Transparenz aus [3] ergänzt wird.

Zuverlässigkeit = Zertifizierung + Transparenz

Um die Zertifizierungs- und Transparenzprozesse zu adressieren, müssen Entwickler eines ML-Systems darlegen, dass ihr System sicher ist, indem sie eine Sicherheitsargumentation bereitstellen. Eine solche Argumentation kann einem Audit-Team zur Verfügung gestellt oder zu einem Benutzerleitfaden erweitert werden, um die Transparenz zu erhöhen.

A Sicherheitsargumentationen

Es existieren zahlreiche Ansätze, um Techniken zu entwickeln, die Argumentationsmuster bezüglich der Sicherheitsvoraussetzungen von ML-Systemen und Methoden zum Nachweis dieser Voraussetzungen definieren (z. B. in [5, 6] sowie eine Übersicht in [7]). Hiervon hebt dieses White Paper die sogenannte AMLAS-Methodik hervor (»Assurance of Machine Learning for use in Autonomous Systems« [6]). AMLAS bietet eine Systematik, um Sicherheitsargumente für ML-Systeme zu erstellen und die Sicherheitsgarantie in den Entwicklungszyklus von ML-Systemen zu integrieren, wie die Abb. S2 zeigt. AMLAS schlägt sechs Stufen vor. Diese bilden definierte Sicherheitsanforderungen auf Systemebene für spezifische Sicherheitsfälle oder -ereignisse für jede Komponente im System ab:

- **Stufe 1** Geltungsbereich für ML-Sicherheit: Dieser definiert den Umfang des Prozesses zur Sicherheitsargumentation und die Sicherheitsfälle für die ML-Komponente. Der relevante Kontext der Sicherheitsargumentation wird spezifiziert basierend auf den Sicherheitsanforderungen.

***Fig. S1**. (Links) Arbeitsplätze der Zukunft am Fraunhofer IPA. (Rechts) Ein intelligenter Sicherheitssensor (PILZ SafetyEYE ©) erkennt, ob sich ein Mensch in der Roboterzelle befindet. Bildquelle: Fraunhofer IPA, Universität Stuttgart IFF/Foto: Rainer Bez.*

- **Stufe 2** Sicherheit in den ML-Anforderungen: Diese Stufe bestimmt und validiert die ML-Sicherheitsanforderungen aus den zugewiesenen Systemsicherheitsanforderungen, wie beispielsweise für die Identifizierung von Menschen aus Bildern mit Begrenzungsrahmen: »Die Begrenzungsrahmen dürfen in keiner Dimension mehr als zehn Prozent größer sein als der kleinste Rahmen, der den gesamten Fußgänger einschließen kann.« [8]
- **Stufe 3** Sicherheit im Datenmanagement: Dieser Punkt stellt sicher, dass die während des Trainings bereitgestellten Daten die Sicherheitsanforderungen beinhalten. Dies wird während des Lernprozesses des ML-Systems mithilfe zusätzlicher Daten, wie Entwicklungs- und Verifikationsdaten, sichergestellt.
- **Stufe 4** Sicherheit im Modell-Lernprozess: Hier werden die Entwicklung von ML-Modellen und deren Validierung durchgeführt. Dabei wird überprüft, ob die entwickelten Modelle in Bezug auf die definierten Sicherheitsanforderungen leistungsfähig genug sind.
- **Stufe 5** Sicherheit in der Modellverifikation: In diesem Schritt wird sichergestellt, dass das ML-Modell die definierten Sicherheitsanforderungen erfüllt, wenn es Daten nutzt, die während der Modellentwicklungsphase nicht verfügbar waren.
- **Stufe 6** Sicherheit in der Modellbereitstellung: Diese letzte Stufe stellt sicher, dass das Zielsystem mit all seinen integrierten Komponenten die definierten Systemsicherheitsanforderungen vor und während des Betriebs in der Zielumgebung weiterhin erfüllt.

In seiner aktuellen Form benennt AMLAS einige State-of-the-Art-Methoden, die das Bereitstellen von Belegen für die Sicherheitsargumentation unterstützen. Eine vollständige Umsetzung von AMLAS auf einen konkreten Anwendungsfall ist jedoch bis zum Zeitpunkt der Erstellung dieses White Papers noch nicht öffentlich verfügbar. Daher hebt diese Arbeit einige der erwähnten State-of-the-Art-Methoden hervor, die nutzbar sind, um Belege für einige Stufen von AMLAS erstellen zu können. Dann werden diese Methoden mit der Sicherheitsargumentation

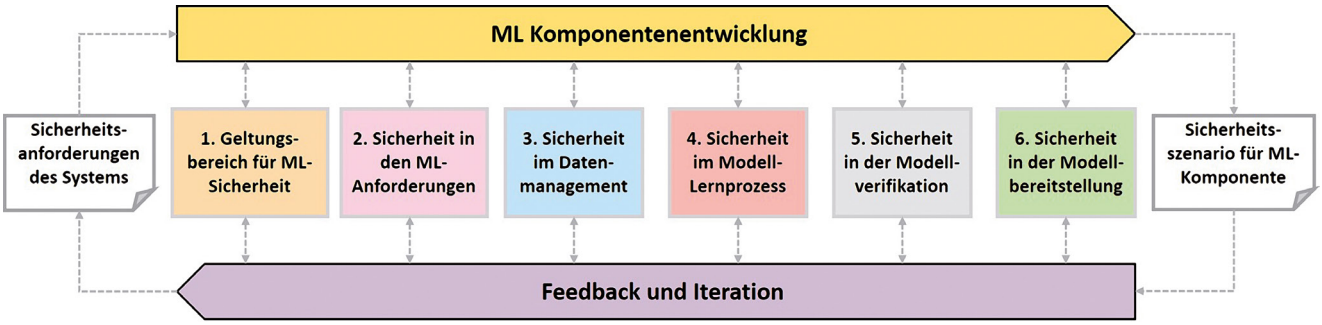

Fig. S2. *Übersicht über den AMLAS-Prozess für den Entwicklungszyklus von ML-Systemen (entnommen aus [6]).*

verbunden und schließlich beides im Rahmen der Zertifizierungs- und Transparenzprozesse als Schritte zur Erreichung von verlässlicher KI verknüpft. Für die Transparenz folgen die Autoren den ethischen Richtlinien, die von der High-Level Expert Group on AI (AI HLEG) der EU-Kommission veröffentlicht wurden [3]. Bei der Zertifizierung werden fünf technische Aspekte berücksichtigt, um die Konstruktion von Sicherheitsfällen und die Bereitstellung von Nachweisen zu unterstützen:

- **Erklärbare KI** (auch Explainable AI oder xAI) versucht, interpretierbare Modelle zu erstellen und Erklärungen für komplexe ML-Modelle zu generieren. Dies geschieht, um ML-Modelle für Menschen verständlich sowie den Einsatz von ML-Modellen vertrauenswürdig zu machen und schließlich die Überprüfung dieser zu vereinfachen.
- Die **formale Verifikation** legt fest, ob bestimmte Eigenschaften für ein vorhandenes Modell innerhalb eines gegebenen Eingabebereichs gelten. Wenn eine Eigenschaft nicht zutrifft, geben die meisten formalen Verifikationsalgorithmen Gegenbeispiele zurück; wenn sie zutrifft, wird das Ergebnis häufig durch einen mathematischen Beweis ergänzt. Während formale Verifikationsmethoden an sich vielversprechend sind, mangelt es ihnen noch häufig an Skalierbarkeit und Kompatibilität mit gängigen ML-Modellarchitekturen.
- **Statistische Validierung** ist eine Alternative und Ergänzung zu formalen Verifikationsmethoden. Sie zielt darauf ab, das ML-System mit einer großen Menge von Fällen (empirisch) zu testen. Eine höhere Fallabdeckung bedeutet, dass mehr vom Verhalten des Modells getestet wird und das Modell robuster in seinen Entscheidungen ist. Diese Methoden sind rechnerisch weniger aufwendig als formale Verifikationsmethoden und können in vielen praktischen Situationen ausreichen. Es muss jedoch sichergestellt werden, dass sie auch Randphänomene abdecken.
- **Unsicherheitsquantifizierung** gibt an, ob die Vorhersagen von ML-Modellen vertrauenswürdig und zuverlässig sind [9]. Übliche ML-Modelle liefern lediglich Punktschätzungen als Vorhersagen, ohne die Vorhersage zu begründen, beispielsweise durch zusätzliche Informationen wie Konfidenzintervalle oder Wahrscheinlichkeiten. In der Literatur werden einige Ansätze vorgeschlagen, damit ML-Modelle die Konfidenz ihrer Vorhersage einschätzen können. Hierfür kommen Methoden aus der Bayes'schen Statistik-Theorie zum Einsatz, die zu einer Wahrscheinlichkeitsverteilung über Modellparameter und Vorhersagen führen.
- **Online Monitoring mit Randbedingungen** erkennt, wenn sich ein ML-System während der Laufzeit falsch verhält. Dazu gehören Fehler, bei denen die Systementscheidung außerhalb spezifizierter Sicherheitseinschränkungen liegt, sowie die Erkennung von Fehlern außerhalb der Verteilung mithilfe von Unsicherheitsquantifizierung.

