Bundesamt
für die Sicherheit
der nuklearen Entsorgung

RESSORTFORSCHUNGSBERICHTE ZUR
SICHERHEIT DER NUKLEAREN ENTSORGUNG

Einsatz und Qualifizierung von Methoden der künstlichen Intelligenz in kerntechnischen Anlagen

Vorhaben FKZ 4722R01290

AUFTRAGNEHMER:IN
Gesellschaft für Anlagen- und Reaktorsicherheit (GRS) gGmbH

Patrick Gebhardt
Hervé Mbonjo
Jaroslaw Shvab

Einsatz und Qualifizierung von Methoden der künstlichen Intelligenz in kerntechnischen Anlagen

BASE-RESFOR-002/25

Bitte beziehen Sie sich beim Zitieren dieses Dokumentes immer auf folgende URN:
urn:nbn:de:0221-2025052052251

Berlin, Juli 2024

Impressum

Bundesamt
für die Sicherheit
der nuklearen Entsorgung
(BASE)

BASE – FORSCHUNGSBERICHTE ZUR
SICHERHEIT DER NUKLEAREN ENTSORGUNG

Auftragnehmer:in
Gesellschaft für Anlagen- und Reaktorsicherheit (GRS) gGmbH

Patrick Gebhardt
Hervé Mbonjo
Jaroslaw Shvab

030 184321–0
www.base.bund.de

Stand: Juli 2024

Gesellschaft für Anlagen- und Reaktorsicherheit (GRS) gGmbH

Einsatz und Qualifizierung von Methoden der künstlichen Intelligenz in kerntechnischen Anlagen

Patrick Gebhardt
Hervé Mbonjo
Jaroslaw Shvab

Juli 2024
4722R01290

Anmerkung:

Dieser Bericht wurde von der Gesellschaft für Anlagen- und Reaktorsicherheit (GRS) gGmbH im Auftrag des Bundesministeriums für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz (BMUV) im Rahmen des Vorhabens 4722R01290 erstellt.

Der Bericht gibt die Auffassung und Meinung des Auftragnehmers wieder und muss nicht mit der Meinung

GRS-A-4147

Deskriptoren

Künstliche Intelligenz; Maschinelles Lernen; Kerntechnik;

Kurzfassung

In den letzten Jahren hat die rasante Entwicklung der künstlichen Intelligenz (KI) das Potenzial von maschinellem Lernen und neuronalen Netzen für sicherheitskritische Anwendungen zunehmend in den Fokus gerückt. Angesichts der wachsenden Bedeutung von KI-basierten Systemen in Automatisierung und Optimierung ist eine fundierte Bewertung ihrer Zuverlässigkeit und Sicherheit unerlässlich, insbesondere für den Einsatz in kerntechnischen Anlagen. Dieses Auftragsforschungsvorhaben zielte darauf ab, eine solide Grundlage für den Einsatz von KI-Methoden in sicherheitsrelevanten Bereichen zu schaffen. Die Analyse umfasst die Bewertung des aktuellen Forschungsstands und die Identifikation relevanter KI-basierten Anwendungen im sicherheitstechnischen und kerntechnischen Umfeld. Dazu wurden aktuelle KI-Methoden, Klassifikations- und Qualifikationsansätze systematisch analysiert, wobei wissenschaftliche Publikationen, Konferenzbeiträge sowie einschlägige Normen und Standards herangezogen wurden. Zudem wurden Ansätze für die Qualifikation von KI-basierten Systemen und eine mögliche Bewertung dieser untersucht. Die Vorhabensergebnisse sollen eine Basis bieten, um eine Brücke zwischen KI-basierten Anwendungen und den hohen Anforderungen sicherheitskritischer Industrien zu schlagen.

Abstract

In recent years, the rapid development of artificial intelligence (AI) has increasingly focussed on the potential of machine learning and neural networks for safety-critical applications. Given the growing importance of AI-based systems in automation and optimisation, a well-founded evaluation of their reliability and safety is essential, especially for use in nuclear facilities. This project aims to create a solid basis for the use of AI methods in safety-relevant areas. The analysis includes the evaluation of the current state of research, the identification of relevant AI-based applications in the safety and nuclear environment. To this end, current AI methods, classification and qualification approaches were systematically analysed using scientific publications, conference papers and relevant norms and standards. In addition, approaches for the qualification of AI-based systems and a possible evaluation of these were analysed. The project results are intended to provide a basis for building a bridge between AI-based applications and the high requirements of safety-critical industries.

Inhaltsverzeichnis

1 Einleitung

Die Methoden der künstlichen Intelligenz (KI) wie beispielsweise das maschinelle Lernen (ML), Expertensysteme oder genetische Algorithmen haben in den letzten Jahren große Fortschritte gemacht. Die erste Hochphase der KI-Entwicklung war in den 50er Jahren nachdem die ersten Neuronalen Netze (Perzeptron) entwickelt wurden. Einige Jahre später erfolgte die erste Tiefphase (sog. KI-Winter) der KI-Entwicklung. Aufgrund der unzureichenden Rechenkapazität und grundsätzlicher Einschränkungen geriet die Weiterentwicklung ins Stocken. Bald darauf folgten eine erneute Hochphase und ein erneuter KI-Winter. Durch den technischen Fortschritt in den letzten Jahren befindet sich die KI-Entwicklung aktuell wieder in einer Hochphase.

Durch die Weiterentwicklung künstlicher Intelligenz hin zu Deep Neural Network (DNN) und Machine Learning (ML) Methoden in den letzten Jahren wurden Dutzende neue Möglichkeiten entwickelt, ein KI-basiertes System in diversen Bereichen einzusetzen, wie beispielsweise die Möglichkeit des autonomen Fahrens. So wurde im Dezember 2021 die Genehmigung für autonomes Fahren Level 3 zum ersten Mal erteilt. Auch in kerntechnischen Anlagen findet der Einsatz von KI-basierten Systemen immer größeren Zuspruch. So hat bei einer Umfrage der U.S. NRC der Hersteller Framatome angegeben, bereits mehrere Veröffentlichungen mit Bezug zu KI-basierten Systemen verfasst zu haben. Andere Hersteller kerntechnischer Anlagen wie Westinghouse verwenden bereits KI-basierte Tools für die Datenanalyse. Besonders interessant ist für einige Hersteller die Verwendung eines digitalen Zwillings (Digital Twin – DT). Hierzu laufen bereits einige Konzeptionsphasen. Mit der Digitalisierung der Leittechnik (LT) besteht auch die Möglichkeit, dass ein KI-basiertes System sowohl beim Design der LT-Architektur als auch während des Betriebs der Anlage unterstützend eingesetzt wird. Diese Beispiele zeigen die vielfältigen Einsatzmöglichkeiten von KI-basierten Systemen in kerntechnischen Anlagen. Da in dem Bereich der KI-basierten Systeme in Anwendungen mit sicherheitstechnischer Bedeutung mit einer fortschreitenden Entwicklung des Standes von Wissenschaft und Technik zu rechnen ist, ist insbesondere vor dem Hintergrund des Einsatzes von KI-basierten Systemen mit sicherheitstechnischer Bedeutung in kerntechnischen Anwendungen ein Kompetenzaufbau hinsichtlich relevanter Fragestellungen zum Thema künstliche Intelligenz auf nationaler und internationaler Ebene dringend notwendig.

Besonders relevant ist in diesem Zusammenhang die Frage nach Qualifikationsansätzen für KI-basierte Systeme mit sicherheitstechnischer Bedeutung im Hinblick auf potenzielle

kerntechnische Anwendungen und deren Bewertung. Die Relevanz ergibt sich u. a. sowohl aus der internationalen Zusammenarbeit auf dem Gebiet der kerntechnischen Sicherheit und der Aktualisierung des internationalen kerntechnischen Regelwerks als auch durch Genehmigungsfragen bei einem möglichen Einsatz KI-basierter Systeme in deutschen Anlagen der nuklearen Ver- und Entsorgung, oder in Kernkraftwerken in Stilllegung und Rückbau.

Das übergeordnete Ziel des vorliegenden Vorhabens ist es deshalb, einen Beitrag zu einem entsprechend fundierten Wissens-/Kompetenzaufbau zu derzeitigen Fragestellungen über den potenziellen Einsatz von KI-basierten Systemen mit sicherheitstechnischer Bedeutung in kerntechnischen Anwendungen zu leisten. Hierzu wird u. a. ein umfassenden Überblick über die Möglichkeiten und Herausforderungen gegeben, welche der Einsatz von KI-basierten Systemen in sicherheitskritischen Bereichen mit sich bringt.

Im Einzelnen sind die Ziele des Vorhabens:

- Die Ermittlung des Standes von Wissenschaft und Technik beim Einsatz von KI-Methoden in Anwendungen mit sicherheitstechnischer Bedeutung, insbesondere im Bereich der funktionalen Sicherheit. (AP 1)
- Die Ermittlung von beispielhaften KI-basierten Anwendungen in der Kerntechnik (in Kernkraftwerken und anderen kerntechnische Anlagen) aus dem In- und Ausland. Dabei sollen sowohl Forschungs- und Entwicklungsarbeiten als auch der produktive Einsatz betrachtet werden. (AP 2)
- Die Ermittlung und Bewertung von Qualifikationsansätzen und -methoden für KI-basierte Anwendungen mit sicherheitstechnischer Bedeutung, und deren Bewertung insbesondere mit Blick auf kerntechnische Rahmenbedingungen.

Die diesen einzelnen Zielen des Vorhabens zugeordneten Arbeitspakete 1 bis 3 werden beschrieben in den nachfolgenden Abschnitten 1.1 bis 1.3.

1.1 Arbeitspaket 1: Ermittlung des Standes von Wissenschaft und Technik beim Einsatz von KI-Methoden in Bereichen mit sicherheitstechnischer Bedeutung

In diesem Arbeitspaket wurde zunächst der Stand von Wissenschaft und Technik zu KI-Methoden in Bereichen mit sicherheitstechnischer Bedeutung ermittelt. Dies erfolgte sowohl durch eine klassische Literaturrecherche als auch durch den Besuch diverser Konferenzen zum Thema KI. Dieses Arbeitspaket teilt sich auf zwei Arbeitspunkte auf. Im ersten Arbeitspunkt (AP 1.1) erfolgte die Recherche zu KI-Methoden und deren Einsatz insbesondere in Bereichen mit sicherheitstechnischer Bedeutung. Hierzu wurden wissenschaftlichen Publikationen, Aktivitäten und Publikationen von Verbänden sowie internationalen Organisationen und Normungs- und Regelsetzungsgremien, auf internationalen und europäischen Konferenzen sowie in weiteren Quellen, falls notwendig, herangezogen.

Im Rahmen dieser Recherche wurde u. a. auf folgende Fragestellungen eingegangen:

- Welche KI-Methoden kommen in Bereichen mit sicherheitstechnischer Bedeutung zum Einsatz?
- Bei welchen Anwendungen mit sicherheitstechnischer Bedeutung kommen KI-Methoden zum Einsatz?
- Welche Anwendungen mit sicherheitstechnischer Bedeutung bestehen, die nicht unmittelbar den Bereich der funktionalen Sicherheit betreffen?
- Wo liegen die Grenzen der Anwendbarkeit von KI-Methoden in Anwendungen mit sicherheitstechnischer Bedeutung?

Für die Ermittlung von KI-Methoden in Bereichen mit sicherheitstechnischer Bedeutung wurden auch Publikationen, Regelwerke und technische Reports, die während der Vorhabenslaufzeit veröffentlicht wurden, berücksichtigt.

Im zweiten Arbeitspunkt (AP 1.2) erfolgte eine für die Zwecke des Vorhabens bedarfsgerechte Klassifizierung der KI-Methoden und eine übersichtliche Darstellung aller KI-Methoden, die sich für einen Einsatz in sicherheitsrelevanten Bereichen potenziell eignen.

Hierfür wurden die erarbeiteten Rechercheergebnisse aus dem AP 1.1 nach anerkannten Klassifikationen (z. B. /LAI 21/) für KI-Methoden eingeteilt und anschließend in übersichtlicher Form dargestellt.

Auf folgende Fragen zur Klassifizierung wurde dabei eingegangen:

- Wie können die KI-Methoden klassifiziert werden?
- Wie kann die sicherheitstechnische Bedeutung einer KI-Anwendung klassifiziert werden?

Die Ergebnisse des Arbeitspakets 1 sind im Kapitel 2 dargestellt.

1.2 Arbeitspaket 2: Ermittlung von beispielhaften KI-basierten Anwendungen in der Kerntechnik und in anderen Bereichen mit sicherheitstechnischer Bedeutung

Im Rahmen des Arbeitspakets 2 wurden Recherchen zu beispielhaften KI-basierten Anwendungen durchgeführt. Es wurden sowohl KI-basierte Anwendungen im Rahmen von Forschungs- und Entwicklungsarbeiten als auch produktive Einsätze von KI-basierten Systemen in der Praxis betrachtet. Ebenso wurden sowohl online arbeitende Systeme, die möglicherweise direkt auf den Anlagenprozess wirken, als auch Systeme, die der Planung, Betriebsoptimierung oder Unterstützung des (Schicht)-Personals dienen, berücksichtigt. Wesentlich für die Ermittlung der KI-basierten Anwendungen sind zum einen Literatur- und Konferenzrecherchen mit dem Thema KI und identifizierte Anwendungen aus dem Arbeitspaket 1 sowie weitere Recherchen zur Ermittlung der KI-basierten Anwendungen aus anderen Quellen. Der Schwerpunkt der Recherche in diesem Arbeitspaket lag insbesondere auf Kernkraftwerke, jedoch wurden auch weitere Bereiche in der Kerntechnik wie z. B. Anlagen der nuklearen Ver- und Entsorgung oder andere Bereiche mit sicherheitstechnischer Bedeutung wie z. B. ausgewählte Anwendungen in der Forschung berücksichtigt.

Die folgenden Aspekte wurden bei dieser Recherche u. a. betrachtet:

- Welche KI-Methoden kommen zum Einsatz?
- In welchem Rahmen erfolgt der Einsatz (Forschungs- oder Entwicklungsarbeit, Konzeptstudie, produktiver Einsatz, etc.)?
- Bei datengetriebenen Methoden/maschinellem Lernen: Welche Datengrundlage benutzt die Methode (z. B. welche Trainingsdaten, Herkunft, Validierung, etc.)?
- Wie wurde die Anwendung verifiziert und validiert?
- Bei Anwendungen mit sicherheitstechnischer Bedeutung: Wie ist die sicherheitstechnische Bedeutung zu bewerten (z.B. anhand einschlägiger Regelwerke, Kategorien nach KTA 3501, etc.)?
- Besteht ein (potenzieller) Sicherheitsgewinn gegenüber konventionellen Methoden?
- Wo liegen potenzielle Sicherheitsrisiken?

Die Rechercheergebnisse wurden zunächst nach anerkannten Methoden klassifiziert und anschließend übersichtlich visualisiert. Ferner wurden KI-basierte Anwendungen aus dem kerntechnischen Bereich mit dem Einsatz von KI-basierten Anwendungen in anderen sicherheitsrelevanten Bereichen verglichen. Dadurch können zukünftige potenzielle KI-basierte Anwendungen für kerntechnische Bereiche abgeschätzt werden.

Die Ergebnisse des Arbeitspakets 2 sind im Kapitel 3 dargestellt.

1.3 Arbeitspaket 3: Ermittlung und Bewertung von Qualifikationsansätzen und -methoden für KI-basierte Anwendungen mit sicherheitstechnischer Bedeutung

Im Rahmen dieses Arbeitspakets wurden zunächst Qualifikationsansätze und -methoden für KI-basierte Anwendungen mit sicherheitstechnischer Bedeutung ermittelt (Arbeitspunkt 3.1) und anschließend bewertet (Arbeitspunkt 3.2).

Im Rahmen dieses Arbeitspakets wurden zunächst Qualifikationsansätze und -methoden für KI-basierte Anwendungen mit sicherheitstechnischer Bedeutung ermittelt. Folgende Bestandteile der ermittelten Qualifikationsprozesse werden dabei untersucht:

- Lebenszyklus von KI-basierten Anwendungen (mit sicherheitstechnischer Bedeutung),
- Spezifikation von KI-basierten Systemen, insbesondere solche in Einsatzgebieten mit sicherheitstechnischer Bedeutung,
- Implementierung und erforderliche Eigenschaften der Datengrundlage, insbesondere die Schwierigkeit der Gewinnung von Trainingsdaten für Anwendungen mit maschinellem Lernen,
- Erforderliche Eigenschaften der Daten und deren Sicherstellung (z. B. Repräsentativität, Unabhängigkeit, Verzerrungsfreiheit (Unbiasedness), etc.),
- Verifikation und Validation von KI-basierten Systemen,
- Weitere relevante Bestandteile des Qualifikationsprozesses, sofern erkennbar und relevant.

Zur Bewertung der Qualifikationsansätze und Methoden wurden u. a. folgende Punkte betrachtet:

- die Forderung von Determinismus vs. die stochastische Natur von KI-Algorithmen,
- die fehlende Nachvollziehbarkeit von KI-Algorithmen,
- mögliche Schwierigkeiten bei der Verifikation und Validation.

Ferner wurde auch anhand des ermittelten Standes von Wissenschaft und Technik darauf eingegangen, bis zu welcher sicherheitstechnischen Bedeutung eine KI-basierte Anwendung in der Kerntechnik denkbar ist.

Zudem wurde die Übertragbarkeit bestehender Anforderungen an Software in softwarebasierter Leittechnik auf KI-basierte Anwendungen bewertet.

Die Ergebnisse dieses Arbeitspakets sind im Kapitel 4 zusammengefasst.

2 Stand von Wissenschaft und Technik beim Einsatz von KI-Methoden in Bereichen mit sicherheitstechnischer Bedeutung

2.1 Zum KI-Begriff, Klassifizierung und Transparenz von KI-Methoden

Die Klassifizierung von KI-Methoden ist für das Schaffen eines systematischen Bewertungs- und Regulierungsrahmens von zentraler Bedeutung. Durch eine Einteilung in klar definierte Kategorien ist es möglich unterschiedliche Methoden systematisch zu organisieren und zu bewerten, zum Bespiel durch das Hervorzuheben spezifischer Vor- und Nachteile verschiedener KI-Methoden bezogen auf die Anwendung.

2.1.1 Klassifizierung von KI-Methoden nach IEC TR 63468

Der Technical Report IEC TR 63468 /IEC 23/ bietet eine Klassifizierung von KI-Methoden nach einer hierarchischen Struktur. Die Klassifikation ist in der Abb. 2.1 dargestellt.

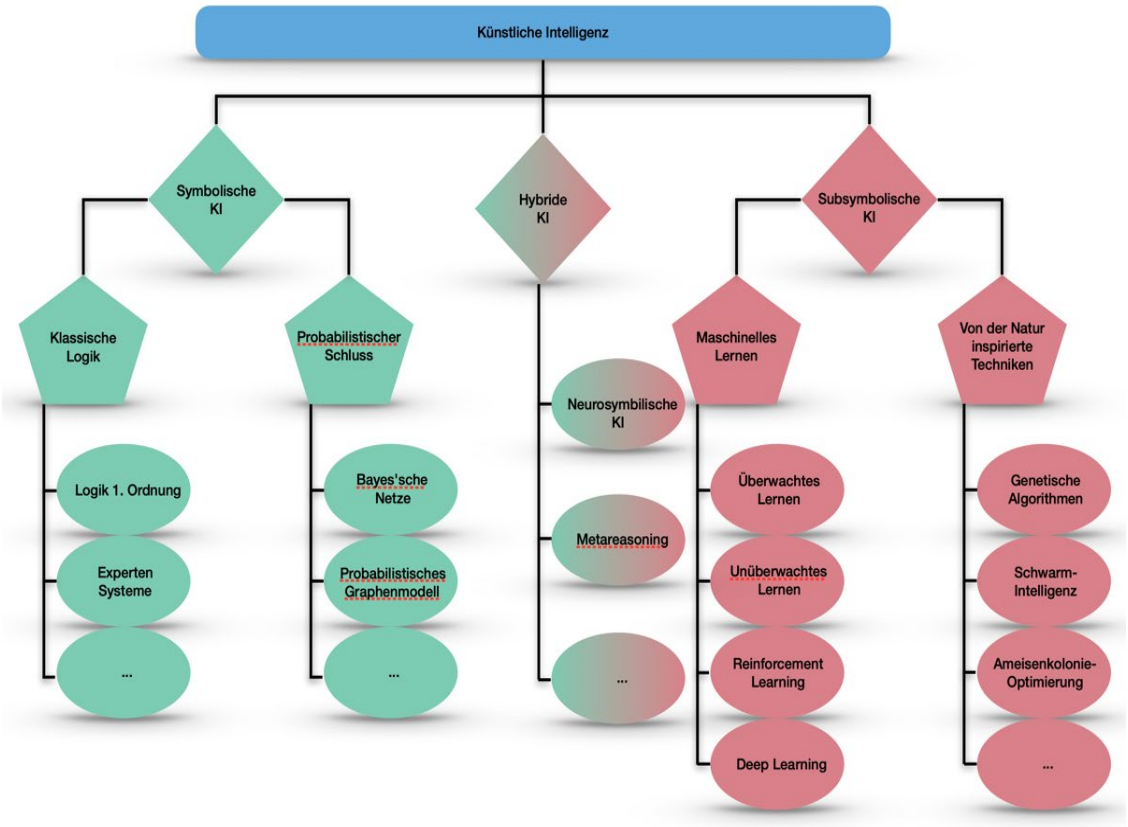

Abb. 2.1 Klassifikation von KI-Methoden (eigene Darstellung) /IEC 23/.

Wie in der hierarchischen Darstellung zu sehen ist, werden die Modelle zunächst in drei Kategorien eingeteilt – Symbolische KI; Subsymbolische KI; Hybride KI. Diese können weiter in Unterkategorien eingeteilt werden. Im Kapitel 2.2 wird im Detail auf einige im hierarchischen Diagramm dargestellte Methoden eingegangen.

Symbolische Künstliche Intelligenz basiert auf der Verwendung von expliziten Regeln und Symbolen, um Wissen darzustellen, zu verarbeiten und Schlussfolgerungen abzuleiten. Diese Form der KI beruht auf mathematischer Logik und der Manipulation symbolischer Repräsentationen der betrachteten Gegenstände, die menschliches Wissen über die Gegenstände explizit modellieren, und den menschlichen Denkprozessen nachempfunden sind. Typische Anwendungen der symbolischen KI umfassen Expertensysteme, die Entscheidungsbäume und logische Schlussfolgerungen nutzen, um Probleme zu lösen. Symbolische KI ist besonders geeignet für Anwendungen, die klare Regeln und strukturiertes Wissen erfordern.

Subsymbolische Künstliche Intelligenz arbeitet auf der Grundlage von neuronalen Netzen und anderen lernbasierten Modellen, die Datenmuster und Korrelationen erkennen können, ohne explizite Regeln zu benötigen. Einige dieser KI-Methoden imitieren die Funktionsweise des menschlichen Gehirns und können komplexe, unstrukturierte Daten wie Bilder, Sprache und Texte verarbeiten. Obwohl die Berechnungsschritte einige dieser Methoden formal nachvollzogen werden können, werden diese nicht als symbolische KI-Methoden angesehen, da die Entscheidungslogik nicht in Form von vordefinierten Regeln, sondern datengetrieben entsteht. Subsymbolische KI eignet sich besonders für Aufgaben, bei denen Mustererkennung und maschinelles Lernen im Vordergrund stehen.

Hybride Künstliche Intelligenz kombiniert die Stärken sowohl der symbolischen als auch der subsymbolischen KI. Durch die Integration von logikbasierten und datenbasierten Ansätzen ermöglicht hybride KI die Entwicklung von Systemen, die sowohl explizite Regeln als auch komplexe Muster verarbeiten können. Dies umfasst komplexe Systeme wie autonome Fahrzeuge, die sowohl symbolische Methoden zur Interpretation von Verkehrsregeln als auch neuronale Netze zur Verarbeitung von Sensor- und Bilddaten verwenden müssen.

2.1.2 Klassifizierung von KI-basierten Anwendungen nach dem risikobasierten Ansatz der EU im EU AI Act

Die Europäische Verordnung über künstliche Intelligenz (EU KI-Verordnung) /EUR 24a/ bietet eine Möglichkeit KI nach Anwendungen zu klassifizieren und beruht auf einem risikobasierten Ansatz. Ein Beispiel für die Klassifizierung nach der EU KI-Verordnung ist in dem Dokument /EUR 24a/ vorzufinden. Die EU KI-Verordnung etabliert einen umfassenden Gesetzesrahmen für die Regulierung der Anwendung von KI innerhalb der Europäischen Union. Dieser Rahmen basiert auf einem risikobasierten Ansatz, bei dem KI-basierte Systeme nach ihrem potenziellen Risiko für die Gesellschaft und die Rechte der Bürger in vier Risikoebenen eingeteilt werden. Diese Klassifizierung reicht von minimalem Risiko, das nur leichte Transparenzanforderungen erfordert, bis hin zu unakzeptablem Risiko, das bestimmte KI-basierte Anwendungen vollständig verbietet. Im Weiteren wird im Detail auf die einzelnen Risikoebenen, deren Eigenschaften und auf Beispiele für diese, eingegangen. In der darauffolgenden Abb. 2.2 ist eine Visualisierung des risikobasierten Ansatzes zu sehen.

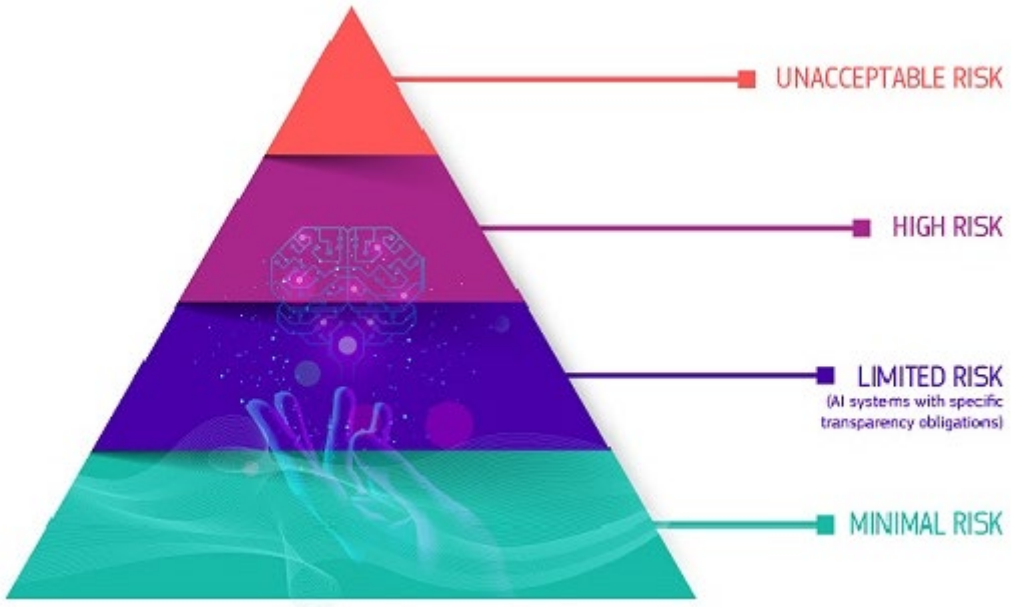

Abb. 2.2 Visualisierung des risikobasierten Ansatzes der EU /EUR 24b/.

Im Rahmen der neuen Regulierung wird besonderer Wert auf die Erklärbarkeit von KI-basierten Systemen und Anwendungen gelegt.

Die Gesetzgebung verlangt von den Entwicklern und Betreibern von KI-basierten Systemen, dass diese Systeme nachvollziehbar und verständlich sind, um Transparenz und Vertrauen zu fördern. Diese Maßnahmen dienen der Erhöhung des Vertrauens in KI-basierte Systeme, indem sie sicherstellen, dass diese fair, transparent und überprüfbar sind. Unternehmen müssen detaillierte technische Dokumentationen bereitstellen, die unter anderem Informationen zur Funktion, Genauigkeit, Qualität der verwendeten Daten, Überwachung und Risikomanagement der KI-basierten Systeme enthalten.

Zudem definiert die EU KI-Verordnung klare Grenzen der Anwendbarkeit für KI-basierte Anwendungen, insbesondere für solche, die als hochriskant eingestuft werden. Anwendungen, die beispielsweise zur sozialen Bewertung von Personen oder zur Massenüberwachung genutzt werden, unterliegen strengen Regulierungen oder sind vollständig untersagt, um die Grundrechte der Bürger zu schützen. Dies sollen sicherstellen, dass KI-basierte Systeme in der EU auf eine Weise entwickelt und eingesetzt werden, die mit den europäischen Werten und dem Schutz der individuellen Rechte im Einklang steht. Auch wenn die E KI-Verordnung primär auf soziale, wirtschaftliche und politische Risiken abzielt, können seine Anforderungen auch auf technische Risiken übertragen werden, da diese in einem Zusammenhang mit sozialen, wirtschaftlichen und politischen Risiken stehen.

Minimal-Risiko-Systeme haben einen sehr geringen oder keinen Einfluss auf die Rechte, die Sicherheit oder die Interessen von Individuen. Solche Systeme unterliegen nur leichten Transparenzanforderungen. Ein Beispiel für ein Minimal-Risiko-System wäre eine einfache KI-basierte Anwendung zur automatischen Rechtschreibkorrektur, die keine sensiblen Daten verarbeitet und keine wesentlichen Entscheidungen trifft.

Begrenztes-Risiko-Systeme stellen ein gewisses Risiko dar, das jedoch als begrenzt angesehen wird. Sie unterliegen Transparenzanforderungen und müssen möglicherweise einer Konformitätsbewertung unterzogen werden, bevor sie auf den Markt gebracht werden dürfen. Ein Beispiel hierfür wäre ein KI-basiertes System zur Empfehlung von Filmen auf einer Streaming-Plattform, das persönliche Vorlieben analysiert, aber keine kritischen Entscheidungen über die Nutzer trifft.

Hochrisiko-Systeme können signifikante Auswirkungen auf die Rechte, die Sicherheit oder die Interessen von Individuen haben. Sie umfassen Anwendungen in Bereichen wie kritische Infrastrukturen, Gesundheitswesen, Transport und Strafverfolgung. Hochrisiko-Systeme müssen strengen Anforderungen hinsichtlich Transparenz, Konformitätsbewertung und spezifischen Anforderungen an Datenqualität, fundamentale Rechte, menschliche Aufsicht und Cybersicherheit entsprechen. Ein Beispiel wäre ein KI-basiertes System zur Diagnose von Krankheiten, das direkt Einfluss auf medizinische Behandlungen und Patientenentscheidungen hat. Hochrisiko-Systeme müssen zwei wesentliche Kriterien erfüllen: Der Einsatz erfolgt in Sektoren, in denen aufgrund der Art der typischen Tätigkeiten erhebliche Risiken zu erwarten sind. Hierbei erfolgt der Einsatz eines KI-basierten Systems in einer Weise, dass mit erheblichen Risiken zu rechnen ist. Beispiele für solche Systeme sind KI-basierte Systeme, die als Sicherheitskomponenten bei der Verwaltung und dem Betrieb kritischer digitaler Infrastrukturen, im Straßenverkehr oder bei der Wasser-, Gas-, Wärme- oder Stromversorgung eingesetzt werden. Weitere Beispiele umfassen die Auswertung biometrischer Daten, Anwendungen in Bildung und Beruf sowie im Bereich der Strafverfolgung.

Unakzeptables Risiko: Diese Kategorie umfasst KI-basierte Systeme, die als zu gefährlich angesehen werden und daher verboten sind. Dazu gehören Anwendungen, die eine soziale Bewertung von Personen durch öffentliche oder private Akteure ermöglichen oder die Menschen ohne ihr Wissen manipulieren. Ein Beispiel wäre ein KI-^basiertes System, das zur Echtzeit-Gesichtserkennung in öffentlichen Räumen durch Strafverfolgungsbehörden eingesetzt wird, außer in sehr spezifischen und begründeten Ausnahmefällen.

2.1.3 Klassifikation nach Anwendungsbereichen im kerntechnischen Bereich

Eine weitere Möglichkeit Klassifikation von KI-basierten Systemen durchzuführen, wurde in der Publikation „Survey on the Use of Artificial Intelligence in Nuclear Power Plants“ von Hyun Seok Noh et al., /HYU 23a/ (A.1.14) welche im Rahmen der PSAM Topical Conference 2023 vorgestellt wurde, präsentiert. Die Publikation befasst sich mit der Anwendung von KI in Kernkraftwerken und kategorisiert bestehende Studien in zwei Hauptbereiche: die spezifischen Anwendungsfelder innerhalb der Kernkraft und die verwendeten Lernalgorithmen. Die Klassifikation nach Anwendungsfeldern der Kernkraft unterteilt sich in vier Kategorien; die Klassifikation ist in Abb. 2.3 dargestellt:

- Diagnose – fokussiert sich auf die Erkennung und Diagnose von Fehlern sowie Abweichungen in sicherheitsrelevanten Komponenten und Systemen. Hier kommen überwiegend überwachte Lernmethoden zum Einsatz, da diese auf vorher klassifizierte Daten zurückgreifen und gezielt auf spezifische Fehler oder Anomalien trainiert werden können. Zu möglichen Anwendungsfällen gehören die Identifikation von Geräteverschleiß, Fehlern im Messkanal, Defekten im Reaktorkühlsystem und Anomalien im Reaktorkern. Beispielsweise würden Methoden wie Long Short-Term Memory, Support Vector Machine und Convolutional Neural Networks verwendet, um Abweichungen in der Leistung von Pumpen oder Korrosion in Sekundärleitungen zu diagnostizieren und somit den Zustand der Anlagen besser zu überwachen und frühzeitig auf potenzielle Probleme aufmerksam machen zu können.

- Prognose – konzentriert sich darauf, künftige Zustände oder kritische Ereignisse, wie Transienten und schwere Unfälle, möglichst präzise vorherzusagen. Dazu wird auf historische Daten und physikalische Modelle zurückgegriffen. Unter den verwendeten Methoden sind Multi-Layer Perceptrons, Recurrent Neural Networks und Random Forests. Diese analysieren große Datenmengen, um Korrelationen und Muster zu identifizieren, die auf bevorstehende Anomalien oder potenzielle Gefährdungen hinweisen. Ein Beispiel wäre die Prognose von Temperaturanstiegen in Wärmetauschern oder die Berechnung der Versagenswahrscheinlichkeit von Ventilen. Diese Vorhersagen können als Grundlage für präventive Maßnahmen dienen und ermöglichen eine gezielte Wartung, was wiederum die Betriebssicherheit und Lebensdauer der Komponenten erhöhen könnte.

- Reaktion - bezieht sich auf die Echtzeitschätzung und das Management von Risikofaktoren, insbesondere im Falle eines Unfalls. Hier sind Methoden gefragt, die in der Lage sind, innerhalb kürzester Zeit Szenarien zu bewerten und Entscheidungshilfen zu bieten. Verstärkendes Lernen, wie Asynchronous Advantage Actor-Critic und Dynamic Graph Neural Networks, wird eingesetzt, um autonome Entscheidungen zu treffen und Reaktionen auf sich schnell entwickelnde Gefahrenlagen zu simulieren. Ein wichtiges Beispiel wäre die Echtzeitschätzung des Unfallverlaufs durch externe Strahlungsdaten, die es ermöglicht, externen Notfällen zu reagieren, selbst wenn Daten aus dem Inneren des Kraftwerks nicht verfügbar sind, oder die Echtzeitanalyse von Netzwerkangriffen, die die Funktionalität der Kernkraftwerksysteme gefährden könnten. Solche Algorithmen sind darauf ausgelegt, den Schutz von Menschen und Umwelt zu maximieren, indem sie bei kritischen Vorfällen schnell eingreifen und Gegenmaßnahmen einleiten können. Ein weiteres Beispiel sind digitale Zwillinge, die ein virtuelles Abbild der Anlage erstellen und diese simulieren. Durch kontinuierliche

Echtzeitdaten aus der Anlage kann ein digitales Modell genutzt werden, um Betriebsabläufe zu simulieren und auf aktuelle Ereignisse entsprechend zu reagieren.

- Prozess - Im Bereich der Prozessoptimierung können KI-basierte Systeme die effiziente Steuerung und Verbesserung des Betriebs unterstützen. Hier geht es darum, den Einsatz von Ressourcen zu optimieren, Ausfälle zu vermeiden und die Lebensdauer der Komponenten zu verlängern. Ein Beispiel wäre die Optimierung des Kühlmittelstroms oder die Optimierung des Brennstoffverbrauchs. Hier kommen Methoden wie genetische Algorithmen und Support Vector Regression zum Einsatz, die den Betriebsprozess durch Simulation und Optimierung so anpassen, dass er sowohl wirtschaftlich als auch sicher bleibt. Auch in diesem Bereich spielen die Entwicklung und der Einsatz digitaler Zwillinge eine wichtige Rolle, da diese virtuellen Abbildungen realer Anlagen schaffen und durch kontinuierliches Monitoring und Lernen Anpassungen und Verbesserungen vorschlagen könnten.

In Bezug auf die Klassifikation der Lernalgorithmen orientiert sich das Paper an den Kategorien des IEC TR 63468-Standards, jedoch nur am subsymbolischen Zweig des Klassifikationsbaums. Dabei besteht die Klassifikation aus der Unterteilung in drei Machine-Learning-Methoden /ISO 22a/, /IEC 23/.

- **Überwachtes Lernen** – Dieses Verfahren lernt anhand von Daten mit bekannten Ergebnissen, um Muster zu erkennen und diese auf neue Daten anzuwenden. Überwachtes Lernen kann zur Optimierung von Betriebsabläufen und zur Fehlererkennung verwendet werden, indem anhand von Messdaten mögliche Fehlfunktionen vorhergesagt werden, bevor sie auftreten.

- **Unüberwachtes Lernen** – Diese Methode kann Muster und Beziehungen in großen Datenmengen ohne vorherige Kennzeichnung identifizieren. In Kernkraftwerken kann unüberwachtes Lernen genutzt werden, um verdeckte Zusammenhänge in Sensordaten zu erkennen und Anomalien zu detektieren, die auf mögliche Anlagenfehler hinweisen könnten.

