Explaining4.0 — Long Report

General Project Information

Project Title: Explaining 4.0
Project Lead: Prof. Dr. Marina Höhne
Project Duration: 01.09.2020 – 31.12.2024
Funding Code: 01IS200551
Lead Institution: University of Potsdam / ATB Potsdam

Förderhinweis

Dieses Projekt wurde vom BMBF unter dem Förderkennzeichen 01IS200551 gefördert. Die Verantwortung für den Inhalt liegt bei der Autorin.

Das diesem Bericht zugrunde liegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS200551 *gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei der Autorin/beim Autor*

1. Einleitung und Zielsetzung

Im Zuge der fortschreitenden Integration Künstlicher Intelligenz (KI) in gesellschaftlich und sicherheitstechnisch kritische Anwendungsfelder wird die Frage nach der Erklärbarkeit und Nachvollziehbarkeit von KI-Systemen immer drängender. Während sogenannte "Black-Box-Modelle" in vielen Aufgabenbereichen beeindruckende Leistungen erzielen, fehlt es bislang häufig an fundierten Methoden, die die internen Entscheidungsmechanismen dieser Modelle transparent und nachvollziehbar machen. Genau hier setzte das BMBF-geförderte Projekt *Explaining 4.0* an.

Ziel war die Entwicklung von globalen Methoden der erklärbaren Künstlichen Intelligenz, die über lokale Visualisierungen (wie etwa Heatmaps) hinausgehen. Im Mittelpunkt stand dabei die Frage, wie sich neuronale Netzwerke nicht nur punktuell, also anhand von einzelnen Datenpunkten, sondern auf globaler und semantischer Ebene interpretieren lassen. Der Projekt Fokus lag daher auf der methodischen Weiterentwicklung von Verfahren, die es ermöglichen, semantische Konzepte, interne Repräsentationen und

funktionale Zusammenhänge in KI-Modellen systematisch offenzulegen – mit dem Ziel, ihre Entscheidungen nachvollziehbar, begründbar und vertrauenswürdig zu machen.

Diese Herausforderung wurde in sechs eng aufeinander abgestimmte Arbeitspakete überführt, die sich über den gesamten Lebenszyklus erklärbarer KI erstrecken: von der Erstellung synthetischer Benchmark-Datensätze über die mathematische Extraktion semantischer Einheiten bis hin zur robusten Evaluation und offenen Bereitstellung praxisnaher Softwarewerkzeuge. Dabei wurde besonderer Wert auf Anwendungsnähe und disziplinübergreifende Anschlussfähigkeit gelegt – insbesondere im Kontext wissenschaftlicher Forschung in Bereichen wie Klimawandel, Bioökonomie und Medizin.

Wissenschaftliche Hauptergebnisse des Projekts *Explaining 4.0*

Dieser Abschnitt fasst die zentralen wissenschaftlichen und systemischen Ergebnisse des Projekts *Explaining 4.0* zusammen, die im Zeitraum von 2020 bis 2025 entstanden sind.

1.1 Von Arbeitspaketen zu thematischen Forschungslinien

Ausgehend von der ursprünglichen Arbeitspaketstruktur entwickelten sich im Projektverlauf drei thematische Forschungsschwerpunkte:
(1) Entwicklung von Methoden für die Evaluierung der erklärbaren KI (XAI),
(2) Entwicklung von robusten XAI Methoden auf Repräsentationsebene,
(3) Domänenspezifische Anwendungen in den Naturwissenschaften, insbesondere der Klimaforschung

Die Projektarbeit adressierte bestehende Lücken in Theorie, Methodik und Werkzeugunterstützung – etwa das Fehlen formaler Evaluationskonzepte, die Notwendigkeit mechanistischer Repräsentationsanalysen sowie die Forderung nach vertrauenswürdiger KI im Kontext wissenschaftlicher Anwendungen wie Klimamodellierung.

1.2 Thematisch geordnete Ergebnisse

1.2.1 Evaluierung ohne Ground Truth: Grundlagen und Metriken

Eine zentrale Herausforderung der XAI-Forschung ist die Evaluation von Erklärungen, da keine Ground Truth Erklärung verfügbar ist. Dazu wurde in *Explaining 4.0* eine Evaluierungsbibliothek Quantus (Hedström et al., 2023) entwickelt, die erstmals ein standardisiertes, softwaregestütztes Framework für den systematischen Vergleich von Erklärmethoden bietet.

Darüber hinaus wurde mit der Generalised Explanation Faithfulness (GEF) (Hedström et al., 2025) ein theoretischer Rahmen geschaffen, der verschiedene bestehende Metriken unter einem geometrisch fundierten Ansatz vereint. Die Metaevaluationsplattform

MetaQuantus (Hedström et al., 2023) ermöglichte erstmals eine systematische Bewertung der Aussagekraft und Robustheit von Evaluationsmetriken selbst – ein entscheidender Fortschritt insbesondere für sicherheitskritische Anwendungsbereiche.

Weitere Beiträge umfassen die Entwicklung robuster Bewertungsverfahren wie dem Mean Relevance Rank, neuartiger Qualitätsmetriken wie eMPRT und QGE, sowie die erstmalige Evaluation von Trainingsdatenattribution (TDA) mit der Methode Quanda (Bareeva et al., 2024).

1.2.2 Repräsentationsbasierte Erklärbarkeit neuronaler Netze

Ein weiterer Forschungsschwerpunkt lag auf der Analyse der internen Repräsentationen neuronaler Netze – also der Frage, wie tiefe neuronale Netze semantische Informationen intern organisieren und verarbeiten. Hierzu wurden mehrere modellagnostische Methoden entwickelt, darunter DORA, INVERT, FSCA, die es ermöglichen, interne Konzepte zu identifizieren, zu benennen und auf Konsistenz zu prüfen.

Diese Werkzeuge erweitern den Stand der Forschung im Bereich der globalen Interpretierbarkeit signifikant und stellen konkrete Werkzeuge für die Analyse der Merkmalsräume bildverarbeitender KI-Systeme zur Verfügung.

1.2.3 Anwendung in der wissenschaftlichen Praxis

Die im Projekt entwickelten Methoden fanden konkrete Anwendung in mehreren wissenschaftlichen Kontexten, insbesondere in der Klimawissenschaft und im Bereich Biodiversitätsmonitoring. Dabei wurden erklärbare Verfahren in bestehende wissenschaftliche Analysepipelines integriert, etwa zur Klassifikation von Temperaturmustern, für subseasonale Wettervorhersagen sowie zur Detektion biologischer Vielfalt.

