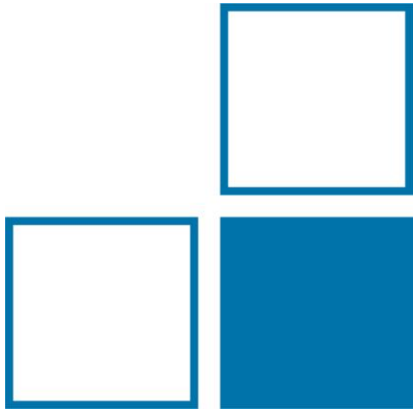
Physikalisch-Technische Bundesanstalt
Braunschweig und Berlin

Technische Universität Berlin

Schritte in Richtung vertrauenswürdiger künstlicher Intelligenz: Forschung und Regulierung

Prof. Dr. Stefan Haufe

QI Digitalforum
11. Oktober 2023



Stefan Haufe

2011: Promotion zum (Dr. rer. nat.) in **Informatik / ML**, TU Berlin

2013-14: Postdoc City College New York

2014-16: Marie Curie Fellow Columbia University und TU Berlin

Seit 2019: Forschungsgruppenleiter (ERC grant) an der Charité Neurologie

Seit Mai 2021:

- **W2 Professor TU Berlin, Fakultät IV EECS**
- **Leiter AG 8.44 (Maschinelles Lernen und Unsicherheit) der PTB**

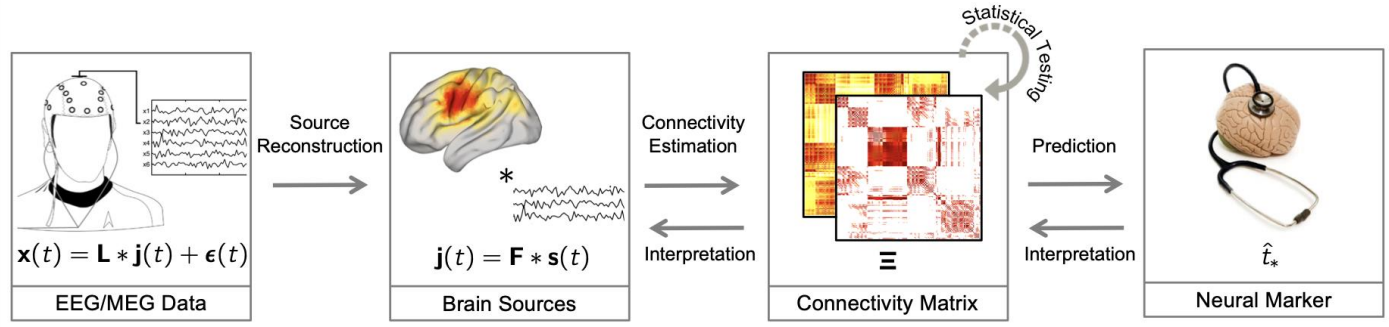
Forschungsthemen

Erklärbarkeit und Qualitätssicherung für Maschinelles Lernen/KI

- Methodenentwicklung und Validierung, Theoriea§
- Use cases: Neuroimaging und NLP

Maschinelles Lernen und Signalverarbeitung in der Neurobildgebung

- MEG/EEG Quellenrekonstruktion
- Zeitreiheninteraktionen und Signallaufzeiten
- ➢ Biomarker zur Charakterisierung neurologischer Krankheiten



