Bei diesem Beitrag handelt es sich um einen wissenschaftlich begutachteten und freigegebenen Fachaufsatz („reviewed paper“).

doi.org/10.37544/1436-4980-2025-06-76

Datensouverän und wettbewerbsfähig: KI verantwortungsvoll im Unternehmen nutzen

KI nutzen, Know-how schützen: Strategien für sichere KI

E. Tinsel, O. Riedel

ZUSAMMENFASSUNG Der Einsatz externer KI-Mehrwertdienste verspricht vielfältige Vorteile für industrielle Produktionssysteme, stellt Unternehmen aber vor besondere Herausforderungen bei Datenschutz, Know-how-Schutz und der Verlässlichkeit ungeprüfter Ergebnisse. Der Beitrag entwickelt ein konzeptionelles Rahmenwerk zum sicheren Umgang mit sensiblen Daten und der kontrollierten Rückführung externer KI-Ergebnisse. Er stellt technische und organisatorische Maßnahmen in den Bereichen Datenverarbeitung, Datensensibilität sowie Zugriffs- und Nutzungsrechte vor. Zudem werden Validierungsmechanismen, simulative Testverfahren und adaptive Selbstanpassung als Strategien zur Absicherung potenziell unzuverlässiger KI-Ausgaben analysiert.

STICHWÖRTER

Industrie 4.0, Künstliche Intelligenz (KI), Sicherheit

How to use AI and protect know-how: Strategies for secure AI

ABSTRACT The use of external AI value-added services promises many advantages for industrial production systems but presents companies with particular challenges in terms of data protection, know-how protection and reliability of unverified results. This article develops a conceptual framework that enables the secure handling of sensitive data and the controlled feedback of external AI results. It presents technical and organizational measures for data processing, data sensitivity. and access and usage rights. In addition, validation mechanisms, simulative test procedures, and adaptive self-adaptation are analyzed as strategies for safeguarding potentially unreliable AI outputs.

1 KI in der Produktion

Die Relevanz der Anwendung künstlicher Intelligenz (KI) in der industriellen Produktion hat in den vergangenen Jahren signifikant zugenommen. Gemäß *Arinez et al.* [1] werden KI-Technologien bereits auf diversen hierarchischen Ebenen eines Fertigungssystems eingesetzt, um Effizienzsteigerungen, Prozessoptimierungen und eine verbesserte Entscheidungsfindung zu ermöglichen. Das Potenzial der KI zeigt sich in einer Vielzahl von Bereichen, angefangen bei der Steuerung komplexer Produktionsanlagen über die vorausschauende Wartung bis hin zur Mensch-Roboter-Kollaboration.

Die vielfältigen Einsatzmöglichkeiten von KI in der Produktion verdeutlichen deren Transformationskraft. Gleichwohl bestehen nach wie vor zentrale Herausforderungen, vor allem in den Bereichen Systemanalyse, Datenqualität, Modellübertragbarkeit und interpretierbare KI-Entscheidungen. Trotz bestehender Forschungslücken lässt sich ein klarer Trend erkennen: Der Einsatz von KI-gestützten Analysewerkzeugen nimmt kontinuierlich zu und führt zu einer nachhaltigen Veränderung des industriellen Umfelds.

Die Systematisierung von künstlicher Intelligenz erfolgt anhand unterschiedlicher Kategorien, die sich in Abhängigkeit des Anwendungsbereichs sowie der Funktionsweise als variabel erweisen. Zu den wichtigsten Arten gehören regelbasierte Systeme, maschinelles Lernen und Deep Learning. Während regelbasierte Systeme feste Entscheidungslogiken nutzen, lernen maschinelle Lernalgorithmen aus Daten und verbessern sich mit der Zeit. Deep Learning hingegen nutzt neuronale Netzwerke, die mehrere Schichten aufweisen, um komplexe Muster zu erkennen und selbstständig Entscheidungen zu treffen.

Unternehmen setzen KI ein, um Prozesse zu optimieren, Maschinen vorausschauend zu warten und Produktionsfehler frühzeitig zu erkennen. Die Kombination von Sensordaten, Algorithmen und Automatisierung führt zu intelligenten Fertigungssystemen, die sich durch erhöhte Flexibilität, Effizienz und Kosteneffektivität auszeichnen. KI avanciert somit zu einem zentralen Treiber der Industrie 4.0 und ermöglicht eine neue Stufe der digitalen Transformation in der Produktion. **Bild 1** zeigt eine Übersicht der Anwendungsfelder von KI in der Produktion nach *Fahle et al.* [2]. Im Folgenden werden ausgewählte Beispiele aus den Bereichen Predictive Maintenance, Qualitätssicherung, Produktionsplanung und kollaborativer Roboter kurz vorgestellt.

Predictive Maintenance ist ein Anwendungsfeld, in dem Unternehmen bereits gegenwärtig Lösungen bereitstellen, die eine Reduktion von Ausfallzeiten und eine Senkung von Wartungskosten ermöglichen. Dazu werden Sensordaten verarbeitet und über maschinelles Lernen sowie Deep-Learning-Algorithmen analysiert. Die Analyse historischer Datensätze sowie die Untersuchung neuer Daten hinsichtlich bekannter Auffälligkeiten sind die Grundlagen dieser Verfahren. Nach *Ucar et al.* [3] ist ein zentraler Fortschritt die Integration erklärbarer und interpretierbarer

KI. Ein weiterer vielversprechender Trend ist die Nutzung generativer KI-Modelle, um die vorausschauende Wartung weiter zu verbessern.

Im Bereich der Qualitätssicherung wird durch den Einsatz von KI-Ansätzen die Fehlererkennung verbessert. Als Beispiele werden von *Fahle et al.* [2] bildbasierte Verfahren, wie Convolutional Neural Networks, genannt, die erfolgreich zur Inspektion von Batterien und Sensoren eingesetzt werden. Neben der visuellen Prüfung optimieren Bayesian Networks und Entscheidungsbäume ganze Fertigungsprozesse, indem sie Ursachen für Qualitätsabweichungen identifizieren. Somit findet KI nicht nur Anwendung in der Endkontrolle, sondern auch in der präventiven Qualitätssicherung.

Kumar [4] präsentiert eine Reihe von Beispielen, welche die Unterstützung der Produktionsplanung durch Expertensysteme und wissensbasierte Systeme beleuchten. Dabei wird menschliches Expertenwissen in digitaler Form erfasst und für Entscheidungsprozesse nutzbar gemacht. Während Expertensysteme auf regelbasierten Methoden basieren, um komplexe Probleme zu analysieren und Lösungen vorzuschlagen, gehen wissensbasierte Systeme einen Schritt weiter, indem sie aus Daten lernen und ihr Wissen dynamisch erweitern können. Die Unterstützung umfasst beispielsweise die Auswahl geeigneter Maschinen, die Planung von Fertigungsschritten und die Ressourcenverteilung. Zu diesem Zweck kombinieren sie historische Produktionsdaten, Expertenwissen und aktuelle Sensordaten.