- **Verstärkendes Lernen** - Hierbei lernt ein Agent, durch Interaktionen mit der Umgebung und Maximierung einer Belohnungsfunktion die bestmöglichen Handlungen zu wählen. In Kernkraftwerken könnte diese Methode z.B. zum Einsatz kommen, um Roboter für Inspektionen und Wartungen zu trainieren, was die Sicherheit der menschlichen Arbeitskräfte erhöht, oder um die Leistungsoptimierung zu unterstützen, indem Kernparameter für eine optimale Leistung gesteuert werden.

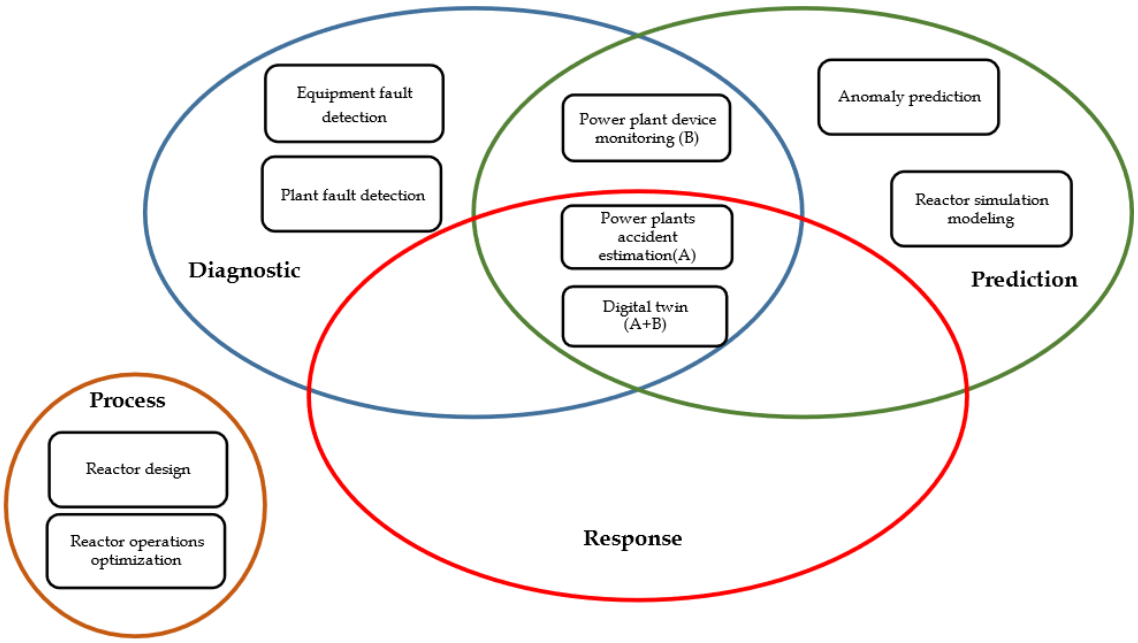

Abb. 2.3 Klassifikation nach Anwendungsfeldern des kerntechnischen Bereiches /HYU 23a/

2.1.4 KI-Begriff

Die im Rahmen des Vorhabens verwendete Definition des Begriffes „Künstliche Intelligenz“ bezieht sich auf Technologien, welche in der Natur vorkommenden Möglichkeiten (kognitive Fähigkeiten, natürliche Entwicklungsprozesse) zur Lösung von Problemen, Entscheidungsfindung und Lernprozesse nachzuahmen versuchen. Als Ansätze zur Umsetzung dienen KI-Methoden z.B. neuronale Netze und Entscheidungsbäume, wie sie in der hierarchischen Darstellung Abb. 2.1 zu sehen ist. Die praktische Umsetzung einer KI-Methode stellt ein KI-Modell dar, welches in Kombination mit anderen nicht KI-basierten Komponenten zu einem KI-basierten System zusammengefasst werden kann. Finalen, direkt nutzbare Produkte, welche KI-Modelle für bestimmte Funktionen einsetzen, stellen KI-basierte Anwendungen, wie beispielsweise Chatbots oder Bilderkennungs-Software, dar.

2.2 Definitionen einiger KI-Methoden

In diesem Abschnitt werden einige KI-Methoden im Detail erläutert, welche als Beispielen von KI-basierten Anwendungen im Kapitel 0 gezeigt werden. In den Erläuterungen wird kurz auf die Funktionsweise eingegangen. Zusätzlich werden allgemeine Beispiele

für mögliche Anwendungen dargestellt. Die Methoden werden nach der in der Abb. 2.1 dargestellten Klassifizierung eingeteilt.

2.2.1 Symbolische KI

2.2.1.1 Expertensysteme

Ein Expertensystem, eine symbolische KI mit klassischer Logik, stellt eine spezifische Art der Künstlichen Intelligenz dar, die darauf abzielt, menschliche Entscheidungsprozesse zu simulieren und spezialisiertes Fachwissen in einem bestimmten Bereich zu replizieren. Der Aufbau eines Expertensystems umfasst eine Wissensbasis, die Fakten und Regeln des Fachgebiets enthält, sowie eine Inferenzmaschine, die diese Informationen nutzt, um logische Schlussfolgerungen zu ziehen und Entscheidungen zu treffen. Ein Beispiel für ein Expertensystem ist ein medizinisches Diagnosesystem, das Ärzten bei der Diagnose von Krankheiten unterstützt, indem es Symptome analysiert und potenzielle Diagnosen vorschlägt.

2.2.1.2 Bayes’sche Netze

Bayes’sche Netze stellen Abhängigkeiten zwischen verschiedenen Ereignissen oder Zuständen in Form eines gerichteten, probabilistischen Graphen dar. Jeder Knoten repräsentiert eine Variable z. B. Kühlmittelverlust und die Kanten geben an, wie wahrscheinlich sich ein Ereignis auf ein anderes auswirkt. Sobald neue Informationen vorliegen, werden die Wahrscheinlichkeiten angepasst. So lassen sich sowohl Wahrscheinlichkeiten für mögliche Ursachen aus beobachteten Wirkungen ableiten als auch Vorhersagen über Konsequenzen bestimmter Ursachen machen. Bayes’sche Netze sind geeignet, Unsicherheiten in komplexen Systemen zu handhaben und liefern dabei erklärbare Ergebnisse, weil ihre Struktur und die bedingten Wahrscheinlichkeiten explizit hinterlegt sind. Ein Beispiel für die Nutzung ist das Rechenprogramm FaSTPro (Fast Source Term Prognosis) zur Quellterm-Prognose /GRS 25/. Damit lassen sich bei einem kerntechnischen Unfall Prognosen für Quellterme anhand aktueller Anlagendaten und Informationen aus einer Probabilistischen Sicherheitsanalyse (PSA), welche mit Hilfe eines Bayes‘schen Netzes zusammengeführt werden, bestimmen.

2.2.2 Subsymbolische KI

2.2.3 Support Vector Machine (SVM)

SVMs sind leistungsfähige Modelle für Klassifikations- und Regressionsaufgaben, die durch die Berechnung von optimalen Trennflächen (Hyperplane) in einem mehrdimensionalen Raum arbeiten. Die Hauptidee hinter SVMs ist, den Abstand (Margin) zwischen den Datenpunkten verschiedener Klassen zu maximieren. Dies wird durch sogenannte „Support-Vektoren“ erreicht, die die Punkte darstellen, die am nächsten an der Entscheidungsgrenze liegen und diese definieren. Bei komplexeren, nicht linear trennbaren Daten verwenden SVM sogenannte Kernel-Tricks, wie den Radial-Basis-Kernel, oder den polynomialen Kernel, um die Daten in einen höherdimensionalen Raum zu transformieren, in dem sie linear separabel werden. Dadurch eignen sich SVMs für vielseitige Anwendungen, von Gesichtserkennung und Krebsdiagnostik bis hin zur Textklassifikation und zur Bildverarbeitung.

2.2.3.1 Random Forest

Der Random Forest ist ein flexibler, robuster Ensemble-Lernalgorithmus, der aus einer großen Anzahl von Entscheidungsbäumen besteht und die Ergebnisse der Bäume aggregiert, um eine verbesserte Genauigkeit und Generalisierbarkeit zu erzielen. Jeder Baum wird auf einer zufälligen Teilmenge der Trainingsdaten und einer zufälligen Teilmenge der Merkmale trainiert, was die Variabilität zwischen den Bäumen erhöht und das Risiko von Overfitting reduziert. Der finale Vorhersagewert wird durch Mehrheitsentscheidungen bei Klassifikationsaufgaben oder durch den Durchschnitt bei Regressionsaufgaben der einzelnen Bäume gebildet. Der Random Forest wird oft zur Betrugserkennung, Diagnose in der Medizin und Feature-Extraktion verwendet.

2.2.3.2 Isolation Forest (iForest)

Der Isolation Forest, eine nach maschineller Lernmethode agierende subsymbolische KI, ist ein Algorithmus zur Anomalieerkennung (d. h. Ausreißer in einem Datensatz), der auf der Prämisse beruht, dass Anomalien in Daten durch Partitionierungen leichter isolierbar sind als normale Datenpunkte. Der iForest besteht aus mehreren Entscheidungsbäumen, die die Datenpunkte zufällig partitionieren.

Jeder Baum isoliert Datenpunkte durch rekursive Partitionierung, wobei die Tiefe der Isolationspfade als Maß für die Größe der Anomalie dient. Ein praktisches Beispiel ist die Erkennung von Kreditkartenbetrug, bei der ungewöhnliche Transaktionen, die leicht isolierbar sind, als potenziell betrügerisch identifiziert werden.

2.2.3.3 Künstliche Neuronale Netze (KNN)

Künstliche Neuronale Netze sind eine digitale Weiterentwicklung der früheren Perzeptronen. Die grundlegende Idee besteht darin, dass ein KNN auf bestimmte Aspekte biologischer Neuronen und ihrer Vernetzung zurückgreift. Elektrische Impulse (Signale) werden in einem biologischen Gehirn nur dann von Neuronen weitergeleitet, wenn ein bestimmter Membranpotential-Schwellenwert – ein bestimmtes Spannungsniveau - überschritten wird. In KNNs übernehmen sogenannte „Nodes“ diese Funktion, indem Eingangssignale mithilfe einer Aktivierungsfunktion und festgelegter Gewichtungen verarbeitet und bis zur Ausgabeschicht weitergeleitet werden. Die Weiterleitung erfolgt hier jedoch durch Berechnung des Eingangswertes mithilfe einer Aktivierungsfunktion, multipliziert mit einer Gewichtung. Dadurch entsteht ein lernfähiges System, welches von dem Prinzip neuronaler Vernetzung inspiriert ist.

Ein einfaches KNN besteht aus einer Eingabeschicht („Input Layer“) mit n Nodes, einer verborgenen Schicht („Hidden Layer“) mit m Nodes und einer Ausgabeschicht („Output Layer“) mit *k* Nodes. Wenn mehrere Hidden Layers verwendet werden, spricht man von einem Deep Neural Network (DNN). Diese Struktur ermöglicht es dem DNN, komplexe Muster und nichtlineare Beziehungen in den Trainingsdaten zu erlernen. Der Prozess des Lernens anhand solcher Daten wird als Maschinelles Lernen bezeichnet und im folgenden Abschnitt näher erläutert.

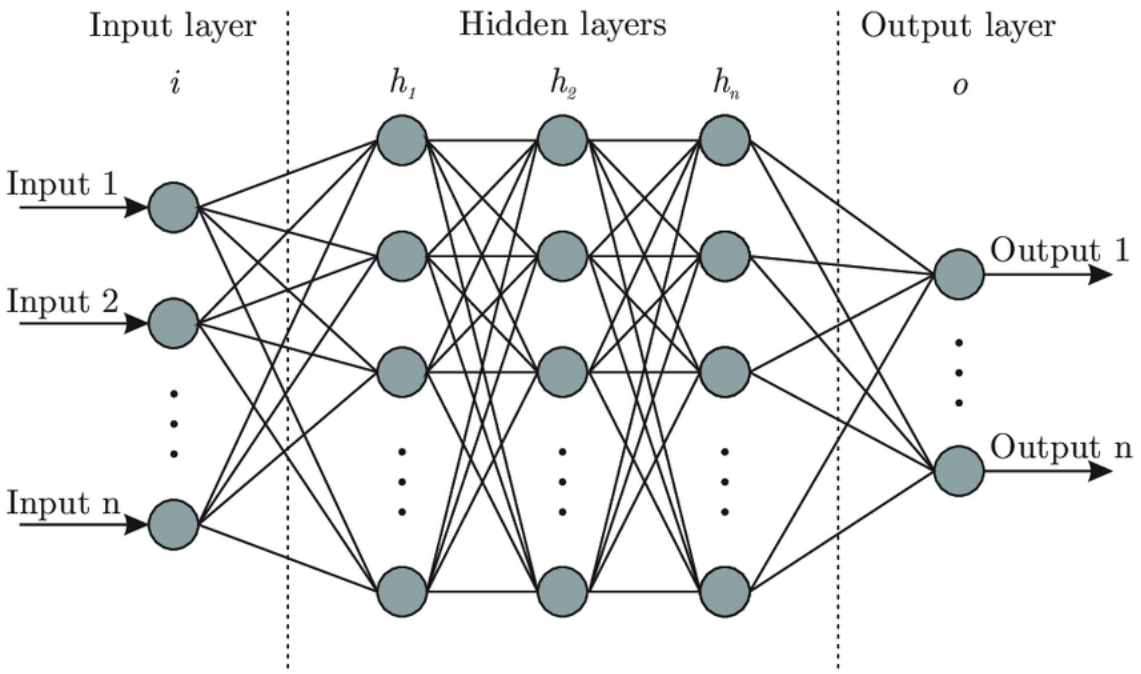

Abb. 2.4 Schematische Darstellung eines KNN /BRE 18/

2.2.3.4 Maschine Learning (ML)

Beim Maschinellen Lernen, oder englisch „Machine Learning“, werden Trainingsdaten genutzt, um die Gewichte eines KNNs anzupassen. Informationen propagieren durch das Netzwerk und werden nach dem Output Layer mit einem Erwartungswert verglichen. Hierbei kommt eine sogenannte Verlustfunktion (engl. *Loss Function*) zum Einsatz, die mithilfe einer Abstandsmetrik die Differenz zwischen erwartetem und tatsächlichem Output berechnet. Anschließend werden durch *Backpropagation* die Gewichte der Nodes so angepasst, dass die Verlustfunktion minimiert wird, wobei das Ziel die Auffindung eines Minimums ist. Beim Training eines KNNs unterscheidet man drei grundlegende Methoden: überwachtes Lernen (*supervised learning)*, unüberwachtes Lernen (*unsupervised learning)* und bestärkendes Lernen (*semi-supervised learning)*.

- **Überwachtes Lernen**: Die Trainingsdaten sind vor dem Training mit Markierungen (Label) versehen. Dadurch erhält das Netz direktes Feedback über die Verlustfunktion. Diese Methode eignet sich besonders, wenn nur wenige Trainingsdaten verfügbar sind oder bereits bekannte Funktionen gelernt werden sollen. Ein klassisches Beispiel ist der binäre Klassifikator in einem Convolutional Neural Network (CNN), das Bilder von Hunden und Katzen unterscheiden soll. Die Bilder werden hierbei im Vorfeld markiert, um das gewünschte Lernziel zu erreichen.

- Unüberwachtes Lernen: Hier sind die Trainingsdaten unmarkiert. Diese Art des Lernens ist sinnvoll, wenn das Netzwerk versteckte oder unbekannte Eigenschaften in den Daten entdecken soll. Unüberwachtes Lernen erfordert in der Regel eine größere Menge an Trainingsdaten. Ein Beispiel wäre ein CNN, das Hundebilder analysiert und in verschiedene Gruppen unterteilt, etwa nach Hunderassen, ohne dass diese Kategorien zuvor bekannt oder vorgegeben sind.
- **Beschränktes Lernen**: Diese Methode kombiniert die beiden zuvor beschriebenen Ansätze, indem sowohl beschriftete als auch unbeschriftete Daten verwendet werden.

2.2.3.5 Autoencoder

Ein Autoencoder, eine nach maschineller Lernmethode agierende subsymbolische KI, ist ein neuronales Netzwerk, das darauf trainiert ist, Eingabedaten in codierter Form darzustellen und anschließend anhand dessen wieder zu rekonstruieren. Der Autoencoder besteht aus zwei Hauptkomponenten: dem Encoder, der die Eingabedaten auf die codierte Darstellung abbildet, und dem Decoder, der diese Darstellung wieder auf die ursprüngliche Darstellung abbildet. Ein typisches Beispiel für die Anwendung von Autoencodern ist die Rauschunterdrückung bei Bildern. Hierbei lernt das Netzwerk, die wesentlichen Merkmale eines Bildes zu extrahieren und das Rauschen zu entfernen, um eine sauberere Version des Bildes zu rekonstruieren. Ein Auto-Associative Neural Network (AANN) hat eine ähnliche Struktur wie ein Autoencoder. Der Unterschied liegt in der spezifischen Zielsetzung, Assoziationen innerhalb der Daten zu erkennen und zu verstärken, anstatt nur eine einfache Komprimierung und Rekonstruktion durchzuführen. Ein Beispiel für die Anwendung von AANNs ist die Fehlererkennung in Maschinen, bei der das Netzwerk ungewöhnliche Muster erkennt, die auf mögliche Fehlfunktionen hinweisen.

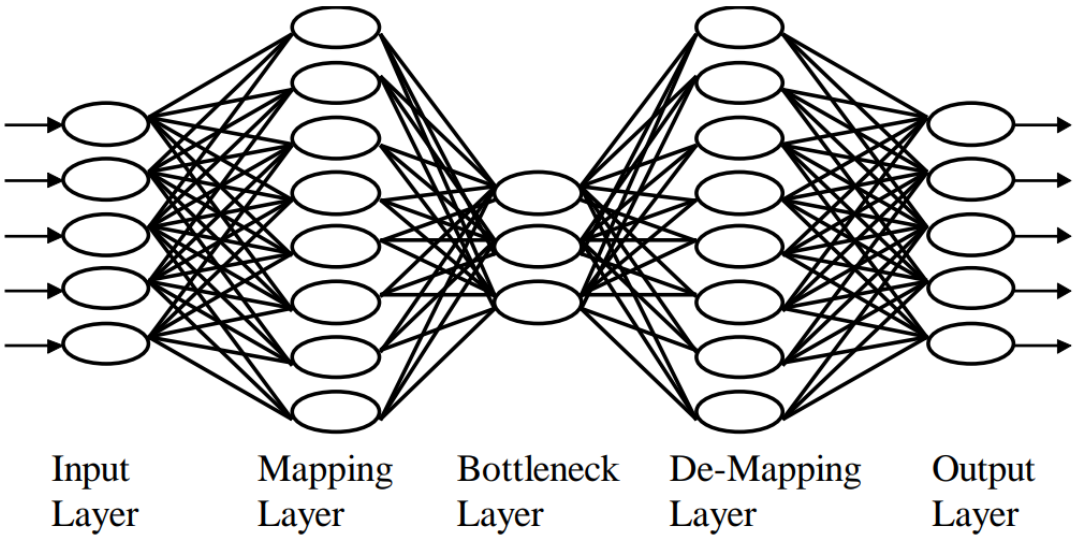

Abb. 2.5 Beispiel eines AANN. /HIN 98/

2.2.3.6 Variational Autoencoder (VAE)

Ein Variational Autoencoder ist ein generatives Modell, welches in der Lage ist, neue Datenpunkte aus einem gegebenen Datensatz zu erzeugen. Das Modell ist eine nach maschineller Lernmethode agierende subsymbolisches KI. Dieses Besteht aus zwei Hauptkomponenten – Encoder und Decoder. Der Encoder nimmt die Eingabedaten und bildet diese auf ein Muster im latenten Raum ab. Der Decoder rekonstruiert die Eingabedaten aus einer Stichprobe des latenten Musters.

Die VAE erlauben es zum ursprünglichen Datensatz ähnliche, neue Datenpunkte zu erzeugen. Zudem besteht die Möglichkeit der Anomalieerkennung, indem der VAE auf einen Datensatz trainiert wird. Im Betrieb wird versucht, Eingabedaten, basierend auf den Trainingsdaten, zu rekonstruieren. Die Erkennung der Anomalien erfolgt durch die Berechnung des Rekonstruktionsfehlers. Dadurch lassen sich Eingabedaten mit einem Rekonstruktionsfehler, welcher über einem festgelegten Schwellenwert liegt, als Anomalien identifizieren.

2.2.3.7 Sparse Autoencoder (SAE)

Ein Sparse Autoencoder (SAE), ebenfalls eine nach maschineller Lernmethode agierende subsymbolische KI, ist eine spezielle Form des Autoencoders, die darauf abzielt, dimensionsreduzierte Repräsentationen der Eingabedaten zu lernen. Auch hier bestehen die Hauptkomponenten aus einem Encoder und einem Decoder. Der entscheidende Unterschied liegt jedoch in der speziellen Regularisierung der versteckten Schicht, um die Aktivierung der Neuronen zu minimieren. Dies führt zu einer spärlichen Darstellung, bei der nur wenige Neuronen aktiv sind, um die wichtigsten Merkmale der Daten zu repräsentieren. Ein Beispiel für die Anwendung eines SAE ist die Bildkomprimierung, bei der die wichtigsten Merkmale eines Bildes extrahiert und gespeichert werden, wodurch Speicherplatz effizient genutzt wird.

2.2.3.8 Convolutional Neural Network (CNN)

CNNs sind neuronale Netzwerke, die speziell für die Verarbeitung von Bilddaten entwickelt wurden, indem sie Merkmale auf verschiedenen Ebenen abstrahieren. CNNs nutzen Faltungsschichten, in denen Filter oder „Kerne“ angewendet werden, um niedrig-, mittel- und hochrangige Merkmale wie Kanten, Formen und komplexe Strukturen zu erkennen. Nach jeder Faltung folgt eine Pooling-Schicht, die das Bild durch Extraktion von Schlüsselmerkmalen verdichtet und so die Anzahl der Parameter und die Rechenkosten reduziert. Diese Struktur macht CNNs besonders effektiv in der Bild- und Videobearbeitung, z. B. für die Gesichtserkennung, Objekterkennung und im autonomen Fahren.

2.2.3.9 Recurrent Neural Network (RNN)

RNNs sind neuronale Netzwerke, die speziell für die Verarbeitung sequentieller Daten, wie Text oder Zeitreihen, entwickelt wurden. Ein entscheidender Aspekt von RNNs ist, dass sie „Gedächtniszellen“ haben, die Informationen von früheren Zeitschritten speichern und für spätere Berechnungen nutzen. Dies ermöglicht es ihnen, zeitliche Abhängigkeiten zu modellieren und kontextuelle Informationen über Sequenzen hinweg zu behalten. Allerdings haben Standard-RNNs Probleme mit dem „Vanishing Gradient“, was das Lernen bei längeren Sequenzen erschwert. RNNs werden häufig für automatische Texterstellung, Sprachsynthese und Finanzmarktprognosen verwendet.

2.2.3.10 Long Short-Term Memory (LSTM)

LSTM ist eine spezielle Art von RNN, die für die Verarbeitung von Sequenzdaten und die Modellierung langfristiger Abhängigkeiten entwickelt wurde. LSTMs bewältigen das Problem des „Vanishing Gradient“, welches in herkömmlichen RNNs auftritt, indem sie interne Zellstrukturen (Speicher-Zelle) nutzen, welche die Aktivierungsfunktion beeinflussen und Informationen über längere Sequenzen hinweg speichern können. Diese Zellstrukturen bestehen aus verschiedenem Gatter: einem Eingangs-, einem Vergessens- und einem Ausgangsgatter. Zudem hat die Zelle multiple Ein- und Ausgänge. Über Eingänge gelangen langfristig gespeicherte Informationen aus dem vorherigen Zeitschritt ($\mathrm{c_{t-1}}$), kurzfristig gespeicherte Informationen aus dem vorherigen Zeitschritt ($\mathrm{h_{t-1}}$) und die Eingabe des aktuellen Zeitschrittes ($\mathrm{x_t}$) in die Zelle. Diese Informationen gelangen zunächst in das Vergessens-Gatter ($\mathrm{f_t}$). Das Vergessens-Gatter entscheidet, welcher Anteil der langfristig gespeicherten Information verworfen werden. Hierfür wird eine Sigmoid-Funktion (σ) verwendet, welche eine Ausgabe in einem Wertebereich von 0-1 liefert. Hierdurch wird entschieden, ob ein großer Anteil (Sigmoid-Ausgabewert nahe 1) des vorherigen Zustandes, oder ein kleiner Teil des vorherigen Zustandes (Sigmoid-Ausgabewert nahe 0), übernommen werde. Als nächsten gelangen die Informationen in das Eingangs-gatter (i_t). Hier wird entschieden inwiefern die neuen Informationen in die langfristig gespeicherten Informationen des derzeitigen Zeitschrittes (c_t), einen Ausgabewert, einfließen. Dies geschieht indem zunächst ein „Kandidat“ ($\tilde{c_t}$) über eine tanh-Funktion gebildet wird. Danach wird basierend auf einer Sigmoid-Funktion bestimmt, zu welchem Anteil der Kandidat in c_t einfließt. Danach wird c_t, dem entsprechenden Anteil nach, mit neuen Informationen aktualisiert. Das Ausgangs-Gatter ($\mathrm{o_t}$) kontrolliert schließlich, welche Informationen an die nächste Zeiteinheit weitergegeben werden. Dies erfolgt, indem c_t, über eine tanh-Funktion begrenzt und mit einer Sigmoid-Funktion gewichtet als Ausgabe ($\mathrm{h_t}$) weitergereicht wird. Durch diesen Aufbau sind LSTMs in der Lage, Muster und Zusammenhänge in Daten wie Sprache, Zeitreihen oder Videosequenzen zu erkennen. Typische Anwendungen umfassen maschinelle Übersetzung, Sprachsynthese und Finanzmarktanalysen.

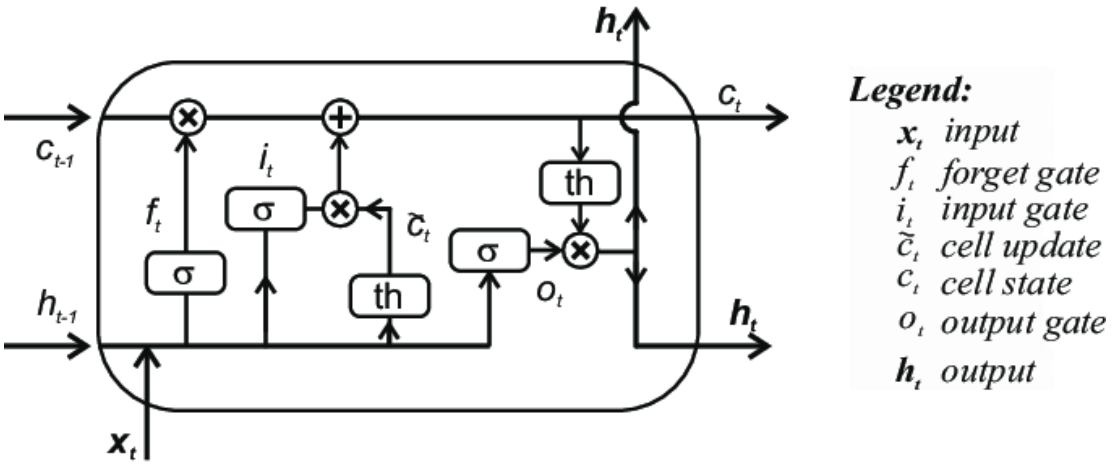

Abb. 2.6 Beispielhafte Darstellung einer LSTM-Zelle. /HRN 19/

2.2.3.11 Genetischer Algorithmus (GA)

Der genetische Algorithmus ist ein evolutionärer Optimierungsalgorithmus, der durch Nachahmung der natürlichen Selektion optimale Lösungen findet. Der Prozess beginnt mit einer Population von zufälligen Lösungen, die über mehrere Generationen hinweg durch Kreuzung und Mutation verändert werden. Bei jeder Generation werden Lösungen nach Fitness-Kriterien bewertet, und die besten Lösungen werden für die nächste Generation ausgewählt. Dieser iterative Prozess ermöglicht es dem GA, optimale Lösungen für komplexe Probleme wie die Layout-Optimierung in Produktionsanlagen oder die Schaltungsoptimierung in der Elektronik zu finden.

2.2.3.12 Asynchronous Advantage Actor-Critic (A3C)

A3C ist ein moderner Reinforcement-Learning-Algorithmus (RL-Algorithmus), der darauf abzielt, die Effizienz und Stabilität des Lernprozesses durch parallele Agenten zu verbessern. Mehrere Agenten interagieren parallel mit unterschiedlichen Kopien der Umwelt, wobei jeder Agent seine eigene Lernstrategie verfolgt und so von individuellen Rückmeldungen profitiert. Dies fördern die Exploration und die Anpassung an unterschiedliche Situationen. Die Agenten teilen ihre Erfahrungen regelmäßig, um die Wissensbasis des zentralen Modells zu aktualisieren. A3C wird in der Robotersteuerung und in Videospielen eingesetzt, um Agenten mit hoher Lernrate und stabilen Entscheidungen zu entwickeln.

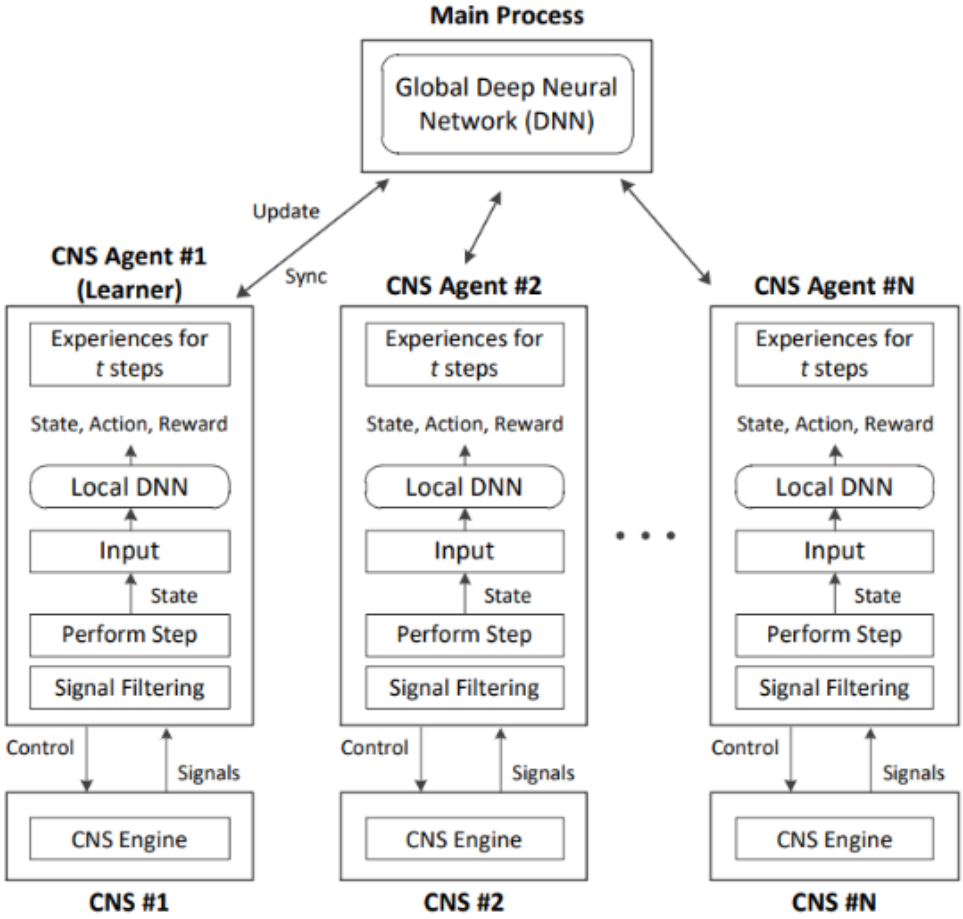

Abb. 2.7 Beispielhafte Lernstruktur multipler Agenten. /PAR 22/

2.2.4 Hybride KI

2.2.4.1 Bidirectional Encoder Representations from Transformers (BERT)

BERT ist ein fortschrittliches Sprachmodell, das auf der Transformer-Architektur basiert. Es verwendet ein bidirektionales Training, um den Kontext von Wörtern in einem Text zu erfassen, indem sowohl die vorhergehenden als auch die nachfolgenden Wörter berücksichtigt werden. Das Modell ist eine nach maschineller Lernmethode agierende subsymbolisches KI. Der Aufbau von BERT umfasst den Encoder-Teil des Transformer-Modells und die Anwendung von Selbstaufmerksamkeit, um Beziehungen zwischen den Wörtern zu erkennen. Ein Beispiel für die Anwendung von BERT ist die Verbesserung der Genauigkeit von Suchmaschinen durch ein besseres Verständnis der Bedeutung von Suchanfragen.

2.3 Einordnung der KI-Methoden nach der Klassifikation nach ISO TR 63468

Die unter die aufgeführte Abb. 2.8 soll eine vereinfachte Darstellung aller im Kapitel 2.2 vorgestellten KI-Methoden bieten. Die Darstellung verwendet die Klassifikation nach /IEC 23/, wie sie in Abb. 2.1 dargestellt ist. Hierbei wird die Klassifikation aus Abb. 2.1 in einer komprimierten Art und Weise dargestellt, um eine grobe und dennoch übersichtliche Darstellung der untersuchten KI-Methoden zu liefern. Die Abbildung ist von innen nach außen hin zu betrachten. Das innere Dreieck stellt die drei grundlegenden Kategorien von KI-Methoden dar. In dem darauffolgenden Bereich (näher zum Rand) werden durch die gestrichelten Linien die Unterkategorien abgebildet. Noch weiter außen befinden sich die KI-Methoden, welche sich der jeweiligen Unterkategorie einordnen lassen.

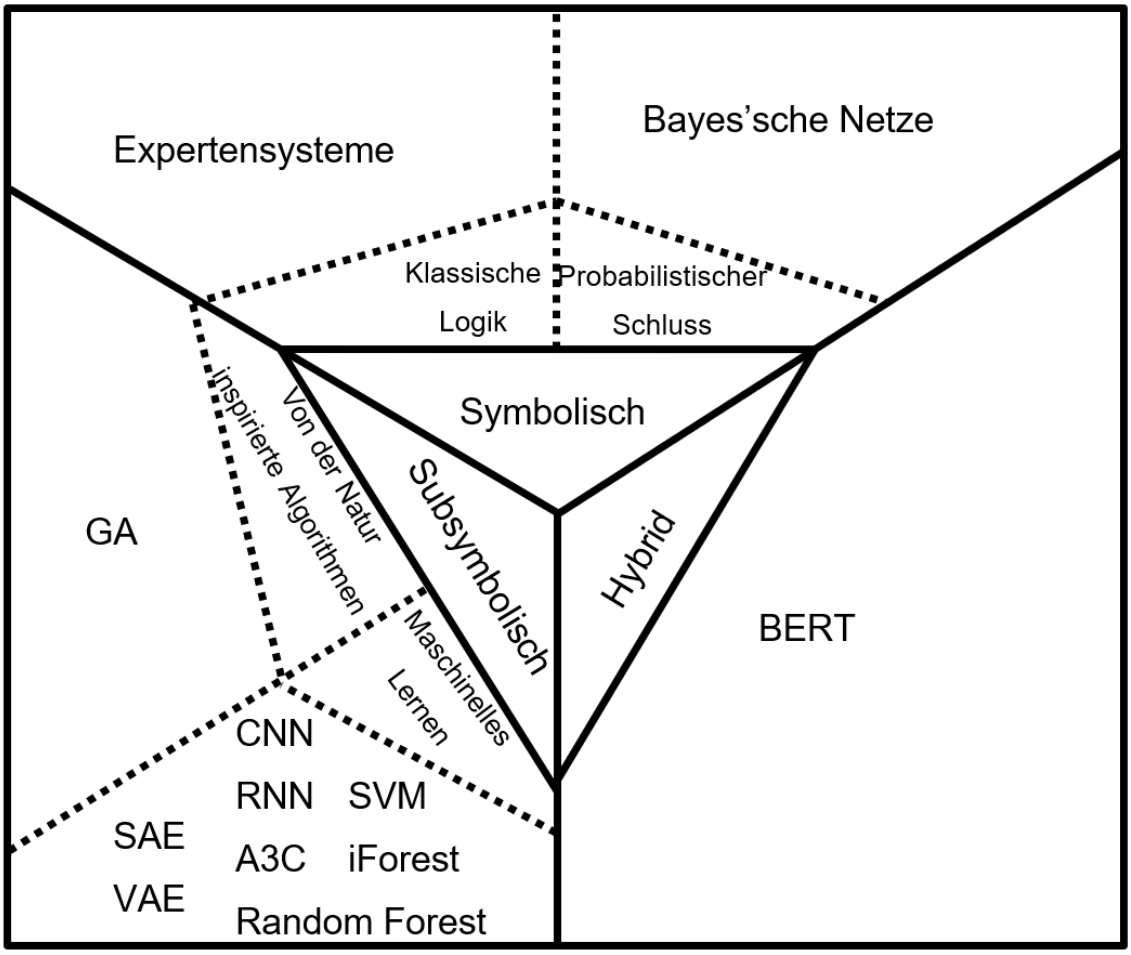

Abb. 2.8 Einordnung der KI-Methoden nach der Klassifizierung nach /IEC 23/.

3 KI-basierte Anwendungen in der Kerntechnik und in Bereichen mit sicherheitstechnischer Bedeutung

3.1 Zusammenfassung der recherchierten KI-basierten Anwendungen

In diesem Abschnitt werden basierend auf die im Rahmen des Vorhabens ausgewerteten Unterlagen (Publikationen, Normen, Tagungen) einige beispielhafte ermittelte KI-basierten Anwendungen erläutert. Dabei beschränken sich die Rechercheergebnisse einerseits auf Anwendungsfälle von KI in Bereichen mit sicherheitstechnischer Bedeutung und andererseits auf Anwendungsfälle in der Kerntechnik. Die in den Anwendungsfällen zum Einsatz kommenden Methoden werden nach der im Kapitel 2.1 dargestellten Klassifikation eingeordnet. Es wird außerdem auf das Einsatzgebiet, die verwendeten Daten und Lernmethoden sowie die Fähigkeiten der jeweiligen Anwendungsfälle eingegangen. Zudem wird die KI-basierte Anwendung im Hinblick auf einen Sicherheitsgewinn und -risiko im Vergleich zu konventionellen Methoden bewertet.