Maschinelles Lernen für Intensivstationsdaten

- Z.b. Vorhersage von postoperativem Delir und Mortalität bei Covid19

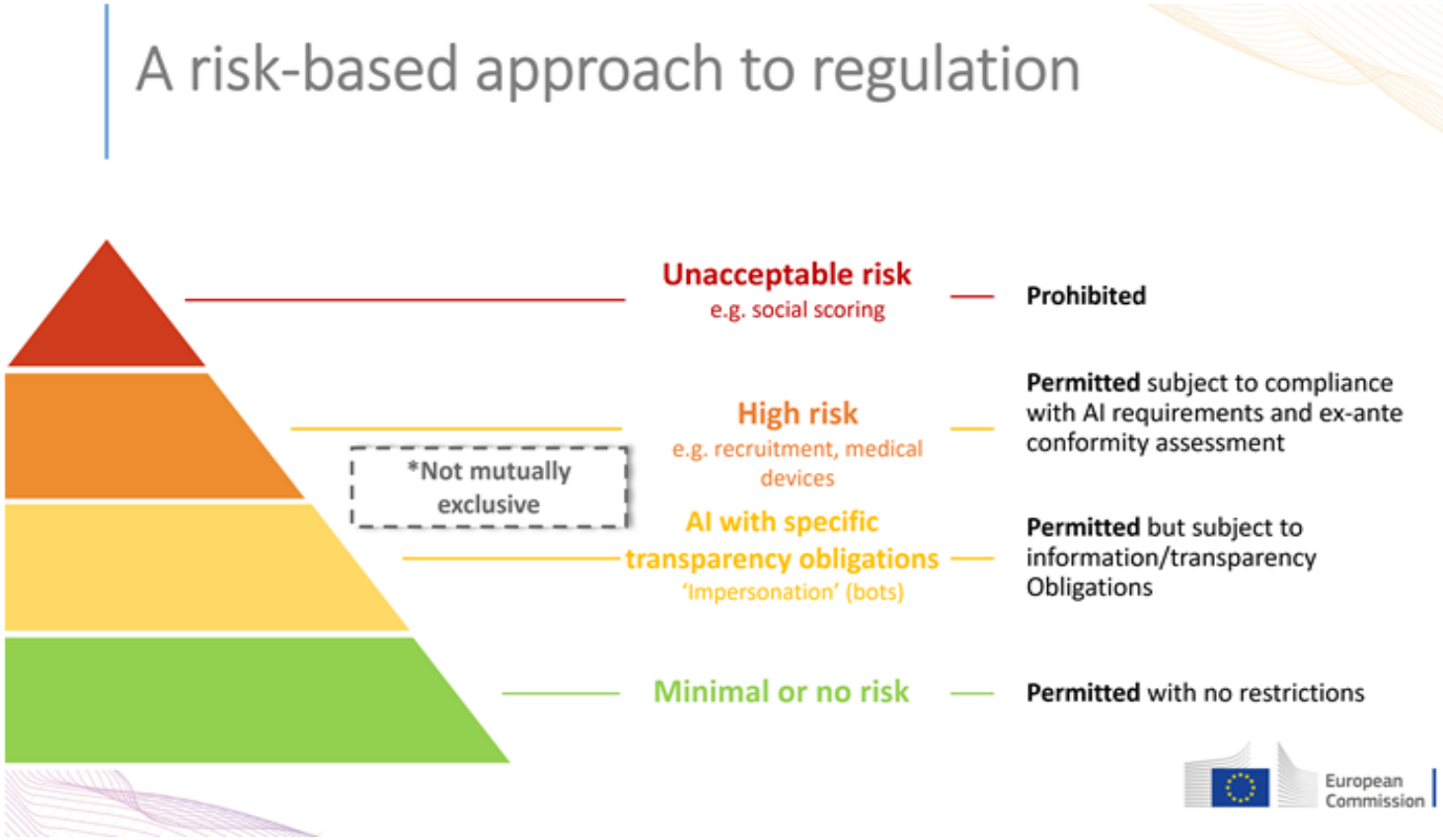
Kritische Anwendungen von KI

KI = transformative Technologie, die in nahezu allen Lebensbereichen gewinnbringend anwendbar ist

- **Katalysatoren:** Digitalisierung, Daten, Rechenleistung, Algorithmen
- **Viele Anwendungen bergen jedoch Risiken**
  - Medizin
  - Finanzen (z.B. Kredite)
  - Recht/Policing (z.b. Gesichtserkennung, Sozialprognose)
  - Mobilität/autonomes Fahren
  - Personalwesen
  - Sicherheitskritische Industrie und Infrastruktur
- **Probleme**
  - Keine formale Verifikation
  - Hohe Komplexität, datengetrieben
  - Fehlverhalten schwer erkennbar/vorhersehbar

Regulierung

- Einsatz von KI in Produkten benötigt Prüfung/Zertifizierung
- AI Act: europäische Gesetzgebung für KI (ab 2025)



AI Act: Anforderungen an Hochrisikosysteme

- Article 9 — Risk management system
- Article 10 — Data and data governance
  - ➢ **Datenschutz**
- Article 11 — Technical documentation
- Article 12 — Record-keeping
- Article 13 — Transparency and provision of information to users
  - ➢ **Erklärbarkeit, Unsicherheitsangabe**
- Article 14 — Human oversight
- Article 15 — Accuracy, robustness and cybersecurity

