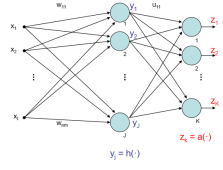



# Explainable Artificial Intelligence

**Felix Neubürger**

**2025**

Fachhochschule Südwestfalen, Ingenieurs- & Wirtschaftswissenschaften

## Abfrage Erwartungen und Vorwissen

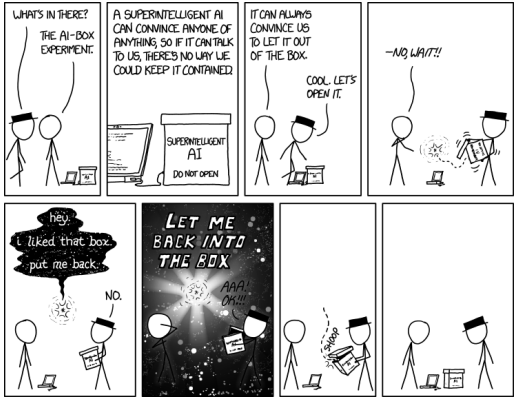
- https://www.menti.com/
- Code: 3972 7236

## Inhalte der Vorlesung

- Begriffsklärungen
- Erkenntnistheoretischer Exkurs
- Methoden der Explainable AI
- Quantitative Methoden
- Anwendung der gelernten Methoden in einem Beispiel

## Ziele der Vorlesung - Welche Fragen sollen beantwortet werden?

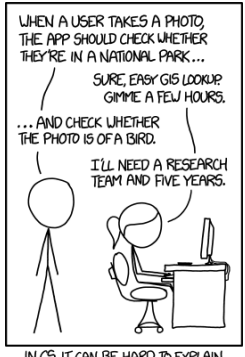
- Wofür Explainable AI?
- Was bedeutet Explainable AI?
- Interpretable AI?
- Trustworthy AI?
- Wie funktioniert das mathematisch?
- Wie schaffe ich Transparenz für Stakeholder?




[https://xkcd.com/1450/]

## Format der Vorlesung - Wie sollen diese Fragen beantwortet werden?

- Theroretischer Teil mit Folien
- Selbststudium mt einem Lehrbuch[a] [1]
- Praktischer Teil in Gruppen an einem Projekt
- Gruppengröße 2 oder 3 Personen
- Einzelarbeit möglich wenn eigenes Thema vorhanden
- Abgabe der Ausarbeitung einen Tag vor der Veranstaltung in der Blockwoche
- Vorstellung der Projektergebnisse in der Blockwoche
- Gewichtung der Bewertung Projektausarbeitung (70%) und Vortrag (30%)




[https://xkcd.com/1425/]

[a] https://christophm.github.io/interpretable-ml-book/

## Künstliche Intelligenz (KI)

- Teilgebiet der Informatik
- Automatisierung intelligenten Verhaltens
- Maschinelles Lernen als Unterbereich

## Maschinelles Lernen (ML)

- Unterbereich der KI
- Algorithmen lernen aus Daten
- Treffen von Vorhersagen oder Entscheidungen
- Deep Learning basiert auf künstlichen neuronalen Netzen

## Explainable Artificial Intelligence (XAI)

- Ansätze zur Verständlichkeit von KI-Entscheidungen
- Wichtig für Vertrauen und Transparenz
- Beispiel: Erklärungen in der medizinischen Diagnostik

## Definitionen und Unterschiede

### Interpretierbarkeit

- Fähigkeit, die internen Mechanismen eines Modells zu verstehen.
- Ermöglicht direkte Einsicht in die Funktionsweise des Modells.
- Beispiel: Lineare Regression, Entscheidungsbäume.

### Erklärbarkeit

- Fähigkeit, die Entscheidungen oder Vorhersagen eines Modells verständlich zu machen.
- Oft durch zusätzliche Methoden bei komplexen Modellen erreicht.
- Beispiel: Neuronale Netze mit Post-hoc-Erklärungen.

## Erkenntnistheoretische Aspekte

- **Wissenserwerb:** Wie tragen Interpretierbarkeit und Erklärbarkeit zum Verständnis von KI-Entscheidungen bei?
- **Vertrauen:** Inwiefern beeinflusst die Nachvollziehbarkeit von Modellen das Vertrauen der Nutzer?
- **Transparenz vs. Komplexität:** Balance zwischen detaillierter Einsicht und praktischer Anwendbarkeit.
- **Ethische Verantwortung:** Bedeutung von Erklärbarkeit für ethisch vertretbare KI-Systeme.

## Bedeutung der Interpretierbarkeit im Maschinellen Lernen

- **Definition:**
  Interpretierbarkeit bezeichnet das Maß, in dem ein Mensch die Ursache einer Entscheidung eines Modells nachvollziehen kann.

- **Warum ist Interpretierbarkeit wichtig?**
  - **Vertrauensbildung:**
    Nutzer vertrauen eher Modellen, deren Entscheidungswege sie verstehen.
  - **Fehleranalyse:**
    Verständnis für Modellentscheidungen erleichtert das Erkennen und Beheben von Fehlern.
  - **Einhaltung gesetzlicher Vorgaben:**
    In sensiblen Bereichen wie Medizin oder Finanzen sind nachvollziehbare Entscheidungen oft gesetzlich vorgeschrieben.