Die darauffolgenden Abschnitte behandeln einige Beispiele für KI-basierte Anwendungen. Weitere und detailliertere Ergebnisse der Recherche können dem Anhang A entnommen werden.

3.1.1 Zusammenfassung der ermittelten KI-basierten Anwendungen in Bereichen mit sicherheitstechnischer Bedeutung

In diesem Abschnitt werden einige Zusammenfassungen der ermittelten KI-basierten Anwendungen in Bereichen mit sicherheitstechnischer Bedeutung, die die funktionale Sicherheit unmittelbar betreffen, erläutert. Anwendungen, welche in Bereichen mit sicherheitstechnischer Bedeutung zum Einsatz kommen, jedoch nicht unmittelbar im Bereich der funktionalen Sicherheit, liegen, beispielsweise Chatbots, Spracherkennungssysteme für Dokumentation oder automatisierte Terminplanung, werden nicht behandelt.

In sicherheitskritischen Bereichen ist die Zuverlässigkeit von Systemen entscheidend, da Fehlfunktionen schwerwiegende Auswirkungen auf Menschen, Infrastruktur oder die Umwelt haben können. Die vorgestellten Anwendungen setzen auf künstliche Intelligenz, um automatisierte Überwachung und Problemerkennung in sicherheitsrelevanten Szenarien zu ermöglichen.

Zum Beispiel trägt die Überwachung von Windturbinenblättern zur Früherkennung von Oberflächenfehlern bei, wodurch Wartungsmaßnahmen zeitnah geplant und potenzielle Ausfälle vermieden werden können. In der Cybersicherheit, etwa in fahrzeuginternen Netzwerken, ermöglichen KI-gestützte Angriffserkennungssysteme (Intrusion Detection System) wie CANBERT die Identifizierung von Netzwerkangriffen, die die Stabilität und Sicherheit vernetzter Fahrzeuge gefährden könnten. Die Einsatzgebiete der analysierten Arbeiten umfassen eine breite Palette sicherheitskritischer Anwendungen /IEE 22/. Dazu gehören unter anderem auch die Vorhersage des Stromverbrauchs im Niederspannungsnetz, datenschutzfreundliche Lösungen für Fahrerlose Transportsysteme sowie die Erkennung von Bedrohungen in der Cybersicherheit /IEE 22/. In der Energieversorgung unterstützt die KI eine präzisere Planung und Stabilität des Netzes, während in der Cybersicherheit die Erkennung unbekannter Bedrohungen und die Optimierung der Überwachungstechniken vorangetrieben werden. Diese vielfältigen Einsatzgebiete verdeutlichen das Potenzial der KI zur Lösung komplexer, sicherheitskritischer Probleme in unterschiedlichsten Branchen. Ein großer Teil der während der Vorhabenslaufzeit recherchierten KI-basierten Anwendungen sind darauf ausgelegt, die Sicherheit in sensiblen Umgebungen zu verbessern, indem sie Risiken frühzeitig identifizieren und Maßnahmen zur Risikominderung ermöglichen.

Die in den Anwendungen genutzten KI-Modelle und -Methoden umfassen ein breites Spektrum an Fähigkeiten, um die jeweiligen Anforderungen an Sicherheit und Genauigkeit zu erfüllen. Convolutional Neural Networks werden beispielsweise für die Bildverarbeitung und Oberflächeninspektion eingesetzt, um Oberflächenfehler präzise zu identifizieren und zu klassifizieren. In der Netzwerksicherheit ermöglicht ein Transformer-basiertes Modell, wie CANBERT, die Analyse von Kommunikationsmustern in Fahrzeugnetzwerken /IEE 22/. Für die Lastprognose im Stromnetz werden Bernstein-Polynom-Normalizing-Flows verwendet, unterstützt durch neuronale Netzwerke wie vollständig verbundene Netze und 1D-CNNs, die die Parametersteuerung übernehmen /ARP 21/. Weitere Methoden umfassen Autoencoder und Reinforcement-Learning-Ansätze für Anomalie- und Angriffserkennung. Viele der beschriebenen Anwendungen nutzen Sensor- und Bilddaten als Grundlage für die Entscheidungsfindung. Die Modelle sind in der Lage, aus Rohdaten wie Bildern, Texten und Netzwerkverkehr Bedeutungen abzuleiten und relevante Muster zu identifizieren. Im Bereich der Infrastrukturüberwachung werden beispielsweise Ultraschall- oder Bilddaten eingesetzt, um den Zustand von Objekten wie Dämmen oder Windturbinenblättern zu überwachen /IEE 22/.

In anderen Anwendungen wie der Cybersicherheit oder der medizinischen Bildverarbeitung, bilden große Datenmengen aus Netzwerken oder bildgebenden Verfahren die Grundlage für die Entwicklung von Modellen. Im Bereich der Lastprognose im Niederspannungsnetz kommen Smart-Meter-Daten zum Einsatz, die den 24-Stunden-Verbrauch von Stromkunden abbilden /ARP 21/. Diese Datenarten ermöglichen die Erkennung potenziell gefährlicher Anomalien und die Optimierung sicherheitskritischer Systeme, indem sie Echtzeitdaten zur Verfügung stellen und Vorhersagen für risikobehaftete Szenarien liefern. Zudem werden in der Cybersicherheit Netzwerk-Telemetriedaten und Textdaten zur Angriffserkennung und Überwachung genutzt.

Ein zentrales Ziel der analysierten KI-basierten Anwendungen ist die Anomalie- und Fehlererkennung. In den vorgestellten Arbeiten werden Anomalien in verschiedenen Kontexten und Datentypen detektiert. Die Fähigkeit, Anomalien frühzeitig zu erkennen, ist in sicherheitskritischen Szenarien unerlässlich, da dadurch vorbeugende Maßnahmen ergriffen werden können, bevor es zu größeren Problemen oder Ausfällen kommt. Anomalieerkennungssysteme, die auf maschinellem Lernen basieren, lernen dabei normale Betriebsbedingungen und identifizieren Abweichungen, die potenziell auf Sicherheitsrisiken hinweisen. Durch die Nutzung von Autoencodern und anderen KI-Modellen lassen sich spezifische Verhaltensweisen oder Merkmale identifizieren, die ein hohes Risiko für das Gesamtsystem darstellen. So ermöglicht die Zero-Day-Bedrohungserkennung mithilfe von Autoencodern die Identifizierung unbekannter Bedrohungen in Echtzeit, was die Cybersicherheit erheblich verbessert und Systeme proaktiv gegen potenzielle Angriffe schützt. Anomalieerkennungssysteme funktionieren oft nur so gut wie die Daten, auf denen sie trainiert wurden. In sicherheitskritischen Anwendungen kann es jedoch vorkommen, dass viele potenziell gefährliche Anomalien selten oder noch nie aufgetreten sind, sodass keine ausreichenden Daten zur Verfügung stehen. Dies erschwert das Training und könnte zu einer erhöhten Rate von Fehlalarmen oder übersehenen Bedrohungen führen. Außerdem besteht das Risiko, dass das System bei neuen, unbekannten Angriffen oder Fehlern keine adäquaten Antworten liefern kann. Solche Anomalieerkennungssysteme müssen daher regelmäßig aktualisiert und neu trainiert werden, was zusätzlichen Aufwand und kontinuierliche Datenquellen erfordert.

In sicherheitskritischen Anwendungen die Nachvollziehbarkeit der Ergebnisse vielfach essenziell, da KI-Modelle Entscheidungen treffen können, welche schwer zu interpretieren sind. Daher spielt die Erklärbarkeit in vielen dieser Arbeiten eine wichtige Rolle. Beispielsweise ist es im medizinischen Bereich entscheidend, dass Ärzte verstehen, warum ein KI-Modell eine bestimmte Läsion als verdächtig klassifiziert hat /ISO 20/.

Post-hoc-Erklärungsmethoden helfen dabei, die Ergebnisse von Deep-Learning-Algorithmen für Radiologen verständlich zu machen, um diese in ihrer Diagnose zu unterstützen. Die Erklärbarkeit fördert das Vertrauen in KI-basierte Systeme, da Benutzer besser verstehen, wie und warum bestimmte Entscheidungen getroffen werden. Dies ist besonders wichtig, wenn Ergebnisse KI-basierter Systeme als Grundlage für sicherheitskritische Entscheidungen dienen, wie beispielsweise die Planung von Wartungsmaßnahmen oder die Erkennung von Cyberbedrohungen. Die Erklärungsmethoden helfen zudem, Schwächen in den Modellen aufzudecken und zu verbessern, was letztendlich die Sicherheit und Zuverlässigkeit der Anwendungen erhöht.

Der Einsatz von KI-basierten Systemen bietet viele potenzielle Vorteile, bringt aber auch Herausforderungen und Risiken mit sich, die bedacht werden müssen. Ein großer Vorteil liegt in der Effizienzsteigerung: KI kann große Datenmengen in kürzester Zeit analysieren und Muster erkennen, die für Menschen nur schwer zu erkennen wären. So können Entscheidungsprozesse beschleunigt und Kosten gesenkt werden. Doch diese Geschwindigkeit und Skalierbarkeit KI-basierter Systeme birgt auch das Risiko, dass Entscheidungen unreflektiert und ohne ausreichende menschliche Kontrolle getroffen werden, was in sicherheitskritischen Bereichen fatale Folgen haben könnte. Ein weiterer Vorteil KI-basierter Systeme liegt in ihrer Fähigkeit zur Automatisierung. Routineaufgaben können durch KI-basierte Systeme übernommen werden, sodass Menschen mehr Zeit für komplexe, kreative Tätigkeiten haben. Gleichzeitig könnte ein weitläufiger Einsatz von KI-basierten Systemen längerfristig zu einem verminderten kritischen Hinterfragen der Ergebnisse führen. Zudem fehlt KI-basierten Systemen oft die Fähigkeit zur Intuition und Kontextverarbeitung, die für viele menschliche Entscheidungen essenziell ist. Ein weiterer Vorteil des Einsatzes von KI-basierten Systemen ist ihre Lernfähigkeit – Machine-Learning-Modelle können kontinuierlich optimiert werden und sich an neue Umstände anpassen, was insbesondere in der Cybersicherheit oder der medizinischen Diagnostik wertvoll ist. Aber auch hier gibt es potentielle Schwierigkeiten. Beispielsweise wenn KI-Modelle auf fehlerhaften oder verzerrten Daten basieren, kann dies zu Fehlentscheidungen und sogar zur Verstärkung von Vorurteilen führen. Dies stellt insbesondere in ethisch sensiblen Bereichen eine große Herausforderung dar, da der „Bias“ in KI-basierten Systemen schwer kontrollierbar ist und langfristige Folgen haben kann.

Die ermittelten KI-basierten Anwendungen werden in Abb. 3.1 nach der Klassifizierung nach /IEC 23/ eingeordnet und visualisiert. Einige der recherchierten Anwendungen wurden im Rahmen der „*IEEE - International Conference on Machine Learning and Applications“* /IEE 22/ identifiziert.

Ähnlich wie in Abb. 2.7 zu sehen ist, befindet sich in der unten aufgeführten Abbildung eine vereinfachte Darstellung der in diesem Unterkapitel vorgestellten KI-basierten Anwendungen im Bereich mit sicherheitstechnischer Bedeutung. Die Darstellung verwendet die Klassifikation nach /IEC 23/, wie sie in der Abb. 2.1 dargestellt ist. Die Abbildung ist von innen nach außen hin zu betrachten. Das innere Dreieck stellt die drei grundlegenden Kategorien von KI-Methoden dar. In dem darauffolgenden Bereich (näher zum Rand) werden durch die gestrichelten Linien die Unterkategorien abgebildet. Noch weiter außen befinden sich die KI-basierten Anwendungen, welche sich der jeweiligen Unterkategorie einordnen lassen.

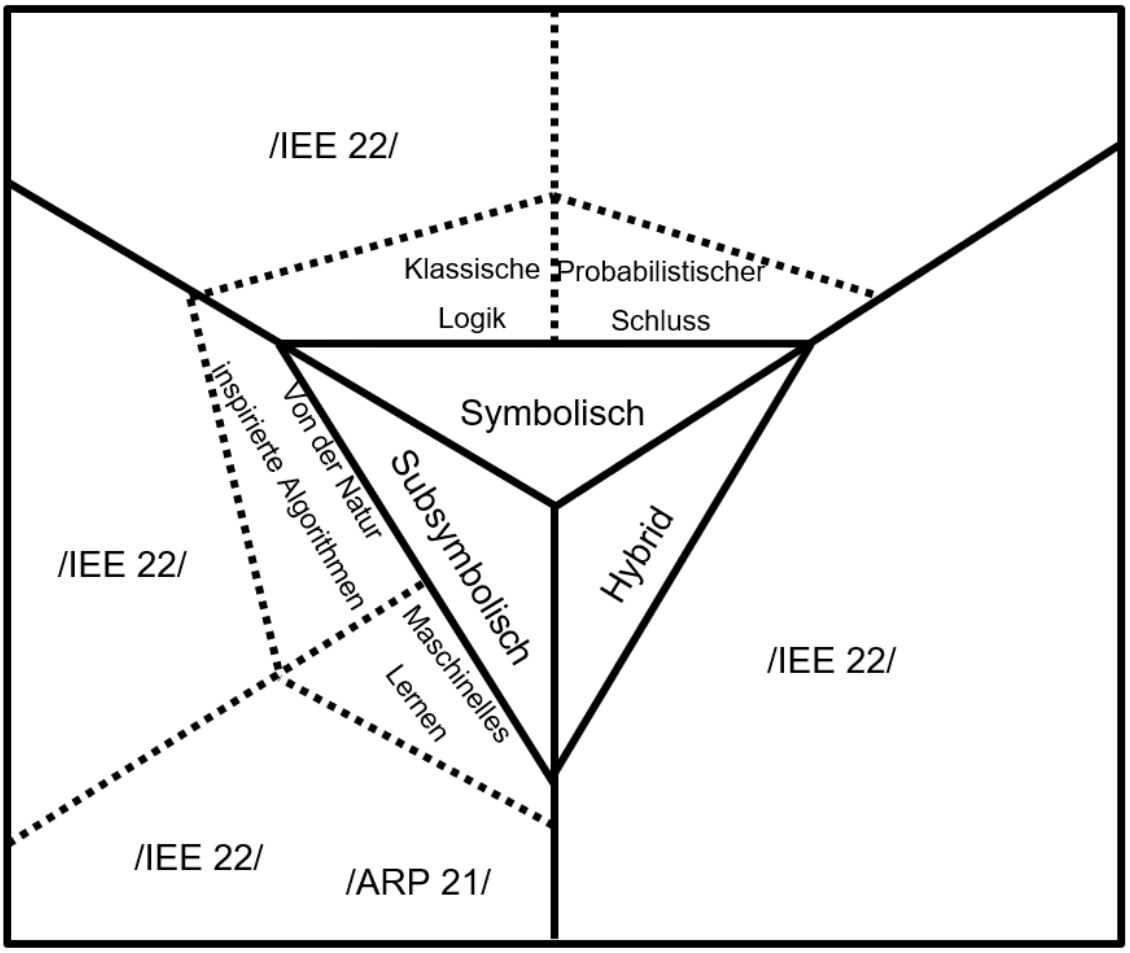

Abb. 3.1 Einordnung von KI-basierten Anwendungen nach /IEC 23/.

3.1.2 Zusammenfassung der ermittelten KI-basierten Anwendungen im kerntechnischen Bereich

In diesem Abschnitt werden einige Zusammenfassungen der ermittelten KI-basierten Anwendungen im kerntechnischen Bereich erläutert.

Der Einsatz von KI-basierten Anwendungen in der Kerntechnik wurde bereits in den 1990er Jahren untersucht und in verschiedenen Bereichen erprobt. In /UHR 93/ wird deutlich, dass damals schon 287 KI-basierte Systeme weltweit in der kommerziellen Energieerzeugung identifiziert wurden, darunter 145 in den USA, 71 in Japan, 29 in Frankreich und 42 in anderen Ländern. Diese frühen Anwendungen konzentrierten sich vor allem auf Expertensysteme, die Entscheidungsunterstützung, Diagnose und Überwachung ermöglichten. Beispiele umfassen Systeme wie das Reactor Emergency Alarm Level Monitor (REALM) zur Klassifikation von Notfällen oder CLEO und CRAW zur Optimierung der Kernbeladung und zur Diagnose von Brennstabfehlern.

Die Ergebnisse dieser frühen Arbeiten bewegen sich hauptsächlich in den Bereichen Diagnose und Entscheidungsunterstützung. Diese Systeme wurden entwickelt, um spezifische, gut definierte Aufgaben zu bewältigen, wie etwa das Erkennen von Fehlern in Sensoren oder das Bereitstellen von Notfallmaßnahmen. Die Anwendungen beschränkten sich jedoch oft auf Tests und Trial-Implementierungen, da die damalige Technologie in puncto Datenverarbeitung und Modellkomplexität begrenzt war. In anderen Feldern wie der Prozessoptimierung oder der automatisierten Steuerung war die Umsetzung damals weniger verbreitet. Diese Bereiche erfordern eine hohe Flexibilität und Echtzeitfähigkeit, was die KI-basieret Systeme der 1990er Jahre aufgrund begrenzter Rechenleistung und algorithmischer Reife oft nicht leisten konnten. Die meisten Anwendungen blieben daher auf Forschung und Testumgebungen beschränkt.

Die Verwendung von KI in der Kerntechnik hat sich in der heutigen Zeit erheblich weiterentwickelt, insbesondere durch Fortschritte in der Rechenleistung, Datenverarbeitung und Algorithmenentwicklung. Während viele frühe Anwendungen auf spezifische Aufgaben beschränkt waren, zeigen die aktuellen Publikationen, dass KI-basierte Systeme mittlerweile in einer Vielzahl von Anwendungsfeldern erprobt und teilweise produktiv eingesetzt werden. Aktuelle Forschungsarbeiten konzentrieren sich auf Diagnose, Anomalieerkennung, Prozessautomatisierung, Materialprognose und Unfallanalyse. Beispiele dafür sind:

Anomalieerkennung: Modelle wie Variational Autoencoder (VAE) kombiniert mit Isolation Forest (iForest) oder Deep Support Vector Data Description (SVDD) können Anomalien in Echtzeit erkennen und ermöglichen so eine Überwachung des Anlagenbetriebs /LI 22a/, /CHO 22/.

Prozessautomatisierung: Deep Reinforcement Learning (DRL) wird für die automatisierte Steuerung komplexer Betriebsphasen wie der Aufheizphase in Kernkraftwerken, eingesetzt und hat bewiesen, dass es mit der Leistung menschlicher Bediener vergleichbar ist /PAR 22/.

Materialprognose: Anwendungen wie PROMAP kombinieren KI mit probabilistischen Ansätzen, um Materialeigenschaften vorherzusagen und Unsicherheiten in den Daten zu berücksichtigen. Diese Methoden reduzieren den Bedarf an teuren Experimenten und verbessern die Vorhersagegenauigkeit für neue Materialien /LYE 22/.

Digital Twins: Hybride Modelle wie Grey-Box-Digital-Twins, die physikalische und datengetriebene Ansätze kombinieren, werden zunehmend eingesetzt, um Risiken zu überwachen und präzisere Echtzeitprognosen zu ermöglichen /MIQ 22/.

Besonders fortschrittliche Technologien wie hybride Digital-Twin-Modelle oder DRL-basierte Steuerungen erfordern umfangreiche Tests und Validierungsprozesse, bevor sie in sicherheitskritischen Umgebungen implementiert werden können. Ein Beispiel ist die Anomalieerkennung durch Deep SVDD, die vielversprechende Ergebnisse in Tests zeigte, jedoch bei Szenarien mit geringem Schweregrad der Anomalie – einer geringen Abweichung vom Normalzustand – noch Optimierungspotenzial aufweist /CHO 22/. In den Bereichen Diagnose und Vorhersage hat KI jedoch bereits praktische Anwendungen gefunden. Systeme zur Erkennung von Materialverschleiß, wie das Condition-Monitoring-Framework für Rohrleitungssysteme, ermöglichen präzisere Wartungsstrategien und verlängern die Lebensdauer von Anlagen /HAR 23/. Auch die Nutzung von anlagenexternen radiologischen Daten zur Unfallanalyse, etwa im Kontext von Station Blackouts (SBO), zeigt, wie KI-Modelle in realen Notfallszenarien wichtige Informationen liefern können /HYU 23b/.

Die Unterschiede zwischen produktiven Anwendungen und reiner Forschung hängen häufig mit der Komplexität der Modelle, den Sicherheitsanforderungen und der Verfügbarkeit von Daten zusammen.

Modelle wie PROMAP oder DRL-basierte Steuerungen, die umfangreiche Datensätze und komplexe Berechnungen benötigen, sind noch in der Forschungsphase, da die Integration in Kernkraftwerke erhebliche regulatorische und technische Herausforderungen mit sich bringt. Anwendungen, die auf klar definierte Aufgaben wie Materialüberwachung abzielen, sind einfacher zu implementieren und haben den Übergang zur Praxis bereits geschafft.

Die Daten, die für diese KI-basierten Systeme verwendet werden, stammen aus verschiedenen Quellen, darunter Echtzeitsensordaten, historische Betriebsdaten und simulationsbasierte Daten. Für die Anomalieerkennung und -vorhersage werden normale Betriebsdaten verwendet, sodass Anomalien durch Abweichungen erkannt werden können /LI 22a/, /CHO 22/. Da Unfallereignisse in der Kernkraft selten auftreten und somit kaum reale Daten verfügbar sind, setzen Ansätze auf Simulationen, um einen breiten Datensatz zu erhalten /HYU 23b/. Zusätzlich werden synthetische Datenerweiterungen und probabilistische Ansätze genutzt, um Variationen zu simulieren und Modelle besser auf seltene Ereignisse vorzubereiten /LYE 22/.

Im Vergleich zu herkömmlichen Methoden bieten diese KI-basierten Systeme klare Vorteile für die Sicherheit, da sie Anomalien und Fehlalarme schneller erkennen, präzisere Diagnosemöglichkeiten bieten und die Betriebseffizienz erhöhen. Sie reduzieren auch das Risiko menschlicher Fehler, besonders bei komplexen oder langen Betriebsprozessen. So zeigt die Kombination von KI und DRL in der Aufheizsteuerung verbesserte Reaktionszeiten und Stabilität /PAR 22/. Dennoch bestehen Risiken: Unzureichend trainierte Modelle oder fehlerhafte Daten können zu Fehlentscheidungen führen. Dies unterstreicht die Notwendigkeit strikter Validierungs- und Verifizierungsprozesse, insbesondere für sicherheitskritische Anwendungen /UHR 93/, /MIQ 22/. Mögliche Schwierigkeiten, welche bei dem Einsatz von KI-basierten Systemen im kerntechnischen Bereich entstehen können, sind:

Datenabhängigkeit und Verzerrungen: KI-Modelle sind stark von der Qualität und Vollständigkeit der Trainingsdaten abhängig. Verrauschte, unvollständige oder verzerrte Daten können zu falschen Vorhersagen oder Anomalieklassifikationen führen. Wenn die Modelle nicht für alle möglichen Betriebszustände oder Notfälle trainiert wurden, besteht das Risiko, dass sie diese nicht korrekt erkennen oder falsch interpretieren /HAR 23/, /CHO 22/.

Black-Box-Charakter der Modelle: Viele moderne KI-basierte Systeme, insbesondere neuronale Netze, sind schwer nachvollziehbar und bieten wenig Transparenz. Bediener können Schwierigkeiten haben, Entscheidungen der KI nachzuvollziehen, was zu Vertrauensproblemen oder falscher Anwendung führen kann. Dies ist besonders kritisch in Notfallsituationen, in denen schnelle und fundierte Entscheidungen erforderlich sind /MIQ 22/, /UHR 93/.

Fehlfunktionen durch unzureichendes Training: Unzureichend trainierte Modelle können unerwartete Entscheidungen treffen, insbesondere bei Betriebszuständen, die außerhalb des Trainingsbereichs liegen. Dies kann in sicherheitskritischen Situationen schwerwiegende Folgen haben, da solche Fehler oft schwer vorhersehbar und zu beheben sind /PAR 22/, /LI 22a/.

Cybersecurity-Risiken: KI-basierte Systeme, die auf vernetzte Datenquellen und digitale Plattformen angewiesen sind, können anfällig für Cyberangriffe sein. Ein gezielter Angriff kann beispielsweise Anomalien vortäuschen oder Fehlalarme auslösen, wodurch die Betriebssicherheit gefährdet wird /HYU 23a/, /MIQ 22/.

Übermäßige Abhängigkeit von KI: Während KI menschliche Fehler minimieren kann, besteht die Gefahr, dass Bediener sich zu stark auf die KI verlassen und ihr eigenes Situationsbewusstsein reduzieren. In Fällen, in denen die KI falsche Entscheidungen trifft, können menschliche Eingriffe zu spät oder ineffektiv sein /UHR 93/.

Eingeschränkte Fähigkeit zur Verifikation: Die Validierung und Verifikation (V&V) von KI-basierten Systemen ist eine große Herausforderung, da es oft unmöglich ist, alle potenziellen Betriebszustände und Kombinationen zu testen. Dies führt zu Unsicherheiten in Bezug auf die Zuverlässigkeit der Modelle, insbesondere in unbekannten oder extremen Szenarien /UHR 93/, /MIQ 22/.

Diese potenziellen Probleme machen deutlich, dass trotz der fortgeschrittenen Fähigkeiten von KI-basierten Systemen ein hohes Maß an Vorsicht erforderlich ist. Sorgfältige Planung, umfassende Validierung und kontinuierliche Überwachung sind entscheidend, um die Sicherheit und Zuverlässigkeit dieser Technologien in kerntechnischen Anwendungen zu gewährleisten.

Im Nachfolgenden geschieht die visuelle Einordnung der recherchierten KI-basierten Anwendungen nach der Klassifikation nach /IEC 23/ und /HYU 23a/.

Ähnlich wie in Abb. 2.7 zu sehen ist, befindet sich in der unten aufgeführten Abb. 3.2 eine vereinfachte Darstellung der in diesem Unterkapitel vorgestellten KI-basierten Anwendungen im kerntechnischen Bereich. Die Darstellung verwendet die Klassifikation nach /IEC 23/, wie sie in der Abb. 2.1 dargestellt ist. Die Abbildung ist von innen nach außen hin zu betrachten. Das innere Dreieck stellt die drei grundlegenden Kategorien von KI-Methoden dar. In dem darauffolgenden Bereich (näher zum Rand) werden durch die gestrichelten Linien die Unterkategorien abgebildet. Noch weiter außen befinden sich die KI-basierten Anwendungen, welche sich der jeweiligen Unterkategorie einordnen lassen. Die darauffolgende Abbildung (Abb. 3.3) zeigt ein Mengendiagramm der in diesem Kapitel ermittelten KI-basierten Anwendungen eingeordnet nach Kategorien, wie sie in /HYU 23a/ zu finden sind. Die Abbildung wird in vier zentrale Anwendungsfelder unterschieden: Diagnose, die der frühzeitigen Erkennung von Defekten und Anomalien dient; Vorhersage, mit der mögliche Betriebsstörungen oder Unfälle prognostiziert werden; Reaktion, die sich auf die Echtzeitunterstützung bei Notfällen konzentriert; und Prozessoptimierung, bei der der Reaktorbetrieb effizienter gestaltet wird.

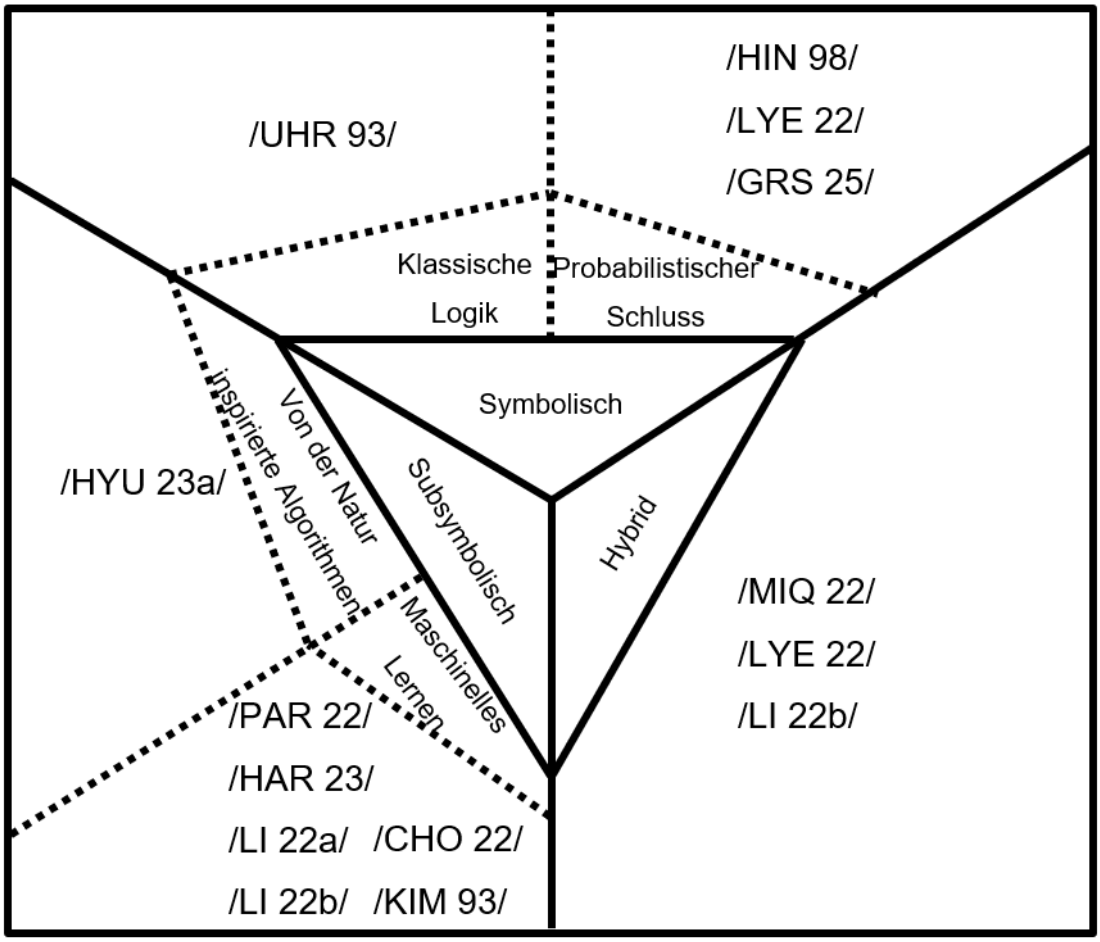

Abb. 3.2 Einordnung von KI-basierten Anwendungen nach /IEC 23/.

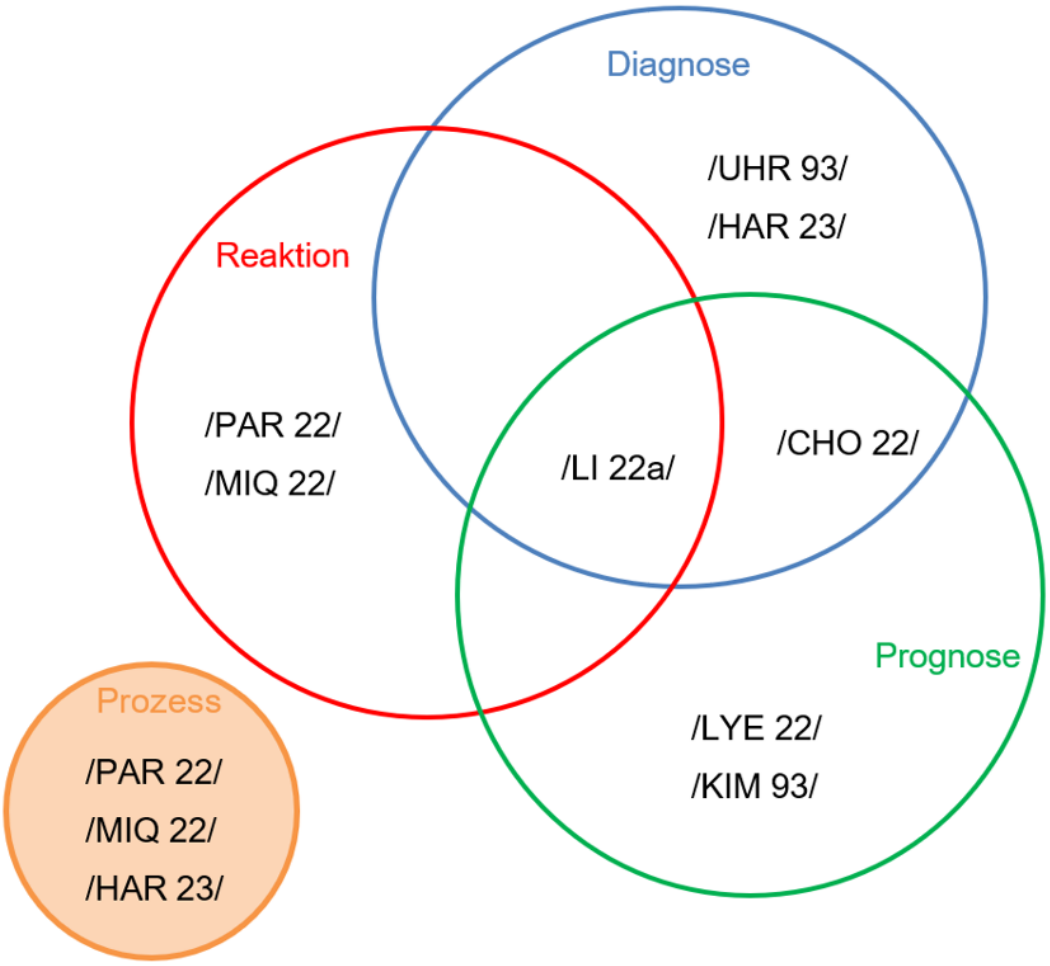

Abb. 3.3 Einordnung von KI-basierten Anwendungen nach /HYU 23a/.

4 Ermittlung und Bewertung von Qualifikationsansätzen und -methoden für KI-basierte Anwendungen mit sicherheitstechnischer Bedeutung

In diesem Kapitel werden die Ergebnisse der Arbeiten zum Arbeitspaket 3 des Vorhabens dargestellt. Der Schwerpunkt des Arbeitspakets 3 lag in der Ermittlung und Bewertung von Qualifikationsansätzen und -methoden für KI-basierten Anwendungen mit sicherheitstechnischer Bedeutung.

Durch die zunehmende Verbreitung und Anwendung von Künstlicher Intelligenz in verschiedenen Branchen werden Qualifikationsansätze und -methoden, um die Sicherheit, Effizienz und ethische Verträglichkeit von KI-basierten Anwendungen sicherzustellen, benötigt und entwickelt. Diese Ansätze und Methoden zielen darauf ab, die Leistungsfähigkeit von KI-basierten Anwendungen systematisch zu bewerten, potenzielle Risiken zu identifizieren und die Einhaltung von Normen und Standards durch kontinuierliche Überprüfung und Validierung zu gewährleisten.

Bei der Auswertung relevanter Dokumente, Normen, Regelwerke und Konferenzbeiträge im Rahmen des Vorhabens wurden drei Qualifikationsansätze und -methoden für KI-basierten Anwendungen mit sicherheitstechnischer Bedeutung ermittelt. Diese Qualifikationsansätze werden im nachfolgenden Abschnitt 4.1 dargestellt und im Hinblick auf einige ausgewählte relevante Aspekte bewertet.

Für die Bewertung eines potenziellen Einsatzes von KI-basierten Anwendungen mit sicherheitstechnischer Bedeutung wurde im Rahmen dieses Vorhabens überprüft, inwieweit bestehende Anforderungen an Software in sicherheitskritischen Bereichen (sowohl im nicht nuklearen als auch im nuklearen Bereich) auf KI-basierten Anwendungen mit sicherheitstechnischer Bedeutung übertragbar sind und somit auch für die Qualifizierung von KI-basierten Anwendungen herangezogen werden können. Die Ergebnisse dieser Untersuchungen sind im Abschnitt 4.2 dargestellt.

4.1 Qualifikationsansätze und -methoden für KI-basierten Anwendungen mit sicherheitstechnischer Bedeutung

4.1.1 Beschreibung der ermittelten Qualifikationsansätze

4.1.1.1 AMLAS-Prozess

Eine Qualifikationsmethode, welche eine umfassende Methodik zur Integration von Sicherheitsaspekten in die Entwicklung von maschinellen Lernkomponenten, insbesondere für autonome Systeme, anbietet, ist der AMLAS-Prozess (Assurance of Machine Learning in Autonomous Systems) /HAW 21/. Der AMLAS-Prozess bietet einen umfassenden Rahmen zur Gewährleistung der Sicherheit von maschinellen Lernkomponenten in autonomen Systemen und soll das Vertrauen in deren Zuverlässigkeit stärken. Der Prozess besteht aus sechs Phasen, die darauf abzielen, Sicherheitsnachweise in den Entwicklungszyklus von ML-Komponenten systematisch einzubinden. Eine Übersicht über den Prozess ist in Abb. 4.1 zu sehen. Der AMLAS-Prozess legt fest, welche Phasen im Entwicklungsprozess zu durchlaufen sind und die jeweiligen Ziele dieser, schreibt jedoch keine konkreten technischen Maßnahmen (Algorithmen, Testmethoden oder Metriken) zum Erreichen der Ziele vor.

In der ersten Phase, der Erarbeitung des Geltungsbereiches für ML-Sicherheit, wird der Umfang der Sicherheitsüberprüfung definiert. Hierbei werden die relevanten Sicherheitsanforderungen und der Kontext der Anwendung festgelegt, um die Risiken für das Gesamtsystem zu identifizieren und zu minimieren.

