

**HOCHSCHULE
MITTWEIDA**
University of Applied Sciences

Kann man KI vertrauen? - Wie trifft KI-Entscheidungen -

Prof. Thomas Villmann und Dr. Marika Kaden

SICIM, Computational Intelligence Group, HS Mittweida



Agenda



Agenda



Einordnung von ChatGPT & CO

... gehören zu dem Sprachmodell-Typ

„Textvervollständigungsapparat“*

Large Language Models (LLMs)

(komplexe Modelle, die mit riesigen Mengen an Textdaten trainiert werden, um natürliche Sprache zu **verarbeiten** und zu **generieren**)

→ Transformers

→ Bild → Text, Text → Bild, Text → Text

→ Spezielle Sprachen, wie z.B. Moleküle, Protein, DNA/RNA

→ Generative KI

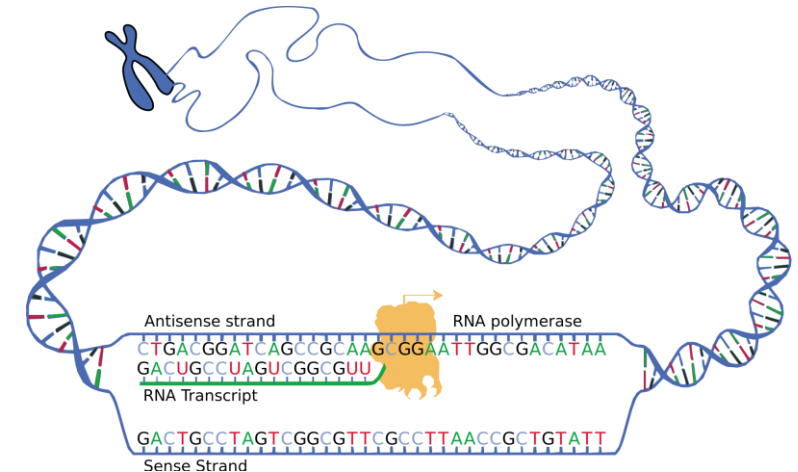


„Dich, Copilot als Bild visualisiert“
erstellt mit Copilot



„Hier ist ein Bild, das ein LLM (Large Language Model) darstellt!“
generiert mit Copilot

*Bezeichnung Kollege Prof. Benjamin Paassen, Bielefeld



<https://humanbiology.pressbooks.tru.ca/chapter/5-6-protein-synthesis/>

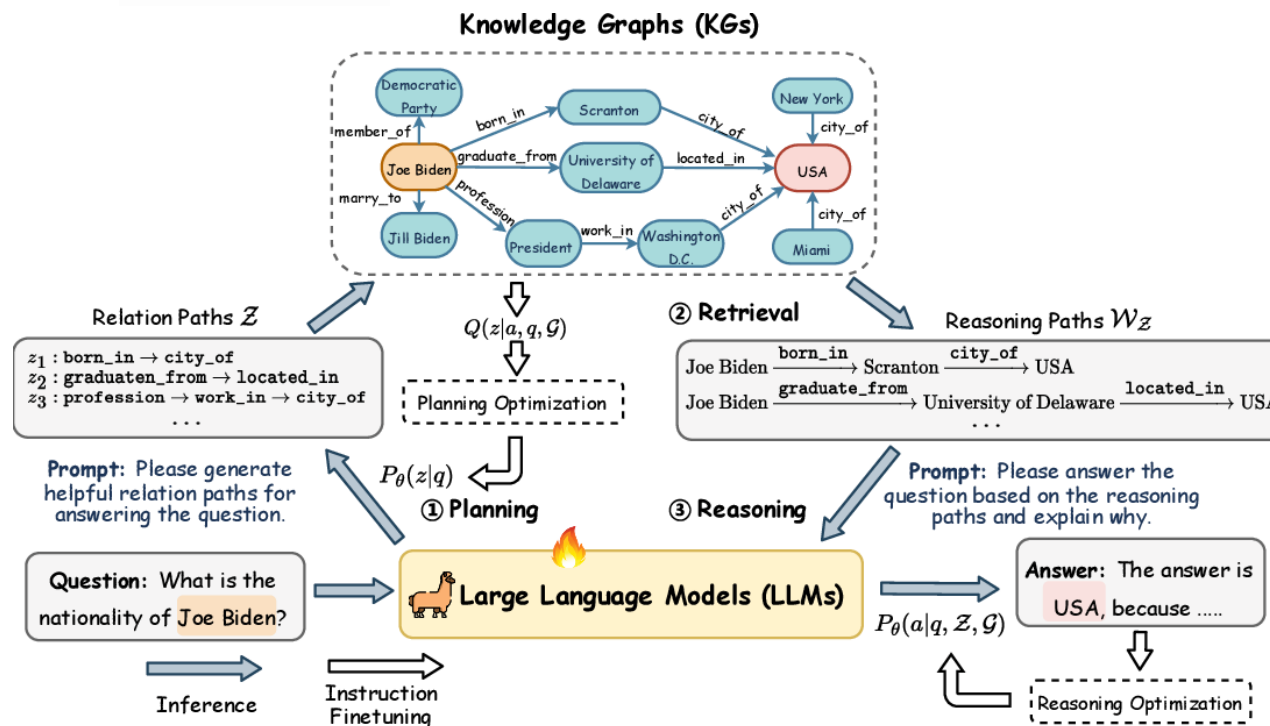
LLMs

sind

- Textverarbeitung- und ~generierungs-Modelle (im allgemeinen für natürliche Sprache)
- Deep-Learning und Transformer-Architekturen
- Training mit GROßEN Datenmengen

sind nicht

- Echte Intelligenz
- Fehlerfrei im faktischen Sinne
- Fachspezifische Datenbanken



Modell

Anzahl an Parametern

GPT-4	1.76 Trillion
Gemini	1.50 Trillion
Bloom	176 Billion
Llama 2	7B, 13B, or 70B
BloombergGPT	50B
Dolly 2.0	12B
GPT-Neo*	2.7B
DeciCoder-1B*	1B
Phi-1.5*	1.5B
Dolly-v2-3b*	3B

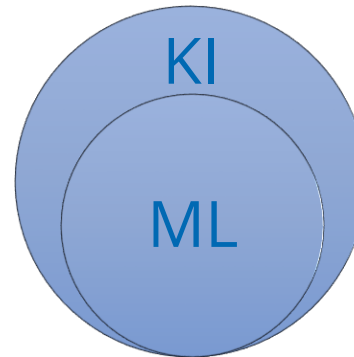
Einordnung von ChatGPT & CO

(allgemeine) Künstliche Intelligenz

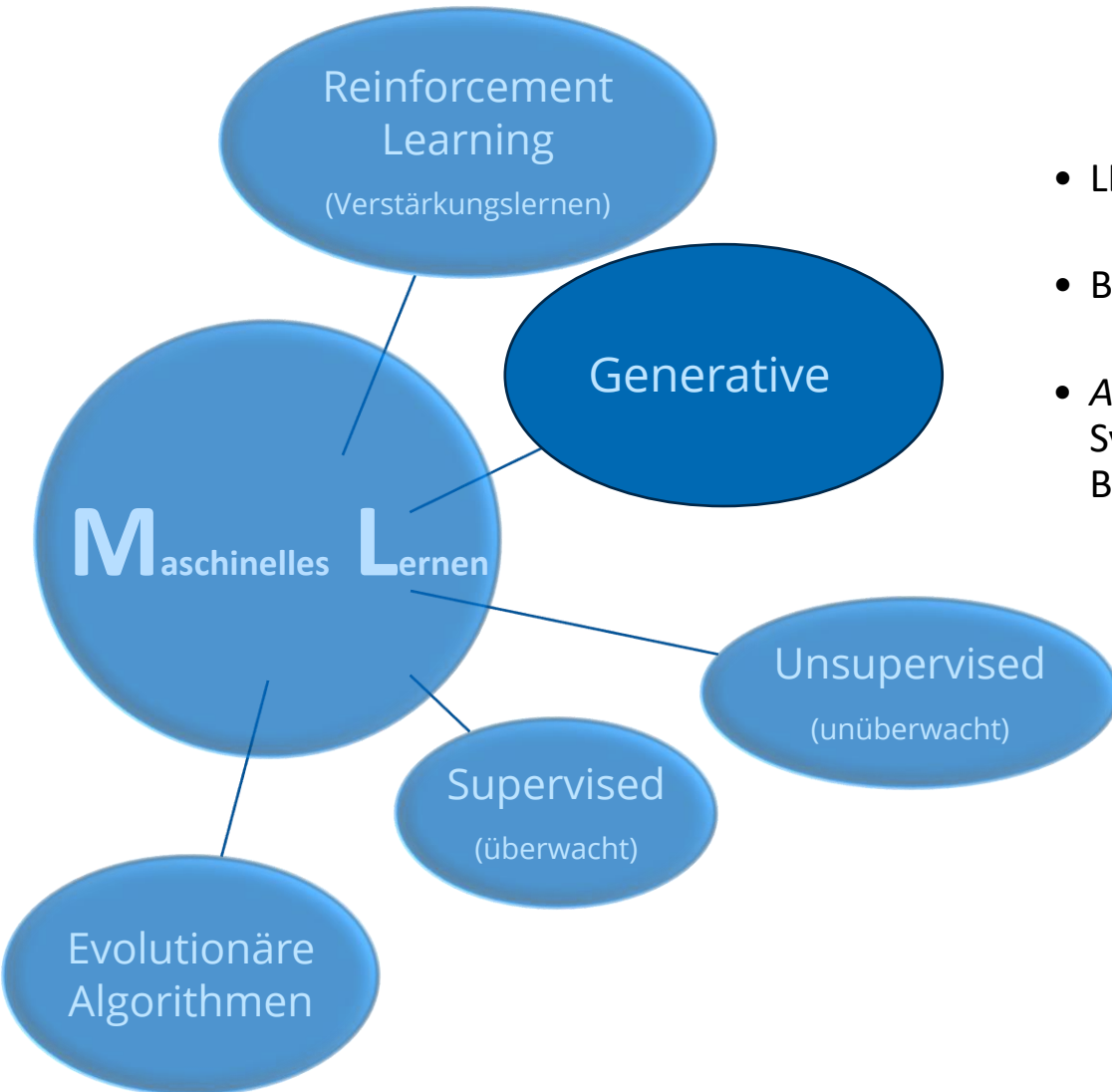
- KI bezeichnet die Fähigkeit von Maschinen, **menschenähnliche Intelligenz** zu simulieren, um komplexe Aufgaben wie Problemlösung und Entscheidungsfindung zu bewältigen → allgemeine KI (unterteilt sich in starke & schwache KI)
- KI umfasst **verschiedene Technologien** und Methoden, einschließlich maschinellem Lernen, Expertensystemen und natürlicher Sprachverarbeitung
- Das **Hauptziel** von KI ist es, Maschinen zu entwickeln, die Aufgaben ausführen können, die normalerweise menschliches Denken erfordern