Ein zentrales Ergebnis ist die Tatsache, dass Domänenwissenschaftler bereits heute Evaluierungstools wie das im Projekt entwickelte Tool Quantus einsetzen, um Deep-Learning-Vorhersagen mit physikalischen und ökologischen Vorwissen zu verbinden und ihre Interpretierbarkeit kritisch zu prüfen.

1.2.4 Stochastische Robustheit und Unsicherheit in XAI

Ein weiterer zentraler Beitrag betrifft die Modellierung von Unsicherheit und Robustheit in Erklärungen. Mit NoiseGrad und FusionGrad wurden zwei neue Methoden entwickelt, die gezielte Stochastizität in die Modelle (und damit auch in den Erklärungsprozess) einführen. Dadurch entstehen robustere und insgesamt verlässliche Erklärungen.

1.2.5 Selbsterklärende Modelle

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).

1.3 Wissenschaftliche Sichtbarkeit und Publikationsleistung

Die Ergebnisse von *Explaining 4.0* wurden erfolgreich in hochrangigen internationalen Fachzeitschriften und Konferenzen publiziert, darunter NeurIPS, ICML, AAAI, ICLR und JMLR. Diese Bandbreite spiegelt sowohl die methodische Tiefe als auch die interdisziplinäre Anschlussfähigkeit der entwickelten Ansätze wider. Die Publikationen adressieren sowohl grundlegende theoretische als auch empirische Fragestellungen und reichen von algorithmischer Entwicklung über Benchmarking bis zur Anwendung in spezifischen Domänen.

1.4 Open-Source-Beiträge und Software-Ökosystem

Ein zentrales Ziel von *Explaining 4.0* war die nachhaltige Bereitstellung der entwickelten Methoden für die Fach-Community. Hierzu wurde ein **modulares, offen lizenziertes Software-Ökosystem** aufgebaut, das bereits heute in Forschung und Industrie breite Anwendung findet. Insbesondere die Werkzeuge **Quantus**, **MetaQuantus**, **CoSY**, **Quanda** und **GEF** haben neue Standards in der Bewertung und Analyse von Erklärbarkeit gesetzt und stoßen kontinuierlich auf internationale Resonanz.

1.5 Dissemination, Betreuung und wissenschaftliche Netzwerke

Das Projekt zeichnete sich durch umfangreiche Disseminationsaktivitäten aus: Neben zahlreichen eingeladenen Vorträgen, Keynotes und Workshops wurden auch Lehrveranstaltungen konzipiert, Abschlussarbeiten betreut und frühkarriere Forschende intensiv begleitet. Die vielfältigen Formen der Einbindung und Sichtbarkeit belegen die Relevanz und Nachhaltigkeit des Projekts – sowohl im akademischen als auch im gesellschaftlichen Kontext rund um vertrauenswürdige KI.

1.1 Überblick über die Arbeitspakete

Das Projekt *Explaining 4.0* war entlang von sechs eng miteinander verknüpften Arbeitspaketen (APs) strukturiert, die jeweils mit eigenen Meilensteinen versehen waren. Gemeinsam verfolgten sie das übergeordnete Ziel, **globale, semantisch interpretierbare und robuste Erklärbarkeitsmethoden** für KI-Systeme zu entwickeln. Wie bereits in der ursprünglichen Projektbeschreibung skizziert, ging *Explaining 4.0* bewusst über klassische lokale Verfahren der Erklärbarkeit, wo die Erklärung für einzelne Datenpunkten generiert wird, hinaus. Stattdessen war das Ziel des Projektes Methoden zu entwickeln, die das Verhalten tiefer neuronaler Netze ganzheitlich anhand von gelernten Konzepten erfassen und nachvollziehbar machen.

Nachfolgend wird der ursprüngliche Arbeitsplan zusammengefasst:

- **AP1 – Testbed:** Aufbau einer umfassenden Testumgebung inklusive Datensätzen, Modellen und Erklärungsverfahren zur Entwicklung und Evaluation von XAI-Methoden.
- **AP2 – Reverse Engineering:** Effizienzsteigerung globaler Erklärungen (MFI) durch gezielte Nutzung von lokalen Wissens zur effizieren Berechnung
- **AP3 – Semantische Elemente:** Definition und Extraktion interpretierbarer, semantischer Einheiten aus den inneren Repräsentationen von KI Modellen.
- **AP4 – Robustheit & Unsicherheit:** Entwicklung robuster Erklärmethoden, die auch unter adversarialen Bedingungen und bei Verteilungsverschiebungen verlässliche Aussagen liefern.
- **AP5 – Evaluation:** Konzeption ganzheitlicher Evaluationsmetriken und -protokolle zur Qualitäts- und Vertrauensbewertung von Erklärungen.
- **AP6 – Toolbox:** Bereitstellung einer offenen, benutzerfreundlichen Software Toolbox inkl. Tutorials für die nutzerfreundliche Anwendung der entwickelten Methoden in Forschung und Praxis.

Diese Struktur war gezielt darauf ausgelegt, die im Antrag formulierten drei Komponenten: *Semantik*, *Robustheit* und *A priori Wissen* methodisch abzubilden und daraus globale Erklärungen zu entwickeln. Jedes Arbeitspaket repräsentierte dabei ein zentrales Puzzleteil zur Lösung der Herausforderungen globaler Erklärbarkeit.

1.1.1 Entwicklung thematischer Forschungsschwerpunkte

Im Verlauf des Projekts kristallisierten sich drei übergreifende **Forschungslinien** heraus, die aus der ursprünglichen AP-Struktur hervorgingen und sich zugleich an dynamische wissenschaftliche und gesellschaftliche Entwicklungen anpassen:

- Evaluationszentrierte XAI (AP1, AP4, AP5, AP6): Entwicklung theoretischer und praktischer Methoden zur systematischen Bewertung und Validierung von Erklärmethoden. Daraus entstanden zentrale Werkzeuge wie Quantus, GEF, CoSy, eMPRT sowie Sanity Checks Revisited. Ergänzt wurde dies durch das Metaevaluation Framework MetaQuantus und das adversariale Bewertungsverfahren MRR, die heute einen etablierten Evaluationsstandard in der XAI-Community bilden.
- Interpretierbarkeit auf Repräsentationsebene (AP2, AP3, AP4): Entwicklung von Methoden zur semantischen Analyse interner Repräsentationen neuronaler Netze. Ergebnisse wie DORA, INVERT und FSCA erweiterten den Stand der Technik zur Identifikation, Benennung und Prüfung semantischer Einheiten innerhalb tiefer neuronaler Netze.
- Anwendungsorientierte XAI in der Klimaforschung (AP-übergreifend): Konkrete Weiterentwicklung von Methoden der erklärbarer KI für hochrelevante Domänen

wie Klimawissenschaft und Biodiversitätsmonitoring Zu den Resultaten zählen u. a. das Quantus x Climate Tutorial, ein Framework zur subseasonalen Vorhersage mit ViT-LSTM-Modellen, sowie Systeme zur Erkennung von Indikatorpflanzen.

