10.20378/irb-110824



# AI Literacy

## Why Basic Understanding of AI Methods is Relevant for Save, Efficient, and Reflected Use of AI-Tools

Ute Schmid
Otto-Friedrich-Universität Bamberg

ECIL 2025

Bamberg Center for Artificial Intelligence

Lernende Systeme
GERMANY'S PLATFORM FOR ARTIFICIAL INTELLIGENCE


https://benjweinberg.com/2021/08/28/a-premonition-from-wall-e/

bidt Bavarian Research Institute for Digital Transformation
AN INSTITUTE OF THE BAVARIAN ACADEMY OF SCIENCES AND HUMANITIES

# AI Literacy

- Knowing and understanding core AI concepts and methods
- Evaluating AI technologies with respect to suitability for specific applications
- Reflecting effects of AI adoption on society and environment

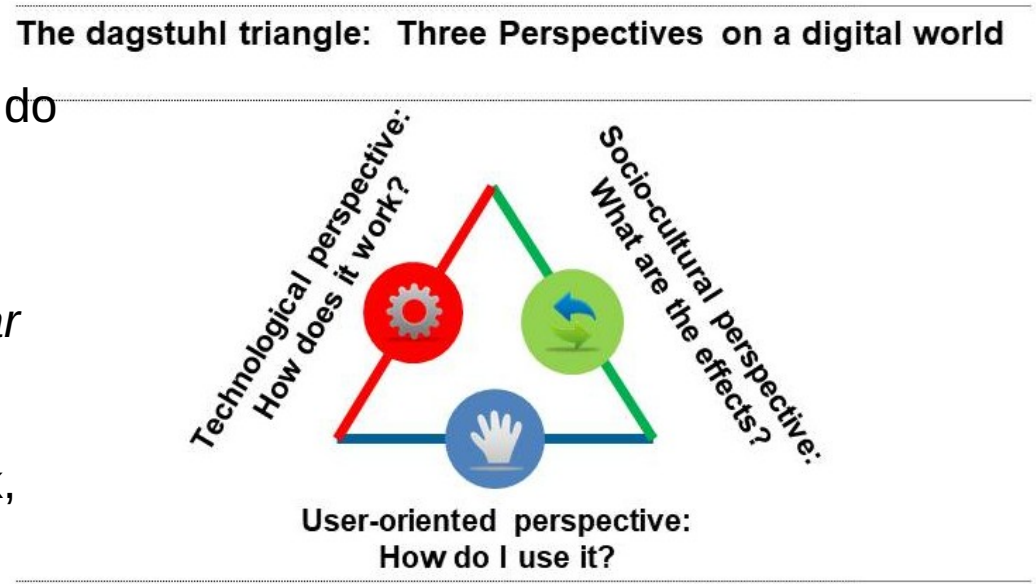
Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. Computers and Education: Artificial Intelligence, 2, 100041.

# Learning about, with, (and despite) AI

- Realistic evaluation of what AI methods can do
- Avoidance of anthropomorphisation
- Calibration of trust in AI systems

`*But one does to need to understand how a car works to be able to drive a car*'

We know much more about cars than we think, e.g. a car cannot fly, it is dangerous to drive against a wall, ...



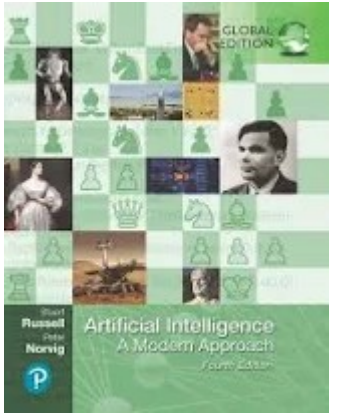

Ute Schmid (2025). Grundkompetenzen im Bereich Künstliche Intelligenz (AI Literacy). In: Gerold Brägger & Hans-Günter Rolff (Hrsg.) Handbuch: Lernen mit digitalen Medien, 3. Auflage. Weinheim: Beltz.

# Artificial Intelligence

- Name given 1956 by John McCarthy (MIT/Stanford)
- As **part of computer scienc**e/informatics
- Based on the assumption that all (many/relevant) aspects of human intelligence can be formalized by algorithms and simulated by computer programs
- *AI is the study of how to make computers perform intelligent tasks that, in the past, could only be performed by humans* (Elaine Rich, 1983)

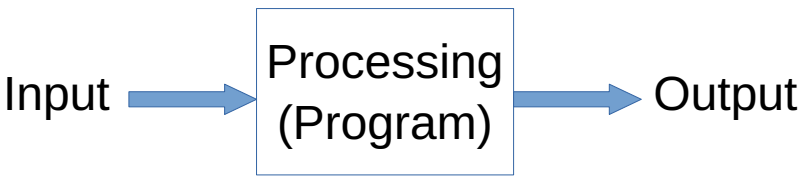
Research in algorithms, formal approaches, and applications addressing