Maschinelles Lernen

- Teilbereich der KI – schwache KI: ML ist eine spezifische Unterkategorie der KI, die sich auf das **Lernen aus Daten** konzentriert, um Muster zu erkennen und Vorhersagen zu treffen
- **Datenabhängig**: ML-Modelle benötigen große Mengen an Daten, um zu lernen und ihre Leistung im Laufe der Zeit zu verbessern
- **Vordefinierte Algorithmen**: ML nutzt Algorithmen, um Muster in den Eingabedaten zu identifizieren und basierend darauf Entscheidungen zu treffen

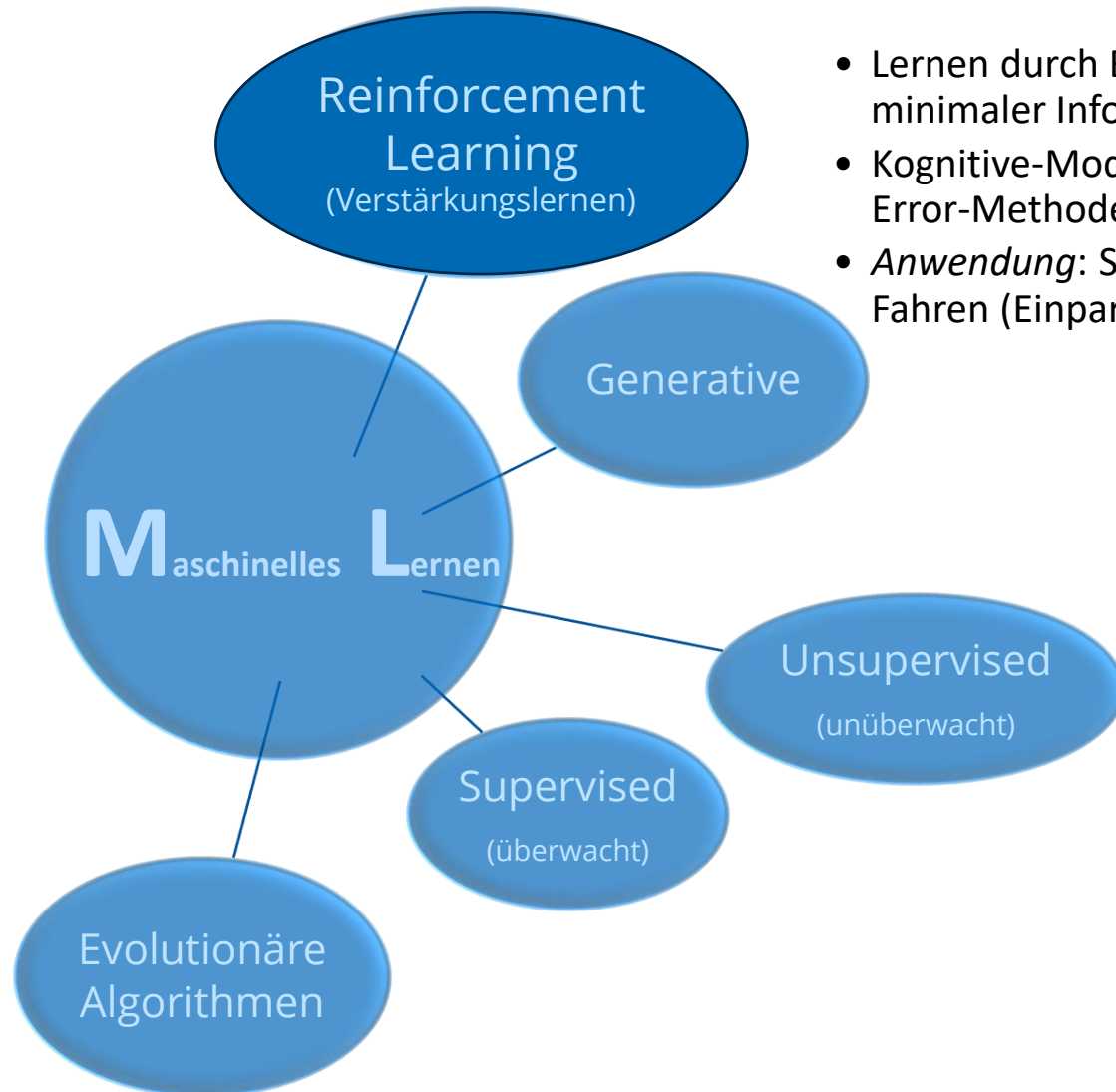


Einordnung von ChatGPT & CO

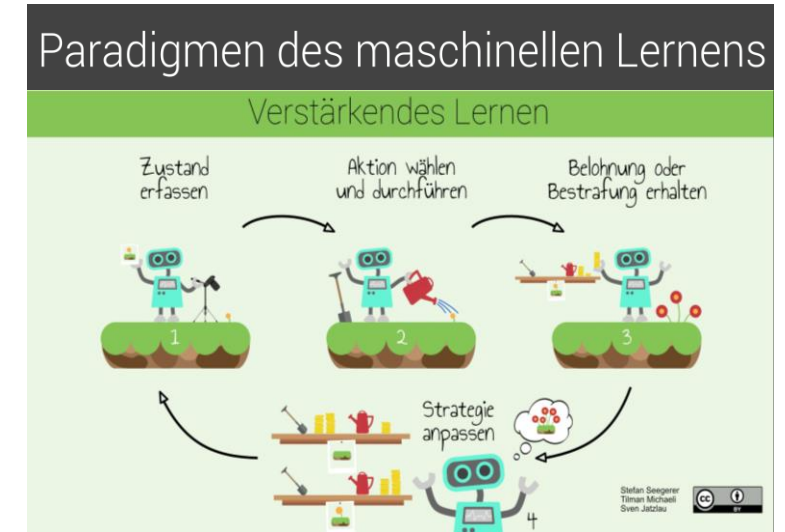


- LLMs
 - Transformers, Autoencoder, (Self-)Attention
- Bild-, Videogenerierung
 - Generative Artificial NN (GAN)
- *Anwendung*: ChatBots, Dokumentenverarbeitung, Paraphrasierung, Synthetische Bilder für Schulungszwecke, Simulation im wissenschaftlichen Bereich

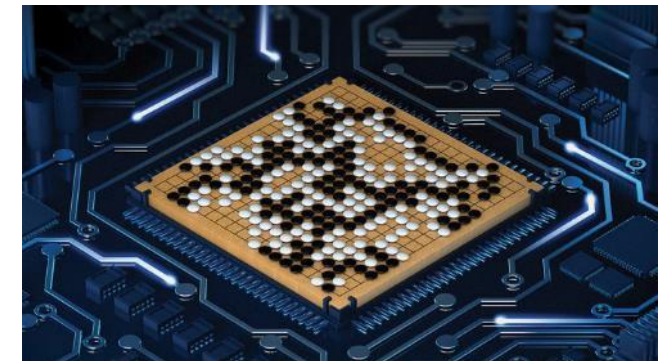
Einordnung von ChatGPT & CO



- Lernen durch Belohnung, Lernen mit minimaler Information → Strategie erlernen
- Kognitive-Modelle, Agentenmodelle, Trail- and Error-Methode
- *Anwendung:* Spiele-KI, Robotik, Autonomes Fahren (Einparkhilfe), ...