Diese Forschungslinien spiegeln die enge Anbindung an die ursprünglichen Arbeitspakete wider, erlaubten aber zugleich Raum für Weiterentwicklungen – insbesondere im Hinblick auf neu aufkommende Herausforderungen wie **Foundation Models**, zunehmende Modellkomplexität und Anforderungen an eine vertrauenswürdige KI im Anwendungskontext.

1.1.2 Entwicklung und Motivation für inhaltliche Anpassungen

Obwohl die meisten Arbeitspakete planmäßig abgeschlossen wurden, kam es im Projektverlauf zu gezielten inhaltlichen Anpassungen einzelner Aufgaben – bedingt durch neue wissenschaftliche Publikation, Methoden und Modelle. Die strategische Schwerpunktverlagerung hin zu evaluationsorientierter Interpretierbarkeit und mechanistischer Analyse großer Modelle erwies sich als fachlich begründet und äußerst fruchtbar. Sie führte nicht nur zur Entwicklung hochwirksamer Frameworks und Methoden, sondern ermöglichte auch neue Kollaborationen mit internationaler Sichtbarkeit. Der Kern der ursprünglichen Projektvision blieb dabei jederzeit erhalten.

1.1.3 Umsetzung und Kooperationsstruktur

Die Arbeiten wurden primär von einer Kernarbeitsgruppe aus Promovierenden unter der wissenschaftlichen Leitung von Prof. Dr. Marina Höhne durchgeführt. Die Umsetzung erfolgte in enger Zusammenarbeit zwischen mehreren führenden Forschungseinrichtungen:

- Technische Universität Berlin (Fakultät für Elektrotechnik und Informatik, sowie BIFOLD)
- Leibniz-Institut für Agrartechnik und Bioökonomie (ATB)
- Universität Potsdam (Institut für Informatik)

Darüber hinaus ermöglichte ein aktives, internationales Partnernetzwerk die Verknüpfung von methodischer Grundlagenforschung mit praktischer Validierung in realweltlichen Szenarien. Kooperationspartner waren unter anderem:

- Fraunhofer Heinrich-Hertz-Institut (Abteilung KI)
- Universität Leipzig
- University of Reading (UK)
- Arctic University (Norwegen)
- Technische Universität Kaiserslautern
- Alan Turing Institute (UK)

- Brown University / ICERM
- Cornell SAIC (USA)
- OIST (Japan)

Diese Zusammenarbeit förderte nicht nur gemeinsame Veröffentlichungen auf höchstem fachlichen Niveau, sondern ermöglichte auch die Validierung der Methoden in verschiedenen Anwendungskontexten.

1.2 Wissenschaftliche Ergebnisse nach Themenfeld

Im Folgenden werden die zentralen wissenschaftlichen Beiträge des Projekts *Explaining 4.0* entlang von drei thematischen Forschungslinien dargestellt, die sich aus der ursprünglichen Arbeitspaketstruktur ergeben haben.

1.2.1 Evaluation-zentrierte Fortschritte bei Erklärbarkeitsmethoden

Ein fundamentales Problem im Bereich der erklärbaren Künstlichen Intelligenz (XAI) ist nach wie vor ungelöst: Wie kann die Qualität einer Erklärung bewertet werden, wenn keine Ground-Truth-Erklärungen existieren? Anders als in der klassischen überwachenden maschinellen Lernpraxis, bei der Vorhersagefehler durch Vergleich mit annotierten Daten quantifiziert werden können, fehlt im Bereich der Erklärbarkeit eine allgemein akzeptierte Referenzerklärung (Benitez et al., 1997; Lipton, 2018). Im Verlauf von *Explaining 4.0* wurde dieses Problem daher zu einem zentralen Thema, das in mehreren Arbeitspaketen adressiert wurde: insbesondere in AP1 (Testbed und experimentelle Umgebung), AP4 (Robustheit und Unsicherheit), AP5(Evaluationsmetriken und Metaevaluation) sowie AP6 (Integration in offene Softwareplattformen wie Quantus, MetaQuantus, GEF, eMPRT). Der methodische Ansatz war dabei vielschichtig und umfasste sowohl neue theoretische Rahmenwerke als auch praxisorientierte Open-Source-Werkzeuge und methodologische Neuentwicklungen.

Im Rahmen des Projekts wurden mehrere interdisziplinär anschlussfähige Publikationen und Software-Tools entwickelt, die systematisch zur Schließung der Evaluationslücke im Bereich XAI beigetragen haben – und die bereits eine weitreichende Anwendung innerhalb und außerhalb der XAI-Community erfahren haben.

Quantus: Eine standardisierte Schnittstelle für XAI-Evaluation

Die Entwickelung und Bereitstellung des Toolkit Quantus hat signifikanten Beitrag in der Standardisierung in der XAI-Evaluierung geleistet. Es handelt sich um eine modulare und erweiterbare Python-Bibliothek, die mittlerweile über 35 Evaluationsmetriken beinhaltet und zugleich eine fundierte Anleitung zur korrekten Parametrisierung und Anwendung bereitstellt. Quantus wurde vollständig öffentlich zugänglich publiziert, ist Bestandteil des renommierten Journal of Machine Learning Research (JMLR) und wurde zusätzlich auf der NeurIPS-Konferenz 2023 vorgestellt.

Das Toolkit steht exemplarisch für die Verzahnung mehrerer Arbeitspakete:

- **AP1**, da es eine programmatische Testumgebung für XAI-Methoden bietet,
- **AP5**, da es eine Vielzahl theoretisch fundierter Metriken enthält, und
- **AP6**, da es aktiv in Vorträgen, Lehrveranstaltungen und Workshops verbreitet wird.

Quantus unterstützt eine breite Palette an Metriken – darunter Faithfulness, Robustheit, Lokalisierung, Komplexität und Zufallsresistenz – und ist mit verschiedenen Deep-Learning-Frameworks (TensorFlow, PyTorch) kompatibel. Die Bibliothek wurde bereits in zahlreichen realen Anwendungsszenarien eingesetzt, etwa in der medizinischen Bildgebung, der Klimamodellierung, sowie in der Objekterkennung und Satellitenbildanalyse. Ziel von Quantus war es nicht nur, ein Werkzeug zu entwickeln, sondern auch einen gemeinsamen Bewertungsstandard und eine wissenschaftliche Sprache für Evaluation zu etablieren. Das Toolkit wurde bislang über 250 Mal zitiert, mehr als 55.000 Mal heruntergeladen und kommt in unterschiedlichen wissenschaftlichen Feldern zum Einsatz – von der Sicherheitsforschung über die medizinische Diagnostik bis hin zur Klimawissenschaft. Mit über 595 GitHub-Sternen zählt Quantus zu den international sichtbarsten und am häufigsten genutzten Open-Source-Werkzeugen im Bereich der XAI-Evaluation. Die Veröffentlichung auf der Plattform Climate Change AI trug wesentlich dazu bei, die Evaluationsmetriken auch in die Fachgemeinschaft der wissenschaftlichen Modellierung einzubringen – insbesondere zur Validierung erklärbarer Vorhersagen in der Klimaforschung.