B Auditing

Unter Auditierung versteht man die Prüfung, ob ein Prozess oder Verfahren bestimmte, in einer Norm formulierte Anforderungen erfüllt. Die Prüfung übernehmen interne oder externe Auditoren, wie zum Beispiel der TÜV Süd oder die DEKRA. Sie erfolgt in der Regel anhand eines Kriterienkatalogs, der aus den zugrundeliegenden Normen abgeleitet wird. Für die Auditierung von KI-Systemen oder KI-basierten Produkten in der industriellen Anwendung gibt es bisher jedoch keinen einheitlichen Standard.

AMLAS ist nicht rechtsverbindlich, sondern eine vorgeschlagene Methode, um die Argumentation für die Sicherheit zu liefern. Letztendlich muss die Auditierung für jeden spezifischen Anwendungsfall durchgeführt werden, um festzustellen, welche Bedingungen für die Zertifizierung oder die Transparenzprozesse des jeweiligen ML-Systems erfüllt sein müssen [10].

C Ziel

Das Ziel dieses White Papers ist es, einen anwendungsorientierten Überblick über ML in sicherheitskritischen Domänen der Industrie zu geben. Aus der Perspektive verschiedener Sicherheitsfälle definieren wir den Umfang der Verlässlichkeit von KI anhand der Aspekte Zertifizierung und Transparenz. Dementsprechend stellt die Arbeit in den Abschnitten 2 und 3 eine nicht vollständige Liste bestehender Methoden vor. Diese unterstützen das Erbringen von Nachweisen für den Aufbau der Sicherheitsfälle sowie die Zertifizierungs- und Transparenzprozesse. Das Paper bietet keine umfangreichen Beschreibungen der Methoden, sondern bietet lediglich einen kurzen Überblick und verknüpft die genannten Methoden mit der Bereitstellung von Sicherheitsnachweisen, um zuverlässige ML-Systeme zu erarbeiten. Abschnitt 4 skizziert, wie eine ML-Pipeline mit dem Fokus auf verlässliche KI im Kontext von industriellen sicherheitskritischen Anwendungsfällen aussehen könnte.

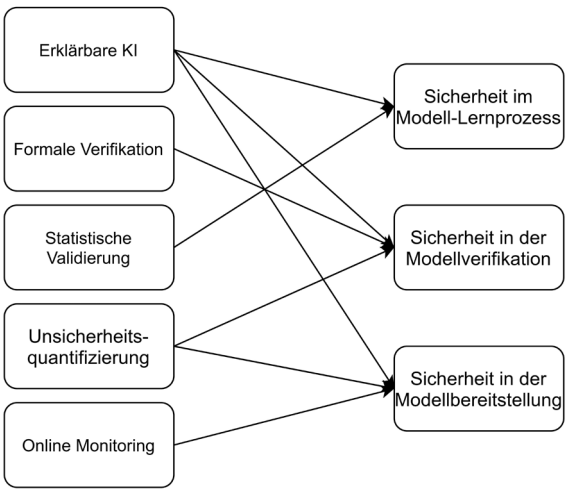

Fig. S3.
Berücksichtigte Methoden und deren Beitrag, um Argumente zu formulieren und Evidenzen hinsichtlich der AMLAS-Stufen zu bieten. Bildquelle: Fraunhofer IPA/Bild: Xinyang Wu.

2 ZERTIFIZIERUNG

Dieses Kapitel diskutiert kurz mehrere Forschungsrichtungen, die die Herausforderungen bei der Zertifizierung von ML-Systemen und das Erfüllen von Sicherheitsanforderungen beschreiben.

A Erklärbare KI

ML ermöglicht, hochpräzise Modelle für verschiedene Anwendungsfälle zu erstellen. Allerdings sind die von ML-Modellen gelernten Zusammenhänge recht komplex und abstrakt, sodass Menschen diese meist nicht nachvollziehen können. Diese nicht interpretierbaren ML-Modelle werden auch als Black-Box-Modelle bezeichnet. In vielen Anwendungsfällen, vor allem in sicherheitskritischen Bereichen in der Industrie, sind allerdings nicht nur die Genauigkeit der Vorhersagen, sondern auch die Transparenz und das Vertrauen in das System von großer Bedeutung. Erklärbare KI (xAI) versucht, Interpretierbarkeit und Erklärbarkeit für Black-Box-Modelle[2] zu erreichen, um Nutzenden Einblicke in den Entscheidungsprozess zu geben. Solche Einsichten vereinfachen die Argumentation für den Einsatz des ML-Modells und ermöglichen die Wartbarkeit. Daher ist erklärbare KI als ein Schlüsselelement zu sehen, um die in Abschnitt 1.A definierten Phasen des Modelllernens, der Modellverifikation und der Sicherheitsgarantie für den Modelleinsatz zu unterstützen.

Dieses White Paper definiert xAI als ein Forschungsfeld, das darauf abzielt, die Ergebnisse von ML-Systemen für den Menschen verständlicher zu machen [11, 63]. In diesem Abschnitt werden zunächst einige einfache ML-Modelle vorgestellt, die selbst auf einer modularen Ebene erklärbar und verständlich sind und daher als White-Box-Modelle gelten können. Danach werden einige populäre modellagnostische und modellspezifische Methoden für Black-Box-Modelle diskutiert. Der Fokus liegt dabei auf der Methode des Deep Learning mit neuronalen Netzen (NN), weil diese aktuell am stärksten verbreitet sind. Die Konzepte sind aber auch auf andere ML-Modelle anwendbar. Unter jeder Kategorie wird eine nicht vollständige Liste von bestehenden Methoden vorgestellt:

2 Eigentlich sind ML-Modelle keine echte Black-Box, da sie untersucht werden können und alle Berechnungen transparent sind, während echte Black-Boxen nicht untersucht werden können und geschlossene Systeme sind. Bei komplizierten ML-Modellen wie tiefen NN ist jedoch die Anzahl der Parameter und Berechnungen so groß, dass kein Mensch in der Lage ist, alle Berechnungen in angemessener Zeit zu verfolgen. Dieser Aspekt wird als Simulierbarkeit bezeichnet und in [65] diskutiert.

- **Erklärbare Modelle:** Während neuronale Netze oder Ensemble-Methoden wie Random Forests immer als Black-Box betrachtet werden, existieren auch relativ einfache und damit erklärbare[3] ML-Modelle. So können beispielsweise numerische Koeffizienten für verschiedene Merkmale in linearen oder logistischen Regressionsmodellen als Indizes für die Wichtigkeit der einzelnen Merkmale verwendet werden. In Entscheidungsbaummodellen liefert das Traversieren vom Wurzelknoten zu einem Blattknoten eine Vorhersage, die dadurch erklärbar wird, dass man den Entscheidungspfad in transparente Entscheidungsregeln zerlegen kann. Wie wichtig ein Merkmal in Entscheidungsbaummodellen ist, kann darüber hinaus dadurch gemessen werden, dass die Varianz der Ausgabe (für Regression) oder der Gini-Index der Klassenverteilung (für Klassifikation) durch den Knoten, der dieses Merkmal verwendet, im Vergleich zu seinem Elternknoten verringert wird [12]. Obwohl ein Mensch die oben erwähnten Modelle verstehen und nachvollziehen kann, ist der Nachteil dieser Modelle relativ deutlich. Nur eine begrenzte Anzahl von ML-Modellen ist ausreichend erklärbar, was die Auswahl der Modelle in der Anwendung einschränkt. Außerdem ist die Leistungsfähigkeit aufgrund der einfachen Architektur und der begrenzten Ausdruckskraft von White-Box-Modellen verglichen mit Black-Box-Modellen meist geringer.

- **Modell-agnostische Methoden:** Um Erklärbarkeit für ein generisches ML-Modell unabhängig von seiner Architektur und dem zugrundeliegenden Algorithmus zu erreichen, wurden in der Literatur mehrere sogenannte modellagnostische xAI-Methoden vorgeschlagen. Zum Beispiel schlägt [13] partielle Abhängigkeitsdiagramme vor, um den marginalen Effekt jedes Merkmals auf die Vorhersagen eines ML-Modells zu zeigen. [14] schlägt eine Methode namens LIME (local interpretable model-agnostic explanations) vor, die sich auf das Training sogenannter lokaler Surrogatmodelle konzentriert. Diese Modelle sind dann interpretierbar und können lokal verwendet werden, um eine einzelne Entscheidung zu erklären. In [15] wird eine Methode vorgeschlagen, die auf der Berechnung von Shapley-Werten [16] basiert, die aus der mathematischen Spieltheorie bekannt sind. Diese Methode namens SHAP gibt die Merkmalsbedeutung für eine gegebene Modellentscheidung aus. [17] schlägt eine Methode namens Anchors vor, die lokale Entscheidungsregeln für jede Vorhersage extrahiert.