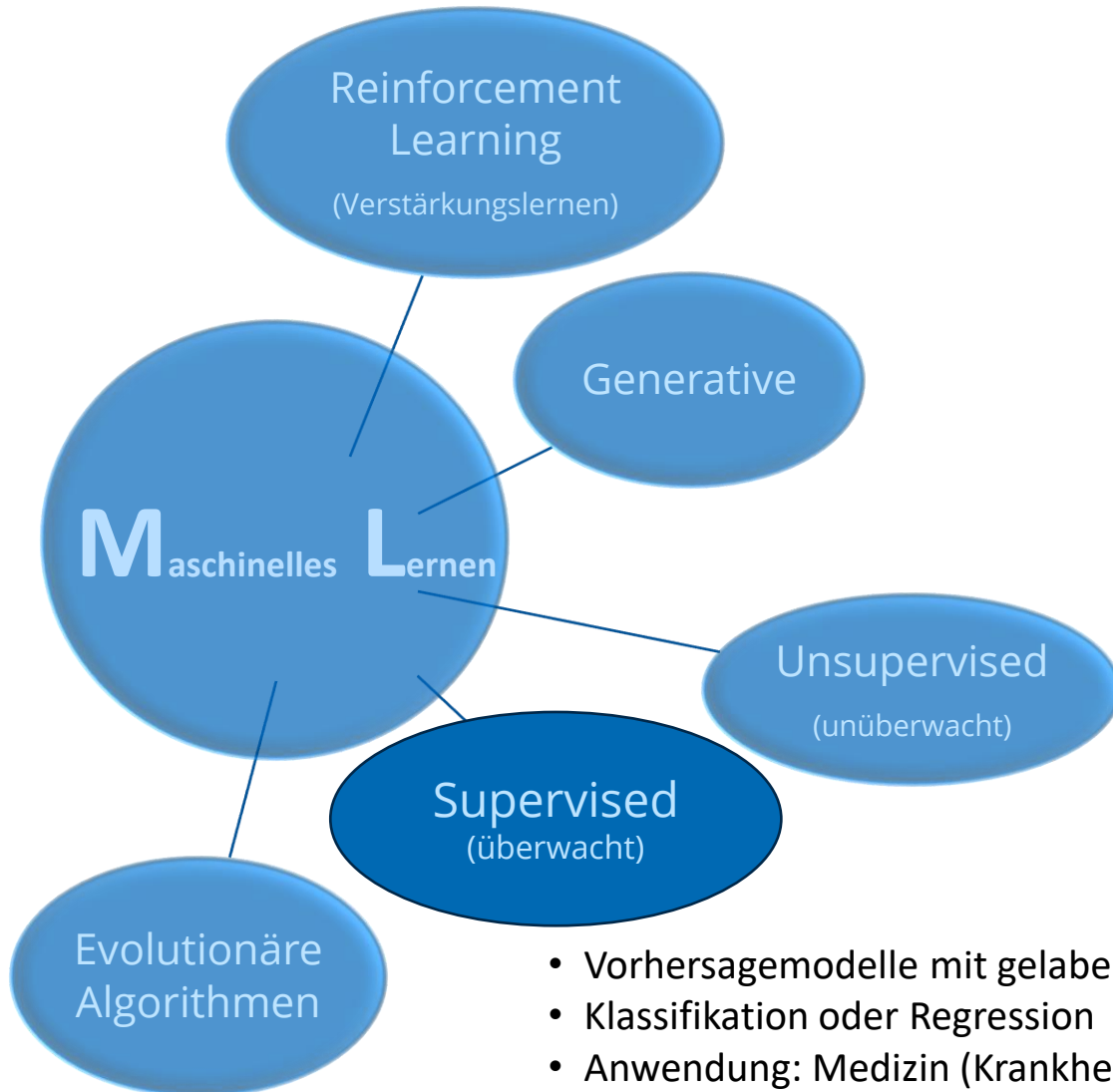


Deep Blue 1996/97

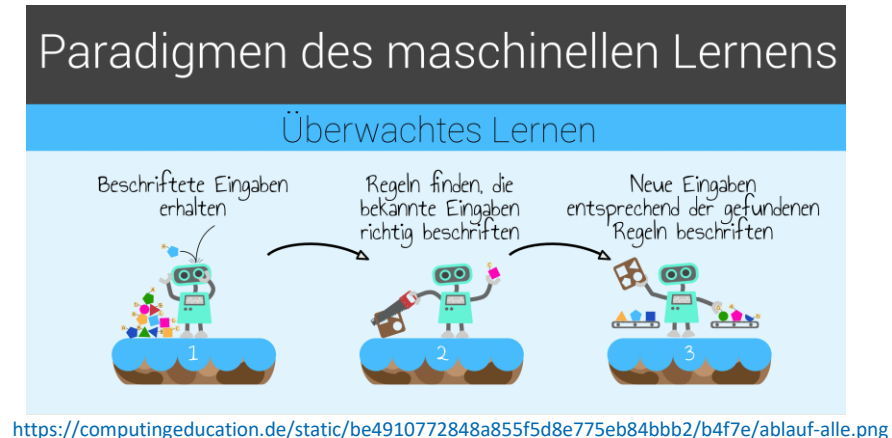


AlphaGo Zero 2017

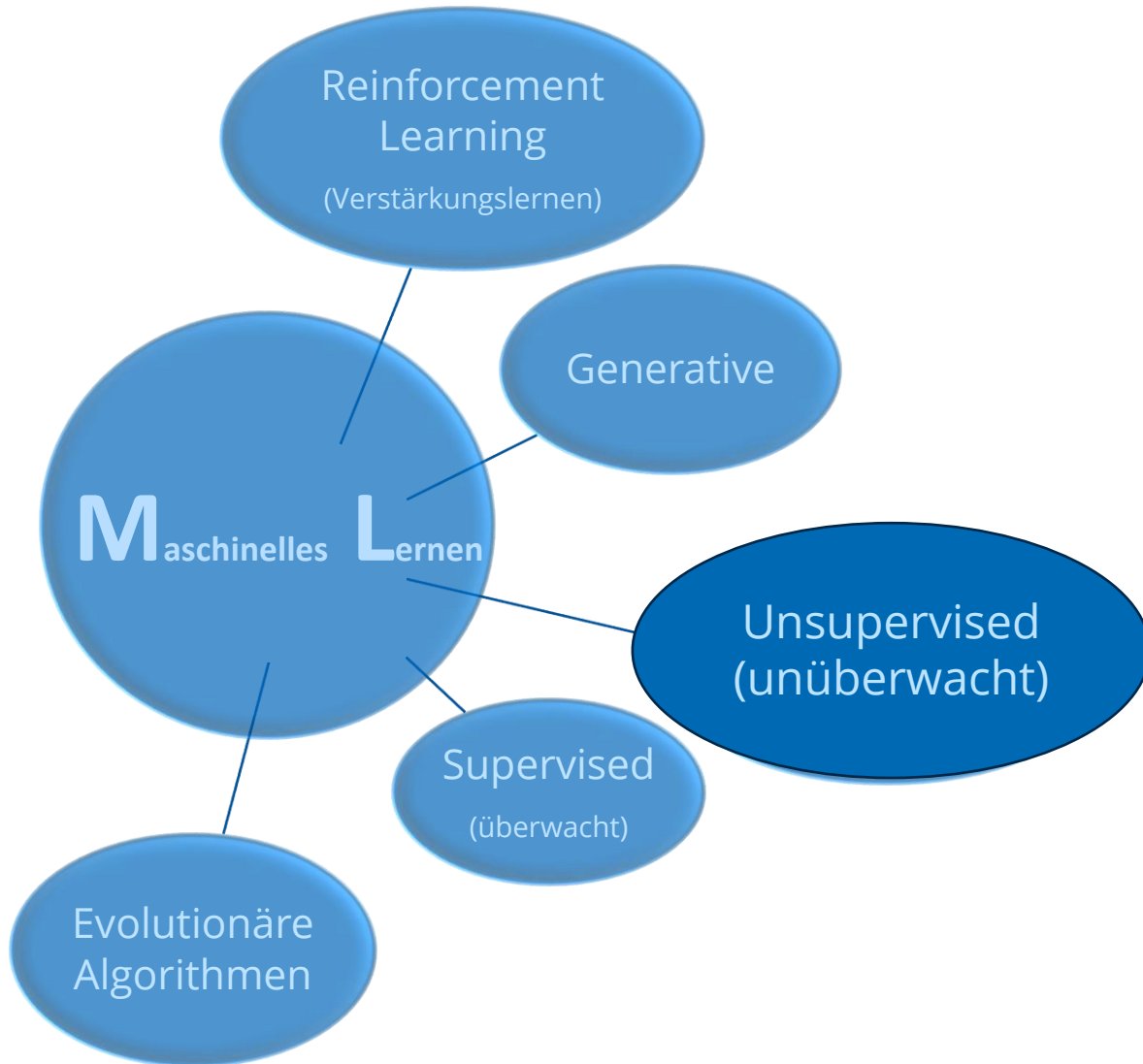
Einordnung von ChatGPT & CO



- Vorhersagemodelle mit gelabelten Daten
- Klassifikation oder Regression
- Anwendung: Medizin (Krankheitsprognose), Autonomes Fahren (Spurassistentz, Fahrbahnbeschaffenheit), Wetterprognose, Predictive Maintenance



Einordnung von ChatGPT & CO



- Lernen durch Analogie
- Aufgaben: Datengruppierung, Datenanalyse, Datenvisualisierung, Dimensionsreduktion/Kompression, Ausreißererkennung
- *Anwendung:* Forensik (Betrugserkennung), Warenkorbanalyse, Datenanalyse