➢ **CEN/CENELEC mit Normung beauftragt**

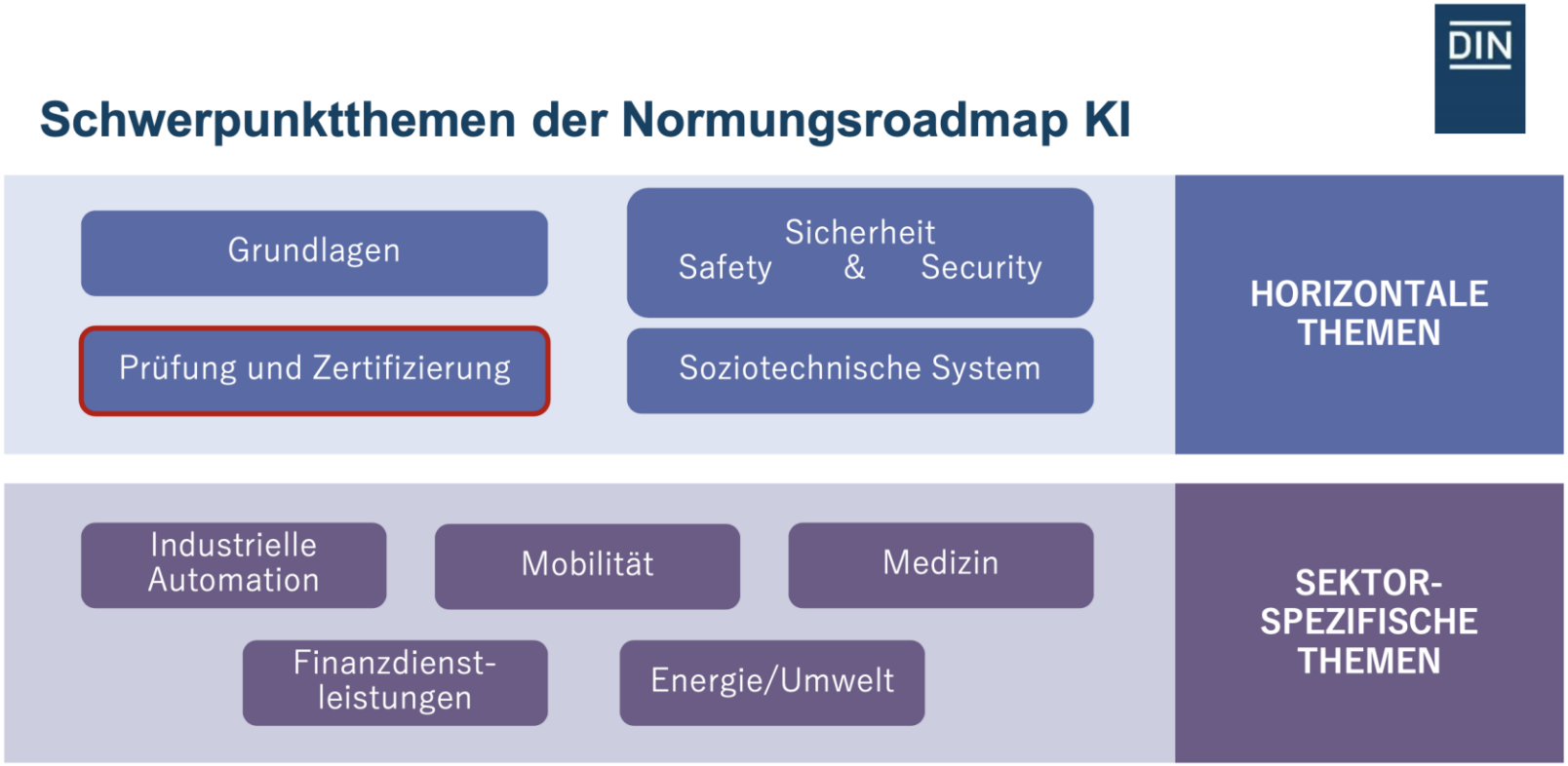
➢ **Aber: nicht nur Aufgaben für die Normung sondern auch die Forschung**

Liste der zu erarbeitenden neuen europäischen Normen und europäischen Normungsunterlagen