In der zweiten Phase, welche sich mit der Sicherheit in ML-Anwendungen befasst, werden spezifische Sicherheitsanforderungen entwickelt. Diese basieren auf den Systemanforderungen und dem operativen Umfeld, um zu bewerten, welche Gefahren die ML-Komponente möglicherweise für das Gesamtsystem darstellen könnte.

Die dritte Phase, welche die Sicherheit im Datenmanagement sicherstellt, sorgt dafür, dass die für das Training der ML-Komponente verwendeten Daten von hoher Qualität und Integrität sind. Es werden Maßnahmen zur Datenaufbereitung und -verarbeitung implementiert, um ein zuverlässiges Training des Modells zu gewährleisten.

In der vierten Phase, welche sich mit der Sicherheit im Modell-Lernprozesses befasst, wird sichergestellt, dass das ML-Modell entsprechend den festgelegten Sicherheitsanforderungen trainiert wird.

Diese Phase umfasst die Überwachung der Lernprozesse und die Bewertung der Modellergebnisse, um die Einhaltung der Sicherheitsstandards zu gewährleisten.

Die fünfte Phase, welche sich mit der Sicherheit in der Modell-Verifikation befasst, beinhaltet die Überprüfung und Validierung des ML-Modells. Durch Tests werden die Leistungsfähigkeit und Robustheit des Modells überprüft, um sicherzustellen, dass es den spezifizierten Anforderungen entspricht.

In der sechsten Phase, welche sich mit der Sicherheit in der Bereitstellung des ML-Modells befasst, werden Richtlinien für den Betrieb und die kontinuierliche Überwachung des ML-Systems definiert. Diese Phase stellt sicher, dass während des laufenden Betriebs fortlaufend Sicherheitsüberwachungsmaßnahmen implementiert werden, um die langfristige Sicherheit und Effektivität des Systems sicherzustellen.

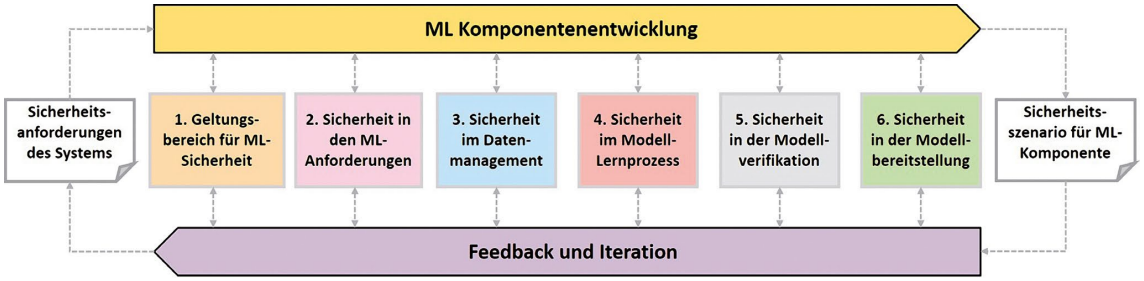

Abb. 4.1 Übersicht über den AMLAS-Prozess /HAW 21/, entnommen aus /WU 21/.

4.1.1.2 Qualifikationsansatz für hochriskante KI-basierte Systeme basierend auf der europäischen Verordnung über künstliche Intelligenz

Das Gesetz schafft einen Rahmen für den Einsatz von hochriskanten KI-basierten Systemen, ohne im Detail auf die einzelnen Qualifikationsmethoden einzugehen. Die EU KI-Verordnung legt keine expliziten Qualifikationsansätze fest, sondern definiert Anforderungen und Standards, die hochriskante KI-basierte Systeme erfüllen müssen. Diese Anforderungen geben jedoch klare Leitlinien vor, aus denen sich Qualifikationsansätze ableiten lassen.

Im Rahmen der EU KI-Verordnung werden hochriskante KI-basierte Systeme durch einen Qualifikationsansatz und -anforderungen bewertet. Für hochriskante KI-basierte Systeme gelten strenge Anforderungen. Diese Systeme müssen ein Risikomanagementsystem vorweisen, das die Ermittlung und Analyse der bekannten und vernünftigerweise vorhersehbaren Risiken umfasst. Zusätzlich müssen die Risiken abgeschätzt und

bewertet sowie andere möglicherweise auftretende Risiken bewertet werden. Die Erprobung solcher KI-basierter Systeme erfolgt bestenfalls zu jedem Zeitpunkt des Entwicklungsprozesses, spätestens jedoch vor dem Inverkehrbringen oder der Inbetriebnahme des Systems. Die Tests basieren auf zuvor festgelegten Messgrößen und Wahrscheinlichkeitsschwellen. Trainings-, Validierungs- und Testdatensätze müssen bestimmte Qualitätskriterien erfüllen, einschließlich der Bewertung ihrer Eignung hinsichtlich Datenlücken, Verzerrung, Kennzeichnung und Herkunft. Eine fortlaufende Überwachung des KI-basierten Systems durch natürliche Personen ist erforderlich, um eine korrekte Interpretation der Ergebnisse zu gewährleisten. Eingriffe in das System müssen durch eine „Stopp"-Taste möglich sein, und das System muss widerstandsfähig gegenüber Fehlern, Störungen und Eingriffen unbefugter Dritter sein.

Der Prozess des Inverkehrbringens eines hochriskanten KI-basierten Systems beginnt mit der Entwicklung des Systems und der Prüfung der Erfüllung der festgelegten Anforderungen. Anschließend muss das Stand-Alone KI-basiertes System in einer EU-Datenbank registriert werden. Eine Deklaration der Konformität ist zu unterzeichnen, und das KI-basierte System muss die CE-Kennzeichnung tragen, bevor es auf dem Markt angeboten werden kann. Sollten während des Lebenszyklus des KI-basierten Systems substanziellen Änderungen vorgenommen werden, ist eine erneute Prüfung der Erfüllung der Anforderungen erforderlich, um die Konformität weiterhin sicherzustellen.

Aus der EU KI-Verordnung könnten sich Qualifikationsansätze für hochriskante KI-basierte Systeme ableiten lassen. Die daraus resultierenden Qualifikationsanforderungen könnten beispielsweise Konformitätsbewertungen und Zertifizierungen beinhalten. Diese würden die Einhaltung rechtlicher Standards sichern, während umfassende Dokumentation die Nachvollziehbarkeit der Systemfunktionen gewährleisten würden. Qualitätsprüfungen der Daten und Bias-Management könnten Verzerrungen reduzieren, und Robustheits- sowie Sicherheitstests Stabilität sicherstellen. Menschliche Aufsicht würden Eingriffe bei Bedarf ermöglichen. Transparenzmaßnahmen würden dafür sorgen, dass Nutzer die Funktionsweise verstehen. Schlussendlich würden Datenschutz und Sicherheitsstandards den Schutz personenbezogener Daten gewährleisten.

4.1.1.3 Qualifikationsansätze des Fraunhofer Institutes für Produktionstechnik und Automatisierung IPA /WU 21/

Das White Paper des Fraunhofer Institutes für Produktionstechnik und Automatisierung IPA /WU 21/ soll einen Überblick über aktuelle Ansätze geben, wie KI-basierte Systeme in Bereichen mit sicherheitstechnischer Bedeutung so gestaltet werden können, dass sie den hohen Anforderungen an Sicherheit und Zuverlässigkeit gerecht werden. Dabei legt es besonderen Wert auf die Zertifizierung von KI-basierten Systemen und die Notwendigkeit, diese Systeme für Menschen transparent und erklärbar zu machen. Das White Paper erwähnt die Transparenz als essenzielles Element für den Einsatz von Künstlicher Intelligenz in sicherheitskritischen Bereichen. Transparenz wird als unverzichtbar betrachtet, um Vertrauen in KI-basierte Systeme zu schaffen und sicherzustellen, dass ihre Entscheidungen nachvollziehbar und überprüfbar sind. Ein wesentlicher Aspekt der Transparenz ist die Erklärbarkeit von KI-basierten Systemen. Das White Paper betont, dass die komplexen Entscheidungsprozesse von KI-Modellen, wie neuronalen Netzen, oft als "Black Boxes" betrachtet werden. Diese Intransparenz stellt ein erhebliches Hindernis für die Bewertung und Zertifizierung KI-basierten Systeme dar. Durch erklärbare KI können die internen Mechanismen dieser Modelle sichtbar gemacht werden, sodass Entwickler, Prüfer und Anwender nachvollziehen können, warum und wie eine Entscheidung getroffen wurde. Dies ist entscheidend, um Vertrauen in die Systeme aufzubauen und ihre Sicherheit zu bewerten. Transparenz betrifft jedoch nicht nur die Entscheidungen eines Modells, sondern erstreckt sich über den gesamten Entwicklungsprozess eines KI-basierten Systems. Von der Datenerhebung über die Modellentwicklung bis hin zur Implementierung und Validierung müssen alle Schritte klar dokumentiert und überprüfbar sein. Diese Nachvollziehbarkeit ist besonders wichtig für unabhängige Prüfstellen (z.B. TÜV Süd, DEKRA), die sicherstellen müssen, dass das KI-basierte System zuverlässig funktioniert und alle regulatorischen Anforderungen erfüllt. Die Dokumentation des gesamten Entwicklungsprozesses ist unerlässlich, um zu gewährleisten, dass das System korrekt entwickelt wurde und dass mögliche Fehlerquellen identifiziert und behoben werden können. Eine transparente Darstellung der Funktionsweise und des Verhaltens von KI-basierten Systemen trägt maßgeblich zur Förderung des Vertrauens in die Technologie bei. Dies gilt nicht nur für Entwickler und Prüfer, sondern auch für die breite Öffentlichkeit, die zunehmend mit KI in alltäglichen Anwendungen in Berührung kommt. Als zentralen Ansatz für die Sicherheitsargumentation von KI-basierten Anwendungen würde in dem White Paper die AMLAS-Methodik hervorgehoben. Grund dafür ist, dass die Methodik eine systematische Herangehensweise bietet, um Sicherheitsnachweise für KI-basierte Systeme zu entwickeln und in den Entwicklungsprozess zu

integrieren. Zudem stellt das Dokument eine Verbindung zwischen mehreren technischen Aspekten und deren Herausforderungen bei der Zertifizierung von KI-basierten Systemen und mehreren Phasen der AMLAS-Methodik her Abb. 4.1. Im Folgenden wird im Einzelnen auf diese Punkte eingegangen.

Erklärbare KI

Die Erklärbarkeit von KI-basierten Anwendungen (Explainable Artificial Intelligence - xAI) spielt eine bedeutende Rolle in der Zertifizierung von KI-basierten Systemen, da diese sich direkt auf das Vertrauen in die Systeme auswirkt. Erklärbare Entscheidungen von KI-basierten Systemen, vor allem in Bereichen mit sicherheitstechnischer Bedeutung, bieten Einblicke in die internen Entscheidungsprozesse dieser und erlauben es dem Menschen, die Vorhersagen zu versehen und das Vertrauen in diese zu stärken. Die Problematik der Transparenz der Entscheidungen entsteht durch den Aufbau tiefer neuronaler Netze, welche oft als „Black-Box“ angesehen werden. Neuronale Netze verarbeiten große Datenmengen und lernen basierend auf Mustern in den Daten. Dies kann zwar zu genauen Vorhersagen führen, jedoch können die Vorhersagen meistens nicht erklärt werden. Die vorgeschlagenen Methoden zur Erklärbarkeit von komplexen KI-Modellen beinhalten unter anderem erklärbare Modelle, modellagnostische Methoden und modellspezifische Methoden. Erklärbare Modelle wie Entscheidungsbäume oder lineare Regression sind von Natur aus verständlich und bieten klare Einsichten in ihre Entscheidungslogik, sind jedoch in komplexen Szenarien oft weniger leistungsfähig. Modellagnostische Methoden, wie LIME und SHAP, bieten Erklärungen unabhängig vom Modelltyp, indem sie lokale Approximationen oder gewichtete Merkmale zur Interpretation von Black-Box-Modellen wie neuronalen Netzen verwenden. Diese Methoden sind flexibel, da sie auf verschiedene Modelle angewendet werden können. Die Local Interpretable Model-Agnostic Explanations (LIME) Methode erklärt die Entscheidungen eines komplexen Modells durch die Konstruktion eines einfachen, lokalen Modells um eine spezifische Vorhersage herum. Dadurch wird ein komplexes Modell durch ein verständliches Surrogatmodell ersetzt, das für die spezifische Situation interpretiert werden kann. Die Shapley Additive Explanations (SHAP) Methode stammt aus der Spieltheorie und misst die Bedeutung einzelner Merkmale für eine Entscheidung. Sie erklärt, wie viel jedes Merkmal zur Vorhersage des Modells beigetragen hat, indem es den Beitrag jedes Merkmals zur Veränderung der Vorhersage quantifiziert.
Modellspezifische Methoden hingegen sind auf bestimmte Modellstrukturen wie tiefe

neuronale Netze oder Entscheidungsbäume zugeschnitten und nutzen deren spezifische Architektur, um effizientere und detailliertere Erklärungen zu liefern, wie beispielsweise DeepLIFT oder Tree SHAP.

Die DeepLIFT Methode, erklärt Vorhersagen von tiefen neuronalen Netzen, indem sie die Aktivierungen auf verschiedenen Ebenen des Netzes verfolgt und aufzeigt, wie viel jede Eingabe zur Ausgabe beigetragen hat. Die dargestellten Methoden lassen sich in drei Kategorien der Erklärbarkeit der KI unterteilen:

- **Instanzweise Erklärungen** zielen darauf ab, die Entscheidungsfindung für eine einzelne Eingabe zu erklären, indem sie aufzeigen, wie ein spezifisches Ergebnis zustande gekommen ist; dies ist besonders nützlich, um individuelle Vorhersagen nachzuvollziehen.
- **Erklärungen auf der Modellebene** bieten hingegen eine Übersicht über das gesamte Modell und versuchen, allgemeine Regeln oder Muster abzuleiten, die beschreiben, wie das Modell typischerweise arbeitet. Diese Ansätze helfen, das grundsätzliche Verhalten des Modells zu verstehen.
- **Erklärungen des Informationsflusses** Erklärungen des Informationsflusses befassen sich damit, wie Daten und Informationen durch die verschiedenen Schichten eines Modells fließen und welche Zwischenschritte zur finalen Entscheidung führen. Diese Methoden sind insbesondere bei tiefen neuronalen Netzen hilfreich, um den Einfluss von Eingabedaten auf die Ausgaben nachzuvollziehen. /EUR 19/

Formale Verifikation

Die formale Verifikation ist eine mathematische Methode, die eingesetzt wird, um die Korrektheit und Sicherheit von softwarebasierten Systemen, in diesem Betrachtungsfall KI-basierten Systemen, nachzuweisen, insbesondere in sicherheitskritischen Anwendungen wie dem autonomen Fahren, der Luftfahrt oder der Medizin. Ziel ist es, formale Beweise dafür zu erbringen, dass ein System unter allen möglichen Bedingungen bestimmte Sicherheitsanforderungen erfüllt. Dies unterscheidet sie von empirischen Testmethoden, bei denen Systeme für eine Vielzahl von Szenarien getestet werden. Stattdessen beweist die formale Verifikation auf Basis logischer Schlussfolgerungen, dass das System keine Fehler macht, die zu gefährlichen Situationen führen könnten. Ein Beispiel ist die Anwendung in autonomen Fahrzeugen. Die formale Verifikation kann hier verwendet werden, um zu beweisen, dass das Fahrzeug unter keinen Umständen eine

Kollision verursacht, unabhängig davon, welche Verkehrssituation vorliegt. Mathematisch wird dabei nachgewiesen, dass bestimmte Bedingungen immer erfüllt werden, wie etwa das Einhalten eines Sicherheitsabstands oder das ordnungsgemäße Erkennen von Hindernissen.

Bei tiefen neuronalen Netzen, die oft als Black-Box-Modelle gelten, stellt die Anwendung der formalen Verifikation jedoch eine Herausforderung dar, da diese Modelle eine enorme Komplexität und zahlreiche Parameter aufweisen. Dennoch werden spezialisierte Methoden entwickelt, um auch bei solchen Modellen formale Beweise zu erbringen, beispielsweise indem bestimmte logische Eigenschaften überprüft werden, wie etwa das Einhalten von Beschränkungen für die Ausgabe des Modells.

Statistische Validierung

Die statistische Validierung ist eine Methode, die verwendet wird, um die Leistung und Zuverlässigkeit eines KI-basierten Systems durch empirische Tests zu bewerten. Im Gegensatz zur formalen Verifikation, die auf mathematischen Beweisen basiert, wird bei der statistischen Validierung ein Modell mit einer Vielzahl von Testdaten und Szenarien getestet, um sicherzustellen, dass es robust und zuverlässig ist. Die statistische Validierung wird oft durch Tests auf Datensätzen durchgeführt, die das Modell während des Trainings nicht gesehen hat. Diese Testdaten sollen reale Szenarien so genau wie möglich widerspiegeln, um sicherzustellen, dass das Modell auch unter echten Bedingungen zuverlässig arbeitet. Für autonome Fahrzeuge könnte dies bedeuten, das Modell auf einer Vielzahl von Verkehrsbedingungen, Wetterlagen, Straßenarten und unerwarteten Ereignissen wie plötzlichen Hindernissen zu testen. Ein KI-basiertes System zur Erkennung von Fußgängern in autonomen Fahrzeugen könnte durch statistische Validierung geprüft werden, indem es auf verschiedenen Datensätzen getestet wird, die unterschiedliche Beleuchtungsbedingungen, Wetterverhältnisse und ungewöhnliche Verhaltensmuster von Fußgängern umfassen. Das Ziel besteht darin, zu validieren, dass das System eine hohe Genauigkeit aufweist und auch in seltenen oder extremen Situationen zuverlässig funktioniert. Die statistische Validierung ergänzt die formale Verifikation, da sie dem KI-Modell eine umfassendere Prüfung unter realen Bedingungen bietet. Sie ist jedoch nicht allumfassend, da sie nicht alle potenziellen Szenarien abdecken kann. Daher wird die statistische Validierung oft in Verbindung mit anderen Methoden, wie der formalen Verifikation und der Unsicherheitsquantifizierung, eingesetzt, um ein vollständiges Bild von der Sicherheit und Zuverlässigkeit eines Systems zu erhalten.

Unsicherheitsquantifizierung

Die Unsicherheitsquantifizierung ermöglicht es, neben der eigentlichen Vorhersage eines Modells auch die Unsicherheit dieser Vorhersage zu quantifizieren und zu kommunizieren.

Dies ist besonders wichtig, da KI-Modelle, insbesondere datengetriebene Verfahren wie neuronale Netze, nicht immer sicher in ihren Entscheidungen sind, was in sicherheitskritischen Umgebungen zu gefährlichen Situationen führen kann. Die Unsicherheitsquantifizierung hilft dabei, das Vertrauen in die Vorhersagen eines Modells zu erhöhen, indem sie nicht nur eine einzelne Entscheidung trifft, sondern auch eine Einschätzung darüber abgibt, wie sicher das Modell in seiner Vorhersage ist. Dies kann besonders nützlich sein, wenn das Modell mit unbekannten oder ungewöhnlichen Daten konfrontiert wird, die nicht im Trainingsdatensatz enthalten waren. In solchen Fällen kann das Modell einen höheren Unsicherheitswert zurückgeben, was es ermöglicht, dass das System entsprechend darauf reagiert, beispielsweise durch den Rückgriff auf menschliche Eingriffe oder vorsichtigere Entscheidungen. Ein Beispiel im autonomen Fahren ist, das einen Fußgänger erkannt werden soll. Zusätzlich zur Vorhersage, dass es sich um einen Fußgänger handelt, könnte eine Unsicherheit über diese Entscheidung liefern. Ist die Unsicherheit hoch, könnte das Fahrzeug durch eine langsamere Fahrt oder Anhalten vorsichtiger agieren. Auf diese Weise kann das System potenzielle Fehler in der Erkennung und Klassifizierung ausgleichen und sicherer agieren. Methoden zur Unsicherheitsquantifizierung umfassen Ansätze wie Bayes'sche Netze, die Wahrscheinlichkeitsverteilungen über die Modellparameter schätzen, oder Monte-Carlo-Dropout, bei dem das Modell während der Inferenz mehrmals mit unterschiedlichen Parametern durchlaufen wird, um die Varianz in den Vorhersagen zu messen.

Online Monitoring

Der Punkt beschreibt die kontinuierliche Überwachung von KI-basierten Systemen während ihres Einsatzes, um sicherzustellen, dass sie auch in realen und möglicherweise unvorhergesehenen Szenarien zuverlässig und sicher agieren. Im Gegensatz zu den vorangegangenen Verifikations- und Validierungsmethoden, die vor der Inbetriebnahme des Systems durchgeführt werden, sorgt das Online Monitoring dafür, dass das System während des Betriebs überwacht wird, um potenzielle Fehler oder Sicherheitsrisiken frühzeitig zu erkennen und darauf reagieren zu können. Die Notwendigkeit des Online

Monitorings ergibt sich aus der Tatsache, dass KI-basierte Systeme in sicherheitskritischen Anwendungen wie dem autonomen Fahren oder der Medizintechnik in dynamischen Umgebungen eingesetzt werden, in denen neue, unerwartete Situationen auftreten können. Das Online Monitoring bietet die Möglichkeit, diese Systeme kontinuierlich zu überwachen und bei Abweichungen von erwarteten Verhaltensweisen sofort Maßnahmen zu ergreifen. Dies kann durch das Zurückgreifen auf vordefinierte Sicherheitsprotokolle, den Übergang in einen sicheren Modus oder durch menschliche Eingriffe geschehen. Ein autonomes Fahrzeug könnte während der Fahrt von einem Monitoring-System überwacht werden, das fortlaufend überprüft, ob das KI-Modell korrekt funktioniert. Sollte das System eine hohe Unsicherheit bei der Erkennung von Objekten feststellen oder auf ungewöhnliche Verkehrssituationen treffen, die es nicht korrekt interpretieren kann, würde das Monitoring eingreifen und das Fahrzeug entweder anhalten oder an einen menschlichen Fahrer übergeben. Dadurch wird verhindert, dass das Fahrzeug in gefährliche Situationen gerät. Das Online Monitoring wird häufig mit Echtzeit-Feedback-Schleifen implementiert, bei denen das KI-basierte System und seine Entscheidungen kontinuierlich mit den aktuellen Umgebungsdaten abgeglichen werden. Überwachungssysteme können Anomalien oder ungewöhnliches Verhalten sofort erkennen und sicherstellen, dass das System entweder seine Leistung korrigiert oder in einen sicheren Zustand wechselt. Dies ist besonders wichtig, um sicherzustellen, dass das System auch in Fällen, die nicht vollständig durch formale Verifikation oder statistische Validierung abgedeckt wurden, sicher bleibt.

4.1.2 Bewertung der ermittelten Qualifikationsansätze

Die GRS hat im Rahmen des Vorhabens für die Bewertung der ermittelten Qualifikationsansätze basierend auf den ausgewerteten Dokumenten ein modulares Bewertungsschema erstellt und die zugehörigen Bewertungskriterien herangezogen. Die Bewertungskriterien entstanden durch die die Analyse der bekannten Unklarheiten, bestehenden und möglichen Schwierigkeiten bei der Qualifizierung und den bisher dokumentierten Erfahrungen beim Einsatz von KI-basierten Systemen in sicherheitstechnisch relevanten Bereichen. Als Herangehensweise wurde ein Top-Down-Ansatz verwendet. Bei diesem Ansatz wurden die im Vorhaben analysierten Qualifikationsprozesse für die Erstellung des Bewertungsschemas herangezogen. Aus den im Rahmen des Vorhabens erarbeiteten Medien (/EUR 24b/, /EUR 19/, /ISO 22a/, /ISO 22b/, /ISO 24/, /HAW 21/, /WU 21/) wurden Informationen extrahiert und hinsichtlich gemeinsamer, sich überschneidender (globaler) Bestandteile analysiert. Basierend auf den ermittelten globalen Bestandteilen wurde die oberste Ebene des Bewertungsschemas erzeugt. Die globale

Ebene beinhaltet grundlegende Faktoren, welche im Rahmen eines Qualifikationsprozesses eines KI-basierten Systems im Bereich mit sicherheitstechnischer Bedeutung behandelt werden müssten und somit ebenfalls in den zu prüfenden Qualifikationsansätzen wiedergefunden werden sollten. Der Top-Down-Ansatz zur Bewertung von Qualifikationsprozessen für KI-basierte Systeme in sicherheitstechnisch relevanten Bereichen ermöglicht eine strukturierte und schrittweise Herangehensweise an den Bewertungsprozess. Basierend auf der Analyse der im Rahmen des Vorhabens identifizierten Qualifikationsprozesse wurden drei übergeordnete Faktoren für das entwickelte Bewertungsschema identifiziert:

- **Menschliche Erklärbarkeit und Benutzerfreundlichkeit**
- **Technische Sicherheit und Robustheit**
- **Mathematische Fundierung und Datenmanagement**

Im ersten Schritt wird geprüft, ob ein Qualifikationsansatz diese drei Hauptfaktoren allgemein adressiert. Dabei wird zunächst auf oberster Ebene eine grobe Einschätzung vorgenommen, um zu bewerten, ob der Qualifikationsansatz eine solide Grundlage zur Bewertung der KI in sicherheitstechnischen Bereichen bietet. In der zweiten Ebene des Schemas werden die drei Hauptfaktoren detaillierter untersucht. Jeder dieser Faktoren wird in Unterfaktoren zerlegt, die spezifische Anforderungen oder Aspekte adressieren. Diese detaillierte Analyse ermöglicht es, die Tiefe und die Präzision eines Qualifikationsansatzes weiter zu bewerten.

Menschliche Erklärbarkeit und Benutzerfreundlichkeit

- **Erklärbarkeit der Entscheidungen**: Hierbei wird geprüft, ob der Entscheidungsprozess der KI so aufbereitet wird, dass der menschliche Nutzer Entscheidungen dieser nachvollziehen kann.
- **Verantwortung und Haftung**: Behandelt die Adressierung von ethischen Fragen, wie die Zuweisung von Verantwortung im Falle von Fehlentscheidungen des KI-basierten Systems.
- **Benutzerfreundlichkeit**: Hierbei wird geprüft, ob darauf geachtet wird, dass die Interaktion mit der KI intuitiv und benutzerfreundlich gestaltet ist, sodass auch in sicherheitskritischen Situationen die Bedienung sicher möglich ist.

Technische Sicherheit und Robustheit

- **Fehlertoleranz und Robustheit**: Die Fähigkeit des Systems stabil zu bleiben, auch wenn Störungen oder unvorhergesehene Situationen auftreten. Dies umfasst die Widerstandsfähigkeit der KI gegenüber Fehlern und ihre Robustheit im Betrieb.
- **Verifikation und Validierung**: Es ist entscheidend, dass umfassende Tests und Prüfungen durchgeführt werden, um sicherzustellen, dass die KI die festgelegten Anforderungen erfüllt und diese adäquat sind und der sicherheitstechnischen Bedeutung der Anwendung entsprechen. Dies dient dazu funktionale und nicht-funktionale Anforderungen an das KI-basierte System zu gewährleisten.
- **Skalierbarkeit und Anpassungsfähigkeit**: Die KI muss in der Lage sein, sich an veränderte Bedingungen oder steigende Komplexität anzupassen, ohne dabei an Sicherheit und Leistungsfähigkeit einzubüßen.

Mathematische Fundierung und Datenmanagement

- **Mathematische Konsistenz und Stabilität**: Die Auswahl von mathematischen Algorithmen und Parametern muss den Anforderungen zur Durchführung der gestellten Aufgaben entsprechen, um eine zuverlässige und stabile Arbeit des KI-basierten Systems zu gewährleisten. Dieser Aspekt stellt sicher, dass keine unerwarteten numerischen Instabilitäten auftreten und die Arbeit des KI-basierten Systems über den gesamten Einsatzzeitraum konsistent erfolgen kann.
- **Datenqualität und -verfügbarkeit**: Die verwendeten Daten werden hinsichtlich ihrer Qualität und Repräsentativität überprüft, um sicherzustellen, dass sie frei von Verzerrungen (Bias) sind, in ausreichender Menge vorliegen und um Under- oder Overfitting zu vermeiden.
- **Wahl der Netzwerkarchitektur und Hyperparameter-Tuning:** Um Under- oder Overfitting zu vermeiden, ist eine sorgfältige Auswahl der Hyperparameter und der Netzwerkarchitektur erforderlich. Das Netzwerk muss so konfiguriert werden, dass es eine ausreichende Generalisierung erreicht und nicht nur die Trainingsdaten lernt. Hierbei spielen insbesondere die Wahl der Netzwerkarchitektur, der Verlustfunktion und der Aktivierungsfunktion eine zentrale Rolle. Diese Komponenten müssen passend zum jeweiligen Anwendungsbereich und den spezifischen Anforderungen ausgewählt werden, um optimale Ergebnisse zu erzielen.

- **Kontinuierliche Überwachung und Anpassung**: Das System wird regelmäßig überwacht, und es existieren Mechanismen zur kontinuierlichen Anpassung und Aktualisierung der KI. Dabei werden neue Daten und Erkenntnisse, u. a. aus möglicher Betriebserfahrung, zurückgeführt und bei Bedarf in einer Anpassung berücksichtigt. Dies stellt sicher, dass neue Daten und Erkenntnisse in das System integriert werden, um es ständig zu optimieren. Eine Änderung in der Peripherie oder der Umgebung (Einbau eines neuen Sensors), welche eine entsprechende Anpassung oder Änderung des KI-basierten Systems zur Folge haben würde, sollte, ähnlich wie bei konventionellen softwarebasierten Systemen, zu einer Bewertung der Änderungen und einer Nachqualifizierung des KI-basierten Systems führen. Hiermit wird sichergestellt, dass das KI-basierte System weiterhin den Anforderungen entspricht, oder ob zusätzliche Maßnahmen notwendig sind, um die Funktionstüchtigkeit und Zuverlässigkeit auch unter den neuen Bedingungen zu gewährleisten.

In der zweiten Ebene des Schemas wird detailliert analysiert, welche spezifischen Aspekte der einzelnen Faktoren im jeweiligen Qualifikationsansatz adressiert werden. Hierbei werden einzelne, im ersten Schritt genannte. Punkte aufgegriffen, vertieft behandelt und präzisiert. Hier werden spezifische Werkzeuge und Techniken untersucht, die zur Umsetzung der in den oberen Ebenen definierten Anforderungen verwendet werden können. Beispiele für solche Methoden in den einzelnen Unterfaktoren könnten sein:

Datenqualität und -verfügbarkeit

- **Datenbereinigungsalgorithmen**: Es werden Verfahren zur automatischen Erkennung und Korrektur von Datenfehlern eingesetzt, wie etwa Algorithmen zur Erkennung von Ausreißern oder fehlenden Werten.
- **Bias-Vermeidungsmethoden**: Es werden spezielle Methoden implementiert, um Verzerrungen in den Trainingsdaten zu minimieren, z. B. durch rebalancing der Daten oder die Nutzung von Fairness-Algorithmen.

Verifikation und Validierung

- **Simulationstests**: Der Qualifikationsprozess könnte auf umfangreichen Simulationen basieren, um das Verhalten der KI unter verschiedenen Bedingungen zu überprüfen, vorausgesetzt die notwendigen Daten stehen zur Verfügung.

Stress-Tests: Dabei wird das KI-basierte System extremen Situationen ausgesetzt, um zu sehen, wie es in sicherheitskritischen Szenarien reagiert.

Obwohl die Cyber-Sicherheit kein ausschließlich KI-spezifisches Thema ist, stellt sie dennoch einen unverzichtbaren Aspekt dar.

Der Schutz von KI-basierten Systemen vor externen Bedrohungen wie Hacking oder Datenmanipulation ist entscheidend, um die Integrität und Sicherheit der Systeme zu gewährleisten. Dazu werden gezielte Maßnahmen ergriffen, die die Sicherheit der KI gegenüber potenziellen Cyber-Bedrohungen stärken. Beispiele solcher Maßnahmen sind:

- **Angriffserkennungssysteme (Intrusion Detection Systeme (IDS)):** Diese Systeme überwachen die Netzwerksicherheit der KI-basierten Systeme kontinuierlich und identifizieren potenzielle Angriffe oder ungewöhnliche Aktivitäten in Echtzeit, sodass frühzeitig Gegenmaßnahmen eingeleitet werden können.
- **Verschlüsselungstechniken:** Um Datenmanipulationen zu verhindern, kommen fortschrittliche Verschlüsselungsmechanismen zum Einsatz, die die Kommunikation und den Datenaustausch der KI-basierten Systeme absichern.

In dieser dritten Ebene wird das Schema so verfeinert, dass spezifische Werkzeuge und Methoden identifiziert werden, die in den Qualifikationsprozessen Anwendung finden. Dies geschieht auf der Basis der im zweiten Schritt aufgeführten Unterfaktoren. Das Ziel ist es, eine tiefgehende Analyse der Implementierungsebenen des Qualifikationsansatzes zu ermöglichen und festzustellen, wie umfassend die verschiedenen Aspekte adressiert werden.

Die Bewertung lässt sich anhand eines mehrstufigen Top-Down-Ansatzes durchführen, der zunächst die globalen Faktoren betrachtet und sich dann auf immer konkretere Aspekte des Qualifikationsansatzes konzentriert. Zu Beginn sollte geprüft werden, ob und in welchem Umfang der Ansatz die drei zentralen Hauptfaktoren – menschliche Erklärbarkeit und Benutzerfreundlichkeit, technische Sicherheit und Robustheit sowie mathematische Fundierung und Datenmanagement – grundsätzlich abgedeckt werden. Auf dieser obersten Ebene reicht es die wesentlichen Themenfelder zu identifizieren und ihre Berücksichtigung grob einzuschätzen, um beurteilen zu können, ob der Qualifikationsansatz eine solide Grundlage für sicherheitstechnische Anwendungen bietet.

Im nächsten Schritt sollte eine genauere Analyse dieser Hauptfaktoren erfolgen, indem sie in spezifische Unterfaktoren aufgeschlüsselt und vertiefend untersucht werden.

So sollte beispielsweise die Erklärbarkeit des KI-basierten Systems daraufhin überprüft werden, wie nachvollziehbar die Entscheidungen aufbereitet werden und inwiefern die Interaktion mit der KI benutzerfreundlich gestaltet ist. Für die technische Sicherheit und Robustheit sollten unter anderem Aspekte wie Fehlertoleranz, Verifikation und Validierung sowie die Skalierbarkeit und Anpassungsfähigkeit des Systems bewertet werden. Bei der mathematischen Fundierung und dem Datenmanagement sollte die Auswahl stabiler, konsistenter Algorithmen im Vordergrund stehen, ebenso wie die Gewährleistung von Datenqualität und der Einrichtung geeigneter Mechanismen für Netzwerkarchitekturen, Hyperparameter-Tuning und die kontinuierliche Überwachung. Diese detaillierte Betrachtung sollte dabei helfen, potenzielle Stärken und Schwächen eines Qualifikationsansatzes gezielt zu erfassen.

In der dritten Ebene sollten schließlich konkrete Methoden, Werkzeuge und Techniken hinzugezogen werden, mit denen sich die in den Unterfaktoren festgelegten Anforderungen umsetzen lassen. Beispiele dafür können Datenbereinigungsalgorithmen zur Erkennung und Korrektur fehlerhafter Datensätze, Bias-Vermeidungsmaßnahmen und Simulationen oder Stresstests, mit denen das Verhalten der KI unter ungewöhnlichen Bedingungen erprobt wird, sein. Auch die Cyber-Sicherheit sollte in dieser Phase eine entscheidende Rolle spielen, sodass Angriffserkennungssysteme Teil des Maßnahmenkatalogs sein sollten. Durch diese detaillierte, transparente Zuordnung sollte deutlich werden, ob der Qualifikationsansatz nicht nur auf dem Papier robust erscheint, sondern auch praktisch umsetzbare und wirksame Schritte vorsieht.

Abschließend bietet es sich an, die Ergebnisse der Bewertung kompakt zu dokumentieren und Handlungsempfehlungen abzuleiten. Diese könnten darauf hinweisen, an welchen Stellen der Qualifikationsansatz bereits gut aufgestellt ist und wo zusätzliche Maßnahmen getroffen werden sollten. Eine regelmäßige Überprüfung des Qualifikationsansatzes könnte dafür sorgen, dass das Vorgehen kontinuierlich verbessert und an den technologischen Fortschritt angepasst wird.

Das Schema zur Bewertung von Qualifikationsansätzen für KI-basierte Systeme lässt sich kontinuierlich erweitern, indem immer mehr Kriterien und Faktoren hinzugefügt werden, um die Komplexität der Anforderungen abzubilden. Diese Erweiterungen können auf Grundlage neuer Informationen, Forschungsergebnisse, anderer Qualifikationsansätze oder sich ändernder Gesetzgebungen durchgeführt werden, was sicherstellt, dass das Schema stets aktuell und relevant bleibt. Theoretisch ist das Schema unbegrenzt

erweiterbar, da es flexibel an neue Erkenntnisse und technologische Entwicklungen angepasst werden kann. Da der Bewertungsprozess eines Qualifikationsansatzes von vielen verschiedenen Faktoren abhängt – wie dem spezifischen Sektor oder der Art der Anwendung – kann sich der Bewertungsrahmen von der globalen Ebene bis hin zu den detaillierten, lokalen Ebenen erstrecken. Dies ermöglicht eine anpassbare und skalierbare Analyse, die sowohl umfassende, allgemeine Kriterien als auch spezifische, auf den Kontext zugeschnittene Anforderungen berücksichtigt. So können auch spezifische Qualifikationsansätze übernommen, in das Bewertungsschema integriert, und zur Bewertung neuer Qualifikationsprozesse angewendet werden. Das Ziel dieses Ansatzes ist es, eine dynamische und erweiterbare Bewertungsstruktur zu schaffen, die bei der Analyse neuer Qualifikationsansätze zur Anwendung kommt. Mit dieser Struktur kann gezielt geprüft werden, welche Faktoren und Unterfaktoren ein spezifischer Qualifikationsansatz abdeckt und wie tiefgreifend dies geschieht. Gleichzeitig ermöglicht das Schema, bewährte Methoden und Ansätze aus vorhandenen Qualifikationsprozessen zu übernehmen und für die Bewertung zukünftiger Prozesse anzuwenden. So entsteht ein flexibles und robustes Analysewerkzeug, das eine kontinuierliche Verbesserung und Standardisierung der Qualifikationsansätze für KI-basierte Systeme in sicherheitskritischen Bereichen fördert.