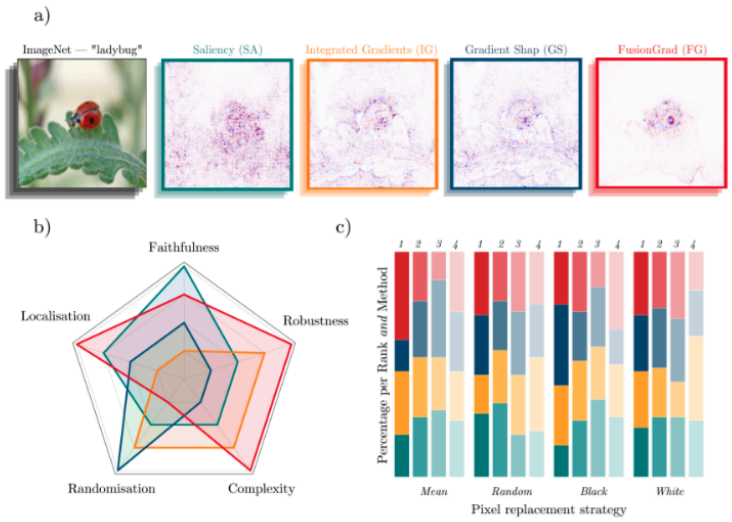

Figure 1. a) Ein rein qualitativer Vergleich von XAI-Methoden – etwa Saliency (Mørch et al., 1995; Baehrens et al., 2010), Integrated Gradients (Sundararajan et al., 2017), GradientShap (Lundberg & Lee, 2017) oder FusionGrad (Bykov et al., 2021a) – ist oftmals unzureichend, um deren relative Eignung zu beurteilen. b) Quantus ermöglicht eine umfassende quantitative Bewertung über verschiedene Kriterien hinweg. c) Darüber hinaus erlaubt das Toolkit Sensitivitätsanalysen, z. B. zur Frage, wie stark die Wahl einer bestimmten *Pixel Replacement Strategy* im Rahmen eines Faithfulness-Tests die Rangfolge der Methoden beeinflusst.

MetaQuantus: Evaluierung der XAI Evaluierungsmetriken

Mit MetaQuantus wurde ein neuartiges Framework entwickelt, das eine Ebene über der herkömmlichen Metrikbewertung ansetzt: Anstatt Erklärmethoden direkt zu bewerten, evaluiert MetaQuantus die Bewertungsmethoden selbst. Ausgangspunkt war die Beobachtung, dass verschiedene Metriken für ein- und dieselbe Erklärung zu widersprüchlichen Bewertungen führen können – etwa weil ihre Skalen und Richtungen nicht vergleichbar sind. So kommt es häufig vor, dass eine Methode (z. B. *Gradient*) von einer Metrik als beste, von einer anderen jedoch als schlechtest eingestuft wird – was in der Praxis zu erheblicher Verunsicherung führt. MetaQuantus wurde in den *Transactions of Machine Learning Research* (2023) veröffentlicht und markiert einen Paradigmenwechsel in der Bewertungspraxis von XAI.

GEF: Eine theoretische Perspektive auf Erklärtreue

Mit GEF (Generalised Explanation Faithfulness) wurde ein mathematisches Rahmenwerk eingeführt, das verschiedene Evaluationskonzepte – wie Robustheit, Sensitivität und Faithfulness – unter einem gemeinsamen theoretischen Horizont vereint. Die Methode basiert auf Differentialgeometrie und erlaubt es, die in der Literatur üblichen, teils vereinfachenden Annahmen (z. B. Linearisierung oder konstante Perturbationseffekte) gezielt zu korrigieren – insbesondere für hochkomplexe Modelle wie Transformer oder tiefe neuronale Netze.

GEF wurde in AP5 konzipiert und als Fast-GEF effizient implementiert, um domänenübergreifende Skalierbarkeit zu gewährleisten. Die Methode ist modalitätsunabhängig und auf Bilder, Texte und strukturierte Daten anwendbar. GEF wurde in *Transactions of Machine Learning Research* veröffentlicht und mit einer „Survey Certification" für ihren grundlegenden Beitrag zur Theorie der Evaluierbarkeit in der XAI ausgezeichnet. Die Entwicklung erfolgte in Kooperation mit der Universität Potsdam, ATB Potsdam, Brown University, TU Berlin, OIST, Cornell und Fraunhofer HHI.

Sanity Checks Revisited: eMPRT und sMPRT

Der in der XAI-Community weitverbreitete Model Parameter Randomisation Test (MPRT) (Adebayo et al., 2018) basiert auf der Annahme, dass Erklärungen auf Modellparametern beruhen, was jedoch als methodisch als unzureichend kritisiert wurde (Binder et al., 2022). *Explaining 4.0* konzipierte daraufhin zwei Weiterentwicklungen:

- sMPRT (Smooth MPRT): reduziert die Rauschanfälligkeit durch Mittelung über leicht veränderte Eingaben.
- eMPRT (Efficient MPRT): verbessert den Test als Messung der Komplexitätsverschiebung bei vollständiger Gewichtsrandomisierung.

Beide Varianten beheben Schwächen der Originalmethode und verbessern deren Aussagekraft. Sie sind mittlerweile Bestandteil des Quantus-Frameworks (AP6) und wurden beim NeurIPS 2023 Workshop „XAI in Action“ sowie auf der XAI World Conference vorgestellt.

Inverse-QGE: Vergleichsmetriken statt Absolutwerte

Die Methode Inverse-QGE verfolgt einen innovativen Ansatz: Statt absolute Scores zu berechnen, wird die Qualität einer Erklärung im Vergleich zu einer absichtlich schlechten Gegenerklärung bewertet. Diese relative Betrachtung erlaubt eine bessere Einordnung, insbesondere dort, wo absolute Werte schwer interpretierbar sind. Inverse-QGE wurde auf der AAAI-Konferenz 2025 in Philadelphia vorgestellt und ergänzt die XAI Evaluierung um einen pragmatischen, vergleichsorientierten Ansatz.