	Referenzangaben
1.	Europäische Norm(en) und/oder europäische Normungsunterlage(n) zu Risikomanagementsystemen für KI-Systeme
2.	Europäische Norm(en) und/oder europäische Normungsunterlage(n) zur Governance und Qualität von Datensätzen, die zur Entwicklung von KI-Systemen verwendet werden
3.	Europäische Norm(en) und/oder europäische Normungsunterlage(n) zur Aufzeichnung durch Protokollierungsfunktionen von KI-Systemen
4.	Europäische Norm(en) und/oder europäische Normungsunterlage(n) zur Transparenz und Information der Nutzer von KI-Systemen
5.	Europäische Norm(en) und/oder europäische Normungsunterlage(n) zur menschlichen Aufsicht über KI-Systeme
6.	Europäische Norm(en) und/oder europäische Normungsunterlage(n) zu Spezifikationen für die Genauigkeit von KI-Systemen
7.	Europäische Norm(en) und/oder europäische Normungsunterlage(n) zu Spezifikationen für die Robustheit von KI-Systemen
8.	Europäische Norm(en) und/oder europäische Normungsunterlage(n) zu Spezifikationen für die Cybersicherheit von KI-Systemen
9.	Europäische Norm(en) und/oder europäische Normungsunterlage(n) zu Qualitätsmanagementsystemen für Anbieter von KI-Systemen, einschließlich Verfahren zur Beobachtung nach dem Inverkehrbringen
10.	Europäische Norm(en) und/oder europäische Normungsunterlagen zur Konformitätsbewertung für KI-Systeme

DIN/DKE Normungsroadmap KI 2.0 (2022)

- Maßnahme der KI-Strategie der Bundesregierung
- Erfassung von Themen und Handlungsbedarfen für die KI-Normung



DIN/DKE Normungsroadmap KI 2.0 (2022)

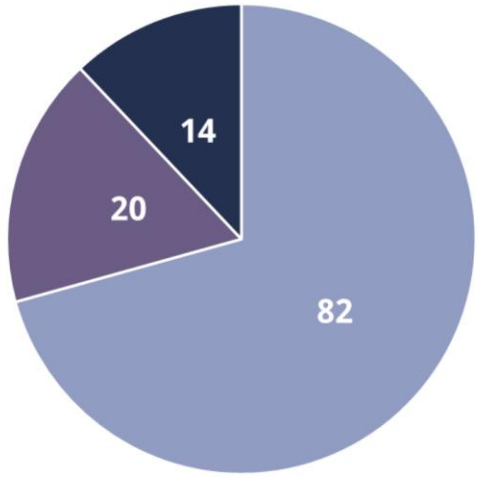
Ergebnisse der NRM KI (Ausgabe 2)

Insgesamt 116 Handlungsbedarfe identifiziert, unterteilt in 3 Kategorien:

KATEGORIE 1:
Bedarf adressiert Normung und Standardisierung

KATEGORIE 2:
Bedarf adressiert Forschung

KATEGORIE 3:
Bedarf adressiert Politik/Gesetzgeber



Stand: Mai 2023

DIN/DKE Normungsroadmap KI 2.0 (2022)

Bedarf 03-… aus Prüfung und Zertifizierung

- 03-01 Spezifikation von formalen Anforderungen an „explainable“ AI („XAI“)-Methoden
- 03-02 Operationalisierung der „Erklärgüte“ von XAI-Methoden
- 03-03 Entwicklung eines Standards mit Guidance-Dokumenten für die Abbildung von Risiken eines Systems in die Funktionalität von KI-Komponenten
- 03-04 Entwicklung von Funktionalitätsklassen für KI-Technologien
- 03-05 Entwicklung von Werkzeugkriterien für die Prüfung von KI-Systemen
- 03-06 Entwicklung von ineinandergreifenden Standards für KI-Systeme und notwendiger Konformitätsbewertungsverfahren
- 03-07 Entwicklung von Qualifikationskriterien für Prüfer und Zertifizierter zu Cybersecurity und Privacy für KI (In Bearbeitung: ISO/IEC 42006)
- 03-08 Vernetzung aller Akteur*innen
- 03-09 Definition von Kontrollpunkten

DIN/DKE Normungsroadmap KI 2.0 (2022)

Bedarf 03-01 (Forschungsbedarf)

Spezifikation von formalen Anforderungen an „explainable“ AI („XAI“)-Methoden

Formulierung konkreter operationalisierbarer/prüfbarer Anforderungen an XAI-Methoden. Welche formalen Aussagen sollen anhand der Ergebnisse einer XAI-Methode möglich sein?

- Die Trainingsdaten betreffend?
- Das Testdatum betreffend?
- Das Modell betreffend?
- Den Zusammenhang zwischen Ein- und Ausgabedaten (Prädiktionen) betreffend?
- Den Zusammenhang zwischen Modell, Ein- und Ausgabedaten betreffend?