Schließlich resultieren aus der Anwendung von künstlicher Intelligenz eine gesteigerte Adaptivität und Sicherheit kollaborativer Roboter (Cobots), die in der Lage sind, flexibel mit Menschen zusammenzuarbeiten. KI-Technologien wie maschinelles Lernen und Computer Vision erlauben es Cobots, menschliche Bewegungen zu erkennen und dynamisch darauf zu reagieren. In diesem Zusammenhang dominieren verschiedene Lernverfahren, die als überwachtes, unüberwachtes und verstärkendes Lernen bezeichnet werden. Diese Lernverfahren tragen zur Verbesserung der Interaktion und der autonomen Entscheidungsfindung bei. Ferner tragen Fortschritte in den Bereichen Sensorfusion und Kraftregelung zu einer erhöhten Präzision und Sicherheit bei [5].

Obwohl das Potenzial von künstlicher Intelligenz zur Optimierung industrieller Prozesse als erheblich eingestuft wird, zögern viele Unternehmen, diese Technologien in ihre Produktionsabläufe zu integrieren. Gemäß einem Bericht des Bundesministeriums für Wirtschaft und Klimaschutz [6] gaben im Jahr

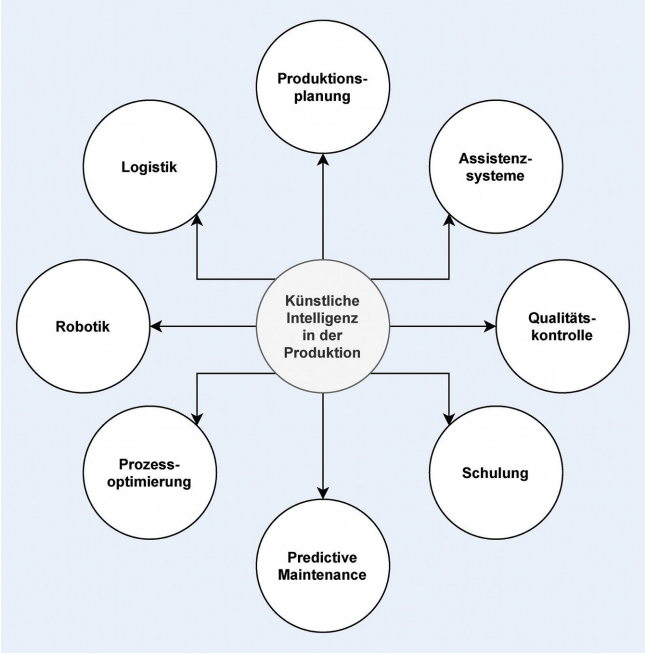

Bild 1. Einsatzbereiche künstlicher Intelligenz nach *Fahle et al.* [2]. *Grafik: ISW, Universität Stuttgart*

2023 nur etwa 12 % der Unternehmen mit mehr als 10 Beschäftigten an, KI-Verfahren einzusetzen.

Somit verteilt sich das genutzte KI-Verfahren, wie in **Bild 2** dargestellt, absteigend auf die Felder Spracherkennung, Automatisierung von Arbeitsabläufen/Entscheidungsfindung, Text Mining, maschinelles Lernen, Erzeugung von natürlicher Sprache, Bilderkennung/-verarbeitung und Bewegung von Maschinen.

Außerdem lässt sich dem Bericht entnehmen, dass nur 15 % der KI-Verfahren in Eigenentwicklung entstanden sind.

Ein weiterer Bericht des Bundesministeriums für Wirtschaft und Energie [7] aus dem Jahr 2021 kann als ein Erklärungsversuch verwendet werden. Für eine weitere Verbreitung der KI-Verfahren werden fünf Punkte als entscheidend aufgeführt:

1. Kostenfaktor: Für Unternehmen, die KI bereits nutzen, sind die hohen Entwicklungs- und Implementierungskosten die größte Herausforderung, insbesondere für KMUs und Unternehmen mit geringem KI-Reifegrad. Zudem fehlt es oft an Investitionsbudgets bei Geschäftspartnern, um KI-Projekte gemeinsam

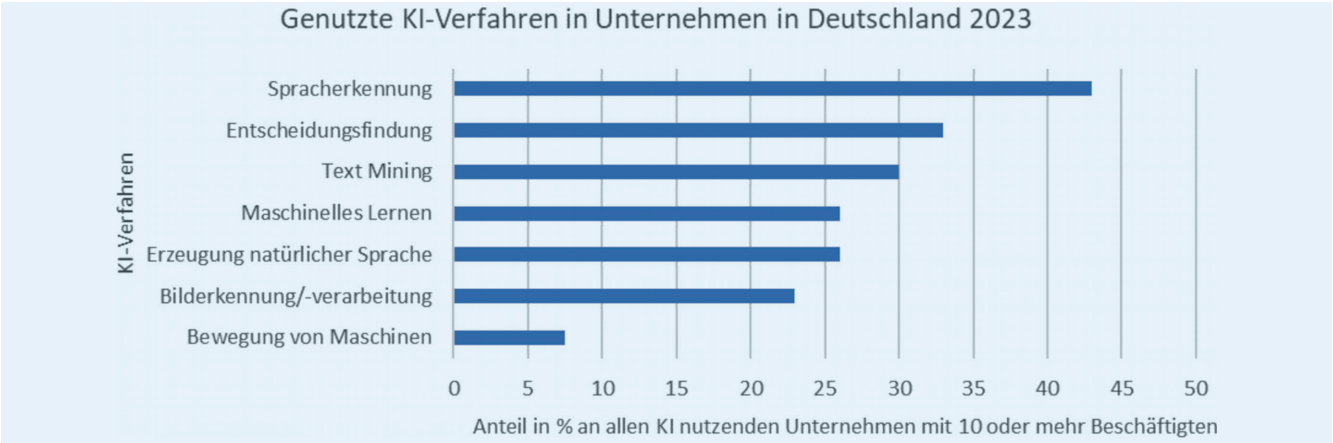

Bild 2. Genutzte KI-Verfahren in Unternehmen in Deutschland 2023 nach [6]. *Grafik: ISW, Universität Stuttgart nach [6]*

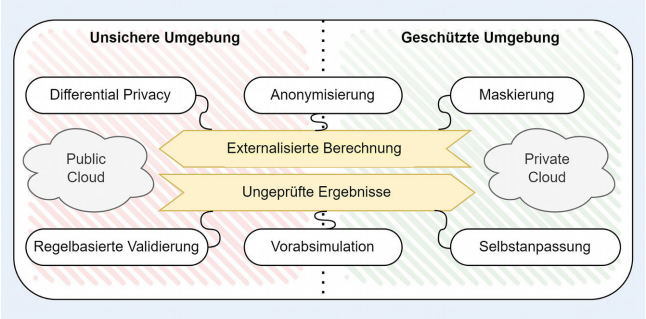

Bild 3. Schematische Darstellung von Handlungsempfehlungen zur Nutzung von KI-Mehrwertdiensten in unsicheren Umgebungen. *Grafik: ISW, Universität Stuttgart*

umzusetzen. Deshalb gilt eine bessere Finanzierung als zentrale Maßnahme zur Stärkung des KI-Standorts Deutschland im internationalen Vergleich.