Adversariale Angriffe auf Evaluationsmetriken

Eine weitere Untersuchung zeigte, wie Evaluationsmetriken gezielt manipuliert werden können – etwa durch subtile Änderungen an Hyperparametern. Je nach Zielsetzung können so bestimmte Methoden künstlich besser oder schlechter erscheinen. Um dem entgegenzuwirken, wurde im Rahmen von AP4 und AP5 eine robuste Gegenmaßnahme entwickelt: die Mean Resilience Rank (MRR). Sie bewertet die Stabilität der Methodeneinstufung über verschiedene Parametereinstellungen hinweg und macht damit Manipulationsversuche sichtbar. Die Ergebnisse wurden auf dem ECCV Workshop 2024 sowie dem NeurIPS Workshop „Interpretable AI“ 2024 präsentiert.

Quanda: Evaluationsframework für Trainingsdaten-Attribution

Mit Quanda wurde ein erstes umfassendes Toolkit zur Evaluierung von Trainingsdaten-Attributionsmethoden (TDA) entwickelt. Während sich in den letzten Jahren zahlreiche Verfahren zur Ermittlung einflussreicher Trainingsbeispiele in Bezug auf das Modellverhalten etabliert haben, mangelte es bislang an Methoden zur Vergleichbarkeit und Validierung dieser Ansätze. Quanda, das in Zusammenarbeit mit TU Berlin, ATB Potsdam, Fraunhofer HHI, Universität Potsdam und BIFOLD entwickelt wurde, schließt diese methodische Lücke.

Das Toolkit bietet eine einheitliche Python-API, unterstützt mehrere Metriken und ist vollständig kompatibel mit bestehenden Attribution-Libraries. Quanda wurde auf dem ATTRIB Workshop der NeurIPS 2024 vorgestellt und hat bereits erste Impulse für neue Benchmarking-Standards im TDA-Bereich gesetzt.

CoSy – Evaluating Text Explanation of Neurons

Mit CoSy (Concept Synthesis) wurde auf der NeurIPS 2024 ein neuartiges, architekturunabhängiges Framework vorgestellt, das die Qualität textbasierter

Erklärungen latenter Neuronen systematisch bewertet. Entwickelt in enger Zusammenarbeit von Forschenden der TU Berlin, ATB Potsdam, Fraunhofer HHI, Universität Potsdam und BIFOLD, adressiert CoSy eine bislang kaum behandelte, aber zentrale Herausforderung: Wie kann überprüft werden, ob eine textuelle Erklärung eines Neurons tatsächlich das Verhalten dieses Neurons widerspiegelt?

Der Ansatz basiert auf einem generativen Modell, das gezielt Eingaben zu vorgegebenen Konzepten synthetisiert und diese anschließend auf ihre neuronale Aktivierung hin untersucht. Auf diese Weise lässt sich quantitativ bestimmen, in welchem Maße ein Konzept mit der tatsächlichen Funktionsweise eines Neurons übereinstimmt.

Die Wirksamkeit von CoSy wurde durch umfassende Metaevaluationsstudien auf gängigen Bildklassifikationsbenchmarks bestätigt. Das Framework bietet damit einen entscheidenden methodischen Fortschritt für die Evaluierung von textuellen Erklärungen neuronaler Netzwerke.

1.2.2 Repräsentationslernen für erklärbare KI (XAI)

Eine zentrale Forschungsrichtung in XAI beschäftigt sich mit der Extraktion der gelernten Repräsentationen neuronaler Netze: Welche semantischen Konzepte codieren einzelne Neuronen, und wie wirken sich diese Konzepte auf die Entscheidungen eines Modells aus? Trotz der Nutzung von Deep Learning Modellen, insbesondere in sicherheitskritischen und gesellschaftlich sensiblen Anwendungsfeldern, ist zum Projektstart wenig bekannt gewesen über die vom Modell gelernten Konzepte und ihrer mechanistischen Funktionsweise.

Im Rahmen der Arbeitspakete WP3 und WP6 hat *Explaining 4.0* einen bedeutenden Beitrag zur mechanistischen Erklärbarkeit geleistet: Es wurden neuartige Frameworks und Metriken entwickelt, um interne Repräsentationen tiefer neuronaler Netze systematisch zu identifizieren, zu benennen und auf ihre Funktion hin zu überprüfen. Damit wurde der Blick in die "Black Box" auf Neuronenebene geöffnet und die innere Struktur des Merkmalsraums präzisiert.

DORA – Datenunabhängige Analyse neuronaler Repräsentationen

DORA (*Data-agnOstic Representation Analysis*) wurde 2023 in *Transactions on Machine Learning Research (TMLR)* veröffentlicht und stellt das erste vollständig datenunabhängige Verfahren zur Erkennung von Ausreißer-Repräsentationen in neuronalen Netzen dar. Entwickelt in Zusammenarbeit mit der TU Berlin, dem ATB Potsdam, dem Max-Planck-Institut für Informatik, der Korea University, BIFOLD, der Universität Potsdam und der UiT (Arctic University of Norway) führt DORA die sogenannte *Extreme Activation Distance (EA)* als zentrale Metrik ein. Diese misst die Aktivierungszustände innerhalb eines Layers und identifiziert dadurch neuronale

Konzepte, die auf unerwünschte gelernte Korrelationen wie Wasserzeichen, Hintergrundtexturen oder Artefakte zurückzuführen sind.

DORA operiert unabhängig von Datensätzen, Labels oder Segmentierungen und vermeidet so datensatzspezifische Verzerrungen in der Erklärung. Die Methode wurde erfolgreich auf gängigen Computervision Modellen angewendet und findet zunehmend Anwendung in der Praxis.

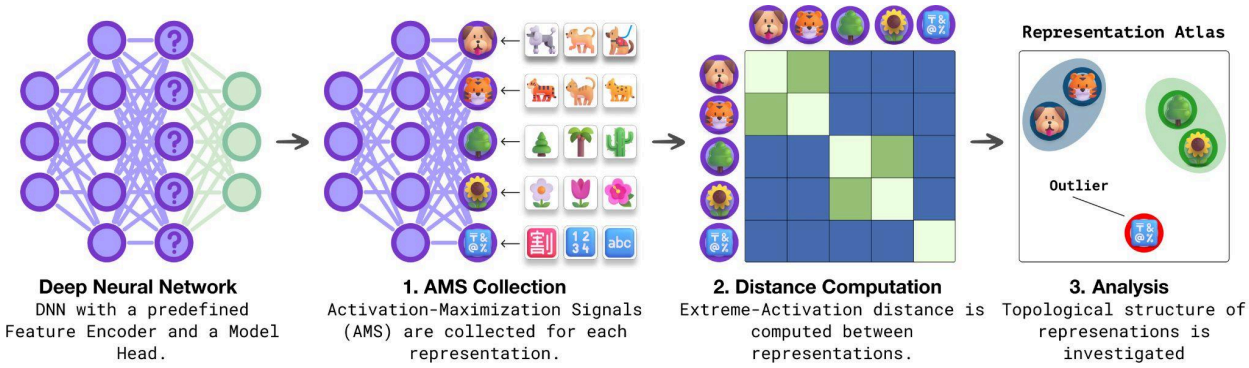

Figure 5. Schematische Darstellung der DORA Methode zur datenunabhängigen Repräsentationsanalyse: Ausgehend von einem vorausgewählten Layer werden (1) Feature Visualisierungsbilder für jede Repräsentation erzeugt, (2) die Extreme Activation Distance berechnet und (3) die Repräsentationen auf Ausreißer-Neuronen hin analysiert.