Welche praktischen Konsequenzen sollen sich sicher aus diesen Aussagen ableiten lassen? Welcher Mehrwert an „Verlässlichkeit“ soll wirklich geschaffen werden, und wie kann er nachgewiesen werden?

Eine sektorübergreifende und auch im Entwurf AI Act verankerte Forderung ist die nach „Erklärbarkeit“, „Interpretierbarkeit“ etc. […] Dementsprechend ist die Validierung/Verifikation dieser Methoden tendenziell oft eher qualitativ, subjektiv und zirkulär. Formale Kriterien sind notwendig, um zu spezifizieren, welche Aussagen / praktischen Konsequenzen auf Basis des Ergebnisses einer gegebenen XAI-Methode korrekt und zulässig sind. Die Einhaltung dieser Kriterien muss formal oder empirisch verifiziert werden. Nur so können Fehlinterpretationen vermieden werden.

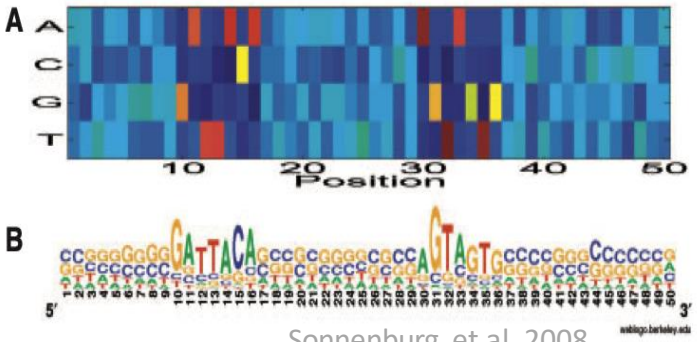
„Erklärbare“ künstliche Intelligenz (XAI)

EU AI Act: KI Entscheidungen in Hochrisikoanwendungen sollen „erklärbar“ sein.

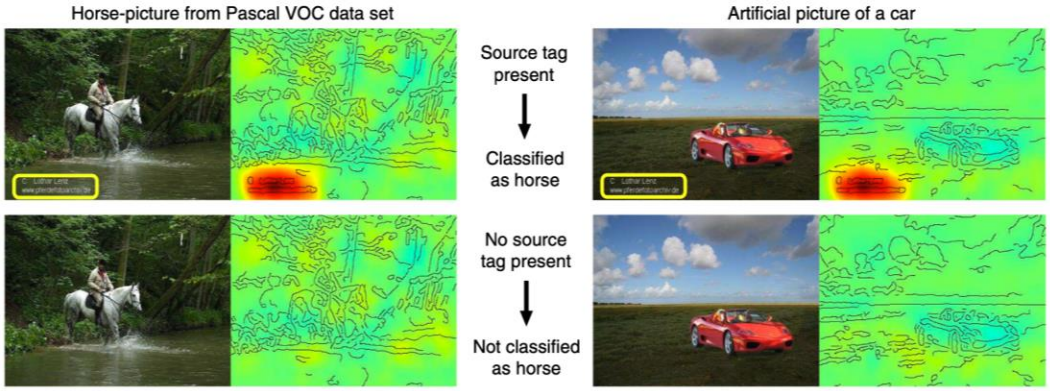
- Seit ca. 15 Jahren zahlreiche XAI Methoden publiziert (teils > 10K Zitationen)

Wesentliche Versprechen:

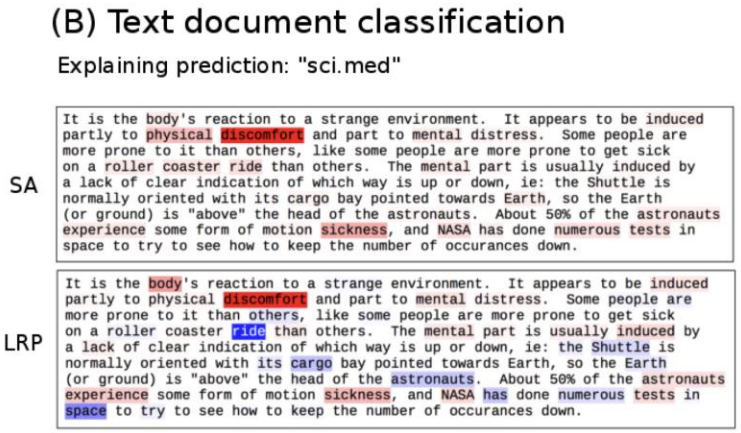
- „Validierung“ oder „Debugging“ von Daten und Modellen
- Scientific Discovery



Sonnenburg et al. 2008



Lapuschkin et al. 2019



Samek et al. 2021

„Erklärbare“ künstliche Intelligenz (XAI)

Aber: fehlende Formalisierung, Validierung

- Welche konkreten Probleme soll XAI lösen?
- Welche Informationen über Modell, Trainingsdaten, Eingabe bereitstellen?