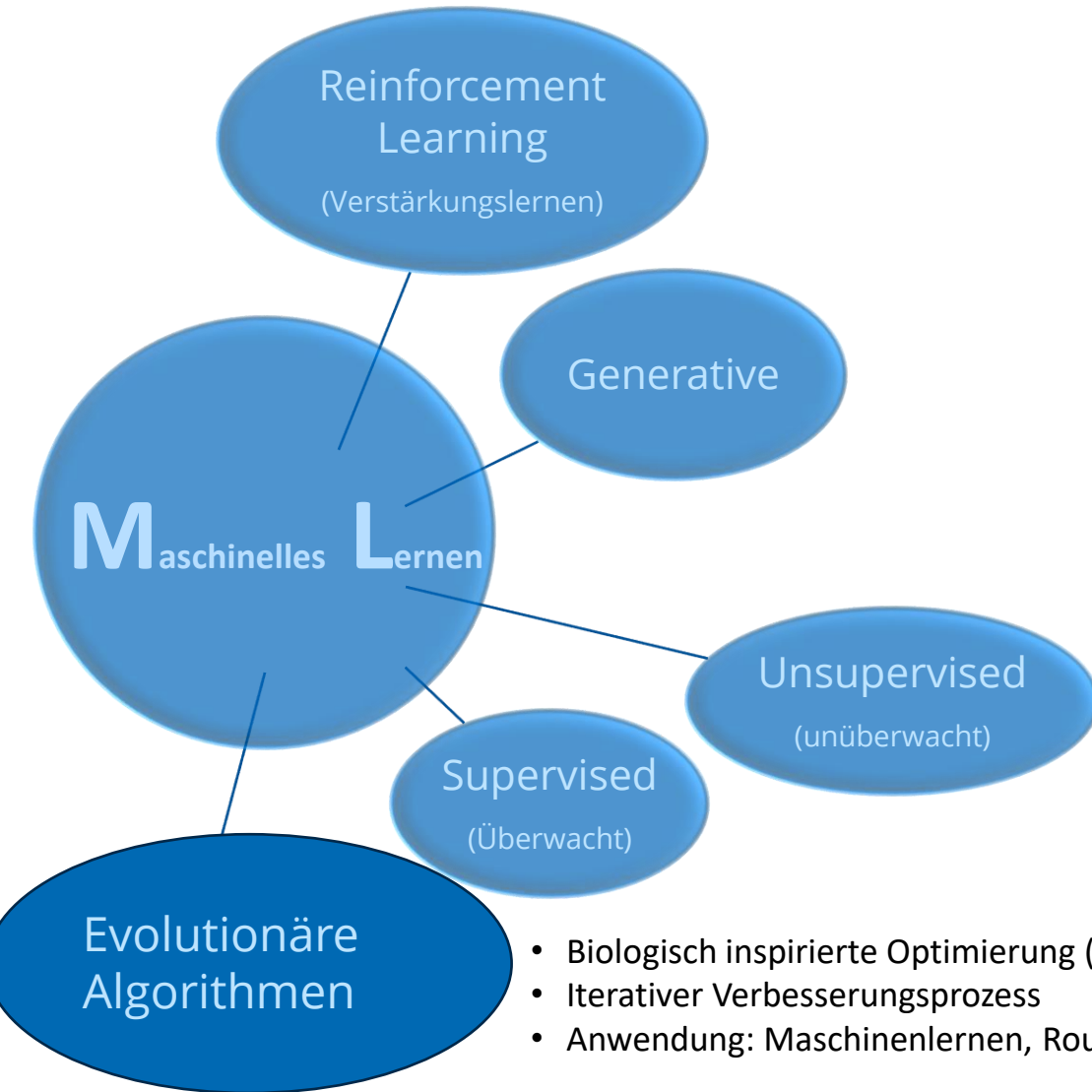
Paradigmen des maschinellen Lernens

Unüberwachtes Lernen

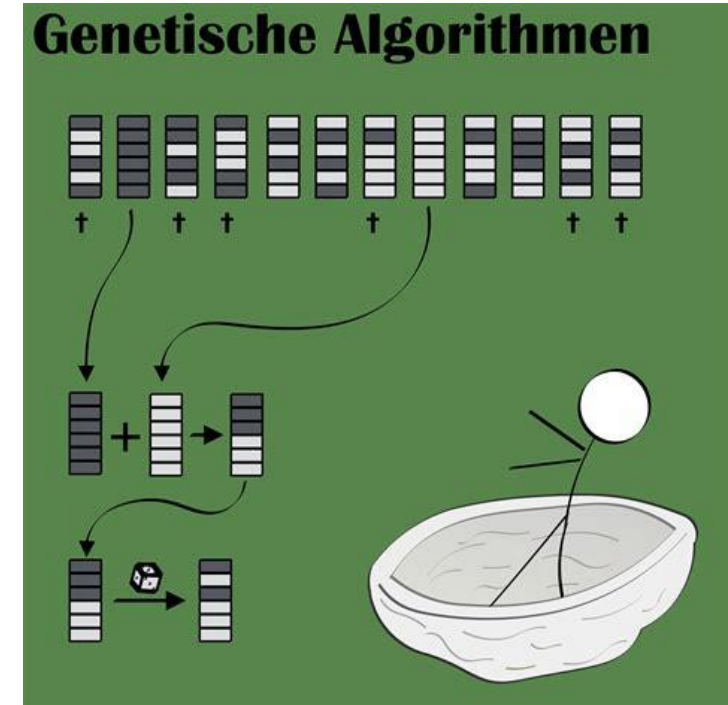


<https://computingeducation.de/static/be4910772848a855f5d8e775eb84bbb2/b4f7e/ablauf-alle.png>

Einordnung von ChatGPT & CO



- Biologisch inspirierte Optimierung (Bionik)
- Iterativer Verbesserungsprozess
- Anwendung: Maschinenlernen, Routenplanung, Designoptimierung, Prozessoptimierung



Agenda

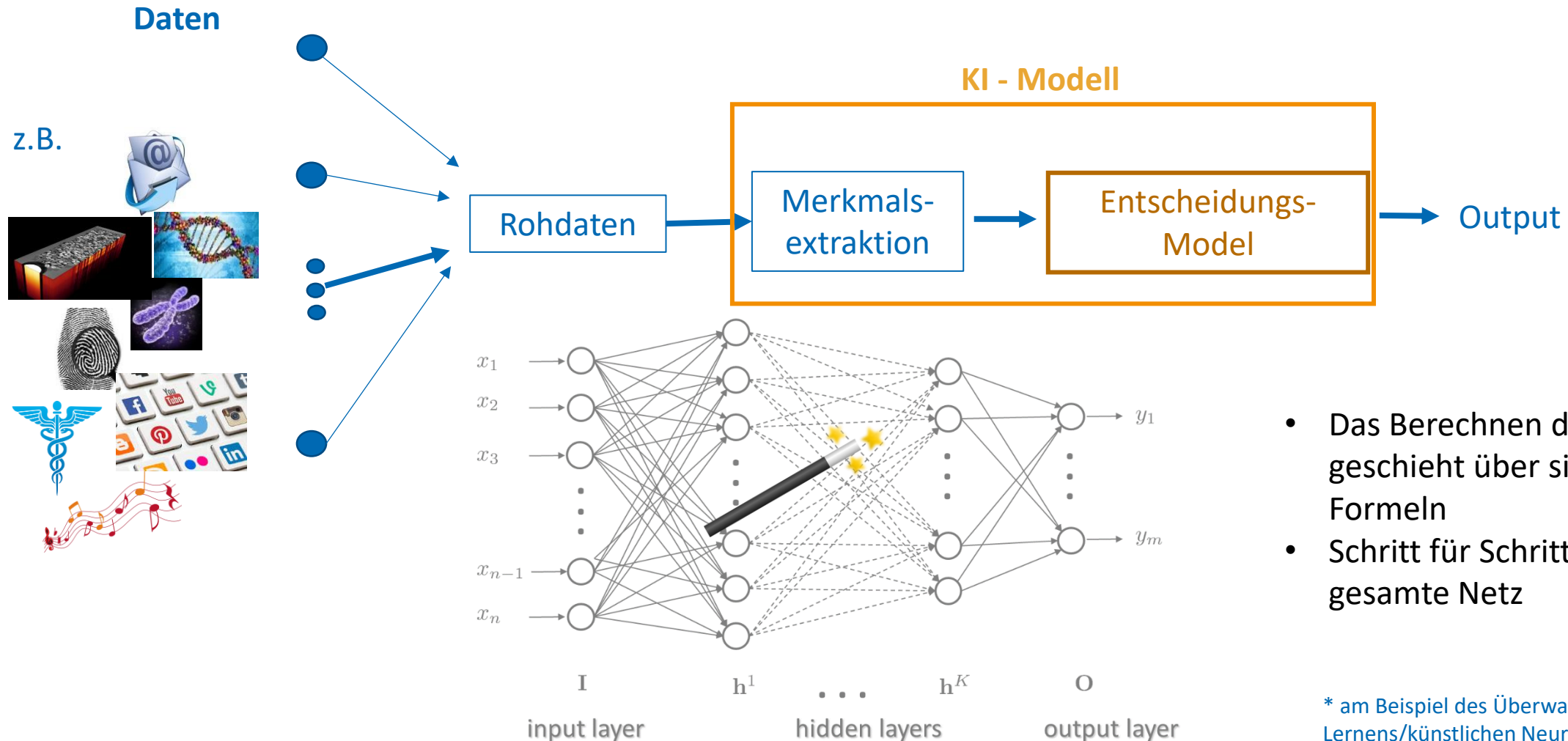


Agenda



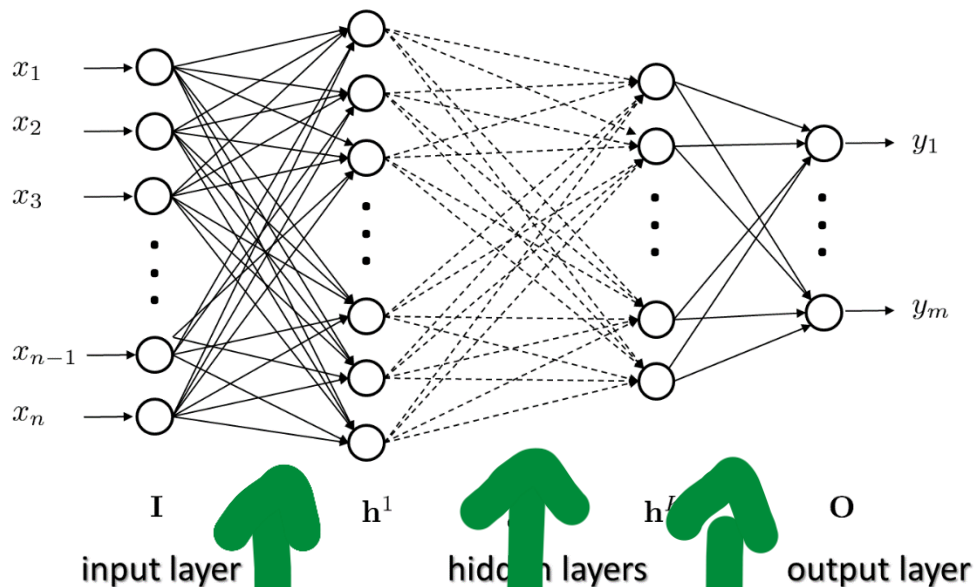
- seit den 1950er wurde der Begriff geprägt und erste Grundlagen gelegt
- KI-Modelle sind bereits seit über 30 Jahren im Einsatz (in Hilfestellungen für medizinischer Diagnosen, Einparkhilfen, Schachprogramme,...)