INVERT – Semantische Benennung von Neuronen durch inverse Erkennung

INVERT (*INVERSE REcogniTion*) wurde auf der NeurIPS-Hauptkonferenz 2023 vorgestellt und verfolgt ein komplementäres Ziel zu DORA: Statt Ausreißer auf Bildern zu detektieren, zielt INVERT darauf ab, Neuronen mit textuellen Erklärungen zu beschreiben – ohne dabei auf Segmentierungsmasken zurückgreifen zu müssen, wie es bei vorherigen Methoden (z. B. Network Dissection) notwendig war. Die Methode entstand in Zusammenarbeit von TU Berlin, ATB Potsdam, RPTU Kaiserslautern-Landau, UiT, Universität Potsdam und BIFOLD.

FSCA – Analyse semantisch-funktionaler Kontraste

Mit FSCA (*Function-Semantic Contrast Analysis*), vorgestellt auf der XAI World Conference 2023 (Lissabon, Vortrag), wurde eine neue Methode präsentiert, um Diskrepanzen zwischen der semantischen und funktionalen Ähnlichkeit von Neuronen aufzudecken. FSCA zeigt insbesondere auf, ob ein Modell semantisch falsche

Korrelationen erlernt hat – etwa wenn ein Netzwerk die Klasse "Monitor" nicht über das Objekt selbst, sondern über häufig auftretende Wasserzeichen im Bild erkennt.

Diese Methodik, entwickelt von einem Konsortium aus TU Berlin, ATB Potsdam, UiT, BIFOLD und der Universität Potsdam, bietet ein generisches Werkzeug zur Aufdeckung interner Verzerrungen, und stellt damit einen bedeutenden Beitrag für die Auditierung sicherheitskritischer Systeme dar.

Mark My Words – Spurious-Konzepte durch Wasserzeichen in ImageNet

Dieses Projekt widmete sich explizit der Herausforderung von der Verbreitung von Wasserzeichen in ImageNet-Bildern. Veröffentlicht auf der ECAI 2023 (XI-ML Workshop, Krakau), zeigte die Studie, dass eine Vielzahl von vortrainierten Modellen nicht nur für bestimmte Klassen (wie „cartoon“) verdächtige Konzepte lernen, sondern systematisch auf Textartefakte wie chinesische Wasserzeichen ansprechen – auch in Kategorien wie „monitor“, „broom“ oder „safe“.

***Gradient Slingshots* – Verwundbarkeit von Feature-Visualisierungen**

Feature Visualisierungen (FV) gelten als eines der Standardwerkzeuge zur globalen Erklärbarkeit neuronaler Netzwerke. Die auf der ICML 2024 (Mechanistic Interpretability Workshop, Wien) vorgestellte Arbeit *Gradient Slingshots* zeigt, dass bereits kleinste Veränderungen im Eingaberaum dazu genutzt werden können um FVs zu manipulieren, ohne das zugrundeliegende Modell selbst zu verändern. Die Studie zeigt nicht nur die Verwundbarkeit synthetischer Visualisierungen auf, sondern entwickelt auch Gegenmaßnahmen, um die Robustheit solcher Methoden zu erhöhen. Damit wurde ein neuer Standard für die Absicherung globaler XAI Visualisierungsverfahren gesetzt.

1.2.3 XAI Anwendungen in den Wissenschaften

Explainable AI (XAI) leistet nicht nur einen fundamentalen Beitrag zum Verständnis von tiefen neuronalen Netzen, sondern entfaltet auch zunehmend transformative Wirkung in angewandten wissenschaftlichen Domänen, wie beispielsweise der Klimaforschung und dem ökologischen Monitoring. Die hier entwickelten Methoden erstrecken sich über alle Arbeitspakete in dem Projekt Explaining4.0. Sie wurden gemeinsam mit führenden europäischen Institutionen realisiert, darunter die TU Berlin, das ATB Potsdam, die Universität Potsdam, die Universität Leipzig, die University of Reading, die UiT Arctic University of Norway sowie das Alan Turing Institute.

Auswahl erklärbarer Methoden mittels XAI-Evaluation in der Klimawissenschaft

Veröffentlicht 2024 auf der AIES-Konferenz, stellt dieses Projekt einen Meilenstein in der wissenschaftlichen Nutzung von XAI in der Schnittstelle Klima und KI dar. In Kooperation zwischen der TU Berlin, dem ATB Potsdam, der Universität Potsdam, der Universität Leipzig, der University of Reading, der UiT und BIFOLD Berlin adressiert die Studie die zentrale Herausforderung der Auswahl und Evaluierung von Erklärmethoden in der Klimawissenschaft. Das Projekt integrierte das Quantus-Toolkit in bestehende Anwendungen der Klimawissenschaften und ermöglichte dadurch eine systematische Evaluierung von Erklärmethoden hinsichtlich Robustheit, Faithfulness, Komplexität, Lokalisierbarkeit und Randomisierung. Es handelt sich um eine der ersten umfassenden XAI-Evaluierungen für klimabezogene KI Anwendungen. Im Fokus standen Methoden wie Integrated Gradients, Layerwise Relevance Propagation (LRP), SmoothGrad sowie verschiedene KI-Architekturen. Diese Arbeit liefert eine fundierte Grundlage für die Auswahl geeigneter XAI-Methoden, stärkt damit die Vertrauenswürdigkeit von KI-Modellen in der Klimaforschung und sensibilisiert für das bislang unterschätzte Risiko potenziell widersprüchlicher Erklärungsergebnisse.

Erklärbare KI für das Biodiversitätsmonitoring in europäischen Grünlandflächen

EU-Richtlinien koppeln Förderungen für FarmerInnen an das Vorkommen von Indikatorpflanzen auf den Feldern. Die manuelle Erhebung von den Pflanzenvorkommen sind teuer, weswegen eine automatische KI gesteuerte Erkennung wünschenswert ist. Dieses Projekt kombiniert das EfficientDet-Objekterkennungsmodell mit XAI-Metriken, um die Zuverlässigkeit von Vorhersagen zu prüfen und robustere Modellentscheidungen zu gewährleisten.