## Herausforderungen und Begriffsabgrenzungen

- **Herausforderungen:**
  - Fehlende einheitliche Definition von Interpretierbarkeit erschwert Kommunikation und Forschung.
  - Kompromiss zwischen Modellkomplexität und Interpretierbarkeit oft notwendig.
- **Abgrenzung zu verwandten Begriffen:**
  - **Erklärbarkeit (Explainability):**
    Fähigkeit, interne Mechanismen eines Modells verständlich zu machen.
  - **Transparenz:**
    Ausmaß, in dem die Funktionsweise eines Modells offenliegt.
  - **Vertrauen:**
    Maß, in dem Nutzer darauf vertrauen, dass ein Modell korrekte und faire Entscheidungen trifft.

## EU-Regulierung und Erklärbare Künstliche Intelligenz

Die Europäische Union hat den Artificial Intelligence Act verabschiedet,[1] der am 1. August 2024 in Kraft trat.[2]

**EU AI Act: Überblick**

- **Ziel:** Einführung eines risikobasierten Klassifizierungssystems für KI-Anwendungen.
- **Risikokategorien:**
  - **Unzulässiges Risiko:** Verbotene KI-Anwendungen.
  - **Hohes Risiko:** Strenge Anforderungen an Transparenz, Sicherheit und Compliance.
  - **Geringes oder minimales Risiko:** Weniger strenge oder keine spezifischen Anforderungen.

**Erklärbarkeit als zentrale Anforderung**

- **Transparenzpflichten:** Anbieter müssen Informationen bereitstellen, die es ermöglichen, die Funktionsweise von KI-Systemen zu verstehen.
- **Vertrauenswürdigkeit:** Erklärbare KI fördert das Vertrauen der Nutzer und erleichtert die Akzeptanz von KI-Technologien.

[1] Regulation (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 über harmonisierte Vorschriften für künstliche Intelligenz. Verfügbar unter: https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[2] Pressemitteilung der Europäischen Kommission: "AI Act tritt in Kraft". Verfügbar unter: https://ec.europa.eu/commission/presscorner/detail/de/ip_24_1234

## Herausforderungen des EU AI Acts für Unternehmen

- **Komplexität der Regulierung:**
  Der AI Act verfolgt einen risikobasierten Ansatz, bei dem KI-Systeme in verschiedene Risikoklassen eingeteilt werden. Unternehmen müssen ihre KI-Anwendungen entsprechend einstufen und die jeweiligen Anforderungen erfüllen.[3]
- **Standardisierung und technische Umsetzung:**
  Die Entwicklung harmonisierter Standards für Hochrisiko-KI-Systeme ist komplex und zeitaufwendig. Verzögerungen können zu Unsicherheiten bei der Implementierung führen und Innovationen hemmen.[4]
- **Vermeidung von Innovationshemmnissen:**
  Es besteht die Sorge, dass strenge Regulierungen Innovationen im Bereich der Künstlichen Intelligenz behindern könnten. Unternehmen müssen Wege finden, um sowohl den gesetzlichen Anforderungen zu entsprechen als auch ihre Innovationsfähigkeit zu bewahren.[5]
- **Wettbewerbsfähigkeit im internationalen Kontext:**
  Unternehmen in Regionen mit weniger strengen Vorschriften könnten schneller Innovationen umsetzen und dadurch Wettbewerbsvorteile erlangen. Europäische Firmen stehen vor der Herausforderung, trotz strengerer Regulierungen konkurrenzfähig zu bleiben.[6]

[3] https://www.it-schulungen.com/wir-ueber-uns/wissensblog/welche-anforderungen-stellt-der-eu-ai-act.html
[4] https://www.connect-professional.de/security/der-ai-act-chancen-nutzen-risiken-managen.332959.html
[5] https://www.dps-bs.de/blog/der-ai-act-weichenstellung-fuer-kuenstliche-intelligenz-in-europa/
[6] https://de.linkedin.com/pulse/der-eu-ai-act-chancen-und-herausforderungen-f%C3%BCr-andreas-quandt-ljxne

## XAI Standardwerke

- **Interpretable ML Book** (Molnar)
  - `https://christophm.github.io/interpretable-ml-book/`
  - Umfassender Leitfaden zu SHAP, LIME, etc.
- **Explainable AI** (Springer 2019)
  - DOI: 10.1007/978-3-030-28954-6
  - Forschungsbeiträge zu Deep Learning Interpretability

## XAI Toolkits & Frameworks

- **Quantus**
  - `https://github.com/understandable-machine-intelligence-lab/Quantus`
  - 35+ Metriken für XAI-Evaluation
  - JMLR Paper: Quantus Paper
- **SHAP**
  - `https://github.com/slundberg/shap`
  - Shapley Values für Feature Importance
- **AI Explainability 360** (IBM)
  - `https://aix360.mybluemix.net/`
  - Enthält ProtoDash, CEM

## XAI Forschungsarbeiten

- **”Explainable AI: A Review”** (Gilpin 2018)
  - `https://arxiv.org/abs/1801.00631`
  - Systematische Klassifikation
- **LIME Paper** (Ribeiro 2016)
  - `https://arxiv.org/abs/1602.04938`
  - Lokale Erklärbarkeit
- **”Sanity Checks Revisited”** (Hedström 2023)
  - Verbesserte MPRT-Metriken

## References I


C. Molnar, Interpretable Machine Learning.
3 ed., 2025.