4.2 Übertragbarkeit von Anforderungen an Software in sicherheitskritischen Bereichen auf KI-basierten Anwendungen mit sicherheitstechnischer Bedeutung

KI-basierte Anwendungen sind als Softwareanwendungen anzusehen. Für die Bewertung eines potenziellen Einsatzes von KI-basierten Anwendungen mit sicherheitstechnischer Bedeutung wurde im Rahmen dieses Vorhabens überprüft, inwieweit bestehende Anforderungen an Software in sicherheitskritischen Bereichen (sowohl im nicht nuklearen als auch im nuklearen Bereich) auf KI-basierten Anwendungen mit sicherheitstechnischer Bedeutung übertragbar sind. Im nachfolgenden Abschnitt 4.2.1 wird die hierfür anzuwendende Bewertungsgrundlage beschrieben. Anschließend wird im Abschnitt 4.2.2 die Ergebnisse der im Rahmen des Vorhabens durchgeführte Übertragbarkeitsprüfung dargestellt.

4.2.1 Bewertungsgrundlage für die Übertragbarkeitsprüfung von bestehenden Softwareanforderungen in sicherheitskritischen Bereichen auf KI-basierte Anwendungen mit sicherheitstechnischer Bedeutung

Basierend aus den gewonnenen Erkenntnissen aus der Bearbeitung des AP 2 (siehe Kapitel 0) ist festzuhalten, dass KI-basierte Anwendungen mit sicherheitstechnischer Bedeutung derzeit beispielsweise in der Leittechnik bzw. für Steuerungsaufgaben eingesetzt werden können. Da KI-basierte Anwendungen als Softwareanwendungen angesehen werden können, können dementsprechend Anforderungen an digitale Leittechnik im nuklearen und im nichtnuklearen Bereich als Grundlage für die Übertragbarkeitsprüfung von bestehenden Softwareanforderungen in sicherheitskritischen Bereichen auf KI-basierte Anwendungen mit sicherheitstechnischer Bedeutung herangezogen werden.

Anforderungen an Software und Hardware digitaler Leittechniksysteme in sicherheitskritischen Bereichen sind in zahlreichen Regelwerken, Normen und Standards sowohl für den nuklearen als auch für den nichtnuklearen Bereich zu finden. In /GRS 16/ findet sich einen umfassenden Überblick über relevante nationale und internationale Regelwerke, Normen und Standards mit Anforderungen an digitale Leittechniksysteme im nuklearen sowie im nichtnuklearen Bereich.

Im nuklearen Bereich finden sich Anforderungen an Software digitaler Leittechniksysteme im deutschen kerntechnischen Regelwerk beispielsweise in Regeln des kerntechnischen Ausschusses /KTA 10/, /KTA 17/ und /KTA 18/. Im internationalen Bereich sind diverse IAEA-Standards vorhanden, welche sich mit Anforderungen an in Leittechniksystemen eingesetzter Software beschäftigen, beispielsweise /IAE 00/, /IAE 94/ und /IAE 99/. In der Publikationsreihe der US NRC NUREG finden sich auch Anforderungen an Leittechniksystemen und zum Einsatz von Software in Leittechniksystemen in Kernkraftwerken wie in /NRC 07/ und in /NUR 01/. Anforderungen an Leittechnikfunktionen, welche durch die Anwendungssoftware des digitalen Leittechniksystems realisiert sind, werden entsprechend ihrer sicherheitstechnischen Bedeutung kategorisiert, wobei gemäß /KTA 10/ zwischen den Kategorien A, B und C unterschieden wird. Die Kategorie A stellt hierbei die höchste und die Kategorie C die niedrigste Einstufung. In den DIN IEC 60880 /DIN 10/ und der DIN EN 62138 /DIN 20a/ werden Softwareaspekte leittechnischer Systeme für Kernkraftwerke entsprechend ihrer Sicherheitsklassifizierung behandelt.

Für sicherheitskritische Bereiche im nichtnuklearen Bereich sind Anforderungen an den Einsatz von Software in Steuerungen von Industrieanlagen beispielsweise in /DIN 11a/, /DIN 23/ und /DIN 13/ angegeben. Es sind ebenfalls diverse IEEE-Standards zu der Thematik Einsatz von Software in sicherheitskritischen nichtnuklearen Bereichen wie z. B. in /IEE 12/ verfügbar.

Diverse andere Normen und Standards beschäftigen sich speziell mit Anforderungen für den Einsatz und die Qualifizierung von Software digitaler Leittechnik- und Steuerungssysteme in der Automobilindustrie, im Schienenverkehr, in der Luft- und Raumfahrt und in der Wehrtechnik. Beispielsweise sind in /DIN 11b/ „Anforderungen an sicherheitsrelevanter Software für Eisenbahnsteuerungs- und -überwachungssysteme“ enthalten. Im Bereich der Automobilindustrie ist die Normenreihe ISO 26262 „Road Vehicles – Functional Safety“ /ISO 18/ mit entsprechenden Anforderungen auf der Softwareebene zu nennen. Für die Luft- und Raumfahrt sind bezüglich Softwareaspekte u. a. die internationalen Standards /RTC 11/, /FED 03/ und /NAT 13/ zu nennen. In /RTC 11/ sind beispielsweise Anforderungen zur Softwareentwicklung im sicherheitskritischen Bereich der Luftfahrt enthalten. Im Bereich der Wehrtechnik enthält beispielsweise der Standard /DEP 10/ entsprechende Anforderungen an die Softwareentwicklung.

Die Softwareanforderungen in den genannten Standards und Normen orientieren sich in der Regel an die Phasen des Software-Lebenszyklus und an die damit verbundenen Tätigkeiten wie beispielsweise in /DIN 20b/ definiert. Der Software-Sicherheitslebenszyklus gemäß /DIN 20b/ umfasst alle notwendigen Tätigkeiten im Zusammenhang mit der Entwicklung und dem Betrieb der Software von der Anforderungsspezifikation der Software in der Konzeptphase bis zu dem Zeitpunkt, in dem die Software nicht mehr für die Nutzung zur Verfügung steht. Zusätzlich zu den typischen Themenbereichen aus dem Software-Lebenszyklus (Anforderungsspezifikation, Entwicklung, Verifizierung, Validierung, Integration, Tests etc.) sind in einigen der genannten Standards und Normen Anforderungen an das Management und an die Organisationsstruktur im Hinblick auf Software sowie zu weiteren Themen wie z. B. der Softwarezugriffschutz zu finden. In /GRS 16/ findet sich eine Übersicht über behandelte Themenbereiche in verschiedenen Normen und Standards zu sicherheitsrelevanter Software im nuklearen und im nichtnuklearen Bereich.

4.2.2 Ergebnisse der Übertragbarkeitsprüfung

Das für die Übertragbarkeitsprüfung entwickelte Konzept sieht vor, dass zunächst die Bewertungsgrundlage festgelegt wird. Hierfür sind entsprechende Normen, Standards und Regelwerke mit Anforderungen an Software digitaler Leittechniksysteme in sicherheitskritischen Bereichen im nuklearen und im nichtnuklearen Bereich wie im Abschnitt 4.2.1 beschrieben heranzuziehen. Anschließend sind in einem zweiten Schritt die bei den Softwareanforderungen im Rahmen der Übertragbarkeitsprüfung zu betrachtende Themenbereiche und Aspekte (Spezifikation, Entwicklung, Verifizierung, Validierung, Prüfung, etc.) zu identifizieren. Im letzten Schritt erfolgt dann die Bewertung der Übertragbarkeit dieser Anforderungen an KI-basierte Anwendungen mit sicherheitsrelevanter Bedeutung unter Berücksichtigung der Bewertungskriterien für Qualifikationsansätze aus Abschnitt 4.1.2. und der ermittelten Eigenschaften der KI-basierten Anwendungen.

Im Rahmen dieses Vorhabens wurden zur Übertragbarkeitsprüfung von bestehenden Softwareanforderungen in sicherheitskritischen Bereichen auf KI-basierten Anwendungen mit sicherheitstechnischer Bedeutung exemplarisch Anforderungen aus den KTA-Regeln /KTA 10/, /KTA 20/, /KTA 17/ und /KTA 18/ im nuklearen Bereich und Anforderungen aus der Luftfahrt /RTC 11/ im nichtnuklearen Bereich zugrunde gelegt. Entsprechende Anforderungen aus anderen zuvor genannten Regelwerken, Standards und Normen im nuklearen und im nichtnuklearen Bereich (siehe Abschnitt 4.2.1) können ähnlich herangezogen werden.

Basierend auf den ausgewerteten Normen und Regelwerken und aus den Ausführungen in /GRS 16/ wurden im Rahmen der Übertragbarkeitsprüfung u. a. Anforderungen an die Software betreffend die nachfolgenden Aspekte betrachtet.

- Entwicklung
- Verifizierung
- Validierung
- Qualifizierung
- Prüfung der Software
- Änderungen

In der KTA-Regel 1401 „Allgemeine Anforderungen an die Qualitätssicherung“ /KTA 20/ wird im Kapitel 3 u. a. die Erstellung einer detaillierten Entwicklungs- bzw. Abnahmespezifikation für die Software gefordert.

Die Durchführung von Tests, darunter Komponententests , Modultests, und Integrationstests, sind gemäß /KTA 20/ nachzuweisen. Bei Änderung der Software sind die erforderlichen Tests und die Dokumentation fortzuführen. Der hierfür erforderliche Detaillierungsgrad der Entwicklungs- bzw. Abnahmespezifikation für die Software ist in /KTA 20/ nicht weiter spezifiziert. Diese Anforderungen stellen daher übergeordnete Anforderungen an die Entwicklung, Prüfung und Änderung der Software dar. Es ist jedoch davon auszugehen, dass eine sinngemäße Umsetzung dieser Anforderungen die Nachvollziehbarkeit, die Verifizierung und die Validierung einer KI-basierten Anwendung mit sicherheitstechnischer Bedeutung unterstützen.

In der KTA-Regel 3501 /KTA 10/ sind im Kapitel 5.1.2 grundsätzliche Anforderungen im Hinblick auf die Softwarequalität für Einrichtungen der Sicherheitsleittechnik in Kernkraftwerken enthalten. Diese betreffen u. a. die Entwicklung und die Qualifizierung der Software, den Aufbau und den Funktionsumfang der Software. Gemäß /KTA 10/ hat die Entwicklung und Qualifizierung der Software so zu erfolgen, dass eine durchgängige Nachweisführung der korrekten Arbeitsweise der Software gewährleistet ist. In diesem Zusammenhang wird in /KTA 10/ gefordert, dass die Software in verifizierbaren Schritten in einem Phasenmodell zu entwickeln ist. Hierbei ist die Anwendersoftware ausgehend von der verfahrenstechnischen Aufgabenstellung zu entwickeln. Hinsichtlich des Aufbaus und des Funktionsumfangs der Software wird in /KTA 10/ gefordert, dass die Software einfach aufgebaut sein soll und dass der Funktionsumfang der Software auf das für die jeweilige Funktion notwendige Maß begrenzt sein soll. Zudem sind gemäß /KTA 10/ die Funktionen der Anwendersoftware und der Systemsoftware in eigenständigen Softwareeinheiten zu realisieren. In der Softwarearchitektur ist die Anwendersoftware von der Systemsoftware zu trennen. Die Nachweisführung der korrekten Arbeitsweise der Software, der einfache Aufbau, die Begrenzung des Funktionsumfangs auf das Notwendige, die Entwicklung der Software in verifizierbaren Schritten, die Entwicklung eigenständiger Softwareeinheiten stellen Aspekte dar, welche bei entsprechender Anwendung die Erklärbarkeit, die Nachvollziehbarkeit, die Verifizierung und Validierung einer KI-basierten Anwendung mit sicherheitstechnischer Bedeutung fördern.

Spezifische Anforderungen für die Typprüfung von elektrischen Baugruppen der Sicherheitsleittechnik und von Messwertgebern und Messumformern der Sicherheitsleittechnik nach KTA 3501 /KTA 10/ sind in /KTA 17/ und /KTA 18/ enthalten. Hinsichtlich der in den Baugruppen, Messwertgebern und Messumformern eingesetzten Software wird im Rahmen der Typprüfung u. a. gefordert, dass der Software-Entwicklungsprozess von rechnerbasierten Baugruppen und der Entwicklungsprozess von programmierbaren Bauelementen durch Unterlagen zu belegen ist. Diese Unterlagen schließen u. a. die Anforderungsspezifikation sowie die Prüf- und Testdokumentation ein. Weiterhin wird gefordert, dass der Aufbau, der Funktionsablauf und das Zeitverhalten des Programmes für rechnerbasierte Baugruppen oder Baugruppen mit programmierbaren Bauelementen zu beschreiben sind. Bei dem Einsatz vorgefertigter Software sind gemäß /KTA 17/ und /KTA 18/ das Qualifizierungs- oder Eignungsnachweisverfahren und die Ergebnisse darzulegen. Die genannten Anforderungen können sinngemäß auf KI-basierte Messwertgeber und Messumformer angewandt werden, welche für einen Einsatz in sicherheitstechnisch wichtigen Systemen vorgesehen sind. Die in diesen Anforderungen enthaltenen Aspekte u. a. bezüglich der Nachweisführung im Software-Entwicklungsprozess von programmierbaren Bauelementen, des Aufbaus und Funktionsablaufs und des Zeitverhaltens des Programmes für rechnerbasierte Baugruppen oder Baugruppen können für die Überprüfung der Erklärbarkeit und Nachvollziehbarkeit der KI-basierten Anwendung herangezogen werden. Sie können zudem den Verifizierungs- und Validierungsprozess der vorliegenden KI-basierten Anwendung unterstützen.

Basierend auf den exemplarisch betrachteten KTA-Regeln ist zusammenfassend Folgendes festzuhalten:

- Übergeordnete Anforderungen an die Entwicklung, Prüfung und Änderung der Software, wie sie in /KTA 20/ enthalten sind, können die Nachvollziehbarkeit, die Verifizierung und die Validierung einer KI-basierten Anwendung mit sicherheitstechnischer Bedeutung unterstützen.
- Grundsätzliche Anforderungen im Hinblick auf die Softwarequalität für Einrichtungen der Sicherheitsleittechnik in Kernkraftwerken, wie in /KTA 10/ angegeben, enthalten Aspekte, welche die Erklärbarkeit, die Nachvollziehbarkeit, die Verifizierung und Validierung einer KI-basierten Anwendung mit sicherheitstechnischer Bedeutung fördern.

- Spezifische Anforderungen für die Typprüfung von elektrischen Baugruppen der Sicherheitsleittechnik und von Messwertgebern und Messumformern der Sicherheitsleittechnik nach KTA 3501 /KTA 10/, wie sie in /KTA 17/ und /KTA 18/ formuliert, können beispielsweise auf KI-basierte Messwertgeber und Messumformer angewandt werden können, welche für einen Einsatz in sicherheitstechnisch wichtigen Systemen vorgesehen sind.

Es ist hierbei zu erwähnen, dass der Grad der Umsetzbarkeit der zuvor genannten Anforderungen bei konkreten KI-basierten Anwendungen mit sicherheitstechnischer Bedeutung u. a. von der verwendeten KI-Methode abhängig wird. Deterministische KI-Systeme, wie z.B. Expertensysteme bringen hierbei aufgrund ihrer klassischen Logik geeignetere Voraussetzungen für eine Umsetzung dieser Anforderungen mit.

4.3 GRS-Fazit zur Bewertung der Qualifikationsansätze und -methoden für KI-Anwendungen und zur Übertragbarkeitsprüfung von Softwareanforderungen

Im Rahmen der durchgeführten Arbeiten konnte gezeigt werden, dass herkömmliche Softwareanforderungen durchaus auch für KI-basierte Systeme relevant sind. Nichtsdestotrotz bleiben einige Herausforderungen bei dem Einsatz von KI-basierten Systemen im sicherheitstechnischen Umfeld, insbesondere im Bezug auf Qualifikationsprozesse solcher Systeme, bestehen. Die Übertragbarkeitsprüfung wirft zusätzlich Fragen im Bezug auf Komponenten-, Modul- und Integrationstests von KI-basierten Systemen auf. Zum einen müssten KI-basierte Systeme grundlegende Anforderungen der klassischen Softwareentwicklung erfüllen, zum anderen sind Kriterien wie Erklärbarkeit, Robustheit und ein durchdachtes Datenmanagement zu Berücksichtigen. Der potenzielle Einsatz von KI-basierten Systemen in Bereichen mit sicherheitstechnischer Bedeutung erfordert neue Ansätze bei Testverfahren und Zertifizierung.

In der klassischen Softwareentwicklung sind Modul- und Integrationstests klar definiert. Modultests zielen darauf ab, einzelne Funktionen oder Klassen isoliert zu überprüfen. Das Verhalten ist dabei vollständig durch den Code definiert, sodass sich Fehler eindeutig identifizieren und reproduzieren lassen. Integrationstests prüfen hingegen das Zusammenspiel mehrerer solcher Module im Gesamtsystem. Sie simulieren realistische Nutzungsszenarien, um sicherzustellen, dass die Komponenten korrekt miteinander interagieren, Schnittstellen stabil funktionieren und keine unerwarteten Effekte auftreten.

Diese Konzepte lassen sich jedoch nicht ohne Weiteres auf KI-basierte Systeme übertragen, insbesondere dann, wenn datengetriebene und subsymbolische Verfahren zum Einsatz kommen. Während Komponententests, etwa zur Überprüfung von Bibliotheken, Datenpipelines oder Framework-Funktionalitäten, weitgehend deterministisch und reproduzierbar durchgeführt werden könnten, ließe sich Prüfbarkeit nur bedingt auf die Modultestebene ausweiten. Schwierigkeit besteht darin, dass obwohl einzelne Komponenten der KI-basierten Software getestet werden könnten, ist das finale Verhalten des Moduls nach dem Training nicht deterministisch, da ihr Verhalten wesentlich durch datenbasierte Gewichtungen, Initialisierungen, und stochastische Prozesse geprägt ist, und kann nicht auf die gleiche Art geprüft werden.

Die Verifikation solcher Systeme könnte daher angepasste, datengetriebene Testansätze, die jenseits klassischer Softwareprüfverfahren liegen, erfordern. Wie sich Modultests und Integrationstests im Kontext von KI-Systemen gestalten lassen ist nicht abschließend geklärt und dürfte in hohem Maße vom jeweiligen Systemaufbau, konkreten Anwendungsfall und den geforderten Eigenschaften wie Robustheit oder Nachvollziehbarkeit abhängen. Eine mögliche Herangehensweise wäre, wenn Modultests sich nur auf die empirische Prüfung der Modellreaktion auf festgelegte Inputs beziehen – etwa durch Vergleiche mit Referenzdatensätzen – ohne eine „richtige“ oder erwartete Ausgabe im klassischen Sinne zu definieren. Aufgrund dieser Nichtdeterministik können wiederholte Testläufe unter identischen Bedingungen nicht immer vollständig reproduzierbar sein. Daher könnte ein stärkerer Fokus auf statistische Auswertungen, mehrfache Testwiederholungen, oder metamorphische Tests nötig sein. Metamorphische Tests sind eine Teststrategie, die es ermöglicht, Systeme zu überprüfen, ohne für jede Eingabe eine exakt definierte Referenzausgabe vorzuschreiben. Stattdessen wird das Verhalten eines Systems unter gezielten Veränderungen oder Umformungen seiner Eingaben untersucht und anhand zuvor festgelegter Relationen zwischen den zugehörigen Ausgaben bewertet. Integrationstests könnten sich im KI-Kontext stärker auf den Datenfluss, die Schnittstellen und die Systemreaktion auf verschiedene Szenarien, wobei das Zusammenspiel zwischen Daten, Modell und Anwendung im Vordergrund steht, konzentrieren.

Hinzu kommt, dass herkömmliche Bibliotheken zur Programmierung von neuronalen Netzwerken (z.B. KERAS und PyTorch) vorerst nicht den Anforderungen nach KTA 3501 /KTA 10/ entsprechen, da der Funktionsumfang der Softwarebibliotheken nicht auf das für die jeweilige Funktion notwendige Maß begrenzt ist, ließe sich diese Problematik u. a. durch Reduzierung des Funktionsumfangs auf die notwendigen Funktionen lösen. Die

Erstellung einer Softwarebibliothek gemäß den Anforderungen aus der KTA 3501, sowie eine Prüfung der einzelnen Komponenten wäre somit gegeben. Während Komponententests deterministisch durchgeführt werden können, verliert sich diese Eindeutigkeit auf der Ebene trainierter Modelle. Hier ist das Verhalten nicht mehr durch expliziten Code, sondern durch gelernte Zusammenhänge bestimmt, deren innere Struktur oft nicht vollständig nachvollziehbar ist.

Neben Fragestellungen rund um das Testen und die Implementierung von KI-basierten Anwendungen erfordern die Bewertung und Qualifizierung von KI-basierten Anwendungen das Berücksichtigen von zusätzlichen Anforderungen. KI-basierte Systeme, insbesondere im sicherheitskritischen Bereich, müssen beispielsweise menschlich erklärbar sein und über robuste Datenmanagement-Strategien verfügen. Eine Integration von KI in sicherheitskritischen Bereichen bringt einige Herausforderungen mit sich, vor allem in Bezug auf die Zertifizierung und Einhaltung von Sicherheitsstandards. Dies führt dazu, dass Zertifizierungsstellen vor der Herausforderung stehen, die Systeme zu validieren, während die Hersteller vor der Herausforderung stehen dieselben Systeme nachweislich sicher zu gestalten. Ein wesentlicher Punkt der Herausforderung bei der Qualifikation von KI-basierten Systemen in der schwierigen Erklärbarkeit ihrer Entscheidungen liegt. Es ist schwierig, Entscheidungen der KI vollständig nachvollziehbar und kontrollierbar zu machen, da viele KI-basierten Systeme auf komplexen, datengetriebenen Modellen basieren, deren Verhalten nicht immer vorhersehbar ist. Hinzu kommt, dass der Begriff „Erklärbarkeit“ in der Forschung häufig unterschiedlich und oft vage verwendet wird. Dabei könnte das Thema adressiert werden, indem zwischen Modellen, die in sich transparent sind, und solchen, die durch nachträgliche Erklärungen interpretierbar gemacht werden, unterschieden wird. Jedoch können nachträgliche Erklärungen gelegentlich irreführend sein, da sie nicht unbedingt das tatsächliche Entscheidungsverfahren des Modells widerspiegeln, sondern nur plausible Gründe für eine Entscheidung angeben. Ein sich bei potentiellen Einsätzen von KI-basierten Systemen im sicherheitstechnischen Bereichen sich ergebenden Zielkonflikt ist die Erklärbarkeit zu verbessern, ohne dabei die Leistung des Modells einzuschränken /LIP 16/. Ein möglicher Ansatz zur Verbesserung der Erklärbarkeit ist das Heranziehen von standardisierten Metriken zur Leistungsbewertung, konsistenten Prüfverfahren, und einer präzisen Dokumentation von Hyperparametern /ISO 22b/. Zwar sind können diese Metriken Entscheidungsprozesse neuronaler Netze nicht direkt erklären, können aber dennoch eine einheitliche Grundlage für die Messung und Erfassung der Leistung von KI-basierten Systemen bringen, was wiederum zu einer verbesserten Vergleichbarkeit und den damit einhergehenden Erkenntnissen über die Funktionsweise des KI-basierten Systems führen könnte.

Zudem stellt die Gewährleistung der Robustheit und Fehlertoleranz eine große Herausforderung dar, da KI-basierte Systeme oft in dynamischen Umgebungen agieren und daher stabil auf verschiedene Veränderungen reagieren müssen. Viele KI-Modelle benötigen große Datenmengen und erhebliche Rechenleistung, was in Echtzeitanwendungen oder auf Geräten mit begrenzten Ressourcen sich als schwierig erweisen kann. Sensor- und Bilddaten können Herausforderungen an die Datenverarbeitung stellen, da sie in großen Mengen und in hoher Frequenz generiert werden. Obwohl datensparsame Methoden Effizienzvorteile bieten, können sie auch zu Informationsverlusten führen, die sich negativ auf die Genauigkeit des Modells auswirken, was sich wiederum auf das Vertrauen auf die Ergebnisse eines KI-basierten Systems auswirkt. Möglichkeiten um den Eintrag von Unsicherheiten in die Ergebnisse eines KI-basierten Systems wären unter anderem die Einhaltung hoher Datenqualitätsstandards, regelmäßige Qualitätsbewertungen und die Vermeidung von Verzerrungen (Bias) stellen anspruchsvolle Anforderungen dar /ISO 24/.

Das Zusammenspiel von diesen Faktoren, welche sich durch die Komplexität dieser Modelle ergeben, führt zu einer Notwendigkeit den Zertifizierungsprozess individuell an den Anwendungsfall anpassbar zu machen. Solche Herausforderungen führen zusätzlich zu einer Notwendigkeit des Einführens umfassender Tests und strenger Kontrollmechanismen für KI-basierte Systeme in sicherheitstechnischen Einsatzgebieten. Um die genannten Schwierigkeiten zu adressieren, könnte die vorgeschlagene strukturierte Bewertungsmethodik, die auf modulare Bewertungskriterien zurückgreift, verwendet werden. Ansätze wie die Nutzung von Datenbereinigungsalgorithmen und Bias-Vermeidungsmethoden könnten helfen, Datenqualität und -verfügbarkeit zu sichern. Zur Sicherstellung der Verifikation und Validierung könnten Simulationen und Stresstests, um das Verhalten der KI unter verschiedenen Bedingungen zu prüfen, getätigt werden. Die sich aus den Schwierigkeiten ergebenden Anforderungen müssten bei einer Zertifizierung eines KI-basierten Systems zusätzlich zu den, beispielsweise an die Sicherheit bei der Datenübertragung, einhergehen Anforderungen berücksichtigt werden.

Insgesamt bietet KI enormes Potenzial, Prozesse zu optimieren und neue Lösungen für komplexe Probleme zu entwickeln. Gleichzeitig ist jedoch eine sorgfältige Überwachung notwendig, um sicherzustellen, dass der Einsatz von KI verantwortungsvoll und sicher erfolgt und vertrauenswürdige Ergebnisse liefert.

5 Zusammenfassung und Ausblick

Dieses Auftragsforschungsvorhaben untersuchte die Einsatzmöglichkeiten und Anforderungen für KI-basierte Systeme in sicherheitskritischen Bereichen, insbesondere der Kerntechnik. In Abhängigkeit von der sicherheitstechnischen Rolle solcher Systeme können unterschiedliche Anforderungen an Zuverlässigkeit, Sicherheit und Nachvollziehbarkeit bestehen. Ziel des Vorhabens ist es daher, ein fundiertes Verständnis für die spezifischen Herausforderungen und Möglichkeiten von KI-Methoden, Klassifikationsansätzen und Qualifikationsprozessen in sicherheitsrelevanten Anwendungen zu entwickeln.

Ein Fokus lag auf der Identifikation und Analyse von Klassifikationsansätze, um einen strukturierten Überblick über die Einsatzmöglichkeiten von KI im sicherheitstechnischen Kontext zu erhalten. Im Rahmen dieser Arbeit wurden mehrere Klassifikationsansätze identifiziert. Darunter sind Klassifikationsansätze, welche auf KI-Modellen und -Methoden, einem risikobasierten Ansatz und einem Ansatz nach kerntechnischem Anwendungsgebiet, beruhen.

Ein weiterer Fokus lag auf der Identifizierung und Analyse existierender KI-basierter Anwendungen in der Kerntechnik und verwandten Industrien. Hierbei wurde eine umfangreiche Recherche in wissenschaftlichen Publikationen, Konferenzen, Konferenzbeiträgen sowie relevanten Normen durchgeführt. Dabei wurden sowohl Forschungsprojekte als auch praxisnahe Einsatzbeispiele betrachtet, um potenzielle Sicherheitsvorteile zu ermitteln und bestehende Risiken besser einschätzen zu können.

Zusätzlich wurde der Stand der Qualifikationsansätze für sicherheitskritische KI-basierte Anwendungen untersucht, um deren praktische Umsetzbarkeit und Regelkonformität zu gewährleisten. Im Rahmen dieser Untersuchung wurde ein modulares Bewertungsschema für Qualifikationsansätze von KI-basierten Systemen in sicherheitskritischen Bereichen entwickelt. Zusätzlich wurden die Anforderungen an konventionelle sicherheitskritische Software auf ihre Übertragbarkeit auf KI-basierte Anwendungen geprüft, um spezifische Anpassungen für KI-basierte Systeme mit sicherheitstechnischer Bedeutung abzuleiten. Außerdem wurden noch ungeklärte Aspekte einer Zertifizierung von KI-basierten Systemen beleuchtet.

Die Vorhabensergebnisse bieten eine umfassende Grundlage zur sicheren und regelkonformen Förderung des Einsatzes von KI in sicherheitskritischen Umgebungen.

Mögliche zukünftige Arbeiten könnten eine fortlaufende Anpassung und Erweiterung des Bewertungsschemas, um den kontinuierlichen Fortschritten im Bereich der KI-Technologien gerecht zu werden. Außerdem könnten zukünftige Forschungsarbeiten der Weiterentwicklung der standardisierten Qualifizierung und Zertifizierung von KI-basierten Systemen und ihre Anwendungssicherheit dienen.

Literaturverzeichnis

/ANT 22/ Antipov, M., Uvakin, M., Nikolaev, A., Makhin, I., Sotskov, E.: Opportunity Analysis of the Machine Learning Technologies Application in VVER RP Safety Asses. In: Leva, M. C., Patelli, E., Podofillini, L., Wilson, S. (Hrsg.): Book of Extended Abstracts for the 32nd European Safety and Reliability Conference. 32nd European Safety and Reliability Conference, S. 2867–2873, ISBN 978-981-18-5183-4, DOI 10.3850/978-981-18-5183-4_S24-01-302-cd, Research Publishing Services: Singapore, 2022.

/ARP 21/ Arpogaus, M., Voß, M., Sick, B., Nigge-Uricher, M., Dürr, O.: Probabilistic Short-Term Low-Voltage Load Forecasting using Bernstein-Polynomial Normalizing Flows. In: Institut für Angewandte Forschung - IAF: ICML 2021, Workshop Tackling Climate Change with Machine Learning, June 26, 2021, virtual. S. #20, 2021.

/BRE 18/ Bre, F., Gimenez, J. M., Fachinotti, V. D.: Prediction of wind pressure coefficients on building surfaces using artificial neural networks. Energy and Buildings, Bd. 158, S. 1429–1441, DOI 10.1016/j.enbuild.2017.11.045, 2018.

/CHO 22/ Cho, S. G., Lee, S. J.: A Deep Support Vector Data Description Model for Abnormality Detection and Application with Abnormality Classification in a Nuclear Power Plant. In: Leva, M. C., Patelli, E., Podofillini, L., Wilson, S. (Hrsg.): Book of Extended Abstracts for the 32nd European Safety and Reliability Conference. 32nd European Safety and Reliability Conference, S. 2889–2896, ISBN 978-981-18-5183-4, DOI 10.3850/978-981-18-5183-4_S24-04-398-cd, Research Publishing Services: Singapore, 2022.

/DEP 10/ Department of Defense: Joint Software Systems Safety Engineering Handbook. 2010.

/DIN 07/ DIN IEC 60880 „Kernkraftwerke – Leittechnische Systeme mit sicherheitstechnischer Bedeutung: Softwareaspekte für rechnerbasierte Systeme zur Realisierung von Funktionen der Kategorie A“

/DIN 10a/ DIN EN 62138: „Kernkraftwerke – Leittechnik für Systeme mit sicherheitstechnischer Bedeutung: Softwareaspekte für rechnerbasierte Systeme zur Realisierung von Funktionen der Kategorien B oder C“

/DIN 50/ DIN EN 50128 „Bahnanwendungen – Telekommunikationstechnik, Signaltechnik und Datenverarbeitungssysteme – Software für Eisenbahnsteuerungs- und Überwachungssysteme“

/DIN 61d/ DIN EN 61508 „Funktionale Sicherheit sicherheitsbezogener elektrischer/elektronischer/programmierbarer elektronischer Systeme“, vor allem Teil 3 „Anforderungen an Software“

/DIN 10/ DIN Deutsches Institut für Normung e.V.: Kernkraftwerke – Leittechnische Systeme mit sicherheitstechnischer Bedeutung – Softwareaspekte für rechnerbasierte Systeme zur Realisierung von Funktionen der Kategorie A. DIN EN 60880, Beuth Verlag: Berlin, 2010.

/DIN 11a/ DIN Deutsches Institut für Normung e.V.: Funktionale Sicherheit sicherheitsbezogener elektrischer/elektronischer/programmierbarer elektronischer Systeme. DIN EN 61508, Beuth Verlag: Berlin, 2011.

/DIN 11b/ DIN Deutsches Institut für Normung e.V.: Bahnanwendungen – Kommunikations-, Signal- und Datenverarbeitungssysteme – Software für Eisenbahnsteuerungs- und -schutzsysteme. DIN EN 50128, Beuth Verlag: Berlin, 2011.

/DIN 13/ DIN Deutsches Institut für Normung e.V.: Sicherheit von Maschinen – Sicherheitsbezogene Teile von Steuerungen – Teil 2: Validierung. DIN EN ISO 13849-2, Beuth Verlag: Berlin, 2013.

/DIN 20a/ DIN Deutsches Institut für Normung e.V.: Kernkraftwerke – Leittechnische Systeme mit sicherheitstechnischer Bedeutung – Softwareaspekte für rechnerbasierte Systeme zur Realisierung von Funktionen der Kategorien B oder C. DIN EN IEC 62138, Beuth Verlag: Berlin, 2020.

/DIN 20b/ DIN Deutsches Institut für Normung e.V.: Kernkraftwerke – Leittechnische Systeme mit sicherheitstechnischer Bedeutung – Softwareaspekte für rechnerbasierte Systeme zur Realisierung von Funktionen der Kategorie A. DIN EN IEC 60880, Beuth Verlag: Berlin, 2020.

/DIN 23/ DIN Deutsches Institut für Normung e.V.: Sicherheit von Maschinen – Sicherheitsbezogene Teile von Steuerungen – Teil 1: Allgemeine Gestaltungsleitsätze. DIN EN ISO 13849-1, Beuth Verlag: Berlin, 2023.

/DOD 10/ Department of Defense: Joint Software Systems Safety Engineering Handbook, 27. August 2010.

/EUR 19/ European Commission, Directorate-General for Communications Networks, Content and Technology: Ethics guidelines for trustworthy AI. DOI 10.2759/346720, Publications Office, 2019.

/EUR 24a/ Europäische Union: Official Journal of the European Union. Amt für Veröffentlichungen der Europäischen Union (Publications Office of the European Union): Luxemburg, 2024.

/EUR 24b/ Europäische Kommission: AI Act. Stand vom 22. Juli 2024, erreichbar unter https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai, abgerufen am 22. Juli 2024.

/FED 03/ Federal Aviation Administration: Software Approval Guidelines. Order 8110.49, 2003.

/FOW 71/ Fowler, T. B., Vondy, D. R., Cunningham, G. W.: Nuclear Reactor Core Analysis Code: CITATION. Oak Ridge National Laboratory, ORNL-TM-2496, Rev. 2 with Supplements 1, 2, and 3: Oak Ridge, Tennessee, 1971.

/GRS 16/ GRS, M. Jopen, C. Quester, S. Römer, D. Sommer, J. Stiller, B. Ulrich: Zuverlässigkeitsbewertung unter neuen Anforderungen an Sicherheitsleittechnik in Kernkraftwerken. Gesellschaft für Anlagen- und Reaktorsicherheit (GRS), 2016.

/GRS 25/ GRS, Manuel Obergfell, Michael Hage, Sören Johst, Jonathan Zert: Erweiterung des Quelltermprognosewerkzeugs FaSTPro zur Planung anlagenexterner Notfallmaßnahmen unter Berücksichtigung aller Radionuklidquellen an einem Kernkraftwerksstandort. Hrsg.: GRS, Gesellschaft für Anlagen- und Reaktorsicherheit (GRS) gGmbH, GRS-782, ISBN 978-3-910548-75-6, 2025.

/HAR 23/ Harleen Kaur Sandhu, Saran Srikanth Bodda, Abhinav Gupta: Digital Condition Monitoring of Nuclear Piping-Equipment Systems using Artificial Intelligence Technology. In: North Carolina State University: Probabilistic Safety Assessment and Management (PSAM) Topical. Virtual, 2023.

/HAW 21/ Hawkins, R., Paterson, C., Picardi, C., Jia, Y., Calinescu, R., Habli, I.: Guidance on the Assurance of Machine Learning in Autonomous Systems (AMLAS). University of York, DOI 10.48550/arXiv.2102.01564, arXiv, 2021.

/HIN 98/ Hines, J. W., Uhrig, R. E., Wrest, D. J.: Merging Process Models with Neural Networks for Nuclear Power Plant Fault detection and Isolation. Journal of Intelligent and Robotic Systems, Bd. 21, Nr. 2, S. 143–154, DOI 10.1023/A:1007981322574, 1998.

/HRN 19/ Hrnjica, B., Bonacci, O.: Lake Level Prediction using Feed Forward and Recurrent Neural Networks. Water Resources Management, Bd. 33, Nr. 7, S. 2471–2484, DOI 10.1007/s11269-019-02255-2, 2019.

/HYU 23a/ Hyun Seok Noh, Jung Soo Kim, Woo Sik Jung: Survey on the Use of Artificial Intelligence in Nuclear Power Plants. In: Sejong University: Probabilistic Safety Assessment and Management (PSAM) Topical. S. 1: Virtual, 2023.

/HYU 23b/ Hyun Seok Noh, Gee Man Lee, Jung Soo Kim, Woo Sik Jung: Application of Artificial Intelligence for Estimating Severe Accidents in Nuclear Power Plants Using Offsite Information. In: Sejong University: Probabilistic Safety Assessment and Management (PSAM) Topical. Virtual, 2023.