Wie trifft 'KI'- nun Entscheidungen*?



* am Beispiel des Überwachten Lernens/künstlichen Neuronalen Netzes

Was und wie wird gelernt*?

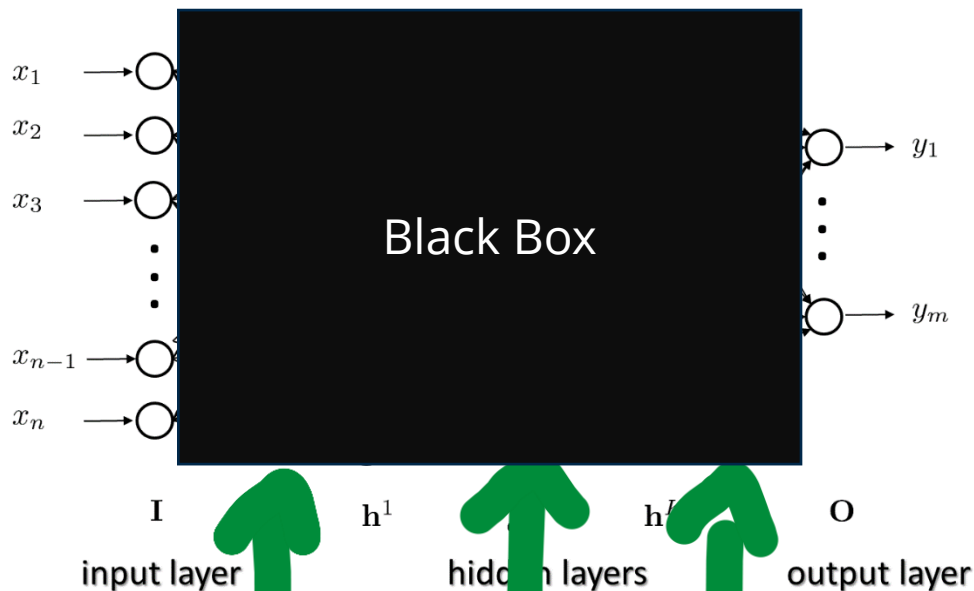


Prinzip: Lernen durch Fehler

- für Daten mit bekanntem Output
- Vergleich von gewünschtem Output mit dem Output vom KI-Modell
- der Fehler wird dann nach hinten an das Netz zurückgegeben und an den *Gewichten geschraubt**

* natürlich stark vereinfacht dargestellt

Was und wie wird gelernt*?



W-Gewichte

Prinzip: Lernen durch Fehler

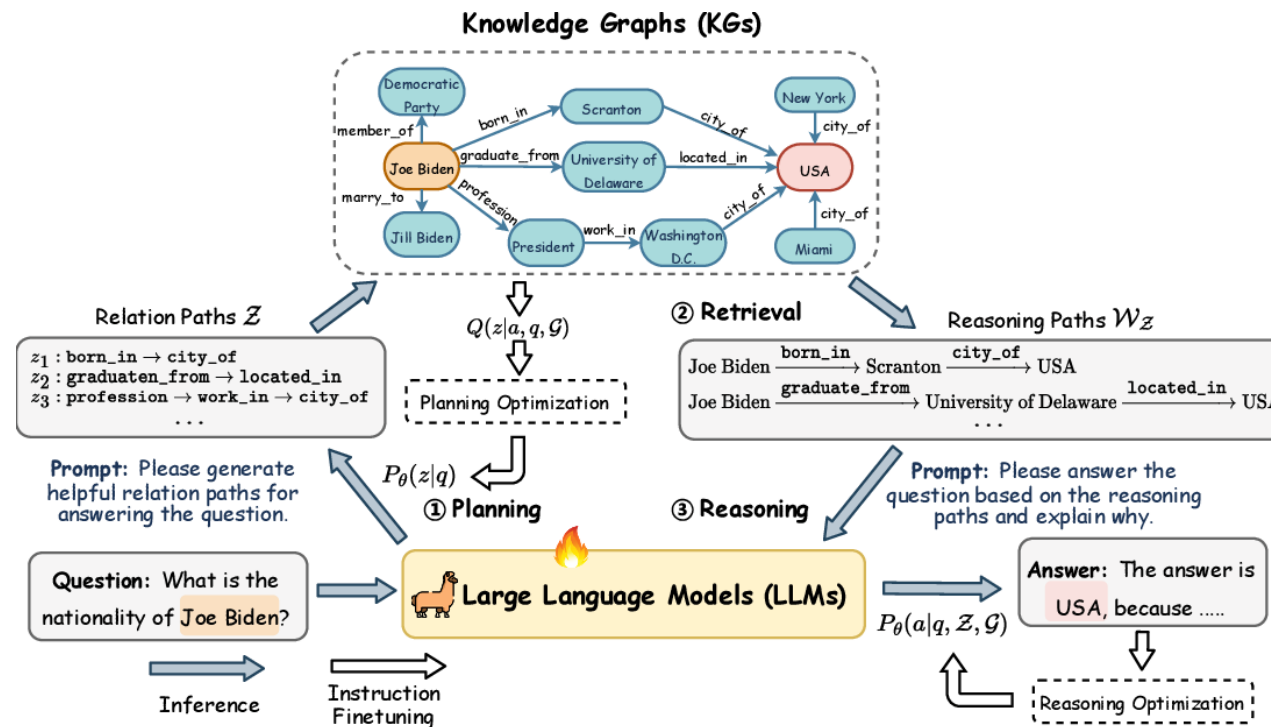
- für Daten mit bekanntem Output
- Vergleich von gewünschtem Output mit dem Output vom KI-Modell
- der Fehler wird dann nach hinten an das Netz zurückgegeben und an den *Gewichten geschraubt**

* natürlich stark vereinfacht dargestellt

Was und wie wird gelernt?



LLMs	Modell	Anzahl an Parametern
	GPT-4	1.76 Trillion
	Gemini	1.50 Trillion
	Bloom	176 Billion
	Llama 2	7B, 13B, or 70B
	BloombergGPT	50B
	Dolly 2.0	12B
	GPT-Neo*	2.7B
	DeciCoder-1B*	1B
	Phi-1.5*	1.5B
	Dolly-v2-3b*	3B



Prinzip: Lernen durch Fehler

- für Daten mit bekanntem Output
- Vergleich von gewünschtem Output mit dem Output vom KI-Modell
- der Fehler wird dann nach hinten an das Netz zurückgegeben und an den *Gewichten geschraubt**

Was ist Vertrauen?

Soziologie/Psychologie

Vertrauen in eine Person



Wissenschaft/Technologie/etc.

Vertrauen auf/in Inhalte, Funktionalität

Zuverlässigkeit,
Präzision,
Nachvollziehbarkeit,
Konfidenz,
Entscheidungssicherheit



„Vertrauenswürdige“ KI

„AI for good“ (ITU)

Transparenz, Erklärbarkeit

Ethische Richtlinien für vertrauenswürdige KI/Trustworthy AI (EU):

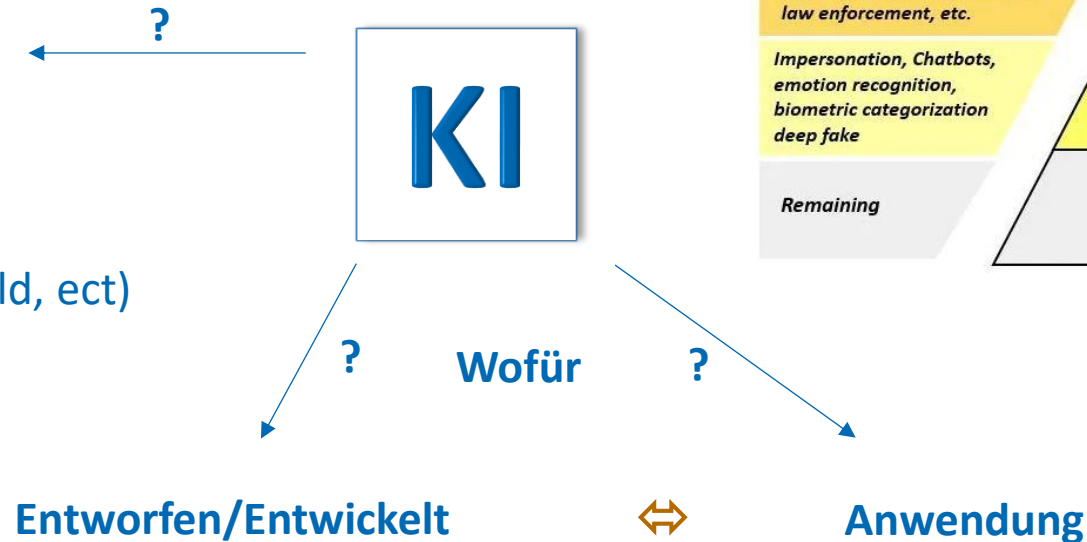
1. Einhaltung von Regelwerk und Gesetzen
2. Ethisch handeln → Fairness (Erklärbarkeit, Transparenz, Verantwortlichkeit, Glaubwürdigkeit)
3. Robustheit innerhalb der eingesetzten sozialen/technischen Umgebung