Quantus-X-Climate: Demokratisierung von XAI in der Klimaforschung

Ein besonderes Highlight stellt das **Quantus-X-Climate-Tutorial** dar, das beim ICLR Climate Change AI Workshop (Kigali, Spotlight Oral + Poster) präsentiert wurde. Anders als klassische Fachbeiträge, die neue Methoden entwickeln oder bewerten, vermittelt dieses Open-Source-Tutorial Klimaforscher:innen konkret, wie man erklärbare KI (XAI) verantwortungsvoll einsetzt.

Vorhersage von Subseasonal-to-Seasonal (S2S) Klimaereignissen

In dem Projekt wurden Modelle für zuverlässige S2S-Vorhersagen für Zeiträume von zwei Wochen bis zwei Monaten erstellt. Veröffentlicht wurde die Arbeit in *Machine Learning for Earth*, und stellt neuartige Deep-Learning-Architekturen vor, die großräumige Klima Muster, sogenannte Telekonnektionen, wie den stratosphärischen Polarwirbel (SPV) und die Madden-Julian-Oszillation (MJO) explizit modellieren. Diese beeinflussen maßgeblich das Wetter im Nordatlantik-Europa-Raum. Das Projekt entspricht direkt den Zielen von WP1 und

WP4, da es auf physikalisch interpretierbare Modellarchitekturen fokussiert. Drei Modelle wurden verglichen: ein reines LSTM, ein SPV/MJO-index-gestütztes LSTM sowie ein hybrides ViT-LSTM, das Vision Transformer mit LSTM kombiniert, um Rohdaten wie zonale Wind- und Strahlungsfelder zu analysieren. Erklärverfahren ermöglichten zudem eine Rückverfolgung der Modellentscheidungen zu bekannten physikalischen Phänomen und schlugen somit eine Brücke zwischen Black-Box-Vorhersagen und prozessorientierter Klimaforschung.

1.2.4 Zusätzliche Grundlagenforschung

Ein zentraler Schwerpunkt von WP4 lag auf der Entwicklung robuster und unsicherheitsbewusster XAI Methoden, woraus die stochastischen Verfahren NoiseGrad und FusionGrad hervorgingen, die Unsicherheiten gezielt in den Erklärungsprozess einführen und damit die bekannte Anfälligkeit gradientenbasierter Methoden beheben. Durch die Kombination von Rauschen in sowohl Gewichts- als auch Eingaberaum liefern sie robuster interpretierbare Erklärungen.
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).

Verwertung der Ergebnisse
Ein zentrales Ergebnis des Projekts *Explaining 4.0* sind die entwickelten Softwaretools, die sich als Standard in der Forschung zu erklärbarer KI (XAI) etabliert haben. Unsere Open-Source-Tools dienen der robusteren Evaluation, vebesserter neuronale Interpretierbarkeit und dem Transfer in praxisrelevante Anwendungen – insbesondere in der Klimaforschung. Mit über 55.000 Downloads und über 500 GitHub-Sternen ist Quantus eines der führenden Toolkits zur Evaluierung von Erklärverfahren. Es wurde durch spezialisierte Module wie MetaQuantus (Evaluation der Evaluatoren), GEF (geometriebasierte Treuebewertung), Quanda (Trainingsdaten-Attribution) und CoSy (textuelle Neuroneninterpretation) ergänzt. Darüber hinaus entwickelten wir robuste Verfahren wie NoiseGrad und FusionGrad sowie Analysewerkzeuge wie DORA und INVERT zur Repräsentationsanalyse. Spezifische Tools wie Quantus X Climate adressieren domänenspezifische Herausforderungen in der Klimamodellierung. Unsere entwickelten Methoden wurden in führenden Konferenzen und Journalen wie JMLR, TMLR, AAAI und NeurIPS veröffentlicht und sind offen zugänglich, dokumentiert und der code auf Reproduzierbarkeit ausgelegt, was einn zentralen Beitrag zu transparenter und vertrauenswürdiger KI liefert.

Ergänzend dazu hat das *Understandable Machine Intelligence Lab (UMI)* an der Universität Potsdam – in enger Kooperation mit dem ATB, der TU Berlin, Fraunhofer HHI und BIFOLD – ein starkes Netzwerk für Forschung, Lehre und Transfer aufgebaut. Mit über 25 eingeladenen Vorträgen, internationalen Kooperationen (u. a. Alan Turing Institute, Cornell, OIST) und aktiver Beteiligung an der akademischen Lehre (z. B. AIDA, Summer Schools, Seminar für KI in der

Bioökonomie) wurden Methoden und Erkenntnisse breit disseminiert. So trägt Explaining4.0 nachhaltig zur Weiterentwicklung erklärbarer, robuster und verantwortungsvoller KI bei.

Nachfolgend sind alle Veröffentlichungen von dem Projekt Explaining4.0 aufgelistet:
bold face = author under "Explaining 4.0" project
(*) shared the first authorship

Hedström A., **Bommer** P.L., Burns T.F., Lapuschkin S., Samek W., **Höhne** M.M.-C. Evaluating Interpretable Methods via Geometric Alignment of Functional Distortions. Transactions on Machine Learning Research (2025).

Bommer P.L., Kretschmer M., Spuler F.R., **Bykov** K., **Höhne** M.M.-C. Deep Learning Meets Teleconnections: Improving S2S Predictions for European Winter Weather. arXiv preprint arXiv:2504.07625 (2025).

Eiras-Franco C., **Hedström** A., **Höhne** M.M.-C. Evaluate with the Inverse: Efficient Approximation of Latent Explanation Quality Distribution. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 27258–27267 (2025).

Gonçalves T., **Hedström** A., de Mortanges A.P., Li X., Müller H., Cardoso J.S., Reyes M. Interpretable AI for medical image analysis: methods, evaluation, and clinical considerations. *Trustworthy AI in Medical Imaging*, 315–346 (2025).

Kopf L., **Bommer** P.L., **Hedström** A., Lapuschkin S., **Höhne** M.M.-C., **Bykov** K. Cosy: Evaluating textual explanations of neurons. *Advances in Neural Information Processing Systems* **37**, 34656–34685 (2024).

Wickstrøm K.K., **Höhne** M.M.-C., **Hedström** A. The Price of Freedom: An Adversarial Attack on Interpretability Evaluation. NeurIPS Workshop on Interpretable AI (2024).

Bareeva D., Yolcu G.Ü., Hedström A., Schmolenski N., Wiegand T., Samek W., Lapuschkin S. Quanda: An Interpretability Toolkit for Training Data Attribution Evaluation and Beyond. arXiv preprint arXiv:2410.07158 (2024).