2. Infrastruktur: Für Unternehmen, die keine KI-Verfahren einsetzen, ist eine leistungsfähige IT-Infrastruktur die zentrale Voraussetzung für den Einstieg.
3. Fachkräfte: Für Unternehmen, die KI-Verfahren einsetzen, sind qualifizierte Fachkräfte essenziell, um diese effektiv zu nutzen. Bei Unternehmen ohne KI-Nutzung ist das Thema vor allem für größere Betriebe relevant, die entsprechende Weiterbildungsangebote als Voraussetzung für den Einstieg sehen.
4. Akzeptanz: Mehr Vertrauen in KI-Lösungen und eine stärkere öffentliche Sichtbarkeit für deren Nutzen sind laut dem Bericht zentrale Handlungsfelder, sowohl für Unternehmen mit KI-Verfahren im Einsatz als auch für KI-affine Nicht-Nutzer. Hier schneidet Deutschland im internationalen Vergleich am schlechtesten ab.
5. Zugänglichkeit: Die begrenzte Verfügbarkeit externer Daten gilt als großer Standortnachteil Deutschlands und ist für Unternehmen mit KI-Verfahren im Einsatz – neben dem Datenschutz – die größte Herausforderung. Auch Datensicherheit ist ein zentrales Thema. Sichere, datenschutzkonforme Cloud-Angebote sind daher essenziell, um den Datenzugang zu verbessern und den Einstieg in KI-Anwendungen zu erleichtern.

Dieser Beitrag rückt die Besorgnis um den Schutz sensibler Daten und des unternehmensspezifischen Wissens von Entscheidungsträgern in Unternehmen in den Fokus, was diese besonders zurückhaltend gegenüber KI agieren lässt.

Obwohl künstliche Intelligenz signifikante Potenziale in Bezug auf die Steigerung der Effizienz und Optimierung von Prozessen bietet, werden zugleich kritische Fragen hinsichtlich des sicheren Umgangs mit Unternehmenswissen aufgeworfen. In datengetriebenen Industrien besteht besonders die Befürchtung, dass durch den Einsatz externer KI-Lösungen wertvolle interne Informationen in die Hände Dritter gelangen oder dass maschinell erlernte Muster unkontrolliert weiterverwendet werden könnten. Diese Bedenken betreffen nicht nur klassische Datenschutzaspekte im Sinne der DSGVO, sondern auch strategische Überlegungen zur Wettbewerbsfähigkeit: Wie kann ein Unternehmen sicherstellen, dass KI-Systeme Wissen aus internen Abläufen nicht ungewollt mit externen Anbietern oder anderen Marktteilnehmern teilen? Ebenso ist zu erörtern, wie proprietäre Daten zur Optimierung generischer Modelle beitragen können, ohne dass diese der Konkurrenz zugutekommen. Für viele Unternehmen stellt sich nicht mehr nur die Frage, ob sie KI einsetzen sollen, sondern auch, wie sie dies tun können, ohne eigenes Firmenwissen zu gefährden.

Zu diesem Zweck stellt dieser Beitrag Handlungsempfehlungen für den Einsatz von KI-Mehrwertdiensten in unsicheren Umgebungen vor. Es werden Mechanismen im Bereich der Datenverarbeitung, der Datensensibilität sowie der Nutzungs- und Zugriffsrechte vorgestellt, um eine externalisierte Berechnung in einer unsicheren Umgebung mit möglichst geringer Preisgabe von unternehmensinternem Wissen zu ermöglichen. Zudem werden Validierungsmechanismen, Einsatzmöglichkeiten von simulativen Tests sowie die Möglichkeit der Selbstadaption für den sicheren Umgang mit ungeprüften Ergebnissen aus unsicheren Umgebungen vorgestellt. Die Kernaspekte zeigt **Bild 3** und werden in den folgenden beiden Abschnitten im Detail vorgestellt.

Der vorliegende Beitrag zielt darauf ab, ein konzeptionelles Fundament zu etablieren, um die nachfolgende Forschungsfrage zu beantworten: „Wie kann der Austausch sensibler Unternehmensdaten mit KI-Mehrwertdiensten in unsicheren Umgebungen sowie der Rückfluss ungeprüfter KI-Ergebnisse abgesichert werden, ohne die Vertraulichkeit und Integrität geschäftskritischer Informationen und Prozesse zu gefährden?“

Die präsentierten Konzepte basieren auf einer themenzentrierten Literaturrecherche, die sich an den Schnittstellen von Industrie 4.0, vertrauenswürdiger KI, IT-Sicherheitsarchitekturen und industriellem Datenmanagement orientiert. Relevante Quellen wurden aus wissenschaftlichen Datenbanken (unter anderem IEEE Xplore, SpringerLink, ScienceDirect) und industriebezogenen Veröffentlichungen (zum Beispiel Fraunhofer-Whitepapers, BMBF-/BMWK-Studien) identifiziert. Darüber hinaus wurden etablierte technologische Muster, beispielsweise aus dem Cloud-Computing-Kontext, herangezogen und auf ihren Transfer in das industrielle KI-Umfeld geprüft.

Der Beitrag verfolgt keinen systematisch-experimentellen, sondern einen konzeptionell-gestaltenden Ansatz. In diesem Rahmen werden existierende Methoden, Muster und architektonische Prinzipien miteinander kombiniert und in ein anwendungsorientiertes Handlungskonzept überführt. Der Fokus liegt auf einem möglichst hohen Automatisierungsgrad und einer gleichzeitig kontrollierten Rückführung von KI-Ergebnissen, um den produktiven Einsatz von KI-Mehrwertdiensten auch in sensiblen Unternehmensbereichen zu ermöglichen.

2 Nutzung von KI-Mehrwertdiensten in unsicheren Umgebungen

Obwohl das Potenzial künstlicher Intelligenz zur Steigerung der Effizienz und Optimierung von Prozessen nachgewiesen wurde, zögern viele Unternehmen, KI-Technologien flächendeckend in ihre Produktionsabläufe zu integrieren. Wie bereits im vorangegangenen Teil dargestellt, liegen die Gründe unter anderem in fehlender Infrastruktur, begrenzten finanziellen Mitteln, einem Mangel an Fachkräften sowie einer geringen gesellschaftlichen Akzeptanz. Ein wesentlicher Aspekt, der in diesem Zusammenhang deutlich wird, ist die Sorge vor dem Verlust oder der ungewollten Weitergabe von unternehmensspezifischem Wissen.