/IAE 94/ IAEA: Software Important to Safety in Nuclear Power Plants. International Atomic Energy Agency, Technical Report Series No. 367, 1994.

/IAE 99/ IAEA: Verification and Validation of Software Related to the Safety of Nuclear Power Plant Instrumentation and Control. International Atomic Energy Agency, Technical Report Series No. 384, 1999.

/IAE 00/ IAEA: Software for Computer Based Systems Important to Safety in Nuclear Power Plants. International Atomic Energy Agency, Safety Guide No. NS-G-1.1, 2000.

/IEC 23/ IEC International Electrotechnical Commission: Nuclear Facilities - Instrumentation and Control and Electrical Power Systems - Artifical Intelligence Applications. 63468:2023, VDE Verlag, 2023.

/IEE 12/ IEEE Computer Society: Standard for System and Software Verification and Validation. IEEE Std 1012-2012, 2012.

/IEE 22/ IEEE, Arif Wani, Mehmed M. Kantardzic, Vasile Palade, Daniel Neagu, Longzhi Yang, Kit Yan Chan (Hrsg.): Proceedings of the 21st IEEE International Conference on Machine Learning and Applications (ICMLA 2022). ISBN 978-1-6654-6283-9, DOI 10.1109/ICMLA55696.2022, IEEE: Nassau, Bahamas, 2022.

/ISO 18/ ISO: Straßenfahrzeuge – Funktionale Sicherheit. International Organization for Standardization, ISO 26262: Geneva, 2018.

/ISO 20/ ISO/IEC TR 24028:2020:2020, 2020.

/ISO 22a/ ISO International Organization for Standardization, IEC International Electrotechnical Commission: Information technology — Artificial intelligence — Artificial intelligence concepts and terminology. International Standard 22989:2022, 2022.

/ISO 22b/ ISO: Information technology — Artificial intelligence — Assessment of machine learning classification performance. International Organization for Standardization, ISO/IEC TS 4213: Geneva, 2022.

/ISO 24/ ISO: Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples. International Organization for Standardization, ISO/IEC 5259-1: Geneva, 2024.

/KAV 94/ Kavaklioglu, K., Upadhyaya, B. R.: Monitoring Feedwater Flow Rate and Component Thermal Performance of Pressurized Water Reactors by Means of Artificial Neural Networks. Nuclear Technology, Bd. 107, Nr. 1, S. 112–123, DOI 10.13182/NT94-A35003, 1994.

/KIM 93/ Kim, H. G., Chang, S. H., Lee, B. H.: Pressurized Water Reactor Core Parameter Prediction Using an Artificial Neural Network. Nuclear Science and Engineering, Bd. 113, Nr. 1, S. 70–76, DOI 10.13182/NSE93-A23994, 1993.

/KTA 10/ KTA: Rechnergestützte Systeme in sicherheitsgerichteten Systemen. Kerntechnischer Ausschuss, KTA 3501: Cologne, 2010.

/KTA 17/ KTA: Leittechnische Systeme mit sicherheitstechnischer Bedeutung in Kernkraftwerken. Kerntechnischer Ausschuss, KTA 3503: Cologne, 2017.

/KTA 18/ KTA: Anforderungen an die Qualifikation von Software in leittechnischen Systemen mit sicherheitstechnischer Bedeutung. Kerntechnischer Ausschuss, KTA 3505: Cologne, 2018.

/KTA 20/ KTA: Allgemeine Anforderungen an die Qualitätssicherung. Kerntechnischer Ausschuss, KTA 1401: Cologne, 2020.

/LI 22a/ Li, X., Cheng, K., Huang, T., Tan, S.: Research on false alarm detection algorithm of nuclear power system based on BERT-SAE-iForest combined algorithm. Annals of Nuclear Energy, Bd. 170, S. 108985, DOI 10.1016/j.anucene.2022.108985, 2022.

/LI 22b/ Li, X., Huang, T., Cheng, K., Qiu, Z., Sichao, T.: Research on anomaly detection method of nuclear power plant operation state based on unsupervised deep generative model. Annals of Nuclear Energy, Bd. 167, S. 108785, DOI 10.1016/j.anucene.2021.108785, 2022.

/LIP 16/ Lipton, Z. C.: The Mythos of Model Interpretability. DOI 10.48550/arXiv.1606.03490, arXiv, 2016.

/LYE 22/ Lye, A., Prinja, N., Patelli, E.: Probabilistic Artificial Intelligence Prediction of Material Properties for Nuclear Reactor Designs. In: Leva, M. C., Patelli, E., Podofillini, L., Wilson, S. (Hrsg.): Book of Extended Abstracts for the 32nd European Safety and Reliability Conference. 32nd European Safety and Reliability Conference, S. 2874–2881, ISBN 978-981-18-5183-4, DOI 10.3850/978-981-18-5183-4_S24-02-306-cd, Research Publishing Services: Singapore, 2022.

/MAT 21/ Matthew Homiack, Giovanni Facco, Michael Benson, Marjorie Erickson, Craig Harrington: Extremely Low Probability of Rupture Version 2 Probabilistic Fracture Mechanics Code. U.S. Nuclear Regulatory Commission, NUREG-2247: Washington, DC, 2021.

/MAT 23/ Matthew Homiack, Jonathan Chien, Adriana Lima, Jason Hales, Kyle Gamble, Kaeli Haverkamp, Tony Valdez, Virginia Wright: Engineering Applications of Artificial Intelligence and Machine Learning for Mechanical Systems and Component Performance. In: U.S. Nuclear Regulatory Commission (NRC): Proceedings of the Probabilistic Safety Assessment and Management Conference (PSAM16). 2023.

/MIQ 22/ Miqueles, L., Ahmed, I., Di Maio, F., Zio, E.: A Grey-Box Digital Twin-based Approach for Risk Monitoring of Nuclear Power Plants. In: Leva, M. C., Patelli, E., Podofillini, L., Wilson, S. (Hrsg.): Book of Extended Abstracts for the 32nd European Safety and Reliability Conference. 32nd European Safety and Reliability Conference, S. 2897–2904, ISBN 978-981-18-5183-4, DOI 10.3850/978-981-18-5183-4_S24-05-579-cd, Research Publishing Services: Singapore, 2022.

/NAT 13/ National Aeronautics and Space Administration: Software Safety Standard. NASA-STD-8719.13C, 2013.

/NRC 07/ NRC: Review Process for Digital Instrumentation and Control Systems. U.S. Nuclear Regulatory Commission, NUREG-0800, Appendix 7.0-A, 2007.

/NUR 01/ NUR: Digital Systems Software Requirements Guidelines. U.S. Nuclear Regulatory Commission, NUREG/CR-6734, 2001.

/PAR 22/ Park, J., Kim, T., Seong, S., Koo, S.: Control automation in the heat-up mode of a nuclear power plant using reinforcement learning. Progress in Nuclear Energy, Bd. 145, S. 104107, DOI 10.1016/j.pnucene.2021.104107, 2022.

/RTC 11/ RTCA: Software Considerations in Airborne Systems and Equipment Certification. DO-178C, 2011.

/UHR 89a/ Uhrig, R. E. (Hrsg.): Use of neurals networks in nuclear power plant diagnostics. Oak Ridge National Lab., TN (USA) Tennessee Univ., Knoxville, TN (USA). Dept. of Nuclear Engineering, 1989.

/UHR 89b/ Uhrig, R. L., Guo, Z.: Use of Neural Networks in Expert Systems for Nuclear Power Plant Diagnostics. In: Proceedings of Applications of Artificial Intelligence VII, SPIE Technical Symposium on Aerospace Sensing. Orlando, Florida, 1989.

/UHR 93/ Uhrig, R. E.: Use of neural networks in nuclear power plants. ISA Transactions, Bd. 32, Nr. 2, S. 139–145, DOI 10.1016/0019-0578(93)90036-V, 1993.

/WU 21/ Wu, X., El-Shamouty, M., Wagner, P.: Zuverlässige KI. White Paper. Hrsg.: Fraunhofer IPA, DOI 10.24406/publica-fhg-300847, Fraunhofer-Gesellschaft, 2021.

Abbildungsverzeichnis

Tabellenverzeichnis

A Anhang

A.1.1 KI-basierte Anwendungen in der Kerntechnik und in Bereichen mit sicherheitstechnischer Bedeutung - Zusammenstellung der recherchierten Arbeitsergebnisse

A.1.2 Use of Autoassociative Neural Networks for Signal Validation /HIN 98/

Die Publikation stellt die Ergebnisse der Anwendung von Autoassoziativen Neuronalen Netzen (AANN) zur Onlineüberwachung von Messumformern in Kernkraftwerken vor. In einem autoassoziativen neuronalen Netz werden die Ausgangsdaten so trainiert, dass sie die Eingangsdaten über einen definierten bzw. geeigneten Wertebereich nachbilden.

Das Ziel der Onlineüberwachung von Messumformern ist es, ein Driften von Messumformern von ihren kalibrierten Referenzwerten bzw. Ausfälle von Messumformern zu erkennen, so dass die Messumformer rechtzeitig neukalibriert bzw. ausgetauscht werden, um so eine Ausbreitung fehlerhafter Messsignale in der Anlage zu verhindern. Das in /HIN 98/ vorgestellte Überwachungssystem für Messumformer besteht aus einem AANN-Modul einem Entscheidungsmodul, einem Korrekturmodul und einem Trainingsanpassungsmodul.

Das AANN-Modul ist als fünfschichtiges neuronales Netz realisiert. Mehrere miteinander korrelierte Anlagenmessdaten aus Sensoren bilden die Eingänge des AANN-Moduls. Das AANN-Modul wird anhand eines robusten Trainingsverfahrens dahingehend trainiert, auf Basis von Informationen, die mit einem bestimmten Sensor korreliert sind, den Messwert dieses Sensors zu schätzen. Bei entsprechendem erfolgreichem Training bleibt der geschätzte Messwert des Sensors (Ausgang des AANN-Moduls) unverändert, wenn der Messwert des Sensors durch Rauschen oder Fehler verzerrt wurde. Diese Eigenschaft ermöglicht es dem AANN-Modul, Sensorabweichungen oder -ausfälle zu erkennen, indem es den Messwert des Sensors am Eingang des AANN-Moduls mit dem entsprechenden geschätzten Messwert des Sensors am Ausgang des AANN-Moduls vergleicht. Die Differenz zwischen geschätztem und tatsächlichem Messwert wird als Fehler oder Residuum bezeichnet. Das Residuum hat in der Regel einen Mittelwert von Null mit einer Varianz, die in der Größenordnung des Rauschanteils im Sensorsignal liegt. Wenn ein Sensor fehlerhaft ist, ändert sich der Mittelwert oder die Varianz des zugehörigen Residuums. Dies kann anhand statistischer Entscheidungslogik ausgewertet werden.

Bei dem vorgestellten Überwachungssystem erfolgt diese Auswertung mittels eines dem AANN-Modul nachgeordneten Entscheidungsmoduls, welches nach dem Sequential Ratio Probability Test (SPRT)-Prinzip realisiert ist. Abhängig des ermittelten Residuums gibt das SPRT-Modul den Zustand des Sensors (korrekt, fehlerhaft, ausgefallen) an. Bei korrekt arbeitendem Sensor hat das Residuum einen zeitlichen Mittelwert von ungefähr Null haben und eine Varianz, die mit der des Sensors vergleichbar ist. Wenn ein Sensor driftet, verschiebt sich der Mittelwert der Residuen. Aufgrund der Verschiebung des Mittelwerts erhöht sich das Wahrscheinlichkeitsverhältnis. Dieses Verhältnis ist ein Maß für die Wahrscheinlichkeit, dass das Residuum gleich Null ist, im Vergleich zu der Wahrscheinlichkeit, dass das Residuum einen anderen vordefinierten Wert hat. Wenn das Wahrscheinlichkeitsverhältnis über einen bestimmten vordefinierten Grenzwert ansteigt, ist es wahrscheinlicher, dass die Residuen aus der fehlerhaften Verteilung als aus der fehlerfreien Verteilung der Sensordaten stammen. Der betreffende Sensor wird als fehlerhaft gekennzeichnet. Anderenfalls (Wahrscheinlichkeitsverhältnis unterhalb des Schwellenwerts) wird vom SPRT-Modul der Sensor als korrekt arbeitend gekennzeichnet.

Mit dem Korrekturmodul werden fehlerhaft erkannte Sensordaten durch ihre geschätzten Werte (best estimate) ersetzt. Die geschätzten Sensordaten werden dann als Eingabe in Steuerungssystemen, zur Anzeige (z. B für das Wartenpersonal) oder für andere Aufgaben verwendet. Die Eingabedaten des fehlerhaften Sensors in das AANN-Modul werden auch durch ihre geschätzten Werte ersetzt, damit die Überwachung der anderen Sensoren durch das AANN nicht beeinträchtigt wird. Der tatsächliche Sensormesswert wird wieder in das AANN-Modul eingespeist, sobald der Fehler behoben ist.

Das Trainingsanpassungsmodul dient dazu, das AANN-Modul an veränderten Betriebsbedingungen der überwachten Sensoren durch erneutes Trainieren anzupassen.

Das vorgestellte Überwachungssystem wurde in Matlab Simulink realisiert (siehe Abb. 5.1).

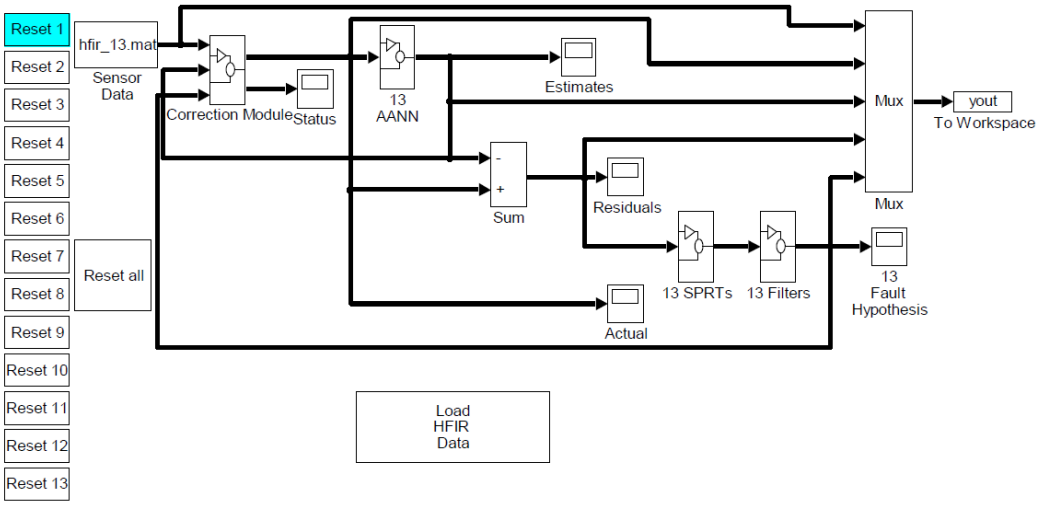

Abb. 5.1 Beispielsrealisierung des AANN/SPRT-Überwachungssystems

Hier: Zur Überwachung von 13 Sensoren im Hochflussisotopenreaktor (HFIR) des Oak Ridge National Laboratory. /HIN 98/

Anwendungsbeispiele des in Matlab Simulink implementierten AANN/SPRT-Überwachungssystems für Sensoren im Kernkraftwerk Crystal River #3 und im Hochflussisotopenreaktor (HFIR) des Oak Ridge National Laboratory sind in /HIN 98/ aufgezeigt. In Crystal River wurde das System zur Überwachung von 22 Sensoren entwickelt, darunter auch Temperatur- und Druckmessungen im Reaktorkühlkreislauf. Im HFIR wurde u. a. die Reaktoreintrittstemperaturmessung überwacht. In beiden gezeigten Fällen sind Drifts der Messumformer in einem Bereich zwischen 0,2 % und 3% ihres gesamten Messbereichs mit dem realisierten AANN/SPRT-Überwachungssystem detektierbar.

A.1.3 Control automation in the heat-up mode of a nuclear power plant using reinforcement learning /PAR 22/

Die Publikation untersucht den Einsatz von Deep Reinforcement Learning (DRL) zur Automatisierung der Aufheizphase in Kernkraftwerken, die bislang überwiegend manuell gesteuert wird. Ziel der Studie ist es, menschliche Fehler zu reduzieren und die Bediener zu entlasten, indem ein autonomer Agent in Echtzeit Druck- und Temperaturparameter überwacht und anpasst.

Dazu wurde der Compact Nuclear Simulator (CNS), ein Modell zur Simulation der Reaktorsysteme, speziell für das DRL-Training angepasst. Die Studie zeigt, dass der Agent in der Lage ist, die Aufheizoperation stabil und effektiv zu steuern.

In dieser Publikation wir der Asynchronous Advantage Actor-Critic (A3C)-Algorithmus verwendet, welcher zur Klasse der Deep Reinforcement Learning (DRL) Algorithmen gehört. Der A3C-Algorithmus kombiniert die Vorteile von Policy-basierter und Wertbasierter Optimierung und ermöglicht so ein schnelles und stabiles Lernen. Im A3C-Modell agieren „Actor“ und „Critic“ als zwei unterschiedliche Komponenten eines neuronalen Netzwerks. Der Actor wählt eine Aktion basierend auf dem aktuellen Zustand, die für den weiteren Verlauf am vorteilhaftesten erscheint. Seine Entscheidung basiert auf einer Wahrscheinlichkeitsverteilung, die durch das neuronale Netzwerk erlernt wird. Der Critic hingegen schätzt den erwarteten Wert des aktuellen Zustands ein und berechnet, wie viel besser oder schlechter der Actor in diesem Moment hätte agieren können. Der so berechnete „Advantage" wird dem Actor als Rückmeldung gegeben und dient dazu, zukünftige Aktionsentscheidungen zu verbessern. A3C ist ein „asynchroner“ Algorithmus – mehrere Kopien des Agenten lernen und trainieren gleichzeitig in verschiedenen simulierten Umgebungen. Die Agenten interagieren parallel mit ihren jeweiligen Umgebungen, und ihre Erfahrungen werden regelmäßig in ein globales neuronales Netzwerk eingespeist. Dieses globale Netzwerk wird kontinuierlich durch die gesammelten Daten aller Agenten aktualisiert, sodass das Modell von einem größeren Spektrum an Situationen und Zuständen lernt und sich effizienter anpasst. Das A3C-Modell vereinfacht den komplexen Prozess der Entscheidungsfindung durch das Verwenden eines sogenannten „Policy-Gradient“-Ansatzes. Der Actor wird anhand der gewonnenen Advantage-Werte trainiert, um die Aktionen auszuwählen, die im Durchschnitt langfristig die höchste Belohnung einbringen. Gleichzeitig passt der Critic das neuronale Netz so an, dass es zuverlässigere Schätzungen des Belohnungswerts für zukünftige Zustände liefert. Im Kontext der Publikation bedeutet dies, dass der A3C-Agent aus einer Vielzahl von Betriebszuständen lernt, die unterschiedlichen Szenarien entsprechen, und mit zunehmender Erfahrung ein immer besseres Verständnis für die optimalen Steueraktionen in der Aufheizphase entwickelt. Die Parallelisierung und Synchronisation zwischen den Agenten ermöglicht, dass der Algorithmus effizient in Echtzeit arbeiten kann und gleichzeitig robust gegen Variationen und Unsicherheiten in den Eingabedaten bleibt.

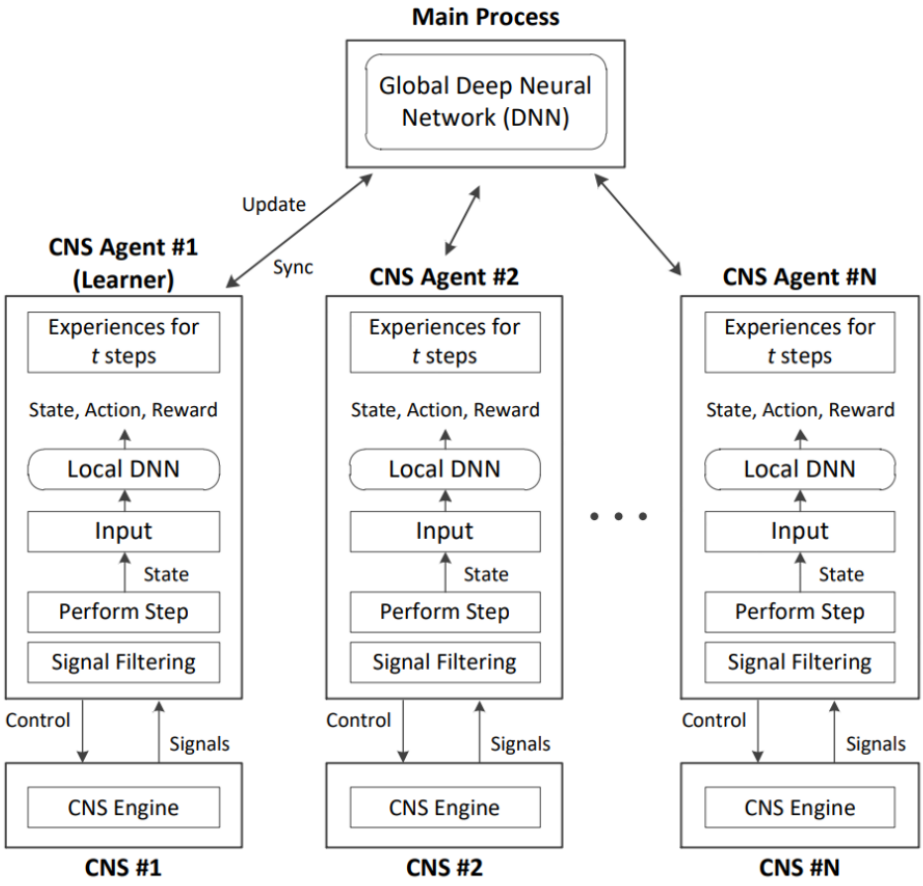

Abb. 5.2 Lernstruktur mehrerer Agenten. /HIN 98/

Die experimentellen Ergebnisse der Publikation zeigen, dass der DRL-Agent nach erfolgreichem Training in der Lage ist, die Aufheizphase stabil zu automatisieren und Parameter wie Druck und Wasserstand im optimalen Bereich zu halten. Der Agent erzielte hohe Leistungen, die mit denen eines erfahrenen Bedieners vergleichbar sind, und zeigte verbesserte Reaktionszeiten auf Veränderungen im Systemzustand. Die Studie belegt, dass DRL-basierte Steuerungssysteme in der Lage sind, sicherheitskritische Aufgaben zuverlässig zu übernehmen. Zukünftige Forschungen könnten auf die Automatisierung weiterer Betriebsphasen abzielen und die Anwendung von DRL in komplexeren Kontrollszenarien vertiefen.

A.1.4 USE OF EXPERT SYSTEMS IN NUCLEAR POWER PLANTS /UHR 93/

Im Kapitel 1 der Publikation wird zunächst erläutert, dass KI-basierte Systeme sowohl zum Einsatz auf der Warte als auch für Planungs-, Wartungs- und Revisionsarbeiten einschließlich BE-Handhabung und zur Fehlerdiagnose von Komponenten eingesetzt

werden können. Zum Zeitpunkt der Veröffentlichung wurden laut /UHR 93/ 287 KI-basierte Systeme identifiziert, welche in der kommerziellen elektrischen Energieerzeugung verwendet werden, 145 davon waren in den USA, 71 in Japan, 29 in Frankreich und 42 Systeme in der übrigen Welt eingesetzt.

Die Anwendungsfelder dieser in der kommerziellen elektrischen Energieerzeugung eingesetzten KI-basierten Systemen werden genannt. Hierzu gehören folgende Bereiche:

- Decision Support Systems
- Real-Time diagnostic systems
- Maintenance applications
- Plant Management
- Control
- Engineering Tools
- Plant Design
- Other: Capturing Human expertise, plant design, emergency response, cognitive model development

Im Kapitel 2 der Publikation folgt eine übergeordnete Unterteilung von KI-basierten Systemen in den folgenden sechs Anwendungskategorien:

- Monitoring Systems:

Sie dienen der Datenerfassung und Analyse über einen bestimmten Zeitraum hinweg. Die gesammelten Werte werden mit Erwartungen/erwartetem Verhalten verglichen, und wenn Diskrepanzen identifiziert werden, generiert das Expertensystem Empfehlungen und/oder benachrichtigt den Betreiber.

- Control Systems

Sie sind Monitoring Systems bei denen aufgrund einer festgestellten Abweichung zum erwarteten Verhalten eine Maßnahme (z. B. Öffnen eines Ventils, Einschalten einer Heizung usw.) ergriffen wird.

- Configuring Systems

Sie behandeln Probleme, bei denen eine endliche Menge von Komponenten in einem von mehreren möglichen Mustern konfiguriert werden soll. Das klassische Beispiel in dieser Kategorie ist z. B. XCON, ein Expertensystem, das von einem großen Computerhersteller eingesetzt wird, um seine Geräte in einer optimalen Konfiguration zu konfigurieren, die mit den Benutzerspezifikationen übereinstimmt.

- Scheduling and Planning Systems

Sie dienen der Koordinierung der Fähigkeiten oder Ressourceninnerhalb einer Organisation, um die Produktion und/oder die Effizienz zu steigern. Der Unterschied zwischen Scheduling und Planning systems besteht darin, dass die zur Verfügung stehenden Ressourcen für eine Aufgabe in Planungssystemen nicht immer bekannt sind.

- Diagnostic Systems

Sie dienen dazu, Daten zu analysieren und zu beobachten und die Analyseergebnisse auf eine Reihe von Problemen zu übertragen. Sobald die Probleme klar identifiziert sind, empfiehlt das KI-basierte System eine Lösung, die auf seiner Wissensbasis und auf den anderen Informationen, die es erwerben kann, basiert.

IM Kapitel 3 der Publikation werden Beispiele von eingesetzten KI-basierten Systemen in Kernkraftwerken angegeben und deren Funktionen jeweils kurz erläutert, wobei der Fokus auf in US-amerikanischen Anlagen eingesetzten Systemen liegt. In Tab. 5.1 sind diese KI-basierten Systeme zusammenfassend dargestellt.

Tab. 5.1: Beispiele von in US-amerikanischen Anlagen eingesetzten KI-basierten Systemen.

KI-Systembezeichnung	**Aufgabe/Funktion des KI-basierten Systems**
Reactor Emergency Alarm Level Monitor (REALM)	REALM (Reactor Emergency Alarm Level Monitor) soll bei einem nuklearen Zwischenfall anhand verfügbarer (teils unvollständiger) Daten schnell die Einstufung des Ereignisses als ungewöhnliches Ereignis, Alarm, örtlichen Notfall oder allgemeinen Notfall unterstützen. Es gleicht beobachtete Symptome mit möglichen Ereignisszenarien in der Wissensbasis ab.
Computerized Tracking System for Emergency Operating Procedure	Die Notfallmaßnahmen sind in etwa 250 Prozeduren geschrieben. Das

KI-Systembezeichnung	Aufgabe/Funktion des KI-basierten Systems
	KI-basierte System sucht nach **Muster-übereinstimmungen** zwischen den diese Prozeduren zugrundeliegenden Voraussetzungen und den Betriebsbedingungen, um dann die zu ergreifenden Maßnahmen zu empfehlen. Es ist als Onlinesystem realisiert, welches keine Eingaben von den Bedienern erfordert. Erläuterungen für seine Empfehlungen werden dem Bediener angezeigt.
CLEO and CRAW	Es handelt sich um die "Klone" CLEO (Clone of Leo, ein Expertensystem für die Beladung des FFTF) und CRAW (Clone of Rawlins, ein Experte für die Diagnose von Brennstab-Hüllrohrversagen im FFTF) von Expertensystemen im Hanford Engineering Development Laboratory und im FFTF (Fast FLUX Test Reactor). CLEO ist ein Expertensystem, welches der Optimierung der Kernbeladung dient. Hierzu wird innerhalb von wenigen Sekunden – im Gegensatz zu herkömmlichen Methoden, die mehrere Tage/Wochen benötigen –, eine Liste der erforderlichen Ladevorgänge beim BE-Wechsel erstellt. CRAW dient der Erkennung von Anzeichen von Brennstoffversagen. Beide Systeme sind direkt auf kommerzielle Kernkraftwerke anwendbar.
INTELLIGENT EDDY CURRENT DATA ANALYZER	Expertensystem zur Analyse der Wirbelstromdaten (aus Wirbelstrommessungen zur zerstörungsfreien Prüfung) der Dampferzeugerheizrohre eines Kernkraftwerks
DIAGNOSIS OF MULTIPLE ALARMS	Das System ist mit der Meldeanlage der Anlage verbunden und verwendet eine Ereignisbaumanalyse, um Muster von Meldungen zu identifizieren, die mit bestimmten Anlagenzuständen zusammenhängen. Die Meldungsmuster wurden für Störungen mit Hilfe von Logikbäumen entwickelt. Jeder erkannte anomale Zustand wird dem Bediener angezeigt und das zu befolgende Verfahren dargestellt.

KI-Systembezeichnung	Aufgabe/Funktion des KI-basierten Systems
IMPROVING NUCLEAR EMERGENCY RESPONSE WITH AN EXPERT SYSTEM	Expertensystem, welches bei der Koordinierung der Notfallorganisation in einem Kernkraftwerk eingesetzt werden soll. Das System wurde auf Basis der New York State (NYS) Procedures für Notfallklassifizierung, dem NYS Radiological Emergency Preparedness Plan, und dem Wissen der Experten der NYS Radiological Emergency Preparedness Group und dem Office of Radiological Health and Chemistry der New York Power Authority Behörde entwickelt.
REACTOR SAFETY ASSESSMENT SYSTEM	Das System wurde für den Einsatz im NRC Operations Center in Bethesda Maryland, im Falle eines schweren Zwischenfalls in einem Kernkraftwerk entwickelt. Dieses Expertensystem liefert eine Lagebeurteilung, woraus anhand von Parameterdaten der Anlage Empfehlungen zur möglichen Verwendung durch das NRC ableitet werden. Es verwendet mehrere Regelbasen und anlagenspezifische Datendateien, um für mehrere lizenzierte US-Kernkraftwerke einsetzbar zu sein. Derzeit deckt es verschiedene Reaktorkategorien und mehrere Anlagen pro Kategorie ab, aber anlagenspezifische Daten sind bislang nur für das Kernkraftwerk Calvert Cliffs verfügbar.
RESIDUAL HEAT REMOVAL EXPERT SYSTEM	Expertensystem, welches zur Überwachung der Nachwärmeabfuhr während des Anlagenstillstands eingesetzt wird. Das System dient der Überwachung von Anlagendaten, der Früherkennung von abnormalen Bedingungen im Nachwärmeabfuhrsystem und der Ursachenanalyse bei Ausfällen des Nachwärmeabfuhrsystems.
HANDLING POTENTIALLY INVALID SENOR DATA	KI-Methoden bzw. KI-basierte Systeme zur Diagnose bei Verwendung potenziell ungültiger Sensordaten. Das entwickelte KI-basierte System kann Diagnosen durchführen, obwohl einige widersprüchliche Daten vorhanden sind.
HALDEN REACTOR PROJECT EXPERT SYSTEMS/DISKETT	Ein regelbasiertes Diagnosesystem zur Unterstützung der Betreiber bei der Analyse von Anlagenstörungen. Es ist ein

KI-Systembezeichnung	Aufgabe/Funktion des KI-basierten Systems
	Expertensystem, welches für die Prozesssteuerung dynamischer Phänomene ausgelegt ist. Es verwendet symbolische Beschreibungen des dynamischen Anlagenverhaltens, in denen charakteristische Änderungen wichtiger Parameter, während einer Transiente als "Fingerabdrücke in einer Wissensbasis gespeichert werden.
HALDEN REACTOR PROJECT EXPERT SYSTEMS/ EARLY FAULT DETECTION	Eine computergestützte Bedienerhilfe bei der Diagnose von Fehlern im Speisewassersystem. Es handelt sich um ein computergestütztes System zur Unterstützung des Bedieners bei der Fehlererkennung, bevor die Alarmgrenzen erreicht werden. Es werden kleine Änderungen der Prozessparameter erkannt, gemessen als Abweichungen zwischen berechneten Referenzmessungen aus mathematischen Modellen und den tatsächlichen Anlagenwerten gemessen und als normal oder abnormal gekennzeichnet.
HALDEN REACTOR PROJECT EXPERT SYSTEMS /COPMA	COPMA ist ein computerbasiertes Verfahren, das darauf abzielt, vorhandene, schriftliche Prozeduren in digitaler Form bereitzustellen und nutzbar zu machen.
ACCIDENT MANAGEMENT	Eine der wichtigsten potenziellen Anwendungen von Expertensystemen ist der Einsatz in der Störfallbehandlung, insbesondere bei seltenen Unfällen, die aus einer Kombination von seltenen Ereignissen resultieren und schwerwiegende Folgen haben können. Expertensysteme könnten die Bewertung von weltweiten Experten zum Zustand einer unfallbedingt isolierten Anlage jederzeit zur Verfügung stellen. Ein Expertensystem kann auch bei schweren Unfällen hilfreich sein. Es ist vernünftig, von Betreibern zu erwarten, mit allen Arten von Störfällen umzugehen, aber es ist vielleicht nicht zu erwarten, dass sie alle Arten von auslegungsüberschreitenden Störfällen behandeln, die über den Rahmen der meisten Schulungen hinausgehen. Expertensysteme können eine Methode sein, um sich auf Ereignisse mit

KI-Systembezeichnung	Aufgabe/Funktion des KI-basierten Systems
	geringer Wahrscheinlichkeit und hohem Schadenspotential vorzubereiten. Beispielsweise kann ein Expertensystem für die Bewertung des Sicherheitsbehälters verwendet werden. Derzeit gibt es nur eine begrenzte Anzahl von Experten, die in der Lage sind, den Zustand eines Containments unter Unfallbedingungen zu beurteilen.
OTHER EXPERT SYSTEMS	Andere Expertensysteme gibt es für verschiedene Bereiche der Kernkraftwerksbetreibung. Einige Systeme, die in den USA entwickelt werden, illustrieren die Vielseitigkeit ihrer Anwendungen, etwa in den Bereichen Ausfallplanung, Wirkungsgradverbesserung, Diagnose von Instrumenten und Ausrüstung, Signalvalidierung, Störfallanalyse, Speisewasserüberwachung, Management von radioaktiven Abfällen, Wasserkontrolle während des Betriebs, Echtzeit-Evakuierungsplanung und Beurteilung von Strahlenexposition in Echtzeit.

Im Kapitel 4 der Publikation werden mögliche Fragestellungen beim Einsatz bzw. Einführung von Expertensystemen in Kernkraftwerken aufgelistet. Hierzu zählen:

- Quantitative und objektive Bewertungsrichtlinien für Expertensysteme:

Neue Methoden und Tools einschließlich objektiver Kriterien quantitativer Art, können erforderlich sein zur Bewertung der Leistungsfähigkeit von Expertensystemen und deren Auswirkungen auf die menschliche Leistung.

- Verifizierung und Validierung (V&V)

In der konventionellen Softwareprogrammierung haben Verifizierung und Validierung wohlbekannte Bedeutungen; Verifizierung ist die Feststellung, dass die Software formal korrekt und in Übereinstimmung mit einer spezifizierten Software-Engineering-Methodik entwickelt wurde, und Validierung bedeutet die Demonstration, dass das fertige Programm die Funktionen innerhalb der Anforderungsspezifikationen erfüllt und für die beabsichtigten Zwecke geeignet ist.

Expertensysteme gehen, ähnlich zu anderen KI-basierten Systemen, über die Verfahren der konventionellen Softwareentwicklung hinaus und können so viele Zustände aufweisen, sodass erschöpfende Tests und andere V&V-Methoden nicht durchführbar sind. Die gilt für viele softwarebasierte Systemene, aber auch für Expertensysteme, insbesondere solche, die unter Unsicherheit oder mit unvollständigen Daten arbeiten. Daher sind **neue Ansätze für V&V für Expertensysteme** erforderlich. Ein Hauptproblem beim Einsatz von Expertensystemen in Kernkraftwerken wird die Angemessenheit der Validierung und Verifizierung dieser Systeme sein.

- User Acceptance

Ein Hauptanliegen der menschlichen Faktoren ist, dass das Expertensystem dem Benutzer Informationen auf einer verständlichen und nachvollziehbaren Weise präsentiert. Die Informationen müssen sich gut in die Perspektive der Benutzer einfügen, und die Art und Weise, wie die Informationen dargestellt werden, sollten den mentalen Modellen des Menschen entsprechen.

Ein weiteres Anliegen ist die Reaktion der Benutzer auf das Expertensystem. Akzeptieren sie das System und nutzen sie es bei Bedarf, und werden sie Vertrauen in die vom Expertensystem dargestellten Informationen haben? Oder wird der Benutzer zu sehr von der Anleitung eines Expertensystems abhängig sein und Hinweise ignorieren, die nicht mit den Empfehlungen des Expertensystems übereinstimmen? Dies sind wichtige Fragen, die zu klären sind.

Die Aufteilung der Aufgaben zwischen dem Expertensystem und dem Benutzer ist eine weitere wichtige Frage. Dem Menschen sollten nur die Funktionen zugewiesen werden, die er entsprechend seiner Fähigkeiten am besten beherrscht.

Expertensysteme sollten die Benutzer physisch und kognitiv entlasten und nicht überlasten. Das System sollte die menschliche Arbeit effizienter machen. Natürlich sollten die Benutzer in diese Analyse einbezogen werden

Im Kapitel 5 wird auf die Anwendung von neuronalen Netzen bei der Entwicklung von Diagnosetools für den Kernkraftwerksbetrieb eingegangen.