Wickstrøm K., **Höhne** M.M.-C., **Hedström** A. From Flexibility to Manipulation: The Slippery Slope of XAI Evaluation. ECCV Workshop (2024).

Hedström A., Weber L., Krakowczyk D., Bareeva D., Motzkus F., Samek W., Lapuschkin S., **Höhne** M.M.-C. Sanity Checks Revisited: An Exploration to Repair the Model Parameter Randomisation Test. NeurIPS Workshop XAI in Action (2024).
Hedström A., Weber L., Lapuschkin S., **Höhne** M. A Fresh Look at Sanity Checks for Saliency Maps. World Conference on Explainable Artificial Intelligence, 403–420 (2024).

Bommer P.L., Kretschmer M., **Hedström** A., Bareeva D., **Höhne** M.M.-C. Finding the right XAI method: a guide for the evaluation and ranking of explainable AI methods in climate science. Artificial Intelligence for the Earth Systems 3, e230074 (2024).

Bareeva D., **Höhne** M.M.-C., Warnecke A., Pirch L., Müller K., Rieck K., **Bykov** K. Manipulating Feature Visualizations with Gradient Slingshots. AI Safety workshop. *International Machine Learing Conference* (2024).

Bommer P.L., Kretschmer M., Boehnke P., **Höhne** née Vidovic M.M.-C. Using spatio-temporal neural networks to investigating teleconnections and enhance S2S forecasts of European extreme weather. *EGU General Assembly Conference Abstracts*, 15174 (2024).

Bommer P.L., Kretschmer M., Boehnke P., **Höhne** M.M.-C. Using spatio-temporal neural networks and XAI to investigating teleconnections in S2S forecasts of European extreme weather. *AGU24* (2024).

Malounas, I. Vierbergen, W. **Kutluk**, S. Zude-Sasse, M. Yang, K. Zhao, M. Argyropoulos, D. Van Beek, J. Ampe, E. Fountas, S. SpectroFood dataset: A comprehensive fruit and vegetable hyperspectral meta-dataset for dry matter estimation (2024) p. 110040

Yu, X., Franzen, J., Samek, W., **Höhne,** M.M.-C., Kainmueller, D. Model guidance via explanations turns image classifiers into segmentation models. In *World Conference on Explainable Artificial Intelligence* (pp. 113–129). Springer Nature Switzerland Cham. (2024).

Kazmierczak, R., Azzolin, S., Berthier, E., **Hedström**, A., Delhomme, P., Bousquet, N., & Franchi, G. (2024). Benchmarking XAI Explanations with Human-Aligned Evaluations. arXiv preprint arXiv:2411.02470. (2024).

Hedström A., **Bommer** P., Wickstrøm K.K., Samek W., Lapuschkin S., **Höhne** M.M.-C. The Meta-Evaluation Problem in Explainable AI: Identifying Reliable Estimators with MetaQuantus. *Transactions on Machine Learning Research* (2023).

Liu S., **Hedström** A., Basavegowda D.H., Weltzien C., **Höhne** M.M.-C. Explainable AI in Grassland Monitoring: Enhancing Model Performance and Domain Adaptability. *GIL 143* (2023).

Grinwald D., **Bykov** K., Nakajima S., **Höhne** M.M.-C. Visualizing the Diversity of Representations Learned by Bayesian Neural Networks. *Transactions on Machine Learning Research* (2023).

(*) Bommer P.L., **(*) Hedström** A., Kretschmer M., Höhne M.M.-C. Tutorial: Quantus x Climate - Applying explainable AI evaluation in climate science. *ICLR 2023 Workshop on Tackling Climate Change with Machine Learning* (2023).

Bykov K., Deb M., Grinwald D., Müller K., **Höhne** M.M.-C. DORA: Exploring Outlier Representations in Deep Neural Networks. *Transactions on Machine Learning Research* (2023).

Hedström A., Weber L., Lapuschkin S., Samek W., **Höhne** M.M.-C. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond. *Journal of Machine Learning Research* 24, 1–11 (2023).

Kopf L., **Bommer** P.L., **Hedström** A., Lapuschkin S., **Höhne** M.M.-C., **Bykov** K. CoSy: Evaluating Textual Explanations of Neurons. *Advances in Neural Information Processing Systems* 37 (2023).

Bykov K., Kopf L., **Höhne** M.M.-C. Finding spurious correlations with function-semantic contrast analysis. *World Conference on Explainable Artificial Intelligence*, 549–572 (2023).

Hanfeld, P., Wahba, K., **Höhne,** M.M.-C., Bussmann, M., Hönig, W. Kidnapping Deep Learning- based Multirotors using Optimized Flying Adversarial Patches. In *2023 International Symposium on Multi-Robot and Multi-Agent Systems (MRS)* (pp. 78–84). IEEE (2023).

Gautam, S., **Höhne,** M.M.-C., Hansen, S., Jenssen R., Kampffmeyer M. This looks more like that: Enhancing Self-Explaining Models by Prototypical Relevance Propagation. *Pattern Recognition*, vol. 136, p. 109172, (2023).

Höhne, M.M.-C., Raut, S., Atzmueller, M., & Sturm, B. Explainable AI Methods for Multimodal Data in the Circular Bioeconomy. In *Presented at the Circular Bioeconomy Systems (CBS) Day, ASABE Annual International Meeting (AIM)* (2023).

Bommer P.L., Kretschmer M., **Hedström** A., Bareeva D., Höhne M.M.-C. Evaluation of explainable AI solutions in climate science. *EGU General Assembly Conference Abstracts*, EGU-12528 (2023).

(*) **Bykov** K., (*) **Hedström** A., Nakajima S., Höhne M.M.-C. NoiseGrad — Enhancing Explanations by Introducing Stochasticity to Model Weights. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 6132–6140 (2022).

Gautam, S., Boubekki, A., Hansen, S., Salahuddin, S., Jenssen, R., **Höhne,** M.M.-C., Kampffmeyer, M. Protovae: A trustworthy self-explainable prototypical variational model. *Advances in Neural Information Processing Systems*, 35, 17940–17952 (2022).

Gautam, S., **Höhne,** M.M.-C., Hansen, S., Jenssen, R., Kampffmeyer, M. Demonstrating The Risk of Imbalanced Datasets in Chest X-ray Image-based Diagnostics by Prototypical Relevance Propagation. In 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) (pp. 1-5). IEEE. (2022).

Abschließend die tabellarische Auflistung der wichtigsten Positionen:

Wichtigste Positionen	
Personal	524.441,43
Studentische HK	45.234,12
Sachausgaben	47.183,00
Reisen	40.487,34
Gesamt	657.345,89