In datengetriebenen Industrien, in denen KI-Anwendungen auf umfangreiche interne Datenbestände zugreifen, stellt sich vielen Unternehmen nicht nur die Frage, ob KI eingesetzt werden soll, sondern vielmehr wie dies geschehen kann, ohne die Kontrolle über sensibles Firmenwissen zu verlieren. Dabei geht es nicht nur

um Datenschutz im klassischen Sinne, sondern auch um strategische Überlegungen zur Wahrung von Wettbewerbsvorteilen und geistigem Eigentum. Daher stellt sich die Frage, wie Unternehmen KI verantwortungsvoll und sicher nutzen können.

Der adäquate Umgang mit KI-Systemen bedingt ein profundes Verständnis der zugrunde liegenden Prozesse sowie der Art und Weise, wie diese Technologien mit Daten interagieren. Moderne KI-Modelle, vor allem solche aus dem Bereich des maschinellen Lernens und Deep Learnings, sind in hohem Maße von den bereitgestellten Trainingsdaten abhängig. Diese Daten können strukturierter Natur sein, beispielsweise aus Maschinen- und Sensormodellen, oder unstrukturiert, beispielsweise aus Dokumenten, Bildern oder E-Mails.

Für die Unternehmen ergeben sich daraus mindestens drei zentrale Fragen:

1. Datenverarbeitung: Wo werden die Daten gespeichert und verarbeitet? Es ist zu erörtern, in welcher Form das Training und die Ausführung von KI-Modellen erfolgt, ob innerhalb der eigenen IT-Infrastruktur oder in der Cloud.
2. Datensensibilität: Weiterhin ist die Sensibilität der verwendeten Daten zu berücksichtigen. Es ist von entscheidender Bedeutung zu ermitteln, ob die Daten Rückschlüsse auf firmeninternes Wissen, Produktionsprozesse oder geschäftskritische Entscheidungen zulassen.
3. Zugriffs- und Nutzungsrechte: Zuletzt bedarf es einer Untersuchung der Zugriffsberechtigung auf die Daten und die resultierenden KI-Modelle. Es ist zu klären, ob die Gefahr besteht, dass externe Anbieter oder andere Unternehmen indirekt von den Erkenntnissen profitieren.

2.1 Datenverarbeitung

Die Frage, an welchem Ort und auf welche Weise Daten für KI-Anwendungen verarbeitet werden, zählt zu den zentralen strategischen Entscheidungen für Unternehmen. Bedingt ist dies dadurch, dass nicht nur die Leistungsfähigkeit, sondern vor allem die Kontrolle über unternehmenskritische Informationen maßgeblich von der gewählten IT-Infrastruktur abhängt. KI-Verfahren, die in der Produktion als Mehrwertdienst eingesetzt werden, müssen als on-demand Selfservice bereitgestellt werden. In diesem Zusammenhang wird die Verarbeitung der Daten in einer Cloud-Infrastruktur untersucht.

Generell lassen sich drei Cloud-Architekturen unterscheiden: Public Cloud, Private Cloud und Hybrid Cloud – ergänzt durch Mischformen wie die Virtual Private Cloud (VPC). Jede dieser Architekturen weist spezifische Vor- und Nachteile auf, die in Abhängigkeit von den jeweiligen Sicherheitsanforderungen und der spezifischen Unternehmensstruktur zu evaluieren sind.

Bei einer Public Cloud stellen Anbieter wie Amazon Web Services, Microsoft Azure oder Google Cloud Platform IT-Ressourcen über das öffentliche Internet bereit. Diese Lösung zeichnet sich durch hohe Skalierbarkeit, Flexibilität und geringen Wartungsaufwand aus, da die Infrastruktur vollständig vom jeweiligen Anbieter gemanagt wird. Für rechenintensive KI-Aufgaben, wie das Modelltraining auf großen Datenmengen, bieten Public Clouds kosteneffiziente Zugriffsmöglichkeiten auf spezialisierte Hardware wie GPUs oder TPUs. Jedoch ist bei der Nutzung von Public Clouds ein Verlust der Datenhoheit für die Unternehmen zu verzeichnen. Zwar erfüllen die meisten Anbieter inzwischen zahlreiche Sicherheits- und Datenschutzstandards, doch bestehen vor allem bei sensiblen Produktionsdaten Bedenken, dass Informationen außerhalb des eigenen Einflussbereichs gelangen, sei es technisch oder durch vertragliche Unklarheiten.

Demgegenüber steht die Private Cloud, bei der die IT-Infrastruktur entweder vollständig im eigenen Haus betrieben oder über spezialisierte Anbieter exklusiv bereitgestellt wird. Bei der Kontrolle über Datenflüsse, Nutzerrechte und Sicherheitsmaßnahmen hat das Unternehmen in diesem Fall die volle Kontrolle, was insbesondere für Branchen mit hohen regulatorischen Anforderungen (zum Beispiel Pharma, Luftfahrt, Verteidigung) von großer Bedeutung ist. Jedoch ist der Aufbau und Betrieb dieser Clouds mit hohen Kosten verbunden und erfordert internes IT-Know-how. Zudem ist eine einfache Skalierung nicht möglich.

Die Hybrid Cloud stellt in diesem Zusammenhang einen praktikablen Mittelweg dar. Dabei erfolgt eine Verteilung der Datenverarbeitung je nach Sensibilität und Performance-Anforderung: Kritische Daten verbleiben in der Private Cloud oder On-Premises, während weniger sensible Aufgaben (wie etwa das Modelltraining mit anonymisierten Daten oder eine Batch-Verarbeitung) in der Public Cloud ausgeführt werden. Dieses Modell erlaubt eine feingranulare Steuerung darüber, wo verarbeitet wird, und kombiniert Skalierbarkeit mit Datensouveränität.

Für Unternehmen, die auf die Infrastruktur großer Cloud-Anbieter setzen, aber dennoch ein erhöhtes Maß an Isolation und Sicherheit benötigen, stellt die Virtual Private Cloud (VPC) eine interessante Option dar. Die VPC stellt einen logisch abgeschlossenen Bereich innerhalb einer Public Cloud dar, in dem Unternehmen eine eigene Netzwerkarchitektur mit Zugriffskontrollen, Firewalls und IP-Adressbereichen definieren können. Diese ist vergleichbar mit einer Private Cloud, aber ohne eigene Hardware. Jedoch ist zu bedenken, dass bei dieser Variante die Daten wie bei der Public Cloud aus dem Unternehmen herausgeleitet werden, was einen vollständigen Schutz vor Datenabfluss nicht gewährleistet.