→ Wichtige Begriffe: Transparenz, Erklärbarkeit, Gesetzeskonform

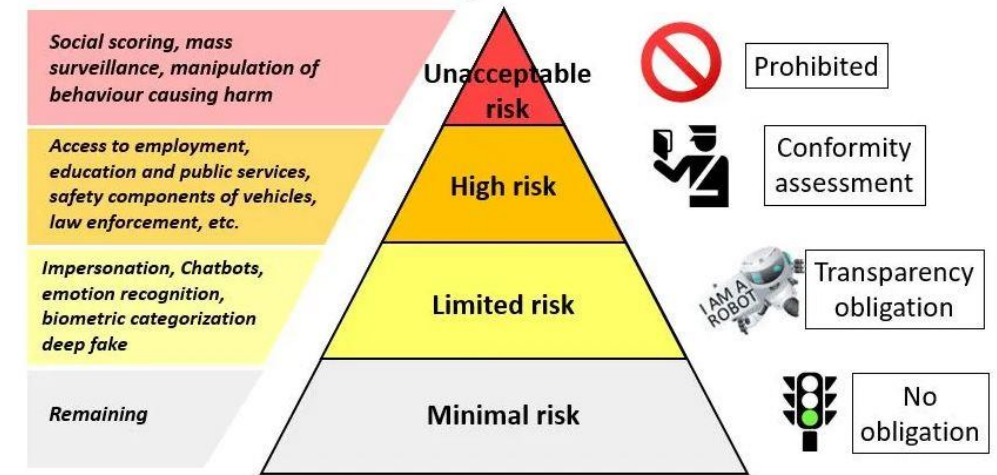
KI-Vertrauen

Welche

- Mustererkennung
- Vorhersagen
- Entscheidungsfindung
- Generierende (z.B. Text, Bild, ect)
- ...



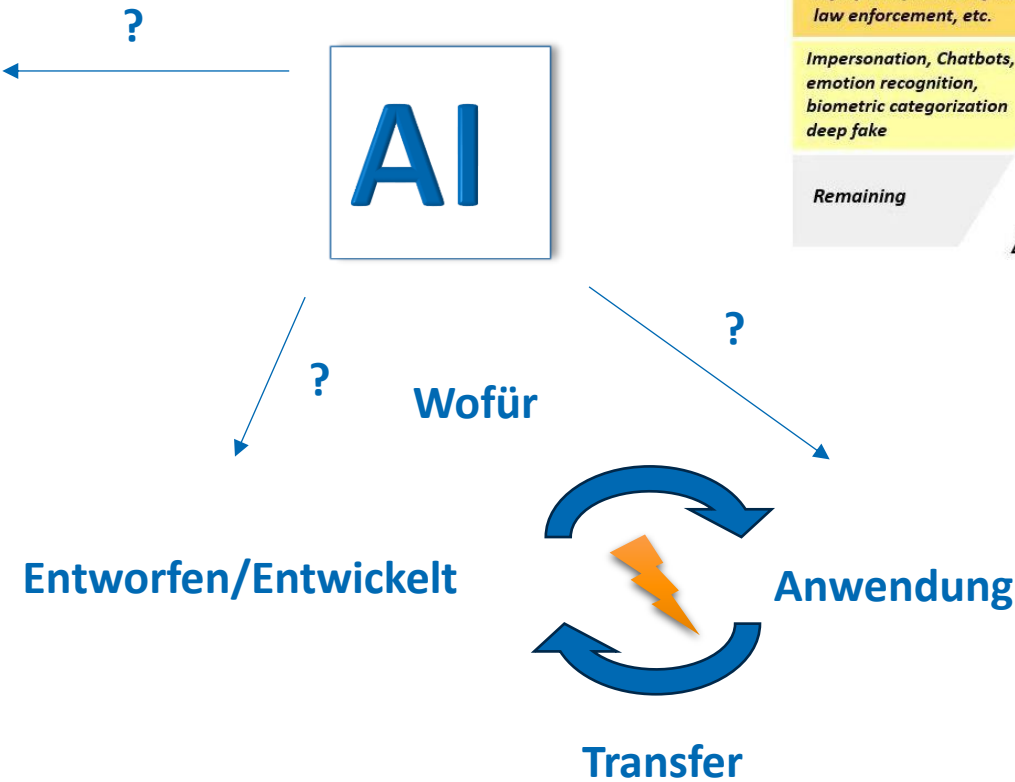
EU Artificial Intelligence Act: Risk levels



- Sicherheitsstufe laut EU-AI-Act
- Just-for-Fun
- Unterstützend
- Entscheidungen treffend
- für Forschung/Informationsgewinnung

KI-Vertrauen

Welche



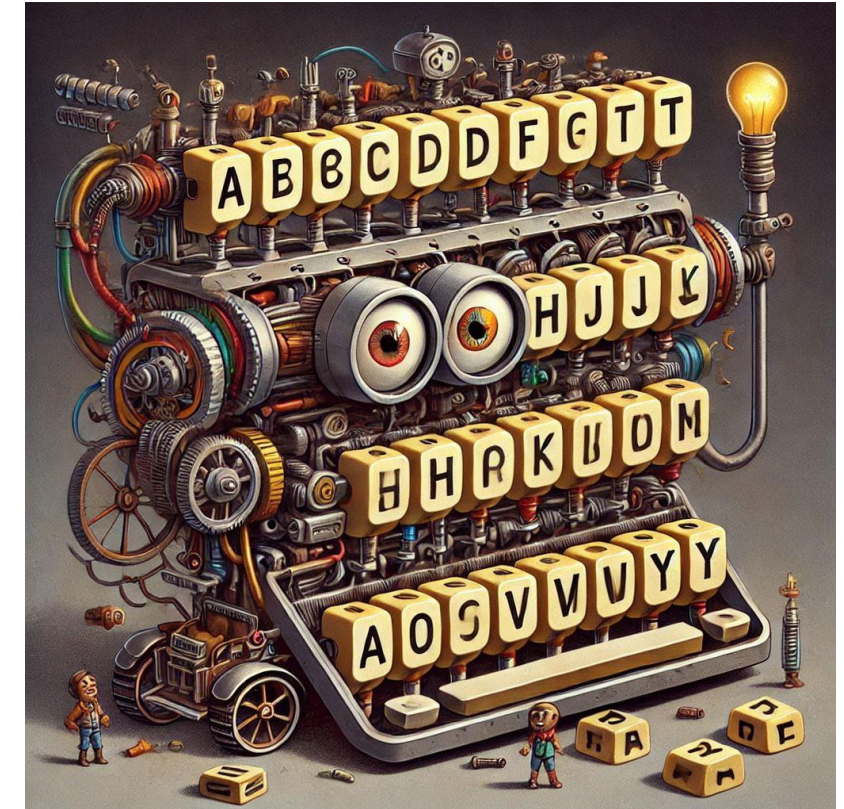
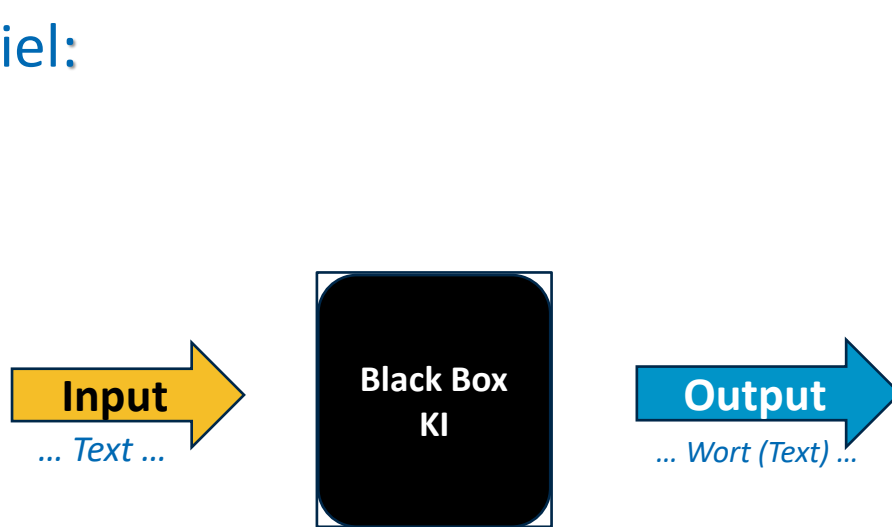
EU Artificial Intelligence Act: Risk levels



Black Box KI – Evaluierung von KI – Entscheidungen ?

Ausgangspunkt: KI liefert zu einer *Eingabe* eine *Ausgabe* (lokale Einzelfallentscheidung)

Beispiel:



"Textvervollständigungsapparat" generiert von Gemini

Black Box KI – Evaluierung von KI – Entscheidungen ?

Was ist, was kann ChatGPT ?



- ChatGPT (Chatbot) ist eine Anwendung, die Künstliche Intelligenz (GPT-4) verwendet, um sich mit Menschen in natürlicher Sprache zu unterhalten. Benutzer können Fragen stellen, auf welche das System in natürlicher Sprache antwortet. Er kann Texteingabe, Audioeingabe oder beides unterstützen.