- **Modellspezifische Methoden:** Viele verfügbare Erklärungsmethoden konzentrieren sich auf spezifische Modellstrukturen, um entweder die Qualität der Erklärung oder die Laufzeitleistung zu verbessern. Basierend auf SHAP schlagen [18, 19] beispielsweise Tree SHAP vor, einen effizienteren Ansatz, um Shapley-Werte für baumbasierte Modelle zu schätzen, sowie Deep SHAP, das Gradienten für Erklärungen von tiefen NN nutzt. DeepLIFT [20] ist eine rekursive Prädiktionserklärungsmethode für tiefe Modelle. [21] erklärt tiefe NN mit einer Entscheidungsbaumdarstellung, indem jede Schicht des Netzwerks zerlegt wird. Ebenso wurde

3 In diesem White Paper werden die Begriffe »erklärbar« und »interpretierbar« synonym genutzt.

die Information-Bottleneck-Methode (IB) aus der Informationstheorie angewandt, um die Black-Box von ML-Modellen zu öffnen. Daraus ergibt sich die Informationsfluss-Perspektive. [22] erklärt den Trainingsvorgang eines tiefen NN, indem es den Fluss der gelernten Informationen in verschiedenen Schichten visualisiert. Basierend auf dieser Arbeit fügt [23] unabhängiges Gauß'sches Rauschen hinzu, um die Leistung zu erhöhen.

Hinsichtlich des Erklärungsumfangs können die oben genannten Methoden in die folgenden drei Kategorien eingeteilt werden [3]:

- **Instanz-weise Erklärung:** Sie zielt darauf ab, die von einem NN auf einer einzelnen gegebenen Eingabe getroffene Entscheidung zu erklären, zum Beispiel LIME, SHAP und DeepLIFT.
- **Modell-Erklärung:** Ermöglicht ein besseres Verständnis des gesamten ML-Modells, zum Beispiel Regel-Extraktion [17] und Entscheidungsbaum-Extraktion [21].
- **Erklärung des Informationsflusses:** Ziel ist es, informationstheoretische Methoden zu verwenden, um den Trainingsvorgang zu erklären, zum Beispiel [22] und [23].

Beispiel für einen Anwendungsfall von xAI: Ein Hersteller von Schüttgut nutzt Anlagen, um das Material durch die Fabrikhalle zu transportieren. Abhängig von verschiedenen Bedingungen (Temperatur, Feuchtigkeit, Materialqualität) verhält sich das Schüttgut leicht unterschiedlich. Um einen nahezu optimalen Materialfluss durch die Schüttgutanlagen zu gewährleisten, wird ein NN trainiert, das deren Parametrierung in Echtzeit steuert. Das NN wird mit Sensordaten trainiert, die Druck, Temperatur, Spannung und Durchsatz an verschiedenen Stellen der Schüttgutanlagen messen. xAI-Methoden, die auf die Berechnung der Merkmalsbedeutung anwendbar sind, können abschätzen, welche Bedeutung die verwendeten Sensoren in verschiedenen Datenregimen haben. Wenn sich das NN überhaupt nicht auf einen installierten Sensor verlässt, kann dies an der Merkmalsbedeutung abgelesen werden und der Sensor kann eventuell entfernt werden, um Kosten und Systemkomplexität zu reduzieren. Die Merkmalsbedeutung kann auch vor und nach dem erneuten Training verglichen werden, um sicherzustellen, dass sich das gelernte Modell immer noch auf ähnliche Sensordaten konzentriert.

Erklärbarkeit von tiefen NN bleibt ein herausforderndes und aktives Forschungsgebiet, und es gibt noch keine funktionierenden Out-of-the-Box-Lösungen. Die große Bandbreite an Methoden sowie die fehlende Standardisierung von Frameworks und Benchmarking für die praktische Umsetzung erschweren die Anwendung von xAI in der Industrie. Eine genauere Untersuchung

des aktuellen Stands der Technik im Bereich xAI sowie eine ausführliche Diskussion verfügbarer xAI-Frameworks findet sich in der Forschungsstudie zu xAI: »Erklärbare KI in der Praxis: Anwendungsorientierte Evaluation von xAI-Verfahren«[4] des Fraunhofer IPA.

B Formale Verifikation

Konventionelle Verifikationstechniken wurden erfolgreich eingesetzt, um die Zuverlässigkeit von realen Software- und Hardwaresystemen zu prüfen (z. B. [24]). Solche konventionellen Techniken sind jedoch nicht auf ML-fähige Systeme anwendbar, weil deren internes Verhalten sehr komplex und die Systemzusammenhänge nicht trivial und schwer zu spezifizieren sind [25, 26].

Die formalen Verifikationstechniken für tiefe NN hingegen sind darauf zugeschnitten, die Verifikation von tiefen NN zu unterstützen und solide Beweise für das tiefe NN in Bezug auf spezifizierte Eigenschaften zu liefern. Solche Beweise können verwendet werden, um eine Argumentation im Rahmen der in Abschnitt 1.A definierten Sicherheitsstufe der Modellverifikation aufzubauen.

Verfahren zur formalen Verifikation von tiefen NN bestimmen, ob eine Eigenschaft für ein tiefes NN gilt, beispielsweise lokale Robustheit gegenüber kleinen Störungen der Eingaben des NN. Wenn diese Eigenschaft gilt, kann ein mathematischer Beweis erbracht werden. Wenn die Eigenschaft nicht gilt, kann ein Gegenbeispiel geliefert werden, das heißt, eine Eingabe mit der gegebenen Störung, bei der das tiefe NN seine Entscheidung sprunghaft ändert. Wenn eine deterministische Antwort rechnerisch nicht machbar ist, kann eine Antwort mit einer Fehlergrenze geliefert werden, die in vielen praktischen Szenarien ausreichen kann. Andernfalls wird das Online-Monitoring nicht verifizierter Regionen verwendet, das in Abschnitt 2.E näher erläutert wird.

Verifikationsverfahren erfordern die Spezifikation der zu verifizierenden Eigenschaften und die mathematische Formulierung des Systems, die die Verifikation dieser Eigenschaften ermöglicht. Die Spezifikation von Eigenschaften auf nicht-trivialen Eingaberaum (zum Beispiel für Bilder oder Text) ist bis heute eine offene Frage. Die Problematik ist hierbei ähnlich gelagert wie im Gebiet xAI.

Es bleibt festzuhalten, dass ein Hauptergebnis von ML-Systemen ist, Muster aus hochdimensionalen Daten zu extrahieren und Funktionen auf der Grundlage dieser Muster zu lernen. Das bedeutet, dass wir, um Verifikationseigenschaften zu extrahieren und ein gegebenes ML-System und Trainingsdaten zu verifizieren, ein anderes ML-System benötigen, das solche Eigenschaften aus den bereitgestellten Daten extrahiert. Wie kann dann das ML-System verifiziert werden, das

4 https://www.ki-fortschrittszentrum.de/de/themen/studien/erklaerbare-ki-in-der-praxis.html

die formalen Verifikationseigenschaften des anderen getesteten ML-Systems extrahiert? Um dies zu ermöglichen, definiert die Forschungsgemeinschaft eine Reihe von Eigenschaften, die systemunabhängig sind, wie zum Beispiel lokale, globale und probabilistische Robustheit gegenüber möglichen gegnerischen Angriffen unter einem gegebenen Angriffsmodell [27]. Ein gängiges Angriffsmodell ist eine Abstandsmetrik l_p zwischen verschiedenen Eingängen oder Ausgängen. Gängige Metriken sind l_1, l_2 und l_∞ [28]. Lokale Robustheit wird basierend auf der Eingabe *x* definiert. Diese besagt, dass ein tiefes NN die gleiche Entscheidung für alle x^0 ausgeben sollte, die sich innerhalb eines Abstands *e* von *x* befinden. Wenn ein tiefes NN einen lokalen Robustheitstest besteht, bedeutet dies, dass das tiefe NN für lokale Robustheit mit Angriffen nach einem bestimmten Bedrohungsmodell um die getesteten Eingabedaten zertifiziert ist. Lokale Robustheit bietet jedoch keine Garantien für Eingaben, die während des Tests nicht berücksichtigt werden. Zugleich verlangt die globale Robustheit, dass die lokale Robustheit für alle Eingaben im Eingaberaum erfüllt wird [27]. Die Autoren in [27] behaupten, dass lokale und globale Robustheitsanforderungen als zu hohe Anforderungen angesehen werden können, insbesondere in Szenarien ohne drohende Fremdeinwirkung. Sie führen einen probabilistischen Robustheitstest mit schwächeren, aber angeblich praktikableren Garantien ein, der – bei einer zugrundeliegenden Verteilung der Eingabedaten – probabilistisch die Lipschitz-Eigenschaft erfüllt. Diese besagt, dass der Abstand zwischen Ausgaben durch ein *k*-faches des Abstands zwischen den jeweiligen Eingaben begrenzt ist.