- Knowledge representation and reasoning
- Automated problem solving and planning
- Machine Learning
- Computer Vision (Object and scene recognition)
- Game Playing
- Natural Language Processing
- AI Robotics

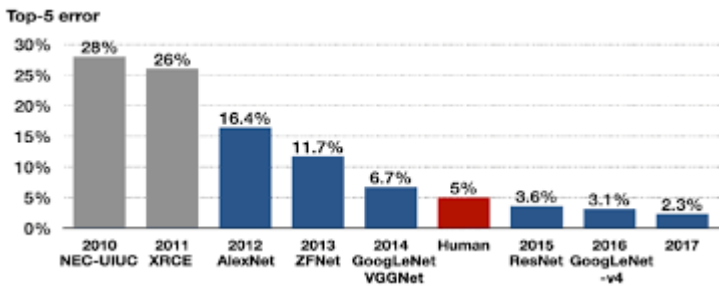


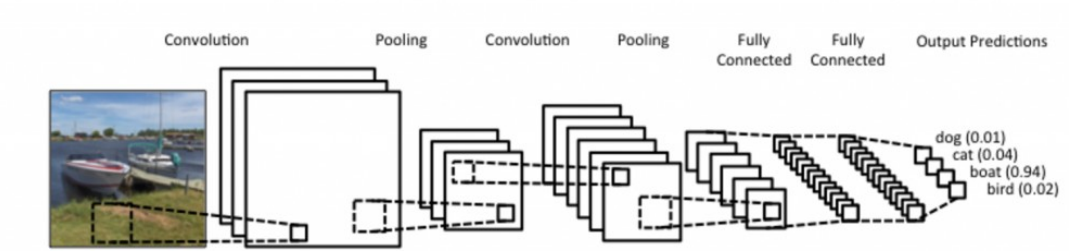
# Standard vs AI Programs

- Most computer programs are not based on AI methods!
- Application of AI methods means to give up requirements concerning correctness and completeness
- AI methods are applied if:
  - A problem is so complex that its (optimal) solution cannot be computed efficiently → **heuristic methods**, approximation
  - A problem involves complex knowledge and (domain/common sense) reasoning → **knowledge-based methods**
  - A problem cannot be described explicitly → **machine learning**, replacement of explicit algorithms by (**black box**) models induced from data



# Learned models cannot be error-free but are nevertheless useful


Convolutional Neural Network CNNs (LeCun, 1998)
Alex Krizhevsky, (PhD student of G. Hinton, 2012)

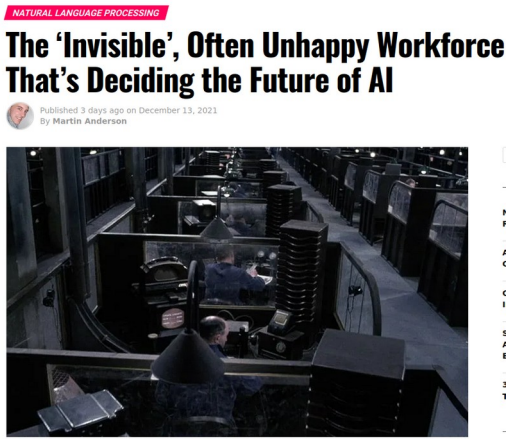
Discriminative vs. generative AI

- Classification of input
- Generation of content (text, image, video, code) from natural language prompts
- Different architectures: e.g. convolutional neural network / transformer

# Anthropomorphization Trap

- AI systems are “self learning”
  - Classifiers need large amounts of training data which have to be labeled by humans
  - Generative AI tools rely on extensive dialog training and are heavily engineered
  - AI companies employ huge amounts of data workers
- If an AI system is good in one domain, it also masters tasks of similar difficulty (AGI)
- AI systems understand what we say  (Eliza-Effect)

https://www.unite.ai/the-invisible-often-unhappy-workforce-thats-deciding-the-future-of-ai/