– Anwenderperspektive

- Es kann
 - Fragen in verschiedenen Sprachen beantworten
 - Text erzeugen, zusammenfassen und evaluieren bezüglich verschiedener Niveau-Level
 - Perspektiven, Disputationen, Vorträge und Präsentationen erzeugen
 - Gedichte schreiben und Computerprogramme, Texte übersetzen und Multiple-Choice-Tests generieren
 - ...
- Es ist in der Lage
 - Beziehungen zwischen Eingabetexten herzustellen und so den Eindruck einer menschen-ähnlichen Konversation zu erzeugen

– Technische Perspektive

- Die KI in ChatGPT
 - ist KI der natürlichen Sprachverarbeitung (GPT-4)
 - ist in der Lage, Text gemäß eines *Wahrscheinlichkeits-Modells* basierend auf einem künstlichen neuronalen Netzes zu erzeugen
 - ist mit Bespieltexten trainiert
 - große, öffentlich zugängliche Textdatenbanken wie z.B. Wikipedia, Literaturdatenbanken, etc.
 - Durch automatisches und assistiertes Lernen (durch Menschen)
- Künstliche neuronale Netze sind
 - biologisch inspirierte mathematische Modelle, welche (begrenzt) natürliche Verhaltensweisen biologischer Neuronen imitieren (für Lernen, Vorhersage, Assoziationen, ...)

Black Box KI – Evaluierung von KI – Entscheidungen ?

Was ist, was kann ChatGPT – *nicht* ?



- ChatGPT ist keine information-beschaffende Computer-Anwendung, welche Fakten, Daten und Wissen aus einer Datenbank extrahiert oder generiert.

– Anwenderperspektive

- Es **ist kein/e**
 - menschenähnlich agierender Agent mit moralischen Konzepten
 - ethische Instanz oder Religion
 - immer korrekt antwortendes System
 - ...
- Es **hat nicht**
 - (kognitives) Bewusstsein
 - irgendein soziales Verständnis
 - ...

– Technische Perspektive

- Die KI in ChatGPT **ist nicht**
 - vorturteilsfrei (Vorurteile sind implizit in den Trainingsdaten enthalten!)
 - basierend auf irgendwelchen logischen oder anderen Implikationen
 - ...
- Künstliche neuronale Netze **sind keine** Modelle für
 - die Immitation des natürlichen Verhaltens von realen Nervenzellen und Hirnarrealen
 - die Erklärung von Bewusstseinto
 - ...

Black Box KI – Evaluierung von KI – Entscheidungen ?



Problem: Diskrepanz zwischen Trainingsziel der KI und Evaluation von bzw. Erwartungen an KI – Entscheidungen ...

Trainingsziel:

- Wahrscheinlichkeitsmodell von Sprache(n) auf Basis *morphologischer und syntaktischer* Analysen (Trainingsdaten)
- Vorhersage von Texten entsprechend des gelernten Wahrscheinlichkeitsmodells

Annahmen:

- Trainingstexte enthalten *implizit faktisches Wissen* (Fakten, Erkenntnisse, Bewertungen) enthalten, welchen *Einfluss auf* die morphologische *Textstruktur* hat

Voraussetzungen:

- KI kann nicht alle Texte lernen sondern soll verallgemeinern, extrahieren, assoziieren, etc. (→ *sonst Datenbanksuche*) - *Generalisierungsfähigkeit*
- Mathematische Modellierung der KI

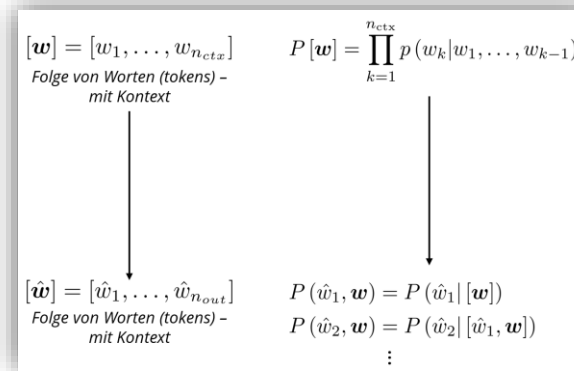
Nutzer-Erwartung:

- KI soll Texte generieren, die wahre Sachverhalte darstellen → *semantische Interpretation und Evaluation*
- KI soll faktisches Wissen bereitstellen

Ursachen:

- Interne Arbeitsweise der KI ist für allgemeine Nutzer weitgehend unbekannt
- Trainingsmethoden und (math.) Zielsetzung nur für wenige Experten zugänglich / verständlich
- Spezifische interne Realisierungen, Berechnungen und Ergebnisevaluierungen nur schwer oder gar nicht nachvollziehbar (im interpretativen Sinn)

KI agiert als **Black Box** für den Nutzer



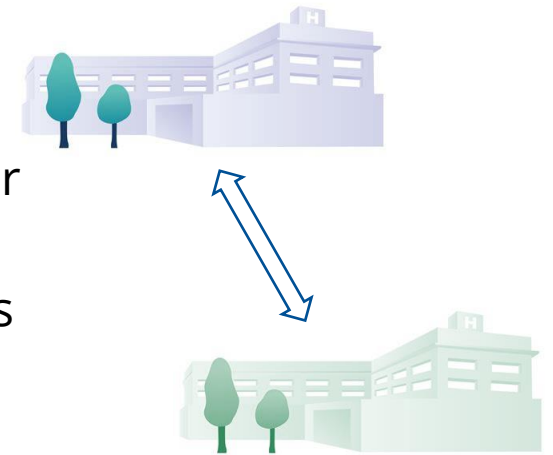
Transfer: Entwicklung ⇔ Anwendung

Zwei Beispiele:

- Open AI: ChatGPT
- Medizinische Diagnose:
 - Diagnosesystem für eine spezifische Krankheit in einem Krankenhaus unter Verwendung einer Messapparatur → KI – Modell A hat hohe Präzession
 - Anwendung dieses KI-Modells A in einem zweiten Krankenhaus mit gleicher Messapparatur → schlechte Vorhersagegenauigkeit
 - Messdaten aus den Krankenhäusern unterscheiden sich → **Transfer** des KI-Modells notwendig



„...kann ChatGPT als
teilweise
vertrauenswürdig
eingestuft werden...“
[GPT-4]



Agenda



Agenda



Es kommt drauf an:

- Für was und wen ist das KI-Modell entwickelt wurden?
- Welche Grenzen sind bekannt?
- Sind alle Voraussetzungen erfüllt?
- Wie kritisch ist der Einsatz? (Einsatzbereich)

Verantwortung



KI* ist

- **keine** Person
- kann **nicht** handeln
- kann **nicht** denken

sind

- *Modelle/Algorithmen/Regelsystem*



Ob KI vertrauenswürdig ist?

ist entscheidend vom

Modell-Entwickler

(Grenzen – mathematischer Natur)

KI-Trainer/Data Scientist

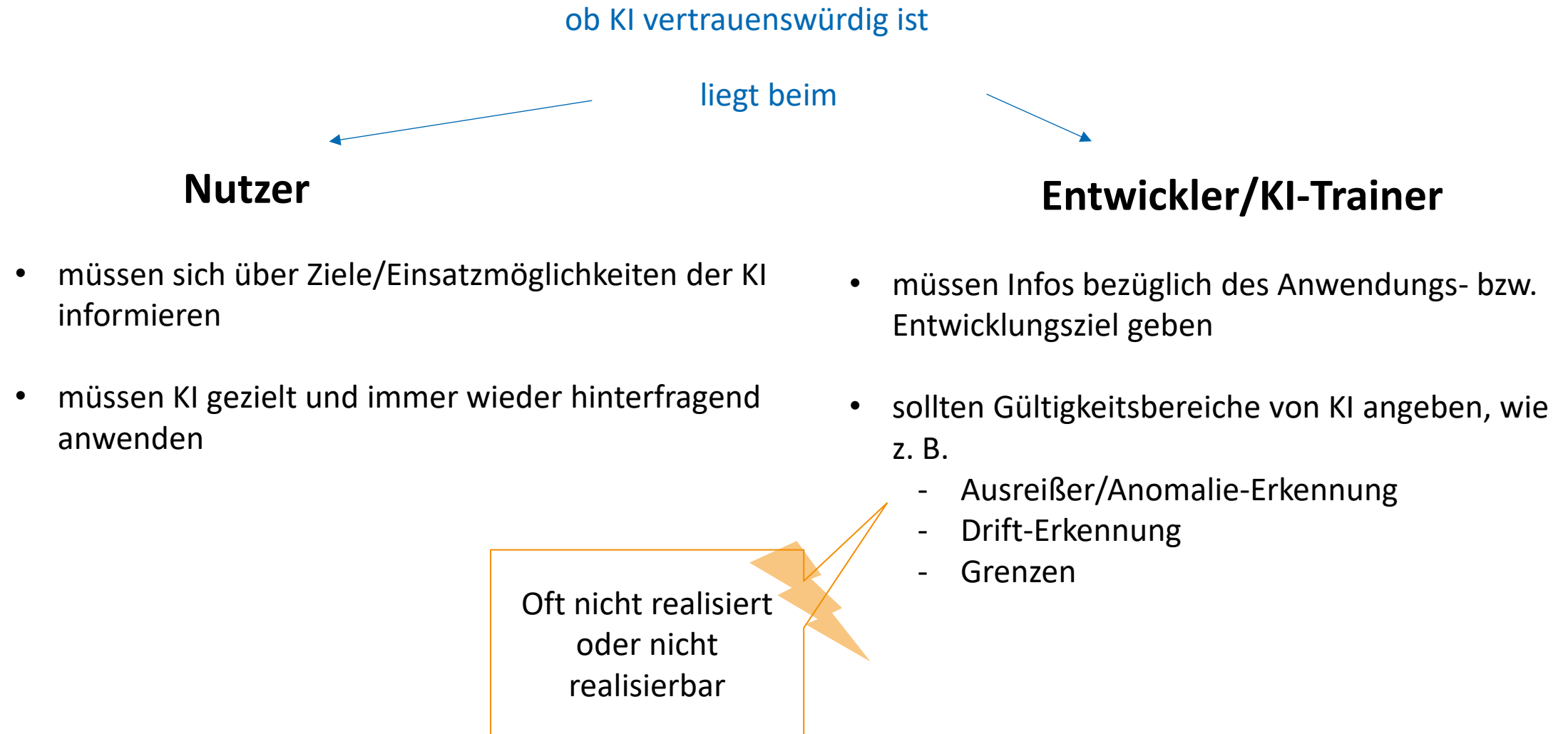
(Daten und Modell)