Dennoch befassen sich viele Methoden mit lokaler Robustheit, wie Reluplex [29], Marabou [30], Neurify [31] und CNNCertify [28]. Nur wenige Ansätze adressieren globale Robustheit, wie beispielsweise [32]. Die neueren Ansätze im Bereich der formalen Verifikation sind sehr vielversprechend und bieten Möglichkeiten, bestimmte Input-Output-Beziehungen auf eine fehlerfreie und/oder vollständige Weise zu verifizieren. Ein Algorithmus wird als fehlerfrei bezeichnet, wenn die vom Algorithmus gefundene Lösung korrekt ist; und er wird als vollständig bezeichnet, wenn der Algorithmus für jede Eingabe *x* eine Lösung liefert, sofern eine Lösung existiert. Dennoch mangelt es aktuellen Methoden an der Skalierbarkeit, wenn sie fehlerfreie und vollständige Ansätze bieten, und sie sind daher nicht in der Lage, die Verifikation für große Modelle und hochdimensionale Eingaben in einer angemessenen Rechenzeit durchzuführen. Zum Beispiel bieten Reluplex [29] und Marabou [30] fehlerfreie und vollständige Lösungen, aber mit begrenzter Skalierbarkeit auf die Größe der tiefen NN. Auf der anderen Seite bieten Neurify [31] und CNNCertify [28] skalierbare Lösungen, aber ohne die Garantie, fehlerfrei (CNNCertify [28]) oder vollständig (Neurify [31]) zu sein. Daher kann die formale Verifikation derzeit nur für sehr spezifische Teile des Datenregimes verwendet werden und die Zertifizierung der ML-Systeme kann sich nicht nur auf die formale Verifikation allein verlassen.

C Statistische Validierung

Die formale Verifikation erfordert eine formale Spezifikation des Systems, das heißt ein vollständiges Verständnis und eine Modellierung des Systems einschließlich seines Anwendungskontexts. Bei komplexen Systemen können nicht alle Eigenschaften konkret spezifiziert werden, um sie formal zu verifizieren. Zum Beispiel kann im medizinischen Bereich die Anwendbarkeit von Medikamenten zur Heilung einer Krankheit nicht zu 100 Prozent formal verifiziert und garantiert werden. Der verfolgte Ansatz in solchen Bereichen sind die statistische Auswertung und Validierung. Die statistische Validierung ist eine quantitative Methode, um die Effektivität eines Systems anhand einer definierten Menge von Metriken zu quantifizieren und zu argumentieren. Diese Methode ergänzt die formale Verifikation und wird verwendet, um eine Argumentation über die globale Robustheit des Systems zu erreichen.

In der ML-Community ist die statistische Validierung eine gängige Praxis, um die Leistung des ML-Modells in Bezug auf bestimmte Anforderungen wie Genauigkeit und Wiedererkennung und anhand von Testdaten zu validieren. Diese Testdaten sind unabhängig von den Trainingsdaten. Die Testdaten können auch verschiedene Arten von Fremdangriffen[5] enthalten, wie beispielsweise das Hinzufügen von Rauschen zu den Eingaben (erwähnt in Abschnitt 2 B), um die Robustheit des Netzwerks gegenüber solchen Angriffen zu testen. Statistische Validierungsmethoden bieten jedoch keine Garantien für die Leistung des tiefen NN. Dennoch sind statistische Validierungsmethoden nutzbar, um eine Argumentation unter einer in Abschnitt 1.A definierten Sicherheitsstufe für das Modelllernen aufzubauen.

D Unsicherheitsquantifizierung

Klassische NN schätzen in der Regel Punkte basierend auf der Maximum-Likelihood-Schätzung, ohne die Unsicherheit in den Daten und dem gelernten Modell zu berücksichtigen. So können beispielsweise Datenabweichungen das Verhalten von ML-Modellen unerwartet ändern [33]. In sicherheitskritischen Domänen sollte ein ML-Ansatz idealerweise ausgeben, wie sicher seine Entscheidung ist, um eine Risikobewertung zu ermöglichen. Der Mangel an Unsicherheit kann hohe Risiken mit sich bringen und schränkt somit die Anwendung von NN ein.

Um das Problem der Unsicherheitsquantifizierung anzugehen, wurden viele Ansätze aus dem Bereich des Bayes'schen Lernens vorgeschlagen. Bayes'sche Methoden bieten Möglichkeiten, Unsicherheiten mithilfe der Bayes'schen Regel und probabilistischem Denken zu quantifizieren, was zu einer Verteilung über Modellparameter und Vorhersagen anstelle von Punktschätzungen führt. Zum Beispiel schlägt [34] vor, Gauß'sche Prozesse zu verwenden, um Regressionsprobleme zu lösen. Eine Quantifizierung der Unsicherheit zeigt an, ob die Vorhersagen vertrauenswürdig und zuverlässig sind [9].

---

5 Englisch: Adversarial Attacks

Im Folgenden konzentriert sich dieses White Paper auf die Verwendung von Bayes'schen Methoden für NN, nämlich Bayes'sche NN (BNN), die die Parameter eines traditionellen NN als Wahrscheinlichkeitsverteilungen modellieren, die über die Bayes-Regel berechnet werden [35]. Diese Forschungsrichtung hat in den letzten Jahren viel Aufmerksamkeit erhalten und es existieren verschiedene populäre Methoden, um BNN zu trainieren. [36] wendet Dropout als Bayes'sche Approximation an, um Modellunsicherheiten zu generieren, während [37] ein Ensemble von Netzen mit unterschiedlichen Initialisierungen trainiert und Mittelwert und Standardabweichung des Ensembles nutzt, um Unsicherheiten zu bewerten. Eine weitere bekannte Idee für probabilistische Inferenz im Allgemeinen und für das Lernen von BNN im Besonderen ist das Markov-Chain-Monte-Carlo-Verfahren (MCMC) [38], das die Approximation von Wahrscheinlichkeitsintegralen mit der Monte-Carlo-Methode über Sampling aus einem Markov-Prozess erlaubt. Ein Nachteil der meisten MCMC-basierten Methoden ist jedoch der hohe Rechenaufwand.

Eine weitere beliebte Methode zum Trainieren von BNN ist die Variationsinferenz (VI) [39]. Sie führt eine einfache Verteilung wie die Normalverteilung ein, um die komplizierte A-posteriori-Verteilung der Gewichte des Netzes approximieren und somit die Berechnungen vereinfachen zu können. Da VI auf Optimierung beruht, kann dieses Vorgehen einfach Methoden wie den stochastischen Gradientenabstieg nutzen, um die Inferenz zu skalieren und zu beschleunigen; allerdings hat VI normalerweise eine geringere Genauigkeit im Vergleich zu MCMC. In einem neueren Ansatz werden BNN über sequenzielle Bayes'sche Filterung trainiert, ohne dass eine gradientenbasierte Optimierung notwendig ist [40]. Das Fehlen von Gradienten verbessert die Dateneffizienz, und die sequenzielle Filtermethode ermöglicht Online-Lernen. Die Unsicherheitsquantifizierung von NN bietet eine Lösung, um zeitnah Veränderungen in der Datenverteilung zu erkennen. Sie kann auch als doppelte Kontrolle in sicherheitskritischen industriellen Anwendungsfällen dienen, bei denen der Vorhersage von ML-Modellen nur dann vertraut wird, wenn das Modell diese Vorhersage sicher trifft. Die Quantifizierung der Unsicherheit kann die Argumentationsbildung unter den Sicherheitsstufen der Modellverifizierung und des Modelleinsatzes in Abschnitt 1.A unterstützen.

E Online-Monitoring

Die formale Verifikation und die statistische Validierung bewerten die Leistung des NN offline vor dem Einsatz. In vielen Fällen jedoch können aktuelle formale Verifikationsmethoden keine Lösung bieten. Daher ist eine Überwachung zur Laufzeit aus dreierlei Gründen erforderlich:

1. Um die Korrektheit der NN-Ausgaben während des Betriebs zu gewährleisten [41, 42].
2. Um den NN-Fehler für ungesehene Daten während des Einsatzes zu begrenzen [41].
3. Oder um dynamische Informationen darüber zu liefern, wie sich das tiefe NN aktuell verhält und ob sich das vom NN abhängige System von dem aktuellen Verhalten erholen kann [43].

Online-Monitoring kann bei der Argumentationsbildung im Rahmen der in Abschnitt 1.A definierten Sicherheitsstufe für den Modelleinsatz unterstützen.

Einige Ansätze kombinieren erfolgreich Online-Monitoring mit der Verifikation während der Modellentwicklung in der sicheren Roboterprogrammierung [44, 45] und in adaptiven Flugsteuerungssystemen [46]. Andere Ansätze entwickeln Laufzeitüberwachung unabhängig von der formalen Verifikation, zum Beispiel für die Korrektheit der menschlichen Posenschätzung [41], die Segmentierung medizinischer Daten [47] oder die Erkennung von Straßenschildern [48]. Online-Monitoring kann verschiedene Formen annehmen, wie für Eingaben, Ausgaben, sowohl Eingaben als auch Ausgaben oder für Fehler während des Einsatzes.