➢ Validierung oft subjektiv, qualitativ, anekdotisch
➢ Fehlende formale Verifikation, Validierung bzgl. geeigneter Kriterien

Bekannte Defizite von XAI

- „Wichtige“ features haben keinerlei Bezug zur vorhergesagten Größe
  - ➢ Nutzen für Debugging/Validierung/Discovery unklar
- Störanfällig: kleinste Änderungen in Daten führen zu beliebigen Erklärungen
- …

➢ XAI Methoden (noch) nicht zur Qualitätssicherung geeignet
➢ Müssen selbst geprüft werden
➢ Anforderungskonforme Methoden müssen noch entwickelt werden

DIN SPEC 92001-3 Artificial intelligence – Life cycle processes and quality requirements – Part 3: Explainability

„This is the third document in a series, and it aims to ensure that AI systems are developed, deployed, and used efficiently, responsibly, and in a trustworthy way. It focuses on "Explainability" – the ability to understand how AI makes decisions. This DIN SPEC 92001-3 provides a domain-independent guide on promoting explainability throughout the AI system's life cycle."

- Teil des Projektes „Zertifizierte KI" , gefördert durch das Land NRW
- Ziel: Umsetzung der Bedarfe der Normungsroadmap
- Enthält keine Empfehlungen bzgl. bestimmter Methoden
- **Stattdessen:** Prozessmodell zur Etablierung von Erklärbarkeit
  - Formulierung use-case und stakeholder-spezif. Anforderungen an XAI
  - Auswahl bzw. Entwicklung geeigneter Verfahren
  - Validierung und Verifizierung im Anwendungskontext

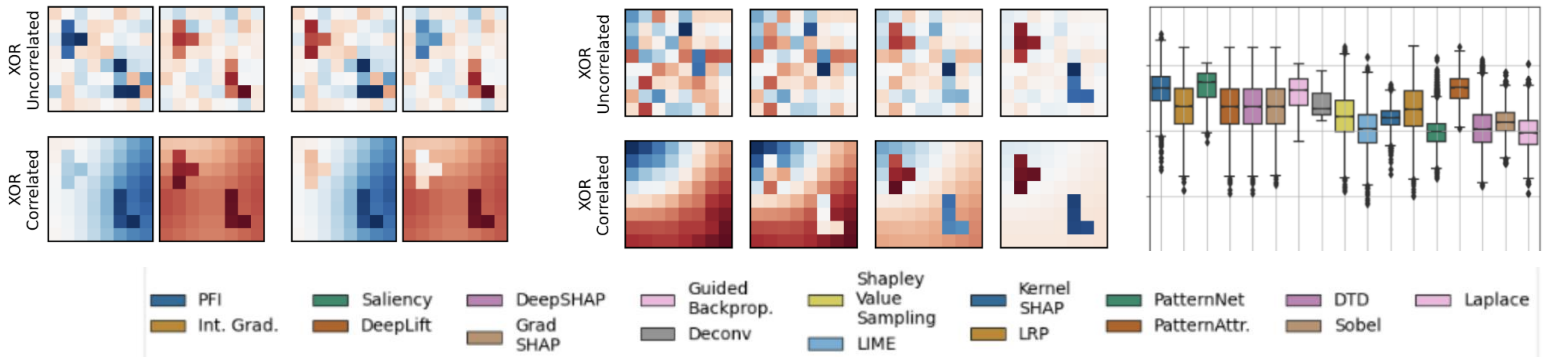
PTB Arbeiten im Bereich „erklärbare“ KI

- Theoretische Analyse der Korrektheit von XAI Methoden
  Haufe et al., 2014; Wilming et al., 2022, 2023
- Entwicklung formaler Definitionen von Erklärbarkeit, Referenzdaten, Gütemaßen, Benchmarks