Nutzer

(entwickelt Vertrauen aus Erfahrungen und
durchdachter/bewusste Nutzung)

* Stand 2025

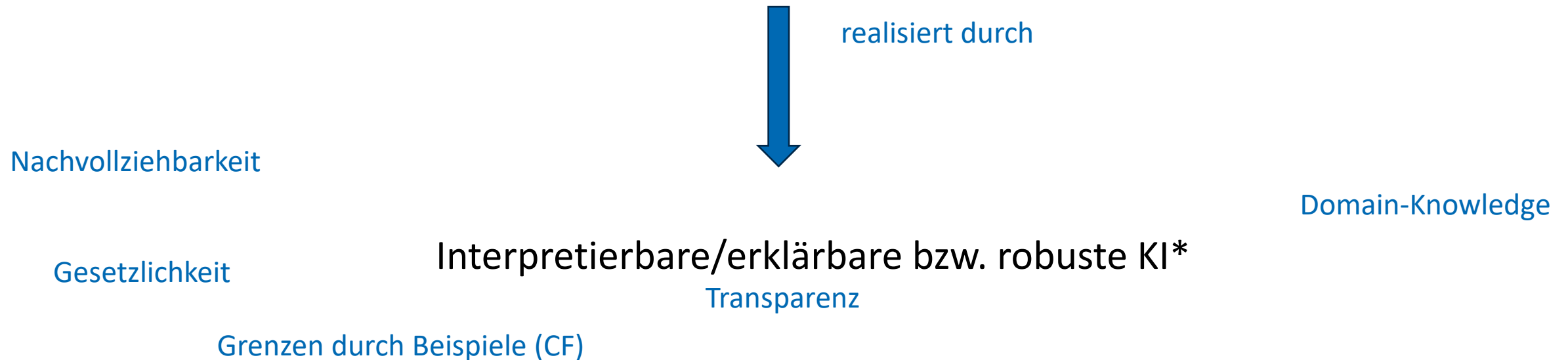
Verantwortung



Vertrauensvolle KI

Für Nutzer/Anwender **wünschenswert**

- Aussage über Grenzen des KI-Modells (Rückweisung bei Unbekanntem/Falscher Anwendung)
- Aussage über Entscheidungs(un)sicherheit



* wird seit 30 Jahren beforscht ... !

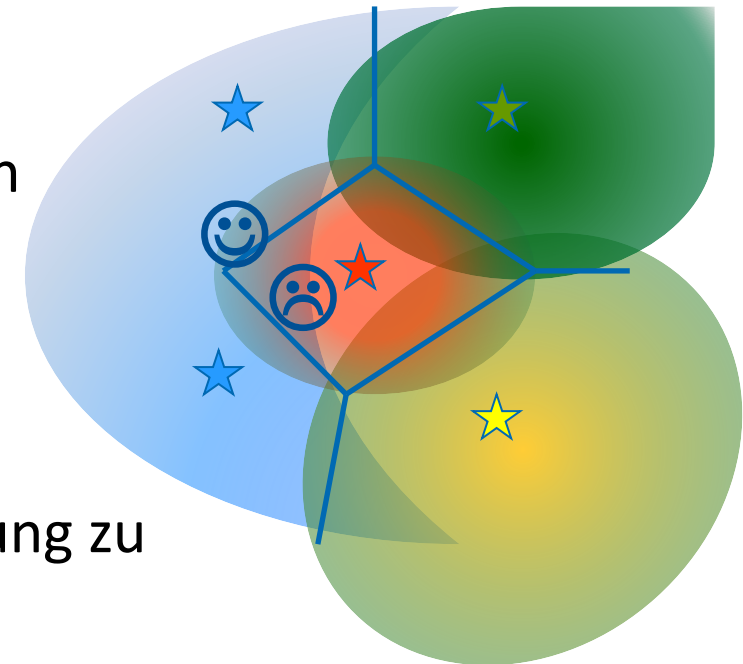
Interpretierbare KI – ein Beispiel

Robustheit und Nachvollziehbarkeit bei Klassifikation

Generalized Learning Vector Quantization

ein Algorithmus

- dessen Entscheidung auf Repräsentanten und Ähnlichkeiten basiert (Prototyp-basiertes Verfahren)
- und dessen Fokus zudem darauf beruht möglich sicher zu diskriminieren (Margin-Optimizer)
- wenn die Sicherheit nicht gewährt ist, auch eine Entscheidung zu *verwehren* (Reject-Option)



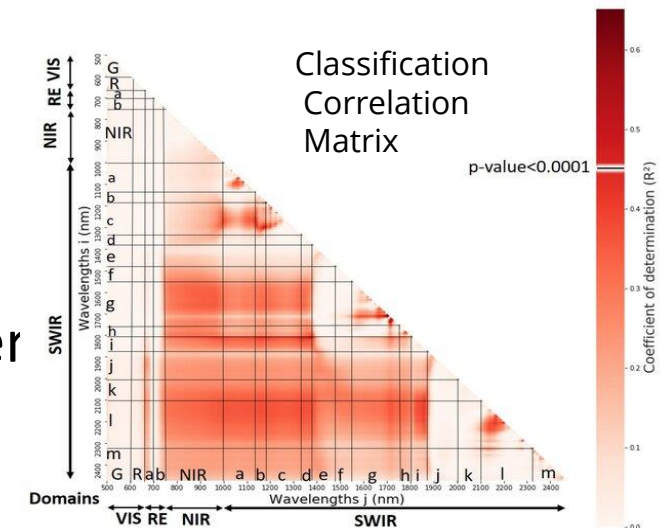
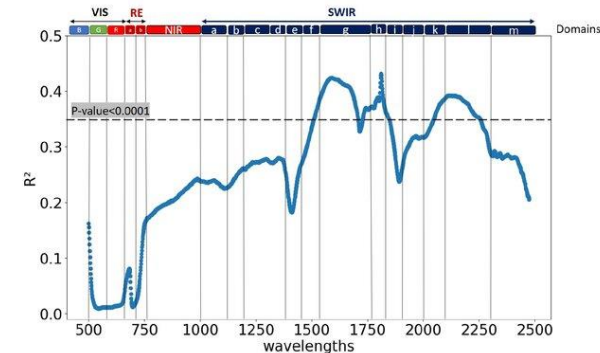
Interpretierbare KI – ein zweites Beispiel

Nachvollziehbarkeit und der Erhalt von zusätzlichen Informationen

Generalized Matrix Learning Vector Quantization

ein Algorithmus

- der zusätzlich die Informationen zurückgibt, welche Merkmale entscheidend zur Diskriminierung beitragen
- Dies gibt sowohl Einsichten in das Problem, kann aber auch auf *Schein*-Verbindungen und somit auf Bias in den Daten aufmerksam machen.



Interpretierbare KI – ein *Bias*-Beispiel

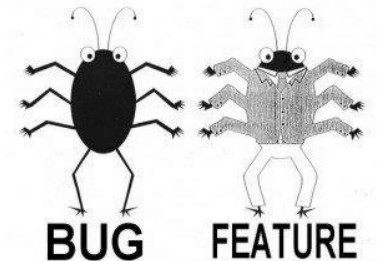
BIAS – was ist das eigentlich

... ist eine **systematische** und **ungewollte Verzerrung**/ Prioritäten
in den Daten

- KI ist vorurteilsfrei, da ein mathematisches Modell
- KI-Anwendung wiederum kann mit *Bias* behaftet sein
 - über das Training mit Bias-behafteten Daten

Bias in Trainingsdaten

- Notwendig: Hypothese, welcher Bias (**genaue Beschreibung** bzgl. einer **konkreten Fragestellung**),
Prüfen der Hypothese
- kann der Bias evaluiert werden (?)
 - nein: nur Philosophie
 - ja: 2 Möglichkeiten
 - 1) Konstruiere ein KI- Modell, was den Bias kompensiert/ignoriert
 - 2) Bereinigung der Daten bzgl. des Bias



Agenda



Agenda



SICIM - Kompetenz

Forschungsschwerpunkt: Entwicklung von Vertrauenswürdige KI- Modellen

Je nach Problem/Aufgabenstellung stehen im Fokus

Transparenz/Nachvollziehbarkeit, Kompakte Modelle/Ressourcenschonend,
Integration und Extraktion von Wissen, Robustheit und Stabilität bei
Entscheidungen, Grenzen und Rückweisungen

Projekte:

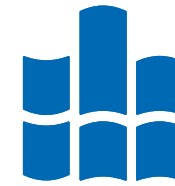
- gemeinsam mit Porsche: Scheinwerfererkennung
- gemeinsam mit LKS: Futtermittelinhaltsstoffvorhersage mittels Spektralaufnahmen
- Medizinprojekt: Krankheitsprognose
- Verbundprojekt: AI meets Space
- ICM (vormals IfM): Predictive Maintenance

Literaturempfehlung



Wenn KI-Systeme sich selbst erklären: Nachvollziehbarkeit, Ko-Konstruktion und Vertrauen.

Von Christiane Attig & Benjamin Paaßen



**HOCHSCHULE
MITTWEIDA**
University of Applied Sciences

Lassen Sie uns nun diskutieren



sicim
Hochschule Mittweida



**HOCHSCHULE
MITTWEIDA**
University of Applied Sciences

Vielen Dank für Ihre Aufmerksamkeit



sicim
Hochschule Mittweida