- **Online-Monitoring für Eingaben:** Eines der möglichen Risiken für die meisten industriellen Anwendungen ist die Datenabweichung , das heißt die Charakteristik der von einem Sensor gesammelten Daten kann sich im Laufe der Jahreszeiten, durch Maschinenabnutzung usw. verändern. [49, 50] beweisen, dass solche Out-of-Distribution-Fehler (OOD) Gefahren verursachen können, da klassische NN die variierende Verteilung der Daten nicht erkennen und trotzdem zuverlässige Vorhersagen machen können. In Kombination mit der in Abschnitt 2.D erläuterten Unsicherheitsquantifizierung können solche Konzeptabweichungen jedoch erkannt werden, indem die Unsicherheitsinformation als zusätzliche Metrik zur Überwachung der Vorhersagen von ML-Modellen eingesetzt wird. Der Vorhersage eines ML-Modells wird nur dann vertraut, wenn dem Modell auch in seiner Vorhersage vertraut wird, sodass OOD-Fehler vermieden werden können. Die Kombination der beiden Bereiche ist vielversprechend und es existieren viele Forschungstätigkeiten. Zum Beispiel schlägt [51] einen OOD-Ansatz vor, um NN mit probabilistischen Informationen robust gegen kleine Konzeptabweichungen zu machen, was auch in den Forschungsbereich der Autoren dieses White Papers fällt.

- **Online-Monitoring für Ausgaben:** Diese Art der Überwachung stellt sicher, dass die Ausgabe der NN korrekt ist. Das Online-Monitoring kann entweder modellbasierte oder lernbasierte Überwachung sein. Im Falle der lernbasierten Überwachung werden die Parameter zusammen mit oder unabhängig vom Training der tiefen NN gelernt. Bei modellbasiertem Monitoring definieren menschliche Experten die Randbedingungen anhand eines vorhandenen Modells der Arbeitsumgebung vor. Das Modell kann entweder erlernt oder vordefiniert sein.

Beispiel eines *Online-Monitoring für Steuerungen:* Derzeit sind die Ausgaben von NN explizite Funktionen in der Form $y = f(x)$, wobei x die Eingabe und y die Ausgabe ist. Die formale Verifikation für lokale Robustheit kann die Glattheit von $f(x)$ garantieren. Allerdings sind die Ausgaben y nicht sicher begrenzt. Eine Forschungsrichtung besteht darin, ein gewünschtes Verhalten hinzuzufügen, anstatt nur explizite Funktionen zu verwenden [52]. Gewünschtes Verhalten kann zum Beispiel in Form von differenzierbaren konvexen Optimierungsschichten am Ausgang von NN formuliert werden [53]. Eine Beispielformulierung kann sein, eine Ausgabe $\hat{y}$ zu wählen, die der aktuellen Ausgabe y des NN am nächsten kommt und zusätzlich die Sicherheitsbedingungen erfüllt. Dies führt zu der Formulierung

$$\min_{y} ||\hat{y} - y||^2$$
$$s.t.\, C \bullet \hat{y} + d \leq e\,,$$

wobei C, d und e die Parameter der Sicherheitsbedingungen sind. Diese besagen, dass die Ausgabe $\hat{y}$ durch einen bestimmten Schwellenwert begrenzt werden sollte. Die Randbedingungen können entsprechend der lernbasierten Überwachung (z. B. [54]) oder der modellbasierten Überwachung (z. B. [55]) definiert werden. Wenn das ML-System die Sicherheitsbedingungen strikt befolgt und sie sich als sicher erweisen, ist auch das Verhalten des ML-Systems garantiert sicher.

Um die Sicherheit von ML-Modellen während der Laufzeit zu zertifizieren, gibt es immer noch Herausforderungen, die gelöst werden müssen. Zum Beispiel müssen die entwickelten Prüfsysteme selbst verifiziert werden. Modellbasierte Prüfsysteme können Annahmen oder eine Überapproximation des ML-Systems verwenden, was seine Sicherheit ungültig machen kann [42]. Lernbasierte Prüfsysteme werden derzeit in verschiedenen Bereichen unterschiedlich eingesetzt, zum Beispiel in der Robotik [54], menschlichen Posenschätzung [41] und Bildsegmentierung [47]. Die Technologien solcher Ansätze, die öffentlich verfügbar sind, befinden sich allerdings noch im Forschungsstadium.

3 TRANSPARENZ

Transparenz bezeichnet ein vielschichtiges, bereichsübergreifendes Konzept [56], zum Beispiel in Wissenschaft, Politik und Wirtschaft; im Kontext der KI gibt es bisher keinen gemeinsamen Standard für die Definition von Transparenz. [57] betrachtet Transparenz als Grundlage für die weitere Nachvollziehbarkeit von ML-Systemen. Während [58] Transparenz als wesentliches Element zur Verbesserung der Reproduzierbarkeit im Forschungsbereich der KI fokussiert, diskutiert [59] die Unterscheidung zwischen algorithmischer Transparenz und der Transparenz in ML-Systemen. Nach den ethischen Richtlinien der AI HLEG der EU-Kommission ist Transparenz eines der sieben Schlüsselelemente für die Realisierung von »Trustworthy AI« [3]. In Abgrenzung zu der in Abschnitt 2.A diskutierten erklärbaren KI ist Transparenz eine Eigenschaft, die sich auf die gesamte Systemebene bezieht, während erklärbare KI den Fokus auf die Algorithmen- und Modellebene legt. Dieses White Paper folgt diesen ethischen Richtlinien und definiert Transparenz als Nachvollziehbarkeit, Erklärbarkeit und Kommunikation, und zwar im Rahmen der Daten, des Systems und der Geschäftsmodelle.

- Die **Nachvollziehbarkeit** bezieht sich auf die Datenaufbereitung und Inferenzprozesse von ML-Systemen in der Entwicklungs- und Einsatzphase. Gut dokumentierte Informationen der verschiedenen Prozesse ermöglichen, eine fehlerhafte Entscheidung des ML-Systems zu begründen, und könnten wiederum helfen, ähnliche Fehler in Zukunft zu vermeiden und die Robustheit des Systems zu verbessern. Andererseits erleichtert ein nachvollziehbarer Inferenzprozess das Verständnis des ML-Systems und verbessert damit die Erklärbarkeit. Allerdings gibt es bisher keinen Standard oder ein standardisiertes Verfahren zur nachvollziehbaren Dokumentation von ML-Systemen.

- Die **Erklärbarkeit** zielt darauf ab, die von ML-Systemen getroffenen Entscheidungen auf einer umfassenden Ebene des menschlichen Verständnisses zu erklären. In den letzten Jahren wurde viel Forschung in diesem Bereich betrieben. Für weitere Informationen sei auf Abschnitt 2.A verwiesen. Neben der Erklärung auf algorithmischer Ebene für Wissenschaftler und Entwickler ist es auch wichtig, die Verfügbarkeit von Erklärungen des ML-Systems für seine menschlichen Nutzer zu gewährleisten und damit die Transparenz des Geschäftsmodells sicherzustellen [3].

- **Kommunikation** bezieht sich auf die Interaktion zwischen Menschen und ML-Systemen. Diese passt nicht in die Paradigmen der Kommunikationstheorie, da diese seit mehr als einem Jahrhundert prägt, wie Menschen mit Menschen kommunizieren [60]. Die Kommunikation zwischen Mensch und ML-Systemen beschränkt sich nicht auf den semantischen Bereich der akustischen und textlichen Kommunikation, wie beispielsweise Amazons Alexa oder Apples Siri, sondern bedeutet auch, dass Endnutzer oder Geschäftskunden ein Recht auf Informationen über die Funktionalitäten und Grenzen von ML-Systemen haben sollten.

Als aufstrebendes Gebiet im Kontext der KI erfordert die Transparenz noch mehr Standardisierung und Regulierung. In [64] wird die Goal-Structuring-Notation-Methode (GSN) vorgeschlagen, die eine klare Art und Weise bietet, die Sicherheitsgarantiefälle zu konstruieren und zu organisieren, um das Sicherheitsniveau des ML-Systems darzustellen. Die klar strukturierten Argumentationen von GSN können die Transparenz des ML-Systems erleichtern und verbessern. Als Querschnittsthema sind aber nicht nur die technische Sichtweise, sondern auch ethische Aspekte notwendig. Die meisten Arbeiten, die die Autoren dieses White Papers begutachtet haben, wurden nach 2019 veröffentlicht. Wir freuen uns auf weitere Forschung und Standard-Frameworks in diesem Bereich.

4 ZUVERLÄSSIGE ML-PIPELINE

In den letzten Jahren lag der Anwendungsfokus von verlässlicher KI hauptsächlich im Bereich des autonomen Fahrens, während andere Branchen nur am Rande berücksichtigt wurden. Im Folgenden werden einige Ansätze zur Sicherheitsüberprüfung genannt, die im Kontext des autonomen Fahrens konstruiert wurden, und deren Anwendbarkeit für die ML-Zertifizierung in anderen Branchen diskutiert.