In kooperativen Szenarien mit Zulieferern, Technologiepartnern oder Logistikdienstleistern bieten VPCs oder dedizierte Mandanten innerhalb einer Hybrid Cloud die Möglichkeit des gemeinsamen Zugriffs auf geteilte KI-Modelle oder Datenräume, ohne dass dabei ein vollständiger Kontrollverlust entsteht. Standards wie GAIA-X oder IDS (International Data Spaces) fördern Ansätze zur kontrollierten Datenökonomie zwischen Partnern auf einer gemeinsamen Plattform.

2.2 Datensensibilität

Der Schutz sensibler Daten ist eine zentrale Herausforderung beim Einsatz künstlicher Intelligenz in der Produktion. Dies ist besonders dann problematisch, wenn Daten Rückschlüsse auf unternehmensspezifisches Wissen, interne Abläufe oder strategisch relevante Entscheidungen zulassen. Dadurch entstehen erhebliche Risiken, die etwa durch ungewollten Datenabfluss, unsachgemäße Verarbeitung oder die externe Nutzung von trainierten Modellen entstehen können. Unternehmen stehen folglich vor der Aufgabe, zu entscheiden, ob und in welcher Form Daten genutzt werden dürfen sowie welche Vorkehrungen zu treffen sind.

Zunächst ist zu ermitteln, wie kritisch ein Datensatz in Bezug auf Know-how, Geschäftsprozesse oder personenbezogene Informationen ist. Es empfiehlt sich eine Klassifikation der Daten durchzuführen nach Sensibilität und Verwendungszweck, beispielsweise in operative Maschinendaten, Produktdaten, Prozess-

parameter oder Kommunikationsprotokolle (vergleiche *Munwar et al.* [8]). In Abhängigkeit der Klassifikation sind spezifische Schutzmaßnahmen zu implementieren.

Ein bewährter Ansatz ist die Maskierung von Daten, bei der identifizierende Merkmale durch neutrale Platzhalter ersetzt werden. Dies können etwa Maschinenkennungen, Produktnummern oder Namen von Mitarbeitenden sein. Eine zusätzliche Filterung kann erfolgen, bei der nicht benötigte oder hochsensible Attribute vollständig aus dem Datensatz entfernt werden. In vielen Fällen ist auch eine Anonymisierung sinnvoll, bei der personenbezogene oder unternehmensspezifische Informationen so transformiert werden, dass keine Rückschlüsse mehr auf die Quelle möglich sind. Eine fortgeschrittene Methode dafür ist Differential Privacy [9], die sicherstellt, dass einzelne Datenpunkte keine signifikanten Rückschlüsse auf das Gesamtmodell ergeben. Alternativ können auch synthetische Daten oder simulierte Werte eingesetzt werden, um reale Muster zu erhalten, ohne auf Originaldaten zurückgreifen zu müssen.

Ein Anwendungskomponenten-Proxy [10] ist der softwaretechnische Baustein dieser Datenvorverarbeitung für die Cloud. Es wird eine vorgeschaltete Anwendungskomponente zwischen Datenquelle und KI-System platziert, die sensible Informationen entweder vollständig blockiert oder vor der Weitergabe transformiert, etwa durch Reduktion, Umkodierung oder Aggregation. Diese Entkopplung erlaubt es, dass die eigentlichen KI-Systeme oder Drittdienste keine direkten Zugriffe auf kritische interne Systeme oder Daten erhalten.

Ein weiterer Baustein ist die konforme Datenreplikation [10], bei der Daten vor ihrer Weitergabe oder Verarbeitung repliziert und gemäß gegebener Datenschutzbestimmungen aufbereitet werden. Diese Methode erlaubt es Unternehmen, beispielsweise die Vorgaben der DSGVO einzuhalten, auch wenn das Training von KI-Modellen grenzüberschreitend erfolgt. In Kombination mit einer geografisch kontrollierten Replikation lassen sich Daten standortkonform nutzen, ohne gegen regulatorische Rahmenbedingungen zu verstoßen.

Diese Ansätze demonstrieren, dass technische Schutzmaßnahmen wie Maskierung, Filterung, Anonymisierung, Proxy-Komponenten und kontrollierte Replikation keine Hindernisse für den Einsatz von KI darstellen. Vielmehr schaffen sie die Voraussetzungen dafür, dass KI-Anwendungen auch im sensiblen Produktionsumfeld rechtskonform, sicher und vertrauenswürdig eingesetzt werden können.

2.3 Zugriffs- und Nutzungsrechte

Im Kontext des Einsatzes von künstlicher Intelligenz sind nicht nur Daten und deren Verarbeitung, sondern auch die Zugriffsberechtigung für diese Daten und die daraus resultierenden Modelle von entscheidender Bedeutung. Der Schutz von unternehmensspezifischem Wissen endet nicht mit der Datenaufbereitung – er setzt sich fort in der Steuerung der Zugriffsrechte und der Sicherstellung, dass weder externe Dienstleister noch interne Akteure unbeabsichtigt oder missbräuchlich Informationen erhalten, die Rückschlüsse auf geschäftskritische Zusammenhänge ermöglichen.

Vor allem in Verbindung mit Cloud-Diensten und Partnernetzwerken kommt es zur signifikanten Zunahme der Komplexität der Zugriffskontrolle. Unternehmen müssen daher sicherstellen, dass ihre Daten sowohl physisch als auch logisch geschützt sind. Dazu empfiehlt sich die Implementierung eines rollenbasierten Zugriffsmanagements (siehe *Sandhu* [11]), das festlegt, welche Nutzer oder Systeme Zugriff auf bestimmte Daten, Modelle oder Funktionen erhalten. In KI-Projekten empfiehlt sich eine feingranulare Trennung zwischen Datenbereitstellung, Modelltraining, Modellnutzung und Ergebnisverarbeitung. Diese sollte technisch unterstützt durch Containerisierung (vergleiche *Pahl* [12]) und Mandantenfähigkeit (siehe *AlJahdali et al.* [13]) erfolgen, sodass einzelne Instanzen strikt voneinander isoliert bleiben.

Der Modellschutz ist ein besonders kritischer Aspekt. KI-Modelle, welche auf unternehmensspezifischen Daten trainiert wurden, erlauben potenziell indirekte Rückschlüsse auf interne Strukturen, Prozesse oder Qualitätsmuster. Werden diese Modelle an Dritte übergeben, beispielsweise an externe Dienstleister oder im Rahmen eines Produktangebots, besteht die Gefahr, dass geschäftsrelevantes Wissen unkontrolliert weitergegeben wird. Um dieser Problematik entgegenzuwirken, sollten Unternehmen ihre Modelle eindeutig versionieren, kontrolliert exportieren und bei Bedarf mit eingebauten Schutzmechanismen versehen. Zu den möglichen Schutzmechanismen zählen etwa Output-Filterung, Nutzungsmonitoring oder die Einschränkung von Modellinferenzfunktionen auf sichere Anwendungsfälle.

