Exposé

Framework für Nachhaltige und Vertrauenswürdige KI

Prof. Dr. Anja Schmiedt, Prof. Dr. Jürgen Frikel, Prof. Dr. Karsten Weber, Nicole Höß, Prof. Dr. Wolfgang Mauerer[1]

[1] *Regensburg Center for Artificial Intelligence (RCAI), Ostbayerische Technische Hochschule (OTH) Regensburg*

Stand: 28. Oktober 2025

Motivation

Seit dem ChatGPT-Moment entwickeln sich multimodale Basismodelle zu universalen Werkzeugen im domänenübergreifenden Einsatz. Veraltete Prozesse sowie neue Produkte und Services werden zunehmend mit Black-Box-Modellen realisiert, statt sie von Grund auf neu zu konzipieren und zu entwickeln – eine Kernkompetenz, die Deutschland einst zum international anerkannten Innovator und Technologieführer machte. Der Fokus verschiebt sich von kausalem, fachlich fundiertem Verständnis zur Komplexitätsabstraktion und Black-Box-Vertrauen, das im starken Kontrast zu klaren regulatorischen Trends zu mehr Vertrauenswürdigkeit und Nachhaltigkeit steht. Dies reduziert nicht nur die eigene Innovationskraft, sondern birgt zudem die Gefahr, dass viele theoretisch visionäre Forschungsansätze nicht in reale Anwendungen überführt werden können. Gefragt sind daher neuartige, mathematisch und technisch tiefgreifende Ansätze, um den vielfältigen Anforderungen an Transparenz, Nachvollziehbarkeit und Energieeffizienz gerecht zu werden. Dafür ist die Perspektive der angewandten Forschung unerlässlich: Spätere Anwendungskontexte und die daraus abgeleiteten technischen sowie sozialen Anforderungen müssen von Beginn an systematisch berücksichtigt werden, um substanzielle Durchbrüche in Wissenschaft und Anwendung zu erzielen.

1. Kontext

Viele der gesellschaftlich relevantesten KI-Systeme wie autonome Verkehrsmittel und medizinische Diagnose- und Präventionssysteme gelten nach dem EU AI Act als Hochrisiko-Systeme und erfordern entsprechend umfangreiche Maßnahmen zu ihrer Absicherung und Garantie von Transparenz, Fairness und Verantwortlichkeit. Derzeit existieren jedoch nur begrenzt technische Ansätze, um diesen Anforderungen gerecht zu werden. Während etwa die multimodale Auswertung von örtlich verteilten Patientendaten enormes Potenzial in der Entdeckung bislang unbekannter Zusammenhänge bietet, ist die Sicherstellung der Nachvollziehbarkeit bei dafür nötigen Deep-Learning-Verfahren derzeit nur eingeschränkt möglich. Zugleich ergeben sich aufgrund ihres Resource Footprints in Hinblick auf Speicher und Rechenlast Limitierungen für den Einsatz in kritischen Umgebungen wie Fahrzeugen und Produktionsanlagen. Ein Wechsel auf einfachere, intrinsisch erklärbare Verfahren ist aufgrund der hohen domänenspezifischen Komplexität oft ebenfalls kontraindiziert.

2. Zielsetzungen

Um dieses Dilemma zu lösen, ist die Entwicklung eines *offenen Frameworks* aus *konkreten, unmittelbar anwendbaren technischen Tools und Methoden* erforderlich, welches alle Ebenen von der Rechentechnologie über KI-Algorithmen bis zum Monitoring und Risikomanagement während des Betriebs adressiert. Neben wissenschaftlich fundierten Forschungsarbeiten muss der langfristige praktische Mehrwert dieses Frameworks bereits während der Entwicklung im Fokus stehen. Für die erfolgreiche Umsetzung arbeiten Anwender und interdisziplinäre Experten aus Mathematik, Informatik, Naturwissenschaft und Technik daher von Beginn an Hand in Hand und evaluieren die Passgenauigkeit und Generalisierbarkeit anhand konkreter Demonstratoren für diverse domänenspezifische Anwendungsfälle.