XAI method	Zero Attribution to Suppressors
Gradient	
Pattern	✓
Faithfulness (Pixel Flipping)	
Permutation Importance	
Partial Dependency Plot	
Marginal Plot	✓
Shapley Values (R^2)	✓
SHAP (Marginal Expectation)	
SHAP (Conditional Expectation)	
Counterfactual	
FIRM	✓
Integrated Gradient	
LIME	
Saliency map	
PatternNet/PatternAttribution	✓

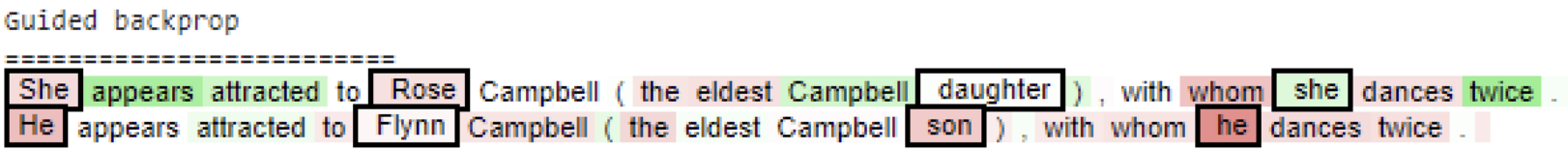
Benchmark Suite zur empirischen Prüfung von XAI Methoden

- **XAI-TRIS** – *Synthetische ground-truth Daten zur* **Messung der Erklärgüte**

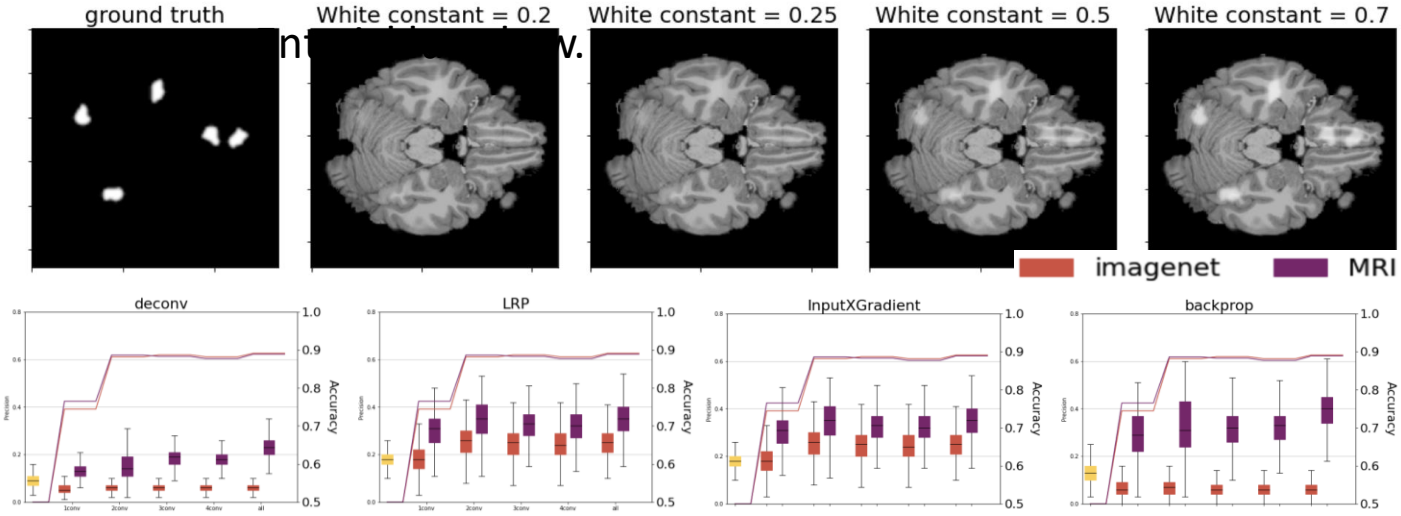


PTB Arbeiten zu „erklärbarer KI“ und Fairness

- Ground-truth NLP Textkorpus zum Studium von Fairness und Erklärbarkeit



- Einfluss von pretraining auf Erklärgüte bei synthetischen MRT Daten



PTB Arbeiten zu Unsicherheitskalibrierung

- **Ziel:** Zuverlässige & skalierbare Methoden zur Quantifizierung von Unsicherheiten für Vorhersagen trainierter neuronaler Netze.
- Kalibrierung bedeutet: Netzwerk lernt korrekte Unsicherheit
- Benchmark mit ground-truth Daten

Schmähling et al., Appl. Intell., 2022

Comparison of different procedures for uncertainty quantification in deep regression with a statistical reference method (BLR).