Eine vielversprechende Architektur zur Wahrung der Datenhoheit ist das föderale Lernen (siehe *Zhang et al.* [14]). Bei dieser Methode verbleiben die Daten vollständig in den lokalen Systemen der jeweiligen Organisation, beispielsweise bei Zulieferern oder Produktionsstandorten. Es werden nur Modellparameter oder Gradienten ausgetauscht. So wird ein gemeinsames Modelltraining ermöglicht, ohne dass zentrale Datenbestände aggregiert oder offengelegt werden müssen. In Kombination mit verschlüsselten Übertragungswegen und vertrauenswürdigen Ausführungsumgebungen (siehe *Sabt et al.* [15]) lässt sich ein kooperatives KI-Ökosystem schaffen.

Der Schutz der Daten erstreckt sich nicht nur auf die Speicherung oder Aufbereitung, sondern auch auf die gesamte Lebensdauer von Modellen und Anwendungen, einschließlich Zugriff, Weitergabe, Nutzung und Integration. Die Berücksichtigung dieser Dimension ist somit von entscheidender Bedeutung, um einen sicheren und rechtskonformen Einsatz von KI-Systemen zu gewährleisten und eine unternehmensweite vertrauensvolle Skalierung zu ermöglichen.

3 Umgang mit ungeprüften Ergebnissen aus KI-Mehrwertdiensten

Obwohl Unternehmen zunehmend beginnen, KI-gestützte Dienste zur Analyse, Optimierung und Steuerung von Produktionsprozessen zu nutzen, bestehen weiterhin erhebliche Unsicherheiten im Hinblick auf die Verlässlichkeit und Einsetzbarkeit der erzeugten Ergebnisse. Besonders bei der Nutzung externer Mehrwertdienste – etwa cloudbasierter Prognosemodelle, KI-Plattformen von Drittanbietern oder vortrainierter Machine-Learning-Modelle – stellt sich die zentrale Frage, inwiefern deren Ausgaben ohne gesonderte Überprüfung direkt in produktive Prozesse zurückgeführt werden dürfen. Die Herausforderung besteht darin, potenziell unvollständige, nicht validierte oder auf abstrahierten Daten basierende Ergebnisse so in die Produktion einzubinden, dass Qualität, Sicherheit und Prozessrobustheit jederzeit gewährleistet bleiben.

Wie im vorangegangenen Kapitel dargelegt, stellt der Schutz sensibler Daten sowie firmenspezifischen Wissens ein zentrales Anliegen vieler Unternehmen dar. Es sei jedoch angemerkt, dass im Zuge dessen auch Informationen entfernt werden, die für die Genauigkeit von KI-Vorhersagen von entscheidender Bedeutung sind. Dies erfolgt durch Anonymisierung, Maskierung oder Datenabstraktion was jedoch in Unsicherheiten hinsichtlich der KI-Ergebnisse, die ein methodisch abgesichertes Rückspiel in operative Systeme erfordern, resultiert.

In Anbetracht dessen ergibt sich die Fragestellung, mit welchen technischen und organisatorischen Konzepten Unternehmen den Umgang mit ungeprüften Resultaten aus KI-Mehrwertdiensten gestalten können, ohne dabei die Vorzüge datengetriebener Intelligenz einzubüßen. In den folgenden Abschnitten werden drei Herangehensweisen präsentiert: Validierungsmechanismen in Form von automatisierten oder regelbasierten Prüfverfahren, die Absicherung durch simulatives Testen in virtuellen Umgebungen sowie die adaptive Selbstanpassung, bei der das System aus Rückkopplung mit der Realität eigenständig zu optimierten Entscheidungen gelangt.

3.1 Validierungsmechanismen

Um KI-Ergebnisse aus Mehrwertdiensten in produktive Abläufe zurückzuführen, bedarf es automatisierter Verfahren zur Plausibilitätsprüfung und Qualitätssicherung. In diesem Kontext bilden Validierungsmechanismen eine zentrale Schnittstelle zwischen der Ausgabe eines KI-Modells und dessen praktischer Anwendung im Produktionsumfeld. Das Ziel besteht darin, Entscheidungen, Empfehlungen oder Prognosen, die aus KI-Systemen hervorgehen, vor ihrer Umsetzung systematisch zu bewerten und potenzielle Fehlfunktionen frühzeitig zu erkennen, ohne dass menschliches Eingreifen erforderlich ist.

Im industriellen Kontext finden insbesondere regelbasierte Validierungssysteme Anwendung (vergleiche Beispiele aus *Naser et al.* [16]). Die Überprüfung der Ausgaben für KI erfolgt anhand definierter Bedingungen, welche sich aus technischen, logischen oder prozessualen Anforderungen ableiten lassen. Es muss sich dabei nicht notwendigerweise um feste Grenz- oder Schwellenwerte handeln. Strukturelle Konsistenzprüfungen, die Erkennung von Abweichungen gegenüber historischen Vergleichsdaten, die Prüfung auf Vollständigkeit von Output-Parametern sowie die Korrelation mit Parallelprozessen können Bestandteile des Validierungsprozesses sein.

Ein häufig genutzter Mechanismus ist der Abgleich mit regelbasierten Heuristiken, bei dem KI-Vorschläge gegen bestehende Entscheidungsbäume oder Prozessmodelle geprüft werden. Abweichungen der KI-Ausgabe von bekannten Mustern oder Erfahrungswerten, die als signifikant einzustufen sind, können zu einer Zurückweisung der Ausgabe oder einer Ausführung unter Einschränkungen führen. In den Validierungsprozess können zudem Kontextbedingungen wie der aktuelle Betriebsmodus einer Anlage oder externe Einflussfaktoren (wie Temperatur, Materialcharge) einbezogen werden. Auf diese Weise kann situationsabhängig entschieden werden, ob eine KI-Empfehlung zulässig ist oder ein Risiko birgt.

Außerdem besteht die Möglichkeit, die Konfidenzbewertung des KI-Modells selbst in die Analyse miteinzubeziehen. Es sei darauf hingewiesen, dass eine Vielzahl zeitgenössischer Modelle Wahrscheinlichkeiten oder Unsicherheitsmaße bereitstellen. Diese ermöglichen die automatisierte Eliminierung von Ergebnissen mit geringer Vertrauenswürdigkeit (vergleiche *Jourdan et al.* [17]) oder die Zuweisung einer spezifischen Behandlung.

Die Stärke dieser Validierungsmechanismen liegt in ihrer Automatisierbarkeit, Integrationsfähigkeit und Geschwindigkeit. Die nahtlose Einbindung in bestehende Prozessleitsysteme, MES oder Edge-Komponenten ermöglicht eine schnelle, konsistente Bewertung von KI-Ausgaben, noch bevor diese produktiv wirksam werden. Damit schaffen sie die Voraussetzung, um auch in hochdynamischen Produktionsumgebungen sicher und kontrolliert von extern generierten KI-Ergebnissen zu profitieren. Dies ist besonders dann relevant, wenn die Eingabedaten unvollständig, abstrahiert oder anonymisiert vorlagen.

