Explaining 4.0 – Abschlussbericht

Projekttitel: Explaining 4.0
Projektleitung: Prof. Dr. Marina Höhne
Laufzeit: 01.09.2020 – 31.12.2025
Förderkennzeichen: 01IS200551
Durchführende Einrichtung: Universität Potsdam / ATB Potsdam

Mit dem zunehmenden Einsatz von Künstlicher Intelligenz (KI) in sicherheitskritischen und gesellschaftlich sensiblen Anwendungsfeldern wächst der Bedarf an transparenten KI Modellen und nachvollziehbaren Entscheidungsprozessen. Das Projekt *Explaining 4.0* verfolgte das Ziel, Methoden zur globalen Erklärbarkeit von KI-Modellen zu entwickeln, die über klassische lokale Visualisierungen (z. B. Heatmaps) hinausgehen. Statt einzelner Erklärungen wurden Methoden entwickelt, die einen semantisch interpretierbaren, robusten und ganzheitlichen Einblick in das Entscheidungsverhalten von KI-Systemen ermöglichen. Damit leistet das Projekt einen wesentlichen Beitrag zur algorithmischen Rechenschaftspflicht und zur Entwicklung ethisch verantwortungsvoller KI.

Zur Verbesserung des Verständnisses interner Repräsentationen neuronaler Netze wurden im Rahmen von Explaining 4.0 mehrere richtungsweisende Methoden entwickelt, die das Feld der repräsentationsbasierten Erklärbarkeit wesentlich vorangebracht haben. Mit DORA (Bykov et al., 2023) wurde ein neuartiger Ansatz etabliert um Ausreißer- und verdächtige Konzepte, wie beispielsweise gelernte Artefakte, innerhalb neuronaler Modelle zu identifizieren – und das ohne jegliche Abhängigkeit von Datensätzen, Labels oder externen Annotationen. Aufbauend darauf ermöglicht INVERT (Bykov et al., 2023) die semantische textuelle Benennung einzelner Neuronen, ohne die Einbeziehung von externen Segmentierungsmasken. Diese Methoden leisten einen entscheidenden Beitrag zur globalen Erklärung der Funktionsweise tiefer neuronaler Netzwerke. Ein weiteres bedeutendes Ergebnis stellt FSCA (Bykov et al., 2023) dar, das funktional-semantische Inkonsistenzen in neuronalen Repräsentationen sichtbar macht und so Fehlanpassungen vom Modell zwischen gelernter Funktionalität und beabsichtigtem, aufgabenbezogenen semantischen Verhalten offenlegt. Mit Gradient Slingshots (Bareeva et al., 2024) konnte darüber hinaus gezeigt werden, dass viele gängige globale Erklärungsansätze systematisch manipuliert werden können, ein wichtiges Ergebnis für die generelle Vertrauenswürdigkeit der Feature-Visualisierungs-Methoden. Ergänzend legte Mark My Words (Bykov et al., 2023) offen, wie Artefakte aus Trainingsdaten, etwa Wasserzeichen in Bildern, unbemerkt in die inneren Repräsentationen von vortrainierten Modellen übergehen und dort sogar nach Finetuning auf anderen Datensätzen bestehen können. Weiterhin leistete das Projekt auch Beiträge zur Entwicklung selbsterklärender Modelle, etwa durch Mechanismen der intrinsischen Interpretierbarkeit (Gautam et al., 2023), sowie zur Robustheitssteigerung prototypischer Erklärungen durch autoencoderbasierter Regularisierung (Gautam et al., 2022a).

Ein weiterer zentraler Beitrag des Projekts Explaining 4.0 liegt in der Entwicklung fundierter Verfahren zur Evaluierung von Methoden der erklärbaren Künstlichen Intelligenz (XAI). Die Bewertung von Erklärungen stellt aufgrund der fehlenden „Ground Truth“, also einer objektiv richtigen Referenzerklärung, eine besondere Herausforderung dar. Im Rahmen des Projekts

wurden mehrere wegweisende Ansätze erarbeitet, die sowohl theoretische als auch praktische Aspekte der Evaluierung adressieren. Mit Quantus (Hedström et al., 2023) wurde das erste standardisierte Software-Toolkit geschaffen, das eine systematische Bewertung unterschiedlicher Erklärmethoden ermöglicht. Aufbauend darauf wurde mit MetaQuantus (Hedström et al., 2023) ein neuartiges Metaevaluationsverfahren entwickelt, das die *Zuverlässigkeit* der eingesetzten Evaluationsmetriken analysiert. Eine theoretische Fundierung liefert das Konzept der Generalised Explanation Faithfulness (GEF) (Hedström et al., 2025), das auf Prinzipien der Differentialgeometrie beruht und eine einheitliche Sicht auf verschiedene Evaluationsansätze ermöglicht. Zur Überprüfung der Robustheit von Erklärungen wurden bestehende Sanity-Check-Protokolle erweitert um den Mean Resilience Rank (Wickstrøm et al., 2024), der gezielt die Widerstandsfähigkeit von Erklärungen gegen Störungen erfasst. Darüber hinaus wurden mit den Quality Gap Estimates (Eiras et al., 2025) probabilistische Verfahren eingeführt, die Qualitätsschwankungen in Erklärungen quantifizieren und Unsicherheiten systematisch erfassen. Ergänzend dazu adressieren die Methoden Quanda (Bareeva et al., 2024) und CoSy (Kopf et al., 2024) die Bewertung der Attribuierung von Trainingsdaten, und die Bewertung von textuellen neuronalen Erklärungen auf ihre Treue und Verständlichkeit hin.

Die im Projekt entwickelten Methoden zur erklärbaren KI wurden erfolgreich in unterschiedlichen wissenschaftlichen Anwendungsfeldern eingesetzt. In der Klimaforschung ermöglichten systematische Benchmark-Studien ein besseres Verständnis darüber, wie stark die Wahl der Erklärmethode wissenschaftliche Modellinterpretationen beeinflussen kann (Bommer et al., 2024). Auch in der Biodiversitätsforschung und der Landwirtschaft kamen XAI-Methoden zur Analyse ökologischer Zusammenhänge und pflanzlicher Gesundheit zum Einsatz (Liu et al., 2023). In der Wettervorhersage, insbesondere im Bereich der Subseasonal-to-Seasonal-Prognosen, führten hybride ViT-LSTM-Modelle zu einer verbesserten Vorhersagequalität, wobei erklärende Verfahren maßgeblich zur Vertrauensbildung in die Modellentscheidungen beitrugen (Bommer et al., 2025).

Die im Projekt *Explaining 4.0* entwickelten Methoden bilden heute eine tragfähige Grundlage für die reproduzierbare und disziplinübergreifende Anwendung und Evaluierung von erklärbarer KI. Sie ermöglichen tiefgreifende Einblicke in die internen Mechanismen moderner Modelle, stärken die Nachvollziehbarkeit neuronaler Entscheidungen und schaffen neue Möglichkeiten zur Validierung sicherheitskritischer Systeme. Durch zahlreiche wissenschaftliche Veröffentlichungen, offene Frameworks wie Quantus, INVERT, DORA, MetaQuantus und CoSy sowie deren Integration in Lehre, Ausbildung und internationale Workshops leistet das Projekt einen nachhaltigen Beitrag zu Forschung, Bildung und Open Science und trägt zu einer verantwortungsvollen Nutzung von von KI bei.