Gemäß /UHR89/ können neuronale Netze (NN) so konzipiert sein, dass sie Muster von mehreren vordefinierten Eingabetypen (z. B. verschiedene Fehler- oder Übergangszustände eines Kraftwerks) klassifizieren oder um je nach Bedarf Kategorien oder Klassen von Systemzuständen zu erstellen, die von einem menschlichen Bediener interpretiert werden können. NN haben die Fähigkeit, in Echtzeit auf die sich ändernden Systemzustände zu reagieren, welche durch kontinuierliche Sensoreingaben bereitgestellt werden. Für komplexe Systemen mit vielen Sensoren und möglichen Fehlertypen (z. B. Kernkraftwerken) ist die Echtzeitreaktion eine schwierige Herausforderung sowohl für menschliche Bediener als auch für Expertensysteme. Dennoch, sobald ein neuronales Netz darauf trainiert wurde, die verschiedenen Bedingungen oder Zustände eines komplexen Systems zu erkennen, braucht es nur einen Zyklus des neuronalen Netzes, um eine bestimmte Bedingung oder einen bestimmten Zustand zu erkennen.

Neuronale Netze haben die Fähigkeit, selbst dann Muster zu erkennen, wenn die Informationen, aus denen diese Muster bestehen, verrauscht, spärlich oder unvollständig sind. Anders als die meisten Computerprogramme sind Hardware-Implementierungen – eine physikalische Realisierung des NN in dedizierter Hardware – von neuronalen Netzwerken sehr fehlertolerant, d. h. neuronale Netzsysteme können auch dann funktionieren, wenn einzelne Knoten des Netzes beschädigt sind. Die Verringerung der Systemleistung ist etwa proportional zu dem Teil des Netzes, der beschädigt ist.

Daher sind Systeme aus künstlichen neuronalen Netzen sehr vielversprechend für den Einsatz in Umgebungen, in denen eine robuste, fehlertolerante Mustererkennung in Echtzeit erforderlich ist, und in denen die eingehenden Daten verzerrt oder verrauscht sein können.

Jüngste Arbeiten an der Universität von Tennessee haben die Eignung neuronaler Netze zur Erkennung sechs verschiedener Transienten, die Teil der Simulation eines Dampferzeugers eines Kernkraftwerks sind, nachgewiesen /UHR 89a/, /UHR 89b/.

A.1.5 Pressurized Water Reactor Core Parameter Prediction Using an Artificial Neural Network /KIM 93/

Die Veröffentlichung befasst sich mit der Anwendung eines Neuronalen Netzwerk, das den Backpropagation-Algorithmus (Backpropagation Neural Network (BPN)) verwendet und zur Vorhersage wichtiger Parameter im Kern eines Druckwasserreaktors dient. Ziel

ist es, die Geschwindigkeit und Genauigkeit bei der Bestimmung von Kernparametern wie dem maximalen lokalen Leistungsfaktor P_{max} und dem effektiven Multiplikationsfaktor k_{eff} um die Brennstoffnutzung zu verbessern. Diese Parameter sind entscheidend für die Sicherheit und Wirtschaftlichkeit des Reaktorbetriebs, da eine präzise Vorhersage dazu beiträgt, optimale Betriebsbedingungen sicherzustellen.

Um diese Herausforderungen zu bewältigen, setzt die Studie auf ein BPN. Dieses Netzwerk wird darauf trainiert, bestimmte Eingabemuster von Kernbeladungsanordnungen zu verarbeiten und daraus P_{max} und k_{eff} vorherzusagen. Durch die Optimierung der Netzwerkstruktur, insbesondere der Anzahl der Neuronen in der versteckten Schicht (auf 500 festgelegt, um den Vorhersagefehler zu minimieren), erreicht das Modell eine Balance zwischen hoher Berechnungsgeschwindigkeit und präzisen Ergebnissen. Das Training des Netzwerks erfolgt durch die Anpassung der Gewichtungen, um den Vorhersagefehler zu minimieren. Die Studie beschreibt die Wahl der Trainingsparameter im Detail, darunter die Trainingsrate und Anpassungsmethoden, die für eine schnelle und stabile Konvergenz sorgen. Durch Experimente werden optimale Werte für verschiedene Parameter, einschließlich der Anzahl der Iterationen und der Anzahl versteckter Einheiten, ermittelt.

Die Ergebnisse zeigen, dass das neuronale Netzwerk die Kernparameter wesentlich schneller als herkömmliche numerische Verfahren berechnen kann und dabei eine zufriedenstellende Genauigkeit erreicht. So liegen 90 % der Vorhersagen für P_{max} innerhalb eines Fehlerbereichs von ±6,0 %, während 95 % der k_{eff} Vorhersagen eine Abweichung von lediglich ±0,3 % aufweisen. Eine vergleichende Analyse mit dem etablierten CITATION-Code – einem deterministischen Reaktorphysik-Code zur Berechnung stationärer Neutronenfluss- und Leistungsverteilung /FOW 71/ – zeigt, wie gut die Vorhersagen des BPN-Netzwerks mit Referenzwerten übereinstimmen und unterstreicht die Effizienzgewinne im Vergleich zu klassischen Methoden. Die Schlussfolgerungen der Studie heben hervor, dass das BPN-basierte Vorhersagesystem nicht nur effizienter ist, sondern potenziell einige der aufwändigeren traditionellen Methoden ersetzen könnte. Die Genauigkeit des Vorhersagesystems kann insbesondere bei stark veränderten Reaktorkonfigurationen mit deutlichen Leistungsabweichungen weiter optimiert werden. Für künftige Forschungen wird empfohlen, den Auswahlprozess der Trainingsmuster zu optimieren, um Vorhersagefehler zu minimieren, und das Modell so zu erweitern, dass es auch Kernparameter am Ende des Zyklus vorhersagen kann, da sich die aktuellen Ergebnisse auf den Beginn des Zyklus konzentrieren.

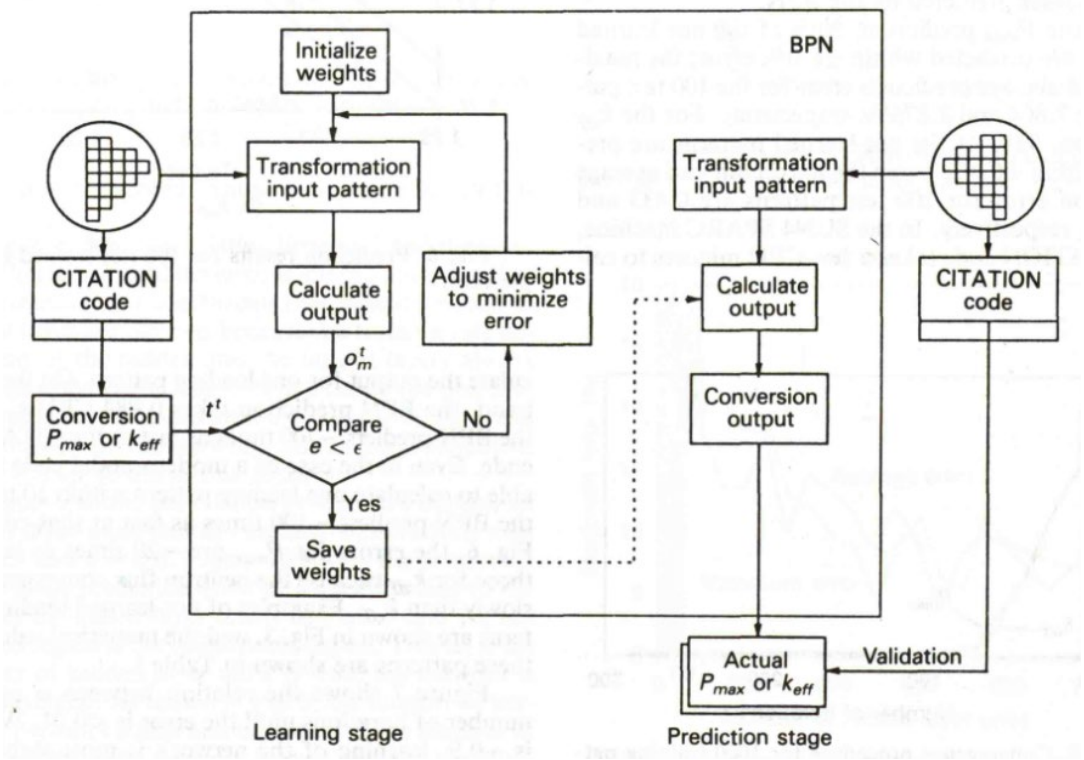

Abb. 5.3 Prozedurale Darstellung des Lern- und Vorhersagealgorithmus. /KIM 93/

A.1.6 Monitoring Feedwater Flow Rate Pressurized Water Reactors by Means of Artificial Neural Networks /KAV 94/

Diese Veröffentlichung befasst sich mit der Verwendung eines Multilayer Perceptron Networks, welches mit Hilfe eines Backpropagating (BP) Algorithmus trainiert wurde, für die Vorhersage der Degradierung von Messinstrumenten im kerntechnischen Bereich. Das verwendete Netzwerk besteht aus einem mehrschichtigen (3 Ebenen), vollständig verbundenen, heteroassoziativen Netzwerk. Die Studie beinhaltet außerdem eine Sensitivitätsstudie zur Analyse der Eingangsparameter, um herauszufinden, welcher den größten Einfluss auf den Kraftwerksparameter „Speisewassermassenstrom“ hat.

Als Ziel der Studie wurde versucht die Schädigung eines Venturi-Rohrs zur Durchflussmessung des Speisewassers im Laufe des Betriebs festzustellen, indem das neuronale Netz (NN) mit zusammenhängenden Eingangsparametern ausgesuchter Komponenten trainiert wurde. Die verwendeten Trainingsdaten basieren auf gemittelten Messwerten, welche kurz nach einer Revision erhalten wurden, um einen möglichst guten Zustand der Komponenten gewährleisten zu können. Damit sollte es möglich sein, im laufenden Betrieb die Schädigung von Komponenten festzustellen, da sich die Ausgabe des NN

durch die Degradierung der Komponenten ändern würde. Die für das Training des NN verwendeten Daten wurden „vorbehandelt“, was bedeutet, dass Werte mit starken Abweichungen herausgefiltert wurden und die Daten in einem Wertebereich von 0.1-0.9/0.2-0.8 transformiert wurden. Dadurch sollte das NN in der Lage sein über den Zustandsraum der Trainingsdaten hinweg zu extrapolieren.

Die Analysen wurde sowohl für die Blöcke 1 und 2 eines Westighouse 1140 MW DWR-Kraftwerks als auch die Loops eines Sechs-Loop 450 MW DWR-Kraftwerks durchgeführt. Die Sensitivitätsanalyse, basierend auf einem NN, ergab, dass folgende Eingangsparameter den größten Einfluss auf das Ergebnis des NN haben (Units – Blöcke des Kraftwerks Westinghouse 1140 MW DWR-Kraftwerks):

Tab. 5.2 Eingabedaten der Priorität nach sortiert (eigene Übersetzung). /KAV 94/

Reihenfolge	Block 1	Block 2
1	Temperatur des heißen Strangs	Temperatur des heißen Strangs
2	Temperatur des kalten Strangs	Speisewassertemperatur
3	Speisewassertemperatur	Druck im Reaktorkühlsystem
4	Druck im Dampferzeuger	Temperatur des kalten Strangs
5	Druck im Reaktorkühlsystem	Speisewasserdruck
6	Speisewasserdruck	Druck im Dampferzeuger
7	Abblasemassenstrom	Abblasemassenstrom

Das Ergebnis der Studie hat gezeigt, dass sich der Speisewassermassenstrom mit Hilfe eines NN abgeschätzt werden kann – die Ergebnisse zeigen eine gute Übereinstimmung für beide Blöcke der Westinghouse DWR-Anlage. Für den Fall des 450 MW Sechs-Loop DWR war die Abschätzung des NN teilweise gut. Jedoch wurde deutlich, dass eine adäquate Quantifizierung der Schädigung des Venturi-Rohrs nicht möglich ist. Ein weiteres NN, welches die durch ein numerisches Tool berechnete Speisewasservorwärmereffizienz abschätzen sollte, hat sehr gute Übereinstimmungen mit den Berechnungen gezeigt.

A.1.7 Research on anomaly detection method of nuclear power plant operation state based on unsupervised deep generative model /LI 22b/

Die Publikation untersucht eine Methode zur Erkennung von Anomalien im Betrieb von Kernkraftwerken. Das Besondere an dieser Methode ist, dass sie ohne die üblicherweise notwendigen Betriebsdaten jenseits des ungestörten Betriebs auskommt und stattdessen nur Daten aus ungestörten Betriebszuständen für das Training nutzt. Dadurch wird das Problem des Datenungleichgewichts adressiert. Die Autoren verwenden ein hybrides Modell, das einen Variational Autoencoder (VAE) und einen Isolation Forest (iForest) kombiniert. Die Herausforderung liegt darin, dass nur wenige Daten für Unfallbedingungen im nuklearen Bereich zur Verfügung stehen. Traditionelle Modelle können daher Schwierigkeiten haben, Anomalien zuverlässig zu erkennen, wenn nur begrenzt Daten für Unfallfälle verfügbar sind. Die Autoren schlagen daher ein Modell vor, das ausschließlich mit Daten für ungestörte Anlagenzustände arbeiten kann und somit das Problem des Datenungleichgewichts umgeht.

Der Variational Autoencoder (VAE) dient dazu, die hochdimensionalen thermohydraulischen Parameter in einen niedrig dimensionalen Raum zu kodieren. Während des normalen Betriebs werden diese kodierten Daten in der Nähe des Ursprungs gruppiert. Bei Unfällen hingegen zerstreuen sich die Datenpunkte weiter vom Ursprung, was eine Trennung zwischen normalen und anormalen Zuständen ermöglicht. Nachdem der VAE die Daten kodiert hat, wird der Isolation Forest (iForest) zur Klassifizierung verwendet. Dieses Modell identifiziert den Bereich ungestörter Betriebszustände basierend auf der Verteilung der VAE-Kodierungen und kennzeichnet Daten, die außerhalb dieses Bereichs liegen, als Anomalien. Das in der Publikation verwendete VAE-Modell verarbeitet eingehende Kraftwerksdaten und das iForest-Modell bewertet dann, ob die Daten innerhalb des ungestörten Bereichs liegen. Wenn Daten außerhalb des definierten Normalbereichs liegen, identifiziert das System diese als anormal und gibt Warnungen an das Kontrollsystem aus.

Das Modell wurde mit einem Simulatordatensatz getestet, der verschiedene Störfallbedingungen wie Kühlmittelverlust, Rohrbrüche im Dampferzeuger und Lecks im Druckhalter umfasst. Die Ergebnisse zeigten, dass das Modell Anomalien innerhalb von etwa 3 Millisekunden erkennen kann und somit den Anforderungen an die Echtzeitverarbeitung entspricht.

Das Modell ist zwar effektiv bei der Unterscheidung zwischen ungestörten und gestörten Betriebszuständen, jedoch fällt es schwer, zwischen den Ereignisarten zu differenzieren, da sich die kodierten Werte der Ereignisse teilweise überschneiden. Dies stellt eine Einschränkung für die Klassifikation der Unfälle dar, beeinträchtigt jedoch nicht das Hauptziel der Anomalieerkennung. Die Kombination von VAE und iForest bietet eine effiziente Methode zur Echtzeitanomalieerkennung in Kernkraftwerken. Die Fähigkeit des Modells, ohne markierte Daten oder Vorwissen über gestörte Anlagenzustände zu arbeiten, vereinfacht seine Implementierung in sicherheitskritischen Anwendungen, bei denen nur Daten aus dem ungestörten Betrieb verfügbar sind. Die Autoren betonen, dass das Modell gut für die allgemeine Anomalieerkennung geeignet ist, jedoch die Präzision zur Identifikation spezifischer abnormaler Ereignisse fehlt, wodurch es sich besser für Frühwarnungen als für genaue Diagnosen eignet.

A.1.8 Research on false alarm detection algorithm of nuclear power system based on BERT-SAE-iForest combined algorithm /LI 22a/

Die Publikation behandelt die Problematik der Fehlalarmerkennung im Betrieb von Kernkraftwerken. Das vorgeschlagene Modell kombiniert mehrere Algorithmen – Bidirectional encoder representations from transformers (BERT), Sparse Auto Encoder (SAE) und Isolation Forest (iForest) – um den Zustand des Kernkraftwerks zu überwachen und Fehlalarme von echten Anomalien zu unterscheiden.

Fehlalarme entstehen häufig durch kurzfristige Abweichungen in den Messdaten, die durch Umwelteinflüsse oder Schwankungen im System verursacht werden. Da Fehlalarme in Kernkraftwerken zu unnötigen Maßnahmen führen können, ist die Erkennung echter Anomalien entscheidend, um die Betriebssicherheit zu verbessern. Der Erkennungsalgorithmus besteht aus zwei Hauptkomponenten:

1. Verarbeitung transienter Betriebsparameter: Hierbei wird BERT eingesetzt, um Abweichungen zwischen den gemessenen und den theoretisch berechneten Werten zu erfassen. In Kombination mit einem Deep Neural Network (DNN) sorgt BERT für die Erfassung relevanter Datenmerkmale und beschleunigt die Analyse der Abweichungen.

2. Anomalie-Erkennung: Der SAE-Algorithmus kodiert die Daten und extrahiert wichtige Merkmale, die dann vom iForest-Algorithmus verarbeitet werden. Der iForest wird genutzt, um anhand der kodierten Werte den aktuellen Betriebszustand zu klassifizieren. Dabei wird dabei wird zwischen normalem Betrieb und Störfall unterschieden.

Sobald der BERT-basierte Algorithmus eine Abweichung entdeckt, bewertet das iForest-Modul, ob diese Abweichung auf eine Anomalie oder einen normalen Betriebszustand zurückzuführen ist. Wenn das System normal arbeitet, aber dennoch Abweichungen auftreten, identifiziert der Algorithmus die Situation als Fehlalarm. Der iForest-Algorithmus benötigt keine umfangreiche Trainingsdatenbank für Unfallzustände. Das Modell wurde mit simulierten Daten für verschiedene Szenarien getestet. In den Tests zeigte sich, dass der Algorithmus in der Lage ist, mit einer hohen Genauigkeit Fehlalarme zu identifizieren. Die Erkennungsgenauigkeit von ungestörten Betriebszuständen und Störfällen war besonders hoch, während die Fehlalarmgenauigkeit für bestimmte Szenarien (z.B. Kühlmittelverluststörfall; Dampferzeugerheizrohrleck) verbessert werden könnte. Die Kombination von SAE und iForest zeigte gute Ergebnisse bei der Trennung von ungestörten Betriebszuständen und Störfällen, insbesondere in den ersten 200 Sekunden nach einem Vorfall.

Laut den Autoren ermöglicht das Modell eine effiziente und schnelle Erkennung von Fehlalarmen, allerdings sinkt die Genauigkeit der Erkennung von Fehlalarmen nach einer Zeitspanne von ca. 200 Sekunden. Die Verbesserung der Erkennungsgenauigkeit über längere Zeiträume hinweg ist laut den Autoren ein möglicher Ansatz für künftige Forschung.

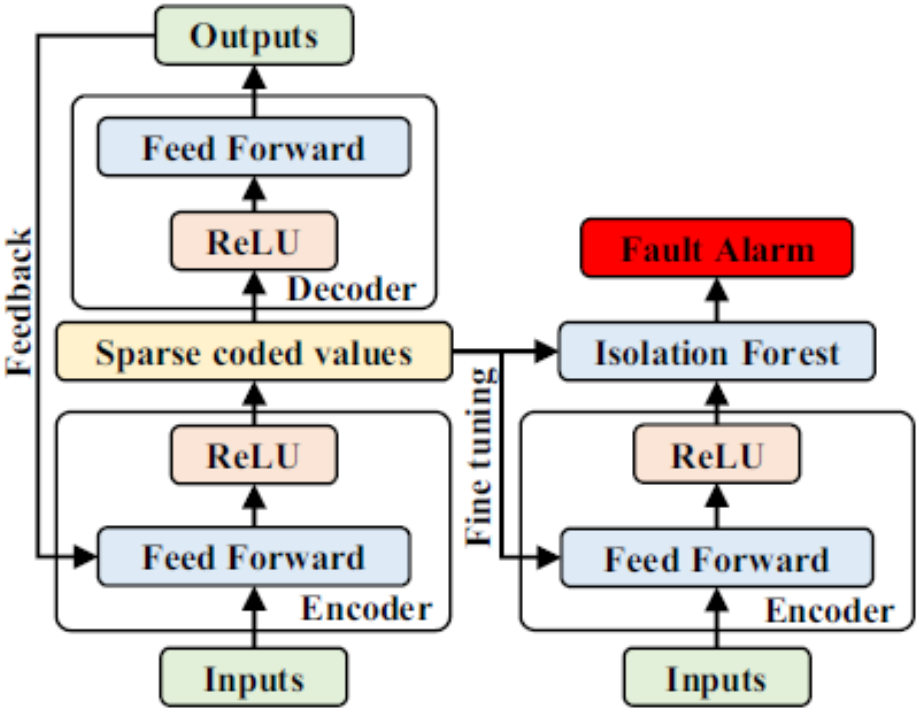

Abb. 5.4 Strukturdiagramm des Algorithmus zur Anomalieerkennung. /LI 22a/

A.1.9 Opportunity Analysis of the Machine Learning Technologies Application in VVER RP Safety Assessment /ANT 22/

Die Publikation untersucht den Einsatz von maschinellem Lernen (ML) zur Sicherheitsbewertung von WWER-Reaktoranlagen. Aufgrund der komplexen Anforderungen an Sicherheitsanalysen für Kernkraftwerke wird eine Methodik vorgestellt, die maschinelles Lernen verwendet, um den Aufwand und die Zeit für solche Berechnungen zu reduzieren. Die Autoren beschreiben ein Vorgehen zur Analyse von Eingabedaten für thermohydraulische Berechnungen und zeigen, wie ML-Modelle die Auswahl repräsentativer Unfall-Szenarien effizienter gestalten können.

Die Sicherheitsbewertung für Kernkraftwerke erfordert oft umfangreiche Berechnungen und die Betrachtung zahlreicher Störfall-Szenarien mit verschiedenen Startparametern. Die konservative Herangehensweise erfordert das Durchspielen der pessimistischsten Szenarien, was eine hohe Anzahl an Rechenläufen bedingt und ressourcenintensiv ist. Maschinelles Lernen wird als vielversprechender Ansatz gesehen, um die Datenmengen besser zu verarbeiten und potenzielle Risikoszenarien schneller zu identifizieren. Die Autoren haben sich für ein Modell namens Deep Neural Decision Forest (DNDF), das Entscheidungsbäume mit neuronalen Netzen kombiniert, entschieden.

Dies ermöglicht die automatische Identifikation der sicherheitskritischen Startzustände. Das Modell teilt die Eingabeparameter in Haupt- und Neben-Sets auf, um die Rechenkomplexität zu reduzieren und gleichzeitig die Relevanz der verschiedenen Parameter für die Analyse zu berücksichtigen. Für die Validierung der Methodik wurde ein Testmuster mit repräsentativen Startparametern und konservativen Annahmen zur Druck- und Temperaturverteilung des Reaktorkerns erstellt. Durch die Aufteilung der Eingabeparameter in Haupt- und Neben-Sets konnte der DNDF verschiedene Kombinationen durchspielen und die sicherheitsrelevantesten Szenarien bestimmen. Erste Ergebnisse zeigen, dass das Modell in der Lage ist, die von menschlichen Experten identifizierten, kritischen Zustände korrekt zu reproduzieren und somit eine deutliche Zeitersparnis im Vergleich zu traditionellen Verfahren ermöglicht. Die Validierung zeigt, dass das DNDF-Modell die Sicherheitsbewertung durch automatisierte, effiziente Berechnungen erheblich verbessern könnte. Für zukünftige Arbeiten schlagen die Autoren vor, das Modell durch zusätzliche Unfalltypen und Daten zu erweitern, um die Flexibilität und Anwendbarkeit auf verschiedene Reaktordesigns zu erhöhen. Das Ziel ist die Schaffung einer umfassenden Bibliothek, die auf allen bisher durchgeführten Sicherheitsanalysen basiert und die kontinuierlich verbessert wird.

A.1.10 Probabilistic Artificial Intelligence Prediction of Material Properties for Nuclear Reactor Designs /LYE 22/

Die Publikation befasst sich mit der Entwicklung eines probabilistischen KI-Ansatzes zur Vorhersage von Materialeigenschaften für den nuklearen Reaktordesign. Die Autoren stellen das Konzept PROMAP vor, das auf der Kombination von Künstlicher Intelligenz (KI) mit probabilistischen Methoden basiert. Ziel ist es, Materialeigenschaften unter Berücksichtigung von Unsicherheiten vorherzusagen und dabei die physikalischen Zusammenhänge (z.B. Temperatur, Legierungselemente) in den Eingangsvariablen realistisch abzubilden, sodass Korrelationen und Abhängigkeiten im Datensatz erhalten bleiben. Der Datensatz umfasst ursprünglich experimentelle Messdaten zu verschiedenen Stahllegierungen und soll durch künstliche erzeugte Datenpunkte erweitert werden. Konkret handelt es sich um Datenpunkte, welche mit Hilfe von stochastischen Verfahren erzeugt wurden, um experimentelle Daten zu erweitern und dabei Korrelationen zwischen Daten (z.B. Temperatur; Materialfestigkeit) beizubehalten.

PROMAP soll Unsicherheiten in einem Datensatz – bestehend aus experimentellen Messdaten zu Stahllegierungen – integrieren und eine robuste Schätzung von Materialeigenschaften ermöglichen. Dazu werden probabilistische Methoden genutzt, um ein synthetischen Datensatz zu erzeugen, das physikalische Abhängigkeiten zwischen Variablen beibehält. Dies erlaubt die Entwicklung zuverlässigerer Vorhersagen ohne aufwändige und kostspielige Experimente. Die Autoren führen aus, dass die Anwendung von KI im nuklearen Bereich bisher hinter anderen Sektoren, wie Luft- und Raumfahrt, zurückliegt und sehen großes Potenzial, diesen Rückstand aufzuholen. Die Grundlage der Analyse bildet eine Datenbank mit Materialeigenschaften von 58 verschiedenen Stahlsorten, die in einem früheren Projekt erhoben wurden. Für die Vorhersage der Kriech- und Zugfestigkeit dieser Materialien werden künstliche neuronale Netze (KNNs) genutzt, die durch ein synthetisches, probabilistisch erweitertes Dataset trainiert wurden. Die Datenerweiterung erfolgte durch die Berechnung der Pearson-Korrelationskoeffizienten zwischen den relevanten Merkmalen, sodass ein multivariates Gaußsches Verteilungsmuster konstruiert werden konnte, welches die physikalischen Zusammenhänge der Daten bewahrt. Daraus wurde ein Satz synthetischer Daten erzeugt, um das Training der KNNs zu verbessern. Da ein mathematisches Modell zur Beschreibung der Beziehung zwischen Eingabe- und Zielmerkmalen fehlt, wurden die KNNs als Ersatzmodelle verwendet. Verschiedene Netzwerkarchitekturen wurden trainiert, um Modellunsicherheiten zu berücksichtigen, und die Vorhersagen dieser Modelle mit den experimentellen Daten validiert. Die Robustheit der Modelle wurde mit Determinationskoeffizienten (R^2-Werten) im Bereich von 92,91 % bis 100 % bewertet, was die Effektivität des synthetischen Datensatzes für das Training der KNNs zeigt. Die Methode zur probabilistischen Vorhersage der Materialeigenschaften erfolgt durch eine adaptive Bayes'sche Modellauswahl (ABMS). Die ABMS kombiniert die Vorhersagen mehrerer KNN-Modelle und ermöglicht es, Konfidenzintervalle für die Vorhersagen zu berechnen. Dies führte zu verlässlichen Ergebnissen, bei denen alle experimentellen Datenpunkte innerhalb der berechneten Konfidenzintervalle lagen. Diese Methode bietet eine robuste Schätzung der Materialeigenschaften und erlaubt eine quantitative Bewertung der Unsicherheiten in den Vorhersagen. Die 95 %-Konfidenzintervalle umschließen alle experimentellen Datenpunkte. Für einige Zielmerkmale der Zugfestigkeit, wie die Streckgrenze, war der Konfidenzbereich jedoch breiter, was auf eine niedrigere Genauigkeit der Vorhersagen hinweist. Die Autoren sehen Potenzial, die Methode für künftige Designs, wie die Entwicklung neuer Materialien für Reaktoren der Generation IV, einzusetzen und dabei den Bedarf an teuren Experimenten zu reduzieren.

Die Publikation zeigt, dass die Kombination von KI mit probabilistischen Methoden eine vielversprechende Möglichkeit zur Vorhersage von Materialeigenschaften unter Berücksichtigung von Unsicherheiten bietet. Ein Nachteil des Ansatzes ist die Annahme normalverteilter Unsicherheiten in den Daten, die möglicherweise nicht in allen Fällen zutrifft. Künftige Arbeiten sollen diese Annahme verfeinern und das Modell auf umfangreichere Datensätze anwendet.

A.1.11 A Deep Support Vector Data Description Model for Abnormality Detection and Application with Abnormality Classification in a Nuclear Power Plant /CHO 22/

Die Publikation stellt ein KI-gestütztes Modell zur Anomalieerkennung in Kernkraftwerken vor. Das Modell verwendet Deep Support Vector Data Description (SVDD) zur Detektion von Zuständen außerhalb des ungestörten Anlagenbetriebs, basiert jedoch nur auf Daten aus dem ungestörten Anlagenbetrieb. Der vorherige Ansatz basiert auf einem Referenzmodell, welches aus einem 2D-CNN (zweidimensionales Convolutional Neural Network) besteht. Das Modell wurde zur Diagnose von Zuständen jenseits des ungestörten Betriebs verwendet, scheiterten jedoch an ausreichender Verfügbarkeit von Daten für diese Zustände.

Diese Publikation schlägt daher ein Modell vor, das allein auf normalen Betriebsdaten basiert und Zustände außerhalb des regulären Betriebs durch Abweichungen von diesen erkennt. Der Deep SVDD-Algorithmus ist ein Ein-Klassen-Klassifikationsmodell, das einen Hyperraum definiert, der die normalen Zustände umschließt. Veränderungen in den Daten werden durch einen Autoencoder und eine 2D-Convolutional Neural Network (CNN)-Struktur erfasst, welche die Distanz zu einem zentralen Punkt im Hyperraum berechnet. Eine erhöhte Distanz signalisiert einen abnormalen Zustand. Die Trainingsdaten wurden durch einen Reaktorsimulator erzeugt, der normale und 15 verschiedene Zustände außerhalb des regulären Betriebs abbildet. Die SVDD-Struktur wurde speziell darauf abgestimmt, die Grenze zwischen normal und abnormal zu bestimmen, ohne abnormale Daten direkt zu verwenden. Die verwendete Architektur umfasst einen Autoencoder, der die relevanten Datenmerkmale extrahiert, und einen SVDD, der die Distanzwerte zur Klassifikation verwendet. Die Daten wurden durch eine Min-Max-Normierung aufbereitet und in ein zweidimensionales Format umgewandelt.

Der Deep SVDD konnte die regulären Betriebszustände zu 100 % korrekt als solche klassifizieren und erzielte bei den meisten nicht regulären Betriebszuständen ebenfalls eine hohe Genauigkeit. Lediglich bei Zuständen, die dich nur wenig von normalen Zuständen unterscheiden, fiel die Erkennungsrate etwas niedriger aus. Das Modell wurde in ein Diagnosesystem integriert, das Anomalien frühzeitig erkennt und klassifiziert. Das System zeigte eine Verbesserung der Gesamtgenauigkeit um 0,2 % gegenüber dem Referenzmodell. Das Modell bewies, dass auch mit ausschließlich normalen Zustandsdaten eine robuste Erkennung und Klassifikation von Anomalien möglich ist, wodurch potenziell weniger Trainingsdaten benötigt werden.

Das vorgestellte Deep SVDD-Modell zeigt großes Potenzial für den Einsatz in der Anomalieerkennung in Kernkraftwerken, insbesondere dort, wo begrenzte Daten zu abnormalen Zuständen vorliegen. Künftige Arbeiten könnten das Modell weiter anpassen, um die Erkennungsgenauigkeit bei Zuständen, die sich nur geringfügig vom Normalzustand unterscheiden, zu erhöhen und die Integration in bestehende Überwachungs- und Diagnosesysteme zu erleichtern.

A.1.12 A Grey-Box Digital Twin-based Approach for Risk Monitoring of Nuclear Power Plants /MIQ 22/

Die Publikation präsentiert einen Ansatz zur Überwachung von Risiken in Kernkraftwerken durch den Einsatz eines „Grey-Box“ Digital Twins (DT). Digital Twins bieten eine vielversprechende Möglichkeit, die Zuverlässigkeit von Anlagen durch Echtzeitüberwachung und risikobasierte Wartung zu verbessern, stoßen jedoch bei sicherheitskritischen Systemen wie Kernkraftwerken auf Herausforderungen, da die Modelle oft als Black-Boxen schwer zu interpretieren sind. Die Autoren schlagen daher einen „Grey-Box“-Ansatz vor, der (physikalisch basierte) White-Box-Modelle mit (datengetrieben) Black-Box-Modellen kombiniert.

Digital Twins (DTs) können Echtzeitüberwachung ermöglichen und somit die Sicherheit von Kernkraftwerken verbessern. Üblicherweise basieren DTs jedoch entweder auf komplexen physikalischen Modellen oder rein datengetriebenen Ansätzen, was zu Kompromissen zwischen Genauigkeit und Interpretierbarkeit führt. Für Kernkraftwerke wird ein hybrider Ansatz bevorzugt, da er die Vorteile beider Modelle vereint und die Einhaltung regulatorischer Sicherheitsanforderungen erleichtert. DTs, die als Grey-Box-Modell auf-

gebaut sind, kombinieren dynamische physikalische Modelle mit datengetriebenen Komponenten zur Verbesserung der Vorhersagegenauigkeit und Risikoeinschätzung in Echtzeit. Das Grey-Box DT-Modell besteht aus drei Teilen: einem physikalischen Objekt (z. B. ein Reaktor, oder Kraftwerk), einem White-Box-Modell, das die Reaktorphysik und Thermohydraulik beschreibt, und einem Black-Box-Modell zur Bereitstellung von Fehlerkorrekturen für die White-Box-Ausgabe. Diese Komponenten sind in einer Echtzeitschleife miteinander verbunden, die Daten vom Kernkraftwerk in das DT-Modell einspeist und auf Basis der Vorhersagen Steuerungsmaßnahmen ergreift. In der Umsetzung wird ein Kleinmodul-Reaktor (SMDFR) als Fallstudie genutzt, um das Potentiel des Modells zur Echtzeitanalyse des Anlagenzustandes zu demonstrieren. Das Grey-Box-Modell ermöglicht eine präzisere Überwachung des Anlagenbetriebs durch den parallelen Einsatz eines White-Box-Modells (zur Reaktorüberwachung) und eines Black-Box-Modells (zur Fehlerkorrektur). Das White-Box-Modell basiert auf vereinfachten thermohydraulischen Gleichungen und überwacht kritische Parameter wie Kühlmittelfluss und Reaktortemperaturen. Das Black-Box-Modell, etwa ein künstliches neuronales Netz (KNN), korrigiert Abweichungen und erhöht die Genauigkeit der Vorhersagen durch Anpassungen auf Basis historischer Daten. Die daraus resultierenden Vorhersagen können, laut den Autoren, als Grundlage für risikoinformierte Entscheidungen dienen – eine vollumfängliche, probabilistische Risikobewertung wird im Rahmen der vorgestellten Arbeit nicht explizit durchgeführt.

Die Ergebnisse zeigen, dass der hybride Ansatz zuverlässig Risiken überwachen und die Betriebssicherheit von Kernkraftwerken durch präzisere Echtzeitvorhersagen verbessern kann. Das Modell kann durch die Kombination der Vorhersagen der beiden Modelle eine hohe Genauigkeit erzielen und ist somit für sicherheitskritische Anwendungen geeignet. Die Fähigkeit des Modells, sowohl physikalische als auch datengetriebene Komponenten zu nutzen, macht es besonders wertvoll für die Überwachung von SMDFR und ähnlichen Reaktortypen. Der vorgestellte „Grey-Box“ DT-Ansatz kombiniert die Vorteile physikalischer und datengetriebener Modelle und bietet damit ein Werkzeug für die Risikoüberwachung und -bewertung in Kernkraftwerken. Künftige Arbeiten sollen sich auf die Verbesserung der Fehlerkorrekturmethoden und die Reduzierung der Rechenzeit konzentrieren, um die Effizienz des Modells weiter zu steigern und es auf andere Reaktortypen anzuwenden.

A.1.13 Engineering Applications of Artificial Intelligence and Machine Learning for Mechanical Systems and Component Performance /MAT 23/

Diese Publikation untersucht die Anwendungsmöglichkeiten von künstlicher Intelligenz (KI) und maschinellem Lernen (ML) in der nuklearen Sicherheitsforschung, speziell im Bereich der mechanischen Systeme und Komponentenleistung. Das Hauptaugenmerk liegt auf der Arbeit der U.S. Nuclear Regulatory Commission (NRC), die drei zentrale Anwendungsfälle von KI und ML-Technologien beleuchtet.

Im ersten Anwendungsfall geht es um die Überwachung von Prozessgrößen eines Siedewasserreaktors (SWRs). Hierbei wird in der Publikation lediglich von „system performance“ gesprochen; es werden dabei keine konkreten Systemparameter oder Prozessgrößen erwähnt. Ziel ist es, Anomalien in den Betriebsdaten des Reaktors zu erkennen, die auf Funktionsstörungen von Komponenten hindeuten könnten. Diese Störungen können, wenn sie unentdeckt bleiben, die Sicherheit gefährden. Dazu wurde ein maschinelles Lernmodell entwickelt, das auf einem Long Short-Term Memory (LSTM) Autoencoder basiert, welcher speziell für das Erkennen von Mustern in Zeitreihendaten entwickelt wurde.

Der LSTM-Autoencoder besteht aus einem Encoder, der die Daten auf eine reduzierte Dimension komprimiert, und einem Decoder, der versucht, die Originaldaten daraus zu rekonstruieren. Während des Trainings lernt das Modell die typischen „normalen“ Betriebsbedingungen des Reaktors, sodass es beim Erkennen von Abweichungen eine Anomalie signalisiert. Diese Methode ermöglicht eine frühzeitige Erkennung von Problemen, etwa bei unkontrollierten Pumpenausfällen, und liefert so wertvolle Zeit für eine rechtzeitige Reaktion des Bedienpersonals.