3.2 Simulatives Testen

Simulatives Testen stellt eine zentrale Methode dar, um KI-Ergebnisse aus externen Mehrwertdiensten sicher zu bewerten, bevor sie in produktive Abläufe integriert werden. Die Grundlage dieses Ansatzes bildet der Einsatz digitaler Modelle oder digitaler Schatten (siehe Definition von *Kritzinger et al.* [18]), welche die realen Produktionsprozesse in einem virtuellen Raum abbilden. Diese Modelle nutzen aktuelle oder historische Produktionsdaten, um die Auswirkungen von KI-generierten Entscheidungen oder Empfehlungen unter kontrollierten Bedingungen nachzustellen, ohne dass reale Prozesse beeinflusst werden.

Das Ziel ist, potenzielle Auswirkungen auf Effizienz, Qualität, Stabilität oder Sicherheit bereits vor der Integration ungeprüfter Ergebnisse zu identifizieren. Dies erfolgt mithilfe automatisierter Auswertung von Kennzahlen und Prozessmetriken. Änderungen an Prozessparametern, vorgeschlagene Optimierungen oder Anomalieprognosen können folglich unter realitätsnahen Bedingungen durchgespielt und anhand definierter Zielgrößen objektiv bewertet werden (siehe Beispiel der virtuellen Inbetriebnahme nach *Wünsch* [19]).

Besonders effektiv ist dieses Vorgehen, wenn es in automatisierte Testframeworks eingebettet ist (siehe *Kübler et al.* [20]). Diese durchlaufen systematisch eine Vielzahl von Testszenarien, variieren relevante Eingabegrößen und analysieren die Reaktion des Systems auf die vorgeschlagenen Maßnahmen. Auf Basis der Simulationsergebnisse lassen sich Aussagen darüber treffen, ob ein KI-Ergebnis in seiner aktuellen Form produktiv nutzbar ist oder einer weiteren Anpassung oder Prüfung bedarf.

Durch diesen methodischen Zwischenschritt entsteht ein valides Entscheidungsfundament, das es ermöglicht, auch unter Unsicherheiten oder bei eingeschränkter Datenbasis mit KI-gestützten Ergebnissen zu arbeiten – jedoch erst, nachdem diese simulativ validiert wurden. Das simulative Testen unterstützt somit eine sichere und automatisierte Rückführung externer KI-Ergebnisse in produktive Umgebungen und bildet eine unverzichtbare Komponente im Rahmen einer kontrollierten KI-Integration.

3.3 Selbstanpassung

Im Falle von Abweichungen, Unsicherheiten oder unzureichenden Ergebnissen bei der Rückspielung von KI-Ausgaben identifizieren Validierungsmechanismen oder simulatives Testen derartige Unzulänglichkeiten. Ein dritter Ansatz, der eine Möglichkeit zur automatisierten Korrektur bietet, ist in diesem Fall die adaptive Selbstanpassung. Das Ziel dieses Mechanismus ist es,

fehlerhafte oder suboptimale KI-Ergebnisse nicht nur abzulehnen, sondern systematisch zu verbessern und zu einem verwertbaren Zustand weiterzuentwickeln, ohne manuelles Eingreifen.

Ein zentraler Ansatz ist dabei der Einsatz von Reinforcement Learning (RL). In diesem Lernverfahren interagiert ein Agent mit einer Umgebung, beispielsweise einem digitalen Modell der Produktion, und erlernt durch Rückmeldungen („Rewards“), welche Handlungen zu den angestrebten Zielgrößen führen (siehe *Jaensch et al.* [21]). In Kombination mit der zuvor beschriebenen simulationsgestützten Testumgebung kann RL genutzt werden, um ein KI-Ergebnis iterativ zu optimieren: Wird beispielsweise eine durch KI erzeugte Bahnplanung für ein NC-Programm als nicht zielführend validiert, kann ein RL-Agent verschiedene Varianten durchspielen, bis eine Lösung gefunden wird, die den Qualitäts- und Effizienzkriterien entspricht. Im Anschluss erfolgt die Korrektur des Lösungswegs, welche anschließend zurückgeführt und bei Bedarf für zukünftige Entscheidungen gespeichert wird.

Eine zweite Möglichkeit der Selbstanpassung besteht in der rückgekoppelten Neuberechnung durch den ursprünglichen KI-Mehrwertdienst. Vor allem bei generativen Modellen wie Large Language Models (LLMs) besteht die Möglichkeit, auf Basis der Ergebnisse aus Validierung oder Simulation gezielte Korrekturanfragen zu stellen. Ein LLM, das beispielsweise Prozessparameter oder Konfigurationsvorschläge generiert hat, kann durch einen Dialogmechanismus zu einer neuen, verbesserten Antwort angeregt werden. Dies kann etwa durch die explizite Rückmeldung erfolgen: „Das empfohlene Temperaturprofil überschreitet die zugelassenen Grenzwerte – bitte berechnen Sie eine alternative Lösung unter Berücksichtigung dieser Einschränkung.“ Der entstehende iterative Optimierungsprozess ermöglicht eine schrittweise Annäherung des Modells an die Produktionslogik unter kontrollierten Bedingungen (siehe hierzu ein Beispiel aus der Fabrikplanung [22]).

Die Analyse von Reinforcement Learning im simulierten Raum und der Einsatz dialogischer Korrektur über adaptive KI-Dienste erlauben nicht nur die Identifizierung inkonsistenter Ergebnisse, sondern auch deren Konvertierung in anwendbare Lösungen. Die Kombination dieser Mechanismen gewährleistet, dass externe KI-Dienste nicht als starre Blackbox agieren, sondern sich dynamisch in industrielle Abläufe einfügen lassen. Die vorliegende Arbeit kommt zu dem Schluss, dass die Selbstanpassung ein entscheidendes Element für eine robuste, automatisierte und lernfähige Integration von KI in die Produktion darstellt.

4 Fazit

Der Einsatz von KI-Mehrwertdiensten in externen, potenziell unsicheren Umgebungen eröffnet signifikante Potenziale für industrielle Anwendungen, besonders in den Bereichen Prozessoptimierung, Entscheidungsunterstützung und Effizienzsteigerung. Gleichzeitig ergeben sich hohe Anforderungen an den sicheren Umgang mit sensiblen Daten sowie an die Verlässlichkeit der zurückfließenden KI-Ergebnisse.