3. Methodik

- Neue Rechenparadigmen: Neben der Miniaturisierung von neuronalen Netzen für Edge-Geräte kann photonisches Rechnen dem steigenden Energieverbrauch durch den immer breiteren Einsatz von Deep-Learning-Modelle entgegenwirken, indem es komplexe Rechenoperationen auf Basis von Licht durchführt. Die Passfähigkeit solcher Ansätze wird für spezifische Anwendungsfälle evaluiert, um generalisierbare Standards zu schaffen.
- Mathematische Ansätze für Deep Learning: Zur Weiterentwicklung des State-of-the-Arts der Erklärbarkeit in neuronalen Netzen werden inhärente Strukturen für die intrinsische Erklärbarkeit von Architekturen entwickelt, die beispielsweise zur Steigerung der Robustheit und Zuverlässigkeit beim Lernen aus unbalancierten oder fehlerbehafteten Daten beitragen. Ergänzend werden Post-Hoc-Methoden zur nachträglichen Analyse komplexer Black-Box-Modelle erforscht und angewandt.
- Transparenz durch Reproducibility Engineering: Um Entscheidungen von KI-Systemen nachvollziehen zu können, ist die akkurate Archivierung von sämtlichen Datensätzen, Quellcode und Modellen erforderlich. Auf Basis von Best Practices aus der Forschung werden Tools und Leitfäden auf Basis von Open-Source-Technologien entwickelt, welche Forschende und Unternehmen bei der langfristigen Ermöglichung dieser Aspekte unterstützen.
- Monitoring und Risikomanagement: Um Entwicklungsprozesse nachvollziehen zu können, technische Risiken abschätzen und die Performanz von KI-Modellen während ihres Betriebs überwachen zu können, werden fortschrittliche Softwareanalysetools entwickelt, die komplexe evolutionäre Aspekte wie externe Abhängigkeiten und architektonische Entscheidungen sowie deren Konsequenzen verständlich und zuverlässig auf Basis der zugehörigen Software Repositorien extrahieren und zeitlich aufgelöst zur Entscheidungsunterstützung aufbereiten. Dabei werden Algorithmen der kausalen Entdeckung und Inferenz eingesetzt, um unmittelbar hilfreiche Handlungsempfehlungen zur Behebung von Problemen an der Wurzel zu geben.
- Human-Centered AI: Begleitende Studien zur Einhaltung ethischer, rechtlicher und weiterer sozialer Aspekte sichern die Ausrichtung sämtlicher Forschungs- und Entwicklungsergebnisse nach den tatsächlichen Bedürfnissen verschiedener Stakeholder und damit die breite Akzeptanz des Frameworks.
- In Zusammenarbeit mit Großkonzernen, kleinen und mittelständischen Unternehmen sowie öffentlichen Einrichtungen, Universitäten und Hochschulen wird das entstehende Framework laufend in verschiedenen Domänen evaluiert, um zu analysieren, welche Methoden in welchen Anwendungsfällen den geeignetsten Lösungsansatz darstellen. Neben Deep-Learning-Verfahren werden anwendungsspezifisch auch konventionelle, inhärent erklärbare Verfahren (z.B. Regressionsmodelle und Entscheidungsbäume) sowie Verfahren zur kausalen Entdeckung und Inferenz evaluiert. Basierend auf den Ergebnissen werden generalisierbare Empfehlungen abgeleitet.

4. Outcomes

Aus den Teilprojekten gehen praktisch unmittelbar einsetzbare Tools hervor, die sämtliche Anwendergruppen domänenübergreifend dabei unterstützen, nachhaltige Rechentechnologien zu erschließen, neue mathematisch komplexe Methoden zielgerichtet und ohne große Umstände anzuwenden, Reproduzierbarkeit für gesetzliche Regelungen sowie zur Effizienzsteigerung in der eigenen Unternehmung sicherzustellen, KI-Systeme laufend zu monitoren und dabei deren Risiken automatisiert zu bewerten sowie rechtzeitig Gegenmaßnahmen beginnend von der Wurzel an einzuleiten.

Durch die Evaluation anhand konkreter Bedarfe von KI-Anwendern ergeben sich (1) einrichtungsspezifische Lösungen, die einen unmittelbaren Mehrwert liefern, (2) generalisierbare Best Practices, die Anwendern direkte Empfehlungen liefern, welches Subset an Methoden des Frameworks für sie am besten geeignet ist und (3) konkrete Tools zur einfachen Nutzung dieser Methoden.

5. Impact

Mit diesem Framework zeigt die Hightech Agenda, dass sie europäische Visionen hinsichtlich Nachhaltigkeit und Vertrauenswürdigkeit nicht nur diskutiert und reguliert, sondern sie in greifbare Realität überführt, konkret *implementiert* und *etabliert* und damit internationale Maßstäbe und Best Practices setzt.

Damit befinden sich Bayern und Deutschland in einer Vorreiterrolle und stärkt ie ökonomische Nachhaltigkeit, da vertrauenswürdige KI am Markt erfolgreich ist und Wertschöpfung in Bayern, Deutschland und Europa ermöglicht. Zudem trägt das Vorhaben zur sozialen Nachhaltigkeit bei, da es zentrale europäische Wert stärkt, ein Gegengewicht zu Black-Box-KI-Plattformen darstellt und sozialverträglich eingesetzt wird. Mit ressourcenschonenden Methoden leistet das Vorhaben außerdem einen Beitrag zur ökologischen Nachhaltigkeit.

Nachhaltige und vertrauenswürdige KI zielt auf digitale Souveränität auf Landes-, Bundes- und EU-Ebene: Ressourcenschonung verringert Abhängigkeit, Vertrauenswürdigkeit beruht auf Datenschutz, Verantwortung und Transparenz – überprüfbar und durchsetzbar. Dabei wirkt sie synergetisch mit anderen Technologien zusammen und ermöglicht Innovationen im Querschnitt der bayerischen, deutschen und europäischen Wirtschaft und Gesellschaft.