F. Schmähling et al.

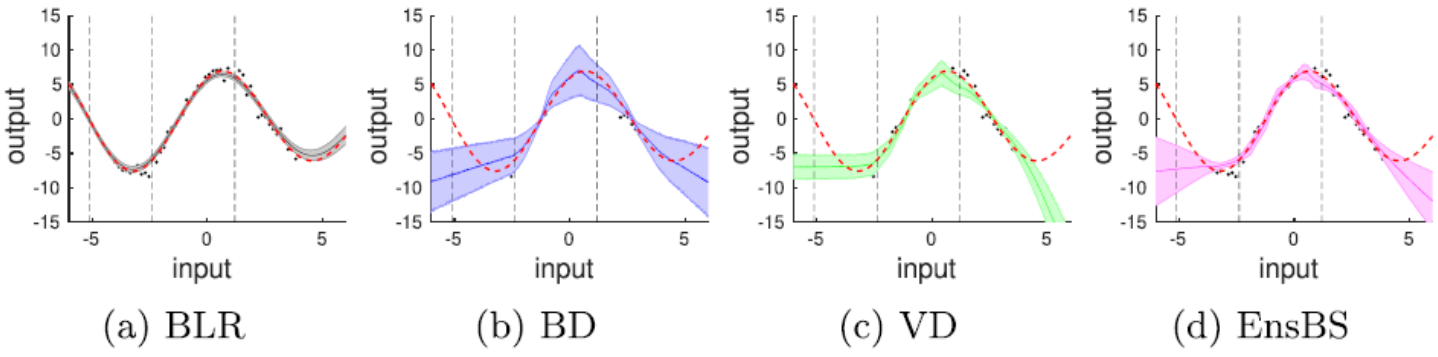


Fig. 1 Experiment E.1 with $f_{\text{main}} = 1$. Predictions (solid lines) and calculated uncertainties times 1.96 (shaded areas) of the anchor model BLR (subplot (a)) and different methods in deep learning (subplots (b)-(d)), together with the used train set (black dots) and ground truth (red dashed line). The abbreviations for the methods are as in the beginning of Section 3

PTB Arbeiten zum Thema Daten und Datenschutz

Herausforderung: Wie önnen Patientendaten zur Verfügung gestellt werden, sodass einerseits Datenschutzkonformität sichergestellt ist und andererseits alle Zugriff auf möglichst diverse Trainingsdaten haben?

Lösung: synthetische Daten

- **EPM Medalcare Projekt:** Datenbank in-silico Herzen (physikalische Modelle)
- **M4AIM P1:** Towards standardized quality control for artificial intelligence systems in critical care (generative KI, datengetriebene Modelle)

TEF Health Projekt beschäftigt sich ganzheitlich mit dem Thema Datenqualität

➢ Daniel Schwabe

Seminar „Quality Assurance for Machine Learning“ (MSc Informatik, TU Berlin)

Themen:

- **Lecture:** introduction to supervised ML and deep learning
- **Student presentations + discussion:**
  - Critical Applications of ML
  - Regulatory frameworks & standardization
  - Good ML practices (performance evaluation, model selection, regularization)
  - Data quality (e.g. missing data, imbalanced data, outliers)
  - Robustness (incl. transfer learning)
  - Uncertainty quantification and propagation (incl. Bayesian inference)
  - „Explainable AI“ and its limitations
  - Ethics and fairness
  - Privacy and data protection
  - Good research practices (software engineering, open science)
- **Homework:** AI forensics clinic
- **Group work:** designing and applying an audit process for AI systems

Zusammenfassung

- Kritische und risikobehaftete KI Anwendungen benötigen Regulierung
- In der EU wird AI Act die Benutzung von KI regeln
  - Verbot ethisch zweifelhafter Anwendungen
  - Hochrisikoanwendungen von KI unterliegen strengen Vorschriften
  - Risikoanwendungen unterliegen Transparenzvorschriften
  - Alle Anwendungen profitieren von Qualitätsstandards
- Normungsarbeit hat begonnen
- Weiter viel Forschungsbedarf
  - Erklärbarkeit
  - Unsicherheit
  - Fairness, Ethik
  - Green AI
  - Datenschutz
- XAI trägt (noch) nicht zur Qualitätssicherung von KI bei