Der vorliegende Beitrag hat dargelegt, auf welche Weise Unternehmen mittels regelbasierter Prüfmechanismen, simulativer Tests und adaptiver Selbstanpassungen KI-Ergebnisse sichern und in bestehende Prozesse integrieren können. Unter gewissen Voraussetzungen erlauben diese Verfahren eine automatisierte Rückführung, beispielsweise in Anwendungsbereichen, die nicht sicherheitsrelevant sind oder in denen eine hinreichende Abgrenzung gegeben ist.

Es ist jedoch festzustellen, dass eine vollständige Automatisierbarkeit aktuell noch deutliche Grenzen hat. Insbesondere, wenn KI-Modelle auf abstrahierten oder anonymisierten Daten basieren, ist eine abschließende Verifikation der Ergebnisse oft nur durch manuelle Überprüfung verlässlich prüfbar.

Die präsentierten Konzepte bilden die Grundlage für den verantwortungsvollen Einsatz von KI-Mehrwertdiensten. Für eine umfassende und sichere Automatisierung sind künftig weiterentwickelte Validierungsverfahren, standardisierte Testumgebungen und vertrauenswürdige Ausführungsinfrastrukturen erforderlich.

LITERATUR

[1] Arinez, J. F.; Chang, Q.; Gao, R. X. et al.: Artificial Intelligence in Advanced Manufacturing: Current Status and Future Outlook. Journal of Manufacturing Science and Engineering 142 (2020) 11, doi.org/10.1115/1.4047855

[2] Fahle, S.; Prinz, C.; Kuhlenkötter, B.: Systematic review on machine learning (ML) methods for manufacturing processes – Identifying artificial intelligence (AI) methods for field application. Procedia CIRP 93 (2020), pp. 413–418

[3] Ucar, A.; Karakose, M.; Kırımça, N.: Artificial Intelligence for Predictive Maintenance Applications: Key Components, Trustworthiness, and Future Trends. Applied Sciences 14 (2024) 2, #898

[4] Leo Kumar, S. P.: Knowledge-based expert system in manufacturing planning: state-of-the-art review. International Journal of Production Research 57 (2019) 15–16, pp. 4766–4790

[5] Taesi, C.; Aggogeri, F.; Pellegrini, N.: COBOT Applications—Recent Advances and Challenges. Robotics 12 (2023) 3, #79

[6] Bundesministerium für Wirtschaft und Klimaschutz BMWK (Hrsg.): KI-Einsatz in Unternehmen in Deutschland. Strategische Ausrichtung und internationale Position. Stand: 2024. Internet: ftp.zew.de/pub/zew-docs/gutachten/KI-Report_07_Strategie2024.pdf. Zugriff am 02.06.2025

[7] Bundesministerium für Wirtschaft und Energie BMWi (Hrsg.): Herausforderungen beim Einsatz von Künstlicher Intelligenz. Ergebnisse einer Befragung von jungen und mittelständischen Unternehmen in Deutschland. Stand: 2021. Internet: www.de.digital/DIGITAL/Redaktion/DE/Digitalisierungsindex/Publikationen/publikation-download-ki-herausforderungen.pdf?__blob=publicationFile&v=3. Zugriff am 02.06.2025

[8] Ali, M.; Tang Jung, L.; Hassan Sodhro, A. et al.: A Confidentiality-based data Classification-as-a-Service (C2aaS) for cloud security. Alexandria Engineering Journal 64 (2023), pp. 749–760

[9] Dwork, C.: Differential Privacy. In: Hutchison, D.; Kanade, T.; Kittler, J. et al. (eds.): Automata, languages and programming. Proceedings of the 33rd international colloquium, ICALP 2006, Venice, Italy, July 10–14, 2006, pp. 1–12

[10] Fehling, C.; Leymann, F.; Retter R. et al.: Cloud computing patterns: fundamentals to design, build, and manage cloud applications. Wien: Springer Vienna 2014

[11] Sandhu, R. S.: Role-based Access Control. Advances in Computers 46 (1998), pp. 237–286

[12] Pahl, C.: Containerization and the PaaS Cloud. IEEE Cloud Computing 2 (2015) 3, pp. 24–31

[13] AlJahdali, H.; Albatli, A.; Garraghan, P. et al.: Multi-tenancy in Cloud Computing. *2014 IEEE 8th International Symposium on Service Oriented System Engineering*, Oxford, UK, 2014, pp. 344–351, DOI doi.org/10.1109/SOSE.2014.50

[14] Zhang, C.; Xie, Y.; Bai, H. et al.: A survey on federated learning. Knowledge-Based Systems 216 (2021),#106775

[15] Sabt, M.; Achemlal, M.; Bouabdallah, A.: Trusted Execution Environment: What It is, and What It is Not. *2015 IEEE Trustcom/BigDataSE/ISPA*, Helsinki, Finland, 2015, pp. 57–64, doi.org/10.1109/Trustcom.2015.357

[16] Masri, N.; Abu Sultan, Y.; Akkila, A. N. et al.: Survey of rule-based systems. International Journal of Academic Information Systems Research 3 (2019) 7, pp. 01–22

[17] Jourdan, N.; Sen, S.; Husom, E. J. et al.: On the reliability of machine learning applications in manufacturing environments. ArXiv 2021, doi.org/10.48550/arXiv.2112.06986

[18] Kritzinger, W.; Karner, M.; Traar, G. et al.: Digital Twin in manufacturing: A categorical literature review and classification. IFAC-PapersOnLine 51 (2018) 11, pp. 1016–1022
[19] Reinhart, G.; Wünsch, G.: Economic application of virtual commissioning to mechatronic production systems. Production Engineering 1 (2007) 4, pp. 371–379
[20] Kübler, K.; Krebser, G.; Verl, A.: Testautomatisierung am Digitalen Zwilling. In: Verl, A.; Röck, S.; Scheifele, C. (Hrsg.): Echtzeitsimulation in der Produktionsautomatisierung. Heidelberg: Springer Vieweg 2024, S. 193–211
[21] Jaensch, F.; Klingel, L.; Verl, A.: Virtual Commissioning Simulation as OpenAI Gym – A Reinforcement Learning Environment for Control Systems. 2022 5th International Conference on Artificial Intelligence for Industries (AI4I), Laguna Hills, CA, USA, 2022, pp. 64–67
[22] Tinsel, E.-F.; Lechler, A.; Riedel, O. et al.: Concept of an Initial Requirements-Driven Factory Layout Planning and Synthetic Expert Verification for Industrial Simulation Based on LLM. 2024 IEEE 22nd International Conference on Industrial Informatics (INDIN), Beijing, China, 2024, pp. 1–6

Erik-Felix Tinsel, M.Sc.
erik-felix.tinsel@isw.uni-stuttgart.de
Tel. +49 (0)711 / 685-84510

Prof. Dr.-Ing. Oliver Riedel

Institut für Steuerungstechnik der Werkzeugmaschinen
und Fertigungseinrichtungen ISW
Universität Stuttgart
Seidenstr. 36, 70174 Stuttgart
www.isw.uni-stuttgart.de