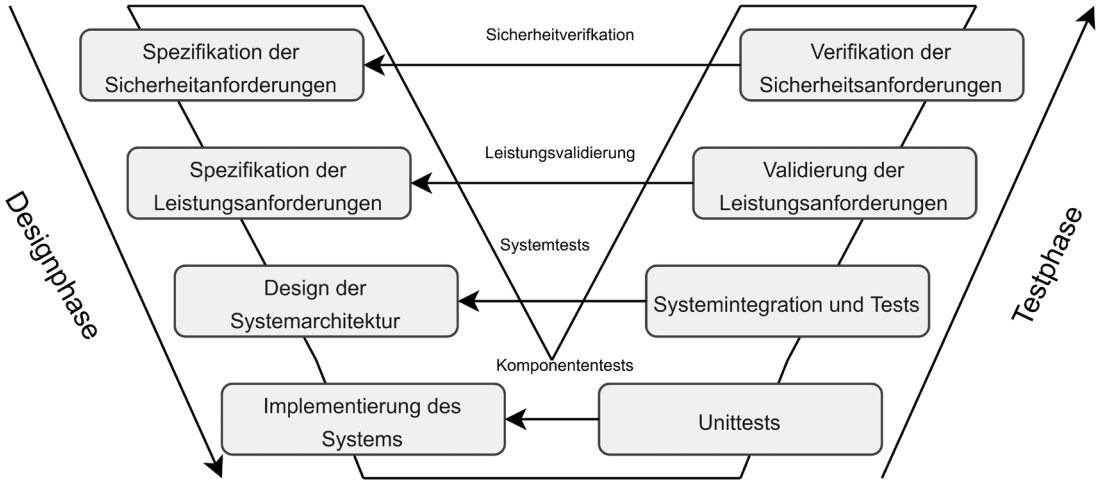

Fig. S4. V&V-Modell in der konventionellen Software-Systementwicklung.
Quelle: Fraunhofer IPA/Bild:Xinyang Wu.

Im öffentlichen Forschungsprojekt PEGASUS wird ein szenariobasierter Ansatz vorgeschlagen, um hochautomatisierte Fahrfunktionen zu bewerten [61]. Die Projektteilnehmer entwickeln eine Pipeline, in der zunächst die Systemanforderungen auf Basis der gewünschten Leistung, Gesetze und anderer Faktoren formuliert werden. Dann werden Datensätze vorbereitet, die sowohl reale als auch simulierte Daten umfassen. Nach der Identifikation möglicher Szenarien durchläuft jedes Szenario mehrere Testfälle, die vor allem auf den Aspekten der formalen und

statistischen Variation sowie der Durchsetzung von Randbedingungen basieren. Dies führt zu einer Sicherheitsargumentation für die bewertete Funktionalität. Während dieser Ansatz für die Entwicklung von ML-Systemen geeignet ist, deckt er den Aspekt des Monitoring während des realen Einsatzes nicht ab. In einem Folgeprojekt basiert die Sicherheitsargumentation auf dem bekannten V-Modell [62], wie es in Abb. S4 dargestellt ist. Für jeden identifizierten Use-Case werden Risikofaktoren und Szenarien ermittelt. Aus diesen Ergebnissen werden Sicherheits- und Testanforderungen abgeleitet. Die Testphase deckt die drei Stufen Simulation, Testanlage und Tests unter Realbedingungen ab, um die Sicherheitsargumentation abzuschließen.

Im Allgemeinen konzentrieren sich die aktuellen Ansätze auf eine Kombination aus realen und simulierten Daten, um eine Argumentation für die Sicherheitszertifizierung zu entwickeln. Die formale Verifikation wird bei diesen Methoden noch nicht als Hauptaspekt betrachtet, da sie im Allgemeinen als nicht gut skalierbar für ML-Ansätze mit vielen Parametern angesehen wird. Außerdem ist es nicht einfach, sehr abstrakte Anforderungen auf eine formale Weise zu formulieren, die für die Verifikation verwendet werden kann. Die Autoren dieses White Papers erwarten, dass die formale Verifikation in Zukunft zumindest für sehr streng definierte Anforderungen eine wichtigere Rolle spielen wird.

Obwohl diese ML-Pipeline-Ansätze noch nicht strikt auf industrielle Anwendungen wie die Mensch-Roboter-Kooperation angewendet werden, bilden sie eine solide Grundlage für eine ähnliche Methodik, die auf Produktionsumgebungen übertragen werden kann. Einerseits ist eine Fabrikumgebung stärker reguliert und weniger vielfältig als Straßen für autonomes Fahren, andererseits ist die Anzahl der potenziellen Anwendungsfälle viel breiter und erfordert sehr individuelle Sicherheitsbewertungen. Außerdem ist noch nicht klar, in welchem Umfang ein solcher Ansatz für die Sicherheitszertifizierung von ML-Systemen im Vergleich zu einem nicht-autonomen Ansatz tatsächlich profitabel ist, da dies sehr stark von der spezifischen Aufgabe, dem Automatisierungsgrad, dem möglichen Risiko und vielen weiteren Faktoren abhängt.

5 FAZIT

Das Thema Zuverlässigkeit im Zusammenhang mit autonomen Systemen ist sehr breit gefächert und beinhaltet mehrere Abstraktionsebenen und unterschiedliche Methoden, um eine bessere Sicherheit zu gewährleisten und die Transparenz zu erhöhen.

Dieses White Paper sieht Zertifizierung und Transparenz als zwei Hauptfaktoren, die helfen können, zuverlässige KI-Systeme zu ermöglichen. Es verweist auf das AMLAS-System zum Aufbau von Argumentationsstrukturen, wenn es um Argumente für die Sicherheit von ML-Systemen geht. Darüber hinaus werden einige vielversprechende Methoden genannt, nämlich erklärbare KI, formale Verifikation, statistische Validierung, Quantifizierung der Unsicherheit und Online-Monitoring, die in den Entwicklungszyklus von ML-Systemen integriert werden können und die Bereitstellung von Beweisen unter den mit AMLAS erstellten Sicherheitsnachweisen unterstützen können. Es werden der aktuelle Forschungsstand solcher Methoden und einige offene Fragen wiedergegeben, die deren Einsatz in der Industrie noch behindern. Abschließend werden einige Projekte genannt, die anstreben, standardisierte Pipelines für zuverlässige KI im Bereich des autonomen Fahrens bereitzustellen. Das Paper kommt zu dem Schluss, dass solche Bemühungen in anderen Branchen noch fehlen, aber eine gute Grundlage für andere Branchen bieten.

Die vorgestellten Methoden und Zertifizierungsansätze sind prinzipiell auf jedes ML-System anwendbar. Allerdings müssen, wie bereits erwähnt, Risikobewertungen und Wirtschaftlichkeitsbetrachtungen für jeden einzelnen Anwendungsfall vorgenommen werden, da standardisierte Prozesse und Best Practices für industrielle Anwendungen bis heute fehlen. Wenn man bereit ist, den Aufwand zu betreiben, sind die vorhandenen Methoden bei der Systementwicklung nutzbar, um performante, sichere und transparente autonome Systeme zu gestalten mit dem Ziel, die Effizienz und Qualität in verschiedenen industriellen Anwendungen zu verbessern. Die bisher in der Forschungsgemeinschaft erzielten Fortschritte sind sehr vielversprechend und deuten darauf hin, dass zuverlässige ML-Systeme in naher Zukunft auf Probleme anwendbar werden, die mit aktuellen Technologien noch nicht zuverlässig gelöst werden können.

LITERATUR

1. PHAM, Duc T.; AFIFY, Ashraf A. Machine-learning techniques and their applications in manufacturing. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture,* 2005, 219. Jg., Nr. 5, S. 395-412.

2. SAE ON-ROAD AUTOMATED VEHICLE STANDARDS COMMITTEE, et al. Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems. *SAE Standard J*, 2014, 3016. Jg., S. 1-16.

3. SMUHA, Nathalie A. The eu approach to ethics guidelines for trustworthy artificial intelligence. *Computer Law Review International*, 2019, 20. Jg., Nr. 4, S. 97-106.

4. HUANG, Xiaowei, et al. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 2020, 37. Jg., S. 100270.

5. MACKALL, Dale; NELSON, Stacy; SCHUMANN, Johann M. *Verification and validation of neural networks for aerospace systems*. National Aeronautics and Space Administration, Ames Research Center, 2002.

6. HAWKINS, Richard, et al. Guidance on the Assurance of Machine Learning in Autonomous Systems (AMLAS). *arXiv preprint arXiv:*2102.01564, 2021.

7. SCHWALBE, Gesina; SCHELS, Martin. A survey on methods for the safety assurance of machine learning based systems. In: *10th European Congress on Embedded Real Time Software and Systems (ERTS 2020)*. 2020.

8. GAUERHOF, Lydia, et al. Assuring the safety of machine learning for pedestrian detection at crossings. In: *International Conference on Computer Safety, Reliability, and Security*. Springer, Cham, 2020. S. 197-212.

9. BEGOLI, Edmon; BHATTACHARYA, Tanmoy; KUSNEZOV, Dimitri. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence,* 2019, 1. Jg., Nr. 1, S. 20-23.

10. Datenethikkommission der Bundesregierung. *Gutachten der Datenethikkommission.* Verfügbar unter: https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.html, 2019. Zugriff am 12.03.2021.

11. ADADI, Amina; BERRADA, Mohammed. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access,* 2018, 6. Jg., S. 52138-52160.

12. MOLNAR, Christoph. *Interpretable machine learning.* https://christophm.github.io/interpretable-ml-book, 2020.

13. FRIEDMAN, Jerome H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 2001, S. 1189-1232.

14. RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos. »Why should I trust you?« Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.* 2016. S. 1135-1144.

15. LUNDBERG, Scott; LEE, Su-In. A unified approach to interpreting model predictions. *arXiv preprint arXiv:*1705.07874, 2017.

16. SHAPLEY, Lloyd S. A value for n-person games. *Contributions to the Theory of Games*, 1953, 2. Jg., Nr. 28, S. 307-317.

17. RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos. Anchors: High-precision model-agnostic explanations. In: P*roceedings of the AAAI Conference on Artificial Intelligence.* 2018.

18. LUNDBERG, Scott M.; ERION, Gabriel G.; LEE, Su-In. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:*1802.03888, 2018.

19. LUNDBERG, Scott M., et al. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, 2020, 2. Jg., Nr. 1, S. 56-67.

20. SHRIKUMAR, Avanti; GREENSIDE, Peyton; KUNDAJE, Anshul. Learning important features through propagating activation differences. In: *International Conference on Machine Learning*. PMLR, 2017. S. 3145-3153.

21. ZILKE, Jan Ruben; MENCÍA, Eneldo Loza; JANSSEN, Frederik. Deepred–rule extraction from deep neural networks. In: *International Conference on Discovery Science*. Springer, Cham, 2016. S. 457-473.

22. SHWARTZ-ZIV, Ravid; TISHBY, Naftali. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:*1703.00810, 2017.

23. SAXE, Andrew M., et al. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019, 2019. Jg., Nr. 12, S. 124020.

24. CLARKE JR, Edmund M., et al. *Model checking.* MIT press, 2018.

25. JOHNSON, C. W. The increasing risks of risk assessment: On the rise of artificial intelligence and non-determinism in safety-critical systems. In: *the 26th Safety-Critical Systems Symposium*. Safety-Critical Systems Club York, UK., 2018. S. 15.

26. SHALEV-SHWARTZ, Shai; SHAMMAH, Shaked; SHASHUA, Amnon. On a formal model of safe and scalable self-driving cars. *arXiv preprint arXiv:*1708.06374, 2017.

27. MANGAL, Ravi; NORI, Aditya V.; ORSO, Alessandro. Robustness of neural networks: a probabilistic and practical approach. In: *2019 IEEE/ACM 41st International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER).* IEEE, 2019. S. 93-96.

28. BOOPATHY, Akhilan, et al. Cnn-cert: An efficient framework for certifying robustness of convolutional neural networks. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019. S. 3240-3247.

29. KATZ, Guy, et al. Reluplex: An efficient SMT solver for verifying deep neural networks. In: *International Conference on Computer Aided Verification*. Springer, Cham, 2017. S. 97-117.

30. KATZ, Guy, et al. Reluplex: An efficient SMT solver for verifying deep neural networks. In: *International Conference on Computer Aided Verification*. Springer, Cham, 2017. S. 97-117.

31. WANG, Shiqi, et al. Efficient formal safety analysis of neural networks. *arXiv preprint arXiv:*1809.08098, 2018.

32. LEINO, Klas; WANG, Zifan; FREDRIKSON, Matt. Globally-Robust Neural Networks. *arXiv preprint arXiv:*2102.08452, 2021.

33. BIFET, Albert; GAVALDA, Ricard. Learning from time-changing data with adaptive windowing. In: *Proceedings of the 2007 SIAM international conference on data mining.* Society for Industrial and Applied Mathematics, 2007. S. 443-448.

34. WILLIAMS, Christopher KI. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In: *Learning in graphical models*. Springer, Dordrecht, 1998. S. 599-621.

35. MACKAY, David JC. A practical Bayesian framework for backpropagation networks. *Neural computation*, 1992, 4. Jg., Nr. 3, S. 448-472.

36. GAL, Yarin; GHAHRAMANI, Zoubin. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: *international conference on machine learning*. PMLR, 2016. S. 1050-1059.

37. LAKSHMINARAYANAN, Balaji; PRITZEL, Alexander; BLUNDELL, Charles. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:*1612.01474, 2016.

38. METROPOLIS, Nicholas, et al. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 1953, 21. Jg., Nr. 6, S. 1087-1092.

39. GRAVES, Alex. Practical variational inference for neural networks. In: *Advances in neural information processing systems*. 2011. S. 2348-2356.

40. HUBER, Marco F. Bayesian Perceptron: Towards fully Bayesian Neural Networks. In: *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 2020. S. 3179-3186.

41. GUPTA, Arjun; CARLONE, Luca. Online monitoring for neural network based monocular pedestrian pose estimation. In: *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020. S. 1-8.

42. SESHIA, Sanjit A.; SADIGH, Dorsa; SASTRY, S. Shankar. Towards verified artificial intelligence. *arXiv preprint arXiv:*1606.08514, 2016.

43. SCHUMANN, Johann; GUPTA, Pramod; NELSON, Stacy. *On verification & validation of neural network based controllers*. EANN'03, 2003.

44. DESAI, Ankush, et al. SOTER: a runtime assurance framework for programming safe robotics systems. In: *2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, 2019. S. 138-150.

45. DESAI, Ankush; DREOSSI, Tommaso; SESHIA, Sanjit A. Combining model checking and runtime verification for safe robotics. In: *International Conference on Runtime Verification*. Springer, Cham, 2017. S. 172-189.

46. SCHIERMAN, John D., et al. *Runtime assurance framework development for highly adaptive flight control systems.* Barron Associates, Inc. Charlottesville, 2015.

47. DEVRIES, Terrance; TAYLOR, Graham W. Leveraging uncertainty estimates for predicting segmentation quality. *arXiv preprint arXiv:*1807.00502, 2018.

48. RAHMAN, Quazi Marufur; SÜNDERHAUF, Niko; DAYOUB, Feras. Did you miss the sign? A false negative alarm system for traffic sign detectors. *arXiv preprint arXiv:*1903.06391, 2019.

49. GOODFELLOW, Ian J.; SHLENS, Jonathon; SZEGEDY, Christian. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:*1412.6572, 2014.

50. NGUYEN, Anh; YOSINSKI, Jason; CLUNE, Jeff. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015. S. 427-436.

51. CHEN, Jiefeng, et al. Robust out-of-distribution detection in neural networks. *arXiv preprint arXiv:*2003.09711, 2020.

52. GOULD, Stephen; HARTLEY, Richard; CAMPBELL, Dylan. Deep declarative networks: A new hope. *arXiv preprint arXiv:*1909.04866, 2019.

53. AGRAWAL, Akshay, et al. Differentiable convex optimization layers. *arXiv preprint arXiv:*1910.12430, 2019.

54. DALAL, Gal, et al. Safe exploration in continuous action spaces. *arXiv preprint arXiv:*1801.08757, 2018.

55. PHAM, Tu-Hoa; DE MAGISTRIS, Giovanni; TACHIBANA, Ryuki. Optlayer-practical constrained optimization for deep reinforcement learning in the real world. In: *2018 IEEE International Conference on Robotics and Automation (ICRA).* IEEE, 2018. S. 6236-6243.

56. HOOD, Christopher; HEALD, David. *Transparency in historical perspective.* Oxford University Press, 2006.

57. BEINING, Leonie. *Wie Algorithmen verständlich werden: Ideen für Nachvollziehbarkeit von algorithmischen Entscheidungsprozessen für Betroffene.* Stiftung Neue Verantwortung, 2019.

58. HAIBE-KAINS, Benjamin, et al. Transparency and reproducibility in artificial intelligence. *Nature*, 2020, 586. Jg., Nr. 7829, S. E14-E16.

59. LARSSON, Stefan; HEINTZ, Fredrik. Transparency in artificial intelligence. *Internet Policy Review*, 2020, 9. Jg., Nr. 2.

60. GUNKEL, David J. Communication and artificial intelligence: Opportunities and challenges for the 21st century. *communication+* 1, 2012, 1. Jg., Nr. 1, S. 1-25.

61. D. Lipinski. *Introduction and overview of 3.5 years of Pegasus.* Verfügbar unter: https://www.pegasusprojekt.de/files/tmpl/Symposium2019/PEGASUS_Symposium_3_5_years.pdf, 2019. Zugriff am 12.03.2021.

62. NEUROHR, Christian, et al. Criticality Analysis for the Verification and Validation of Automated Vehicles. *IEEE Access*, 2021, 9. Jg., S. 18016-18041.

63. BURKART, Nadia; HUBER, Marco F. A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research*, 2021, 70. Jg., S. 245-317.

64. KELLY, Tim; WEAVER, Rob. The goal structuring notation–a safety argument notation. In: *Proceedings of the dependable systems and networks 2004 workshop on assurance cases.* Citeseer, 2004. S. 6.

65. LIPTON, Zachary C. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 2018, 16. Jg., Nr. 3, S. 31-57.

KI-FORTSCHRITTSZENTRUM

Das KI-Fortschrittszentrum »Lernende Systeme« unterstützt Firmen dabei, die wirtschaftlichen Chancen der Künstlichen Intelligenz und insbesondere des Maschinellen Lernens für sich zu nutzen. In anwendungsnahen Forschungsprojekten und in direkter Kooperation mit Industrieunternehmen arbeiten die Stuttgarter Fraunhofer-Institute für Arbeitswirtschaft und Organisation IAO sowie für Produktionstechnik und Automatisierung IPA daran, Technologien aus der KI-Spitzenforschung in die breite Anwendung der produzierenden Industrie und der Dienstleistungswirtschaft zu bringen. Finanzielle Förderung erhält das Zentrum vom Ministerium für Wirtschaft, Arbeit und Wohnungsbau Baden-Württemberg.