Der zweite Anwendungsfall beschreibt die Erweiterung von Monte-Carlo-Simulationen zur Analyse der Integrität von Rohrleitungen in Druckwasserreaktoren (DWRs). Diese Simulationen werden in der probabilistischen Bruchmechanik verwendet, um die Integrität von Rohren, die unter Druck und Belastung stehen, zu bewerten. Ein zentrales Ziel war die Vorhersage der Rissausbreitung und die Analyse des „Leck vor Bruch“- Verhaltens. Dafür wurde das Extremely Low Probability of Rupture (xLPR)-Programm /MAT 21/ durch maschinelles Lernen ergänzt, wobei Random Forest Regression zur Sensitivitätsanalyse eingesetzt wurde. Diese Methode ermöglicht die Bestimmung der einflussreichsten Variablen für die Simulationsergebnisse, wie etwa Schweißspannungen und

Materialeigenschaften. Darüber hinaus wurde ein Ersatzmodell entwickelt, das Zeitreihendaten über die Rissausbreitung erzeugt und dadurch hilft, die Ressourcen effizienter zu nutzen, da nicht jedes Mal eine vollständige Simulation notwendig ist. Die Random Forest Methode erwies sich hierbei als besonders geeignet, da sie die komplexen Abhängigkeiten der Daten gut abbildet und so genauere und schnellere Analysen ermöglicht.

Im dritten Anwendungsfall wird untersucht, wie mit KI-Modellen Datenlücken in der Vorhersage der Materialverträglichkeit über lange Zeiträume überwunden werden können, besonders in korrosiven Umgebungen wie in Salzschmelzereaktoren. Da Langzeitdaten für diese Umgebung nur begrenzt verfügbar sind, wurde ein Stückweises Gauß-Prozess-Regressionmodell entwickelt, das für hohe Genauigkeit und Unsicherheitsabschätzung bekannt ist. Dieses Modell teilt die Daten mit einer k-means Clusteranalyse in Teilbereiche auf, was die Skalierbarkeit des Modells verbessert und es effizienter für große Datensätze macht. Innerhalb jedes Clusters wird ein Gauß-Prozess-Modell trainiert, das die thermodynamische Aktivität von Elementen wie Chrom und Eisen in Legierungen vorhersagen kann – wichtige Indikatoren für Korrosionsverhalten. Der Gauß-Prozess bietet dabei nicht nur exakte Vorhersagen, sondern schätzt auch die Unsicherheit der Ergebnisse ab, was besonders bei Sicherheitsfragen entscheidend ist.

So konnte das Modell mit hoher Genauigkeit Vorhersagen treffen, die auf wenigen experimentellen Daten basieren und eine zuverlässige Grundlage für die langfristige Materialbewertung liefern.

A.1.14 Survey on the Use of Artificial Intelligence in Nuclear Power Plants /HYU 23a/

Diese Publikation untersucht die Anwendungsmöglichkeiten von künstlicher Intelligenz (KI) in Kernkraftwerken.Sie bietet einen umfassenden Überblick über bestehende Forschungsarbeiten zur Integration von KI in den Kerntechnischen Sektor. Es wurden die bisherigen Erfolge und Grenzen der KI-basierten Anwendungen analysiert und klassifiziert. Die Klassifizierung der Forschung erfolgt nach zwei Hauptkriterien: den spezifischen Anwendungsfeldern der Kernenergie und den verwendeten Lernalgorithmen.

Die erste Dimension der Klassifizierung unterteilt die KI-basierten Anwendungen in der Kerntechnik in vier zentrale Anwendungsfelder:

- **Diagnose**: Hierbei geht es um die Erkennung von Defekten und Anomalien in Kraftwerkskomponenten. KI hilft dabei, Störungen in kritischen Systemen wie Pumpen oder Kühlleitungen frühzeitig zu identifizieren, was für die Vorbeugung von Ausfällen und die Planung von Wartungsmaßnahmen entscheidend sein kann.
- **Vorhersage**: In diesem Feld wird KI genutzt, um mögliche Betriebsstörungen und schwere Unfälle vorauszusagen. So können Muster in Sensordaten erkannt und potenzielle kritische Zustände vorzeitig prognostiziert werden, wodurch präventive Maßnahmen ergriffen werden können.
- **Reaktion**: KI unterstützt in Echtzeit die Risikobewertung und Notfallmaßnahmen im Fall eines Unfalls. Besonders digitale Zwillinge, virtuelle Abbilder des Kraftwerks, spielen eine große Rolle, indem sie laufend mit Echtzeitdaten gefüttert werden, um im Notfall die beste Reaktionsstrategie bereitzustellen zu können.
- **Prozessoptimierung**: Dieser Bereich umfasst die Effizienzsteigerung und Optimierung der Reaktorkern- und Anlagenplanung sowie die Prozesssteuerung. Durch KI können die Betriebsabläufe und der Energieertrag optimiert, menschliche Fehler minimiert und die Anordnung von Reaktorkomponenten verbessert werden.

Die zweite Dimension der Klassifizierung basiert auf den verschiedenen Typen von Lernalgorithmen, die in der KI-Forschung eingesetzt werden:

- **Überwachtes Lernen**: Diese Algorithmen lernen aus einem Trainingsdatensatz mit bekannten Ergebnissen und werden häufig für Diagnose und Vorhersage eingesetzt. Sie eignen sich besonders für Aufgaben wie die Detektion von Fehlern oder die Vorhersage von Systemzuständen anhand von Zeitreihendaten.
- **Unüberwachtes Lernen**: Diese Algorithmen erkennen Muster oder Gruppen in Daten ohne vorgegebene Ergebnisse und werden typischerweise für die Anomalieerkennung und die Analyse großer Datenmengen verwendet, wie sie in Kraftwerksbetrieben anfallen. Dies hilft beispielsweise, versteckte Zusammenhänge und Anomalien in den Betriebsdaten aufzudecken.
- **Verstärkendes Lernen**: Verstärkendes Lernen basiert auf Belohnungen, die der Algorithmus für bestimmte Aktionen in einer simulierten Umgebung erhält. Dies ist besonders nützlich für autonome Systeme, die eigenständig lernen sollen, wie sie auf

bestimmte Betriebsbedingungen reagieren können. Digitale Zwillinge nutzen verstärkendes Lernen, um Reaktionen in Notfällen zu simulieren und Optimierungen durchzuführen.

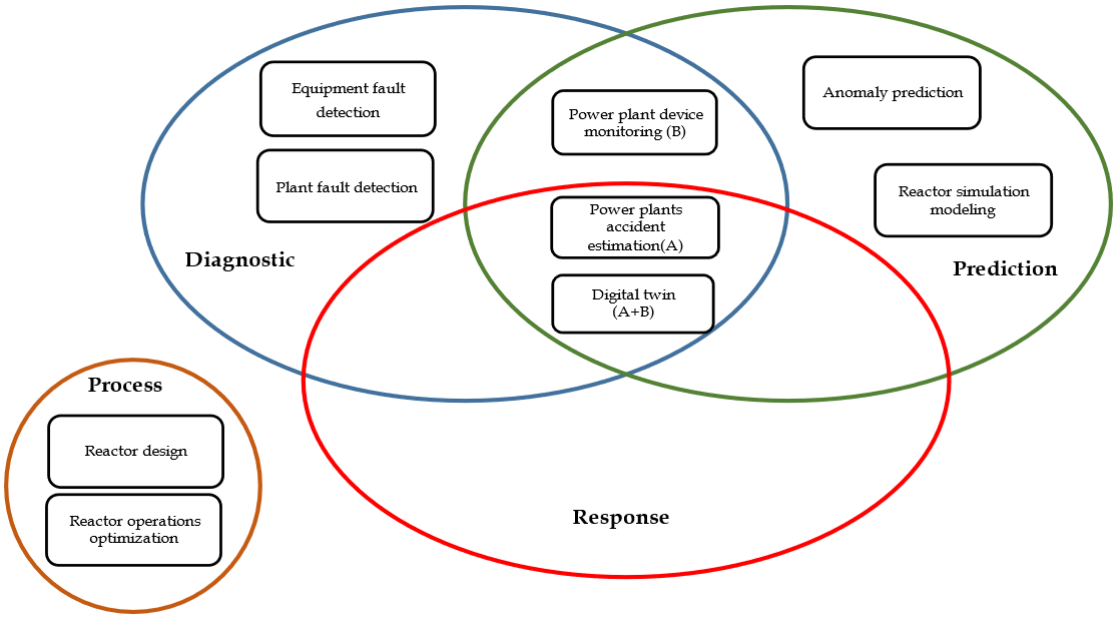

Abb. 5.5 Klassifizierung nach den spezifischen Anwendungsfeldern der Kernenergie. /HYU 23a/

Die Publikation bietet zudem einen umfassenden Überblick über nach Anwendungsgebieten klassifizierten Publikationen mit den zugehörigen KI-Modellen.

Tab. 5.3 KI-Anwendungsbeispiele für das Anwendungsfeld Diagnose /HYU 23a/

Classification	Purpose	AI
Equipment fault detection	Propose a machine learning model with a coherence index for sensor error detection in nuclear power plant emergencies	LSTM
	Estimate sensor error using physics-based machine learning methods that use measurements collected under plant conditions	SVM Random Forest
	Apply fault detection and diagnostics based on neural networks and k-nearest neighbor algorithms to pressurized water reactors	K-NN
	Apply for ensemble learning with bagging and boosting methods for failure diagnosis of a rotating machine in a nuclear power plant	Decision Tree
	Detect corrosion in nuclear power plant secondary piping using machine learning methods	Logistic Regression KNN SVM Random Forest
	Detect cracks in asphalt using convolutional neural network (CNN) learning methods	CNN
Plant fault detection	Use machine learning and deep learning to localize and estimate the mass of loose parts in a finite element analysis-based database	SVM, GP, KNN, CNN
	Propose an algorithm to improve the mass estimation performance of the metal fragment detection system for primary pipes in nuclear power plants	HMM KNN

Classification	Purpose	AI
	Combine OFM and SVM to develop an alert algorithm for loose parts	SVM
	Detect anomalies in pressure tube ultrasonic inspection data of CANDU-type reactor fuel using CNN models	CNN
	A CNN Model for nuclear abnormality diagnosis is proposed	CNN KNN
	Diagnose reactor transients using genetic algorithms	GA
	Diagnose multiple accidents in nuclear power plants using GNN	GNN

Tab. 5.4 KI-Anwendungsbeispiele für das Anwendungsfeld Vorhersage. /HYU 23a/

Classification	Purpose	AI
Anomaly prediction	Compare the trend predictions of nuclear parameters based on device control using different algorithms	LSTM RNN
	Predict an interpretable time series of NPP parameters in accident scenarios	Transformer
	Predict steam and water flow characteristics of NPP transients using SVM learning algorithms	LSTM SVM MLP
	Predict heat exchanger temperature changes using machine learning algorithms and a plotter database	XGB
	Predict core damage time in accident scenarios	Random Forest
	Predict steam and water flow in nuclear power plant transients	SVM
	Predict decomposition by neutron irradiation of reactor pressure vessels	SVM Random Forest XGB Decision Tree
Predicting severe accidents	Predict the critical heat flux for fuel bundles with non-uniform heat fluxes	RNN
	Predict failure thresholds for gate valves in nuclear power plants	RNN
	Predict accidents by changing parameters in dry wells	LSTM

Tab. 5.5 KI-Anwendungsbeispiele für das Anwendungsfeld Reaktion. /HYU 23a/

Classification	Purpose	AI
Accident estimation	Real-time accident source term estimation using internal plant data	Transformer
	Identify underwater source term	DGNN
	Accident prediction using external radiation dose in nuclear accidents	Decision Tree
Digital twins	Developing a digital twin model to predict power distribution in a nuclear reactor	SVM AE
	Suggests how to self-calibrate digital twin models	BPNN
	Automate nuclear power plant commissioning and output operations	A3C LSTM

Tab. 5.6 KI-Anwendungsbeispiele für das Anwendungsfeld Prozessoptimierung. /HYU 23a/

Classification	Purpose	AI
Nuclear Power Control	Quantitate HEP to reduce HRA in nuclear power plant startup and shutdown operations	BBN
	Propose method to extend PWR's core equilibrium cycle using a genetic algorithm	GA
	Develop advanced fuel management tools for heavy water reactors using AI technology	CARS-KNN
	Simple model predictive control for autonomous operation of nuclear power plants	SVR GRU LSTM
	Automatic recognition system for digitizing nuclear power plant documents	Cascade R-CNN
	Calculate the friction coefficient of sump filters and pipelines for long-term cooling of a pipe break scenario at the ACP100 nuclear power plant	Random Forest
Reactor Design	Predict key parameters for reactor core design	Decision Tree SVM Random Forest ANN
	Assess seismic effects on core design for advanced gas-cooled reactors	CNN DNN

A.1.14.1 Application of Artificial Intelligence for Estimating Severe Accidents in Nuclear Power Plants Using Offsite Information / HYU 23b/

Die Publikation untersucht die Anwendung von künstlicher Intelligenz (KI) zur Schätzung und Klassifizierung schwerer Unfälle in Kernkraftwerken, insbesondere in Situationen, in denen interne Daten aufgrund von Stromausfällen oder anderen Betriebsunterbrechungen nicht verfügbar sind. Als Beispiel dient der Fukushima-Unfall, bei dem ein Erdbeben und ein anschließender Tsunami zu einem vollständigen Stromausfall (Station Blackout, SBO) führten. Infolge dieser Katastrophe waren wichtige interne Daten des Kernkraftwerks nicht abrufbar.

Die Arbeit schlägt vor, stattdessen externe Strahlungsmessungen zu nutzen, um Aussagen über den Zustand der Anlage abzuleiten und so den Unfallhergang und Informationen über Schäden am Reaktorkern zu schätzen.

Das Hauptziel der Arbeit ist es, ein System zu entwickeln, welches anhand von radiologischen Messdaten, die außerhalb des Kraftwerks gesammelt wurden, Rückschlüsse auf den Typ und das Ausmaß von Unfällen innerhalb der Anlage ermöglicht. Die Methode bietet eine Notfalllösung für Szenarien, in denen interne Daten aufgrund von Kommunikationsausfällen nicht verfügbar sind. Solche Daten könnten Behörden und Betreibern als Grundlage für schnelle Entscheidungen und effektive Reaktionsstrategien dienen. Da reale Daten für solche schweren Unfälle selten sind, wurden radiologische Daten durch Simulationen mit verschiedenen Unfallanalyse-Codes generiert. Die Ergebnisse dieser Simulationen liefern die benötigten Trainingsdaten für die KI-Algorithmen.

RASCAL (Radiological Assessment System for Consequence Analysis): Ein von der US Nuclear Regulatory Commission entwickeltes Tool, das die Menge und Art der freigesetzten radioaktiven Stoffe berechnet. RASCAL generiert Freisetzungsmengen in 15-Minuten-Intervallen und berechnet die gesamte Strahlungsquelle, die dann für nachfolgende Simulationen genutzt wird.

MACCS (MELCOR Accident Consequence Code System): Dieses Modell ergänzt RASCAL, indem es die Verteilung der Nuklide in der Atmosphäre berechnet und die Strahlungsdosis entlang der Wolkenachse abschätzt. Es verwendet Umgebungs- und Wetterinformationen, um die Verteilung der Strahlung in der Umgebung des Kraftwerks zu simulieren.

MURCC (Multi-unit Radiological Consequence Calculator): Entwickelt an der Sejong Universität, berechnet dieses Modell die zweidimensionalen Konzentrationen von Nukliden an bestimmten Orten über der Erdoberfläche, basierend auf den von MACCS und RASCAL erzeugten Daten.

Die Autoren verglichen daraufhin verschiedene KI-Algorithmen zur Klassifikation und Vorhersage der Unfallart und des Schadensgrades.

XGBoost (Extreme Gradient Boosting) war der leistungsfähigste Algorithmus und erzielte die höchste Genauigkeit (97%) bei der Klassifizierung von Unfalltypen, insbesondere bei der Unterscheidung von Kühlmittelverlust (Loss of Coolant Accident, LOCA) und Ausfall der Notstromversorgung bei Anforderung (Station Blackout. SBO).

Deep Neural Networks (DNN) zeigten die beste Leistung bei der Klassifizierung des Grades der Kernschädigung, mit einer Genauigkeit von 92%.

Lineare Regression schnitt am schlechtesten ab, sowohl bei der Klassifikation der Unfalltypen (79%) als auch bei der Schätzung des Schadensgrades (67%).

Die Daten wurden in einem Verhältnis von 7:3 in Trainings- und Testdaten aufgeteilt, um die Genauigkeit und die Leistung der Algorithmen zu bewerten. Die Forscher berechneten den F-Score als zusätzliche Bewertungsmetrik, um die Klassifikationsleistung der Modelle zu prüfen. Der F-Score ist eine Metrik zur Bewertung der Genauigkeit von Klassifikationsmodellen bei unausgewogenen Verteilung inerhalb von Datensätzen. Die Ergebnisse zeigen, dass KI-Modelle effektiv zwischen verschiedenen Unfalltypen und Schädigungsgraden unterscheiden können. XGBoost erwies sich als besonders präzise bei der Identifizierung des Unfalltyps, während DNN die höchste Genauigkeit bei der Vorhersage des Schädigungsgrades des Reaktorkerns zeigte. Die Publikation zeigt, dass KI-Modelle wertvolle Informationen über den Zustand eines Kernkraftwerks liefern können, selbst wenn direkte Messdaten aus dem Inneren des Reaktors fehlen. Die Autoren empfehlen, die Methode in zukünftigen Forschungen auf Unfälle in Mehrblockanlagen auszuweiten, bei denen mehrere Blöcke gleichzeitig betroffen sind.

A.1.14.2 Digital Condition Monitoring of Nuclear Piping-Equipment Systems using Artificial Intelligence Technology /HAR 23/

Die Publikation stellt ein KI-basiertes Condition-Monitoring-Framework für nukleare Rohrleitungssysteme vor, dass auf der frühzeitigen Erkennung von Materialschädigung wie Korrosion und Erosion basiert. Ziel dieses Systems ist es, die Lebensdauer der Rohrleitungen in Kernkraftwerken zu verlängern und die Wartungskosten zu reduzieren, indem potenzielle Schwachstellen im Material erkannt werden, bevor es zu schwerwiegenden Schäden oder Leckagen kommt. Die Studie zielt darauf ab, sowohl die Genauigkeit als auch die Effektivität der Überwachung und Wartung solcher sicherheitskritischen Systeme zu verbessern.

Als Fallstudie wählten die Autoren das Z-Rohrleitungssystem des Experimental Breeder Reactor II (EBRII), das durch den hohen Schwingungsbelastungen aufgrund von Pumpenbetrieb ausgesetzt ist. Diese Schwingungen erhöhen das Risiko von Materialermüdung und Schädigungen. Sensoren wurden an kritischen Punkten des Systems, wie Rohrbögen und Düsen, angebracht, um die Schwingungs- und Belastungsdaten in Echtzeit zu erfassen. Die gesammelten Beschleunigungsdaten wurden anschließend mittels Power Spectral Density (PSD) analysiert, um frequenzspezifische Merkmale der Schädigungen zu identifizieren. Die Merkmalsextraktion spielt eine zentrale Rolle im Framework. Die Autoren vergleichen zwei Ansätze. Ein einfacherer Ansatz extrahiert lediglich die maximale PSD-Amplitude als degradationssensitiven Kennwert; ein umfassenderer Ansatz erstellt jedoch einen Vektor aus vier degradationssensitiven Größen, der mehr Informationen über die verschiedenen Schwingungsmodi des Systems liefert und eine detailliertere Analyse der Schädigungsmuster ermöglicht. Die Ergebnisse zeigen, dass der umfassendere Ansatz eine deutlich höhere Genauigkeit bei der Identifizierung der Orte und Schwere der Schädigungen erzielt. Während der einfache Ansatz eine Genauigkeit von 86 % erreichte, ermöglichte der umfassende Ansatz eine Genauigkeit von 97 %. Kernstück des Frameworks ist ein Multilayer Perceptron (MLP), ein künstliches neuronales Netzwerk, das auf den extrahierten Merkmalen trainiert wurde, um degradierte Stellen im Rohrleitungssystem zu identifizieren und die Schwere der Schädigung in Kategorien wie gering, moderat und stark zu klassifizieren. Diese tiefere Einteilung der Degradationsschwere ermöglicht es dem Modell, fundierte Wartungsempfehlungen zu geben, die auf spezifische Bedarfe und Risiken der Anlage abgestimmt sind. Die Autoren schlagen außerdem eine Strategiebewertung zur Reduzierung von Ermüdungsrisiken vor. Basierend auf den ASME-Richtlinien zur Materialermüdung berechnen die Autoren die maximal zulässigen Betriebsstunden für bestimmte Pumpgeschwindigkeiten, um die Lebensdauer der Komponenten zu verlängern. Beispielsweise wird für eine bestimmte Pumpgeschwindigkeit eine maximale Betriebszeit von 5,6 Stunden vorgeschlagen, um die Entstehung von Rissen und Leckagen zu vermeiden.

Die Ergebnisse zeigen, dass das vorgeschlagene Framework eine präzisere Vorhersage der Orte und der Schwere von Materialschädigungen ermöglicht. Das System bietet potenziell erhebliche Kosteneinsparungen, da es frühzeitig auf Materialschwächen hinweist und so gezielte und rechtzeitige Wartungsmaßnahmen ermöglicht. Zudem wird die Lebensdauer der Anlagen durch Empfehlungen zu sicheren Pumpgeschwindigkeiten und maximalen Betriebsstunden verlängert, was das Risiko von Ermüdungsschäden minimiert.

Künftige Arbeiten sollten sich auf die Anwendung des Frameworks auf weitere Systeme und Reaktortypen konzentrieren und die Effizienz der Sensorplatzierung und Datenverarbeitung optimieren.

A.1.15 Einsatz von KI-Methoden in Bereichen mit sicherheitstechnischer Bedeutung - Zusammenstellung der recherchierten Arbeitsergebnisse

A.1.16 Probabilistic Short-Term Low-Voltage Load Forecasting using Bernstein-Polynomial Normalizing Flows /ARP 21/

Die Publikation untersucht einen neuen Ansatz zur probabilistischen Vorhersage des kurzfristigen Stromverbrauchs im Niederspannungsnetz. Angesichts der zunehmenden Nutzung erneuerbarer Energien und der Elektrifizierung des Energiesystems ist eine präzisere Vorhersage des elektrischen Energiebedarfs notwendig, um die Netzstabilität zu gewährleisten. Die Autoren betonen die Bedeutung probabilistischer Lastprognosen, die Unsicherheiten besser abbilden können als einfache Punktprognosen. Solche Prognosen helfen dabei, die Energieversorgung zuverlässig zu planen und die Netzlast zu optimieren, was insbesondere im Kontext von Spitzenlastreduzierungen und Spannungsregelung wichtig ist. Im Kern des Ansatzes steht die Nutzung von Bernstein-Polynom-Normalizing-Flows (BNF) zur flexiblen Modellierung von Wahrscheinlichkeitsdichtefunktionen. Normalizing Flows (NF) sind Transformationen, die komplexe Verteilungen auf einfache Verteilungen wie die Normalverteilung abbilden können. Die Autoren verwenden eine Kombination von Transformationen, bei denen Bernstein-Polynome die zweite Transformation darstellen. Die Forscher führten eine empirische Studie mit Daten von 363 Stromkunden durch und verglichen ihren BNF-Ansatz mit herkömmlichen Methoden, wie dem Gaußschen Modell und dem Gaußschen Mischmodell (GMM). Die Daten stammten von einer irischen Energieregulierungsbehörde und wurden für die 24-Stunden-Vorhersage des Stromverbrauchs herangezogen. Neben dem BNF wurden auch neuronale Netz-Architekturen wie vollständig verbundene Netze und ein 1D-CNN für die Vorhersagen eingesetzt. Die Ergebnisse zeigten, dass der BNF-Ansatz sowohl in der Genauigkeit als auch in der Stabilität den traditionellen Modellen überlegen war. Der BNF lieferte verlässlichere Schätzungen und konnte die Unsicherheiten im Verbrauchsverhalten besser darstellen. Vor allem das 1D-CNN in Kombination mit dem BNF erzielte die besten Werte in mehreren Bewertungsmetriken, darunter der Continuous Ranked Probability Score (CRPS) und die Negative Logarithmic Likelihood (NLL).

Die Autoren führen den Erfolg des BNF auf die Fähigkeit zurück, die Verteilungen auf eine kontinuierliche Weise darzustellen und die Unsicherheiten detaillierter zu erfassen. Die Autoren schlagen vor, in zukünftigen Arbeiten die multivariate Natur der Prognosen noch stärker zu berücksichtigen und möglicherweise autoregressive Architekturen zu entwickeln.

A.1.17 Zusammenfassung von Vorträgen im Rahmen der „IEEE - International Conference on Machine Learning and Applications“, Nassau, 2022, /IEE 22/

Structural health and intelligent monitoring of wind turbine blades with a motorized telescope:

In diesem Vortrag wurde eine Methode für eine vorausschauendes Wartungssystem für die Oberflächeninspektion von Windturbinenblättern vorgestellt. Dieses System basiert auf ML. Konkret wurden Convolutional Neural Networks (CNNs) eingesetzt, um Turbinen und ihre Blätter sowie die Oberflächenfehler, die an ihnen auftreten können, zu erkennen und zu klassifizieren. Das System besteht aus einer mobilen Anwendung, welche ein Teleskop nutzt, um Bilder mit einer gewissen Präzision aufzunehmen. Zudem wird einem Rechenknoten, der für die Verarbeitung der aufgenommenen Bilder verantwortlich ist, und einer motorisierten Halterung, die die Bewegung des Teleskops ermöglicht verwendet.

CANBERT: A Language-based Intrusion Detection Model for In-vehicle Networks:

Controller Area Networks (CAN) sind ein Standardmittel, ein serielles Bussystem, für die Kommunikation zwischen elektronischen Steuergeräten in Fahrzeugen ohne eine zentrale Recheneinheit oder eine komplexe Punkt-zu-Punkt Verkabelung. Trotz der Vorteile, die fahrzeuginterne Netzwerke bieten, sind CAN-Busse anfällig für Netzwerkangriffe wie Replay-, Fuzzing- und Denial-of-Service-Angriffe. Die zunehmende Verbreitung von mit dem Internet verbundenen Fahrzeugen macht es außerdem erforderlich, ein robustes und sicheres Fahrzeugnetzwerksystem aufzubauen. Auf Deep Learning basierende Sprachmodelle, wie z. B. BERT-Modelle (Bidirectional Encoder Representations from Transformers), haben bewiesen, dass sie bemerkenswerte Ergebnisse für natürlichsprachliche Aufgaben liefern. BERT-Modelle bieten ein tiefes Verständnis der zugrunde liegenden Semantik in Textdaten. In dieser Arbeit wurde CANBERT vorgestellt, ein sprachbasiertes Modell zur Erkennung von Eindringlingen in CAN-Busse.

So wurde die Leistungsfähigkeit von Transformatormodellen genutzt, um die Erkennung von bösartigen Angriffen auf CAN-Busse zu ermöglichen.

AI privacy preserving robots working in a smart sensor environment:

Um die Sicherheit zu erhöhen, ergänzen autonome Fahrerlose Transportsysteme (FTS) in Lagern ihre lokalen Sensoren mit Daten von Umgebungssensoren, um die Erkennung von Menschen zu verbessern, die sich nicht im Blickfeld der FTS befinden. Das Problem bei vielen bestehenden Ansätzen ist, dass sie alle Umgebungsdaten im Klartext übertragen, was zu Problemen beim Datenschutz und der erforderlichen Übertragungsbandbreite führt. In diesem datenschutzfreundlichen Ansatz des maschinellen Lernens führen die Umgebungssensoren eine Vorverarbeitung der Daten durch und senden nur abstrahierte Merkmale an die FTS, wodurch sowohl das Datenschutz- als auch das Bandbreitenproblem gelöst werden.

Are Post-Hoc Explanation Methods for Prostate Lesion Detection Effective for Radiology End Use?

Deep Learning hat bei medizinischen Aufgaben wie der Klassifizierung von Krebs und der Erkennung von Läsionen beeindruckende Leistungen gezeigt. Trotz dieser beeindruckenden Leistung handelt es sich um einen Black-Box-Algorithmus, der daher schwer zu interpretieren ist. Die Interpretation ist besonders in risikoreichen Bereichen wie der Medizin wichtig. In letzter Zeit wurden verschiedene Methoden zur Interpretation von Deep-Learning-Algorithmen vorgeschlagen. Es gibt jedoch nur wenige Studien, die diese Erklärungsmethoden in klinischen Umgebungen wie der Radiologie bewerten. Zu diesem Zweck wurde in dieser Arbeit eine Pilotstudie durchgeführt, die die Effektivität von Erklärungsmethoden für die Radiologie untersucht. Es wurde untersucht, ob Erklärungsmethoden die Diagnoseleistung verbessern und welche Methode von Radiologen bevorzugt wird.

Zero Day Threat Detection Using Metric Learning Autoencoders:

Die Verbreitung von Zero-Day-Bedrohungen (ZDTs) in den Netzwerken von Unternehmen ist immens kostspielig und erfordert neuartige Methoden, um den Datenverkehr in großem Umfang auf bösartiges Verhalten zu untersuchen. In dieser Arbeit wurde demonstriert wie eine eingeführte Methodik, die einen Dual-Autoencoder-Ansatz zur Identifizierung von ZDTs in der Netzwerkfluss-Telemetrie verwendet, behilflich sein kann.

Der Ansatz ermöglicht es bislang unbekannte Bedrohungen auch ohne gelabelte Trainingsdaten zu erkennen, reduziert die Anzahl falsch-positiver Alarme und verbessert die Unterscheidbarkeit zwischen normalem und bösartigem Verhalten.

Feature Reduction Method Comparison Towards Explainability and Efficiency in Cybersecurity Intrusion Detection Systems:

Im Bereich der Cybersicherheit erkennen und verhindern Angriffserkennungssysteme (IDS) Angriffe auf der Grundlage gesammelter Computer- und Netzwerkdaten. In der jüngsten Forschung wurden IDS-Modelle mit Methoden des maschinellen Lernens und des Deep Learning wie Random Forest und DNNs erstellt. Mit Hilfe der Merkmalsauswahl (FS) können schnellere, besser interpretierbare und genauere Modelle erstellt werden. Es wurden drei verschiedene FS-Techniken untersucht: RF-Informationsgewinn (RF-IG), Auswahl von Korrelationsmerkmalen mit dem Bat-Algorithmus (CFS-BA) und CFS mit dem Aquila-Optimierer (CFS-AO).

Autoencoder Feature Residuals for Network Intrusion Detection: Unsupervised Pre-training for Improved Performance:

In diesem Beitrag wurde ein nicht-pervisierter Pre-Training-Schritt verwendet, um die Vorteile von Autoencoder-Merkmalsresten zu nutzen. Es wurde gezeigt, dass Residuen von Autoencoder-Merkmalen anstelle oder zusätzlich zu einem ursprünglichen Merkmalssatz als Eingabe für einen Klassifikator eines neuronalen Netzwerks verwendet werden können, um die Klassifizierungsleistung zu verbessern.

Knowledge Guided Two-player Reinforcement Learning for Cyber Attacks and Defenses:

Übungen sind ein wichtiger Weg, um die technischen Kapazitäten von Organisationen im Umgang mit Cyber-Bedrohungen zu verstehen. Die aus diesen Übungen gewonnenen Informationen führen oft zur Entdeckung von bisher unbekannten Methoden zur Ausnutzung von Schwachstellen in einer Organisation. Dies führt oft zu besseren Verteidigungsmechanismen, die bisher unbekannte Angriffe abwehren können. Dank der jüngsten Entwicklungen bei Simulationsplattformen für Cyber-Bedrohungen konnte eine Übungsumgebung für die Verteidigung generiert und auf Reinforcement Learning (RL) basierende autonome Agenten trainiert werden, um das durch die simulierte Umgebung beschriebene System anzugreifen.

In diesem Beitrag wurde eine spielbasierte RL-Umgebung für zwei Spieler, die gleichzeitig die Leistung der Agenten von Angreifer und Verteidiger verbessert gezeigt.

Exposing Surveillance Detection Routes via Reinforcement Learning, Attack Graphs, and Cyber Terrain:

Diese Arbeit erweitert frühere Bemühungen zur Entwicklung von Reinforcement-Learning-Methoden (RL-Methoden) für die Pfadanalyse in Unternehmensnetzwerken. Diese Arbeit konzentriert sich auf den Aufbau von Überwachungserkennungsrouten, bei denen sich die Routen auf die Erkundung der Netzwerkdienste konzentrieren, während versucht wird, Risiken zu umgehen. RL wird eingesetzt, um die Entwicklung dieser Routen zu unterstützen, indem ein Belohnungsmechanismus entwickelt wurde, der bei der Realisierung dieser Pfade helfen würde.

Joint Sub-component Level Segmentation and Classification for Anomaly Detection within Dual-Energy X-Ray Security Imagery:

Die Röntgenkontrolle von Gepäckstücken ist weit verbreitet und für die Aufrechterhaltung der Transportsicherheit zur Erkennung von Bedrohungen und Anomalien unerlässlich. Die automatische Erkennung von Anomalien in unübersichtlichen und komplexen elektronischen oder elektrischen Geräten, welche mit Hilfe von 2D-Röntgenbildern aufgenommen wurden, ist von größtem Interesse. Dies wurde durch die Einführung einer gemeinsamen Segmentierungs- und Klassifizierungsstrategie auf der Ebene der Teilkomponenten eines Objekts unter Verwendung einer DNN-Architektur eines CNNs in dieser vorgestellten Arbeit gelöst.

Explainable Unsupervised Multi-Sensor Industrial Anomaly Detection and Categorization:

Die Erkennung von Anomalien in Echtzeit ist bei industriellen Anwendungen von großer Bedeutung, um eine qualitativ hochwertige Produktion zu gewährleisten und Ausfallzeiten oder Systemfehler zu vermeiden. In diesem Beitrag wurde die Anwendung der Anomalieerkennung bei multivariaten Daten aus der Glasproduktion untersucht. Dafür wurden verschiedene unüberwachte multivariate Zeitreihen-Algorithmen zur Erkennung und Lokalisierung von Anomalien eingesetzt und verglichen, die bereits signifikante Ergebnisse in den aktuellen Datensätzen gezeigt haben.

Identifying Metering Hierarchies with Distance Correlation and Dominance Constraints:

In diesem Beitrag wurden Beobachtungen aus einer Reihe von intelligenten Zählern, deren Messwerte entweder vollständig oder teilweise zusammengefasst sind, betrachtet. Es wurde vorgeschlagen, diese wichtigen Metadaten durch eine neuartige Anpassung des Chow-Liu-Baum-Lernverfahrens zu rekonstruieren. Dieser Ansatz berücksichtigt das Vorwissen aus einer Reihe von Dominanzbedingungen, die sich leicht aus den Verbrauchsdaten ableiten lassen und die Strukturverfahren robuster gegenüber nichtlinearen Zusammenhängen machen.

Transfer Learning on Phasor Measurement Data from a Power System to Detect Events in Another System:

Die Methoden zur Erkennung von Ereignissen in Stromversorgungssystemen unter Verwendung von vor Ort aufgezeichneten Daten von Phasormessgeräten (PMUs) erfordern oft viele gekennzeichnete Ereignisse, deren Beschaffung kostspielig oder undurchführbar sein kann. In dieser Arbeit wurde gezeigt, dass Ereignisse in einem Stromversorgungssystem genau erkannt werden können, indem eine kleine Anzahl sorgfältig ausgewählter markierter PMU-Daten aus einem anderen System wiederverwendet wird, ohne dass eine zusätzliche Markierung erforderlich ist.

VDGraph2Vec: Vulnerability Detection in Assembly Code using Message Passing Neural Networks:

Die Erkennung von Schwachstellen in Software ist eine der schwierigsten Aufgaben beim Reverse Engineering. In letzter Zeit hat die Erkennung von Schwachstellen aufgrund des drastischen Anstiegs des Volumens und der Komplexität von Software viel Aufmerksamkeit erregt. Reverse Engineering ist ein zeit- und arbeitsintensiver Prozess zur Erkennung von Malware und Softwareschwachstellen. Mit dem Aufkommen von Deep Learning und maschinellem Lernen ist es den Forschern jedoch möglich geworden, den Prozess der Identifizierung potenzieller Sicherheitslücken in Software durch die Entwicklung intelligenterer Technologien zu automatisieren. In dieser Forschungsarbeit wurde VDGraph2Vec vorgestellt, eine automatisierte Deep-Learning-Methode zur Erzeugung von Repräsentationen von Assembler-Code für die Erkennung von Sicherheitslücken.

Machine learning protocol from ultrasound data for monitoring, predicting, and supporting the analysis of dam slopes:

Die bei der Überwachung von Staudämmen gewonnenen Daten können als wichtiger Indikator für das Risikomanagement von Staudämmen verwendet werden. In dieser Studie wurde eine auf maschinellem Lernen und Ultraschall basierende Methodik zur Überwachung der Sicherheit von Staudämmen vorgestellt. Zunächst wurde ein Prototyp eines Staudamms gebaut, um verschiedene Umweltbedingungen zu simulieren. Zweitens wurden Ultraschallbilder in verschiedenen Bereichen eines Prototyp-Damms aufgenommen. Schließlich wurden verschiedene Algorithmen des maschinellen Lernens angewandt, um die verschiedenen Bereiche des Prototyps zu unterscheiden.

Verteiler

Bundesministerium für Umwelt, Klimaschutz, Naturschutz und nukleare Sicherheit (BMUKN)	
Referat S I 3	1x PDF
Bundesamt für die Sicherheit der nuklearen Entsorgung (BASE)	
Fachgebiet F 4	1x PDF
Forschungsmanagement	1x PDF
Gesamtauflage	3x PDF

GRS (pdf-Datei)

Geschäftsführer (sen, stj)

Bereichsleiter (san, flu, kun, thi)

Projektleitung (shv)

Projektcontrolling (zie)

Abteilungsleiter (mif, bbj, psi)

Autoren (gep, mbo)

TECDO (wev)