Europas größte Forschungskooperation auf dem Gebiet der KI

Das KI-Forschungszentrum ist Forschungspartner des Cyber Valley, einem Konsortium aus den renommierten Universitäten Tübingen und Stuttgart, dem Max-Planck-Institut für intelligente Systeme und einigen führenden Industrieunternehmen. In gemeinsamen Forschungslabors werden Grundlagenforschung und anwendungsorientierte Entwicklung zu aktuellen wie auch zukünftigen Bedarfen behandelt und vorangetrieben.

Menschzentrierte KI

Alle Aktivitäten des Zentrums verfolgen das Ziel, eine menschzentrierte KI zu entwickeln, der die Menschen vertrauen und die sie akzeptieren. Nur wenn Menschen mit neuen Technologien intuitiv interagieren und vertrauensvoll zusammenarbeiten, kann deren Potenzial optimal ausgeschöpft werden. Daher konzentrieren sich die Forschungsaktivitäten unter anderem auf die Themen Erklärbarkeit, Datenschutz, Sicherheit und Robustheit von KI-Technologien.

Studienreihe »Lernende Systeme«

Die Studienreihe »Lernende Systeme« gibt Einblick in die Potenziale und die praktischen Einsatzmöglichkeiten von KI. Nähere Informationen und die aktuellen Versionen der Studien finden Sie unter: www.ki-fortschrittszentrum.de/studien

FRAUNHOFER-GESELLSCHAFT

Die Fraunhofer-Gesellschaft mit Sitz in Deutschland ist die weltweit führende Organisation für anwendungsorientierte Forschung. Mit ihrer Fokussierung auf zukunftsrelevante Schlüsseltechnologien sowie auf die Verwertung der Ergebnisse in Wirtschaft und Industrie spielt sie eine zentrale Rolle im Innovationsprozess. Sie ist Wegweiser und Impulsgeber für innovative Entwicklungen und wissenschaftliche Exzellenz. Mit inspirierenden Ideen und nachhaltigen wissenschaftlich-technologischen Lösungen fördert die Fraunhofer-Gesellschaft Wissenschaft und Wirtschaft und wirkt mit an der Gestaltung unserer Gesellschaft und unserer Zukunft.

Interdisziplinäre Forschungsteams der Fraunhofer-Gesellschaft setzen gemeinsam mit Vertragspartnern aus Wirtschaft und öffentlicher Hand originäre Ideen in Innovationen um, koordinieren und realisieren systemrelevante, forschungspolitische Schlüsselprojekte und stärken mit werteorientierter Wertschöpfung die deutsche und europäische Wirtschaft. Internationale Kooperationen mit exzellenten Forschungspartnern und Unternehmen weltweit sorgen für einen direkten Austausch mit den einflussreichsten Wissenschafts- und Wirtschaftsräumen.

Die 1949 gegründete Organisation betreibt in Deutschland derzeit 75 Institute und Forschungseinrichtungen. Rund 29 000 Mitarbeiterinnen und Mitarbeiter, überwiegend mit natur- oder ingenieur-wissenschaftlicher Ausbildung, erarbeiten das jährliche Forschungsvolumen von 2,8 Milliarden Euro. Davon fallen 2,4 Milliarden Euro auf den Leistungsbereich Vertragsforschung. Rund zwei Drittel davon erwirtschaftet Fraunhofer mit Aufträgen aus der Industrie und mit öffentlich finanzierten Forschungsprojekten. Rund ein Drittel steuern Bund und Länder als Grundfinanzierung bei, damit die Institute schon heute Problemlösungen entwickeln können, die in einigen Jahren für Wirtschaft und Gesellschaft entscheidend wichtig werden.

Die Wirkung der angewandten Forschung geht weit über den direkten Nutzen für die Auftraggeber hinaus: Fraunhofer-Institute stärken die Leistungsfähigkeit der Unternehmen, verbessern die Akzeptanz moderner Technik in der Gesellschaft und sorgen für die Aus- und Weiterbildung des dringend benötigten wissenschaftlich-technischen Nachwuchses.

Hochmotivierte Mitarbeiterinnen und Mitarbeiter auf dem Stand der aktuellen Spitzenforschung stellen für uns als Wissenschaftsorganisation den wichtigsten Erfolgsfaktor dar. Fraunhofer bietet daher die Möglichkeit zum selbstständigen, gestaltenden und zugleich zielorientierten Arbeiten und somit zur fachlichen und persönlichen Entwicklung, die zu anspruchsvollen Positionen in den Instituten, an Hochschulen, in Wirtschaft und Gesellschaft befähigt. Studierenden eröffnen sich aufgrund der praxisnahen Ausbildung und des frühzeitigen Kontakts mit Auftraggebern hervorragende Einstiegs- und Entwicklungschancen in Unternehmen.

Namensgeber der als gemeinnützig anerkannten Fraunhofer-Gesellschaft ist der Münchner Gelehrte Joseph von Fraunhofer (1787–1826). Er war als Forscher, Erfinder und Unternehmer gleichermaßen erfolgreich.

Fraunhofer IPA

Das **Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA**, kurz Fraunhofer IPA, ist mit annähernd 1000 Mitarbeiterinnen und Mitarbeitern eines der größten Institute der Fraunhofer-Gesellschaft. Der gesamte Haushalt beträgt 76 Millionen Euro. Forschungsschwerpunkte des Instituts sind organisatorische und technologische Aufgaben aus der Produktion. Methoden, Komponenten und Geräte bis hin zu kompletten Maschinen und Anlagen werden entwickelt, erprobt und umgesetzt. 15 Fachabteilungen arbeiten interdisziplinär, koordiniert durch 6 Geschäftsfelder, vor allem mit den Branchen Automotive, Maschinen- und Anlagenbau, Elektronik und Mikrosystemtechnik, Energie, Medizin- und Biotechnik sowie Prozessindustrie zusammen. Das Fraunhofer IPA orientiert seine Forschung an der wirtschaftlichen Produktion nachhaltiger und personalisierter Produkte.

»Zentrum für Cyber Cognitive Intelligence«

Das **Zentrum für Cyber Cognitive Intelligence CCI** des Fraunhofer IPA ist ein industrienaher Forschungs- und Entwicklungspartner für die Umsetzung von Applikationen im Bereich Künstliche Intelligenz (KI) und insbesondere Maschinelles Lernen (ML) in der produzierenden Industrie. Ziel des Zentrums ist es, sowohl die KI-Forschung als auch den Technologietransfer von KI und ML in die Anwendung voranzutreiben.

Durch die Vernetzung von Produktionsanlagen und die fortschreitende Digitalisierung werden Daten in großen Mengen verfügbar. Diese Daten werden zunehmend mit ML- bzw. KI-basierten Verfahren ausgewertet und nutzbar gemacht. Dies bietet beachtliche Vorteile für die Industrie: Zum einen sind Leistungssprünge in der Nutzung von Maschinen und Anlagen in Bezug auf Qualität, Flexibilität und Effizienz zu erwarten. Zum anderen entstehen neue Automatisierungslösungen. Hierbei werden nicht zuletzt mit ML ausgestattete Roboter vermehrt Einzug in alle Arbeits- und Alltagsumgebungen halten.

Team »Zuverlässige KI-Systeme«

Der vollumfängliche Einsatz von KI-Funktionen in immer mehr Anwendungsbereichen wie z. B. kollaborierenden Robotersystemen, autonomem Fahren, Medizintechnik oder digitalisierter Produktion stellt vermehrt hohe Anforderungen an die funktionale Sicherheit, Nachvollziehbarkeit und Akzeptanz. Die Gruppe »Zuverlässige KI-Systeme« des Fraunhofer IPA forscht und entwickelt Methoden zur Bewerkstelligung erklärbarer, verifizierbarer und robuster maschineller Lernverfahren mit dem Ziel, Vertrauen in KI-Lösungen zu stärken und Unternehmen bei der Umsetzung, Implementierung und Absicherung von KI-Funktionalitäten zu helfen.

IMPRESSUM

Kontaktadresse
Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA
Nobelstraße 12, 70569 Stuttgart

Mohamed El-Shamouty
Telefon +49 711 970-1660
mohamed.el-shamouty@ipa.fraunhofer.de

Philipp Wagner
Telefon +49 711 970-1988
philipp.wagner@ipa.fraunhofer.de

Xinyang Wu
Telefon +49 711 970-3673
xinyang.wu@ipa.fraunhofer.de

Herausgeber
Thomas Bauernhansl, Marco Huber, Werner Kraus

Titelbild
© zenzen – stock.adobe.com / Fraunhofer IPA

Satz und Gestaltung
Armin Zebrowski, komwerb Agentur

URN-Nummer
urn:nbn:de:0011-n-6306687

Online verfügbar als Fraunhofer-ePrint
http://publica.fraunhofer.de/dokumente/N-630668.html

Gefördert durch das Ministerium für Wirtschaft, Arbeit und Wohnungsbau Baden-Württemberg

Gefördert durch

Baden-Württemberg
MINISTERIUM FÜR WIRTSCHAFT, ARBEIT UND WOHNUNGSBAU

CyberValley

S·TEC