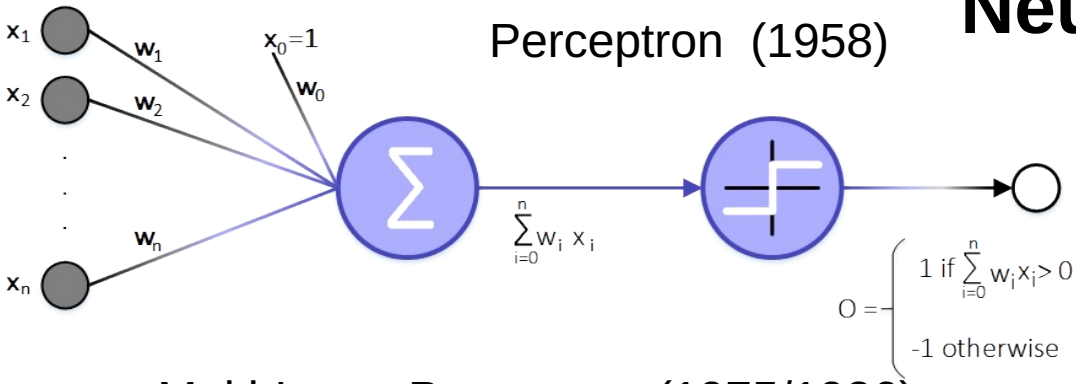
# Human Learning


Josh Tenenbaum

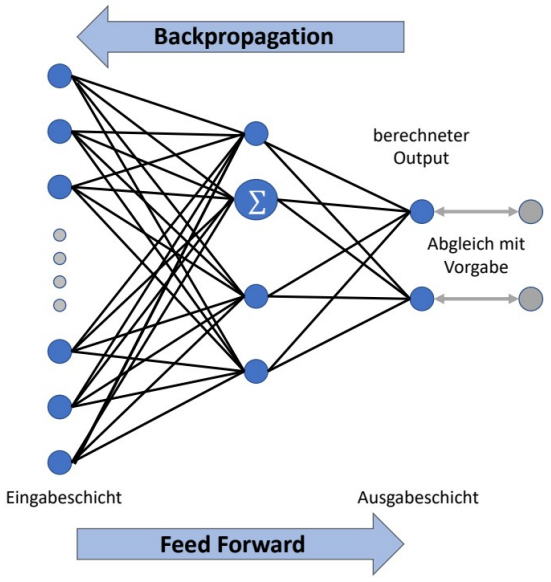
- Inductive Bias (*do not confuse with sampling bias!*)
- Generalization over data is only possible with inductive bias, otherwise one could only store information (rote learning)
- Over-generalization: goed (instead of went)

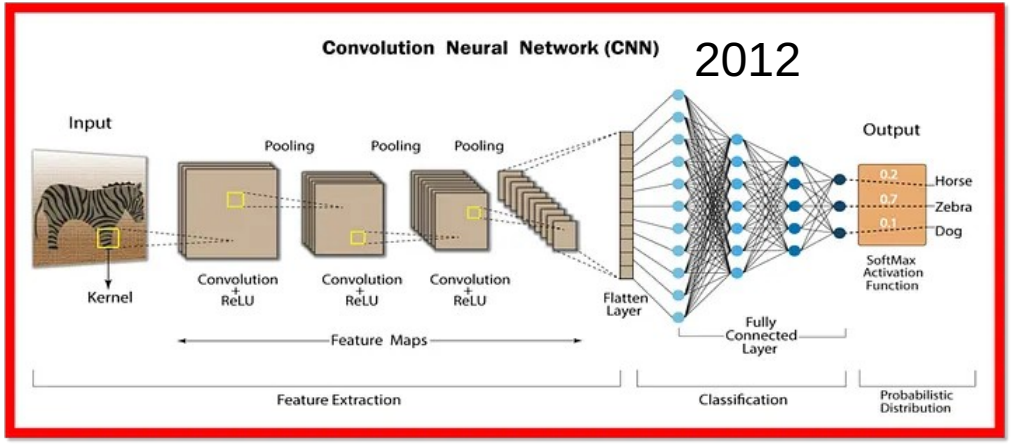
- Dark side of inductive generalization: Stereotypes and prejudice (girls are not good in math, boys are not good in interpreting poems)

# Neural Networks




Multi-Layer Perceptron (1975/1986)




https://medium.com/tech-ai-made-easy/what-are-convolutional-neural-networks-cnns-456175b5691c






MINT-LinK

BMBF Projekt
KI-Kompetenzen
im außerschulischen
Kontext

klaro!KI

BMUV-Project klaro!KI

KI-Campus BETA

KI-Campus-Original

**Data Literacy für die Grundschule**

5 Module à 60 Minuten

Leistungsnachweis

# Generative AI and Understanding

> Gary: *What does it refer to in this sentence? The trophy doesn't fit into the brown suitcase because it is too small.*
>
> ChatGPT: *In the given sentence, "it" refers to the trophy.*

Winograd Challenge

https://www.salon.com/2023/04/30/chatgpt-chatbots-artificial-intelligence-llms/

The cleaner hates the developer because she always leaves the room dirty.

DeepL translation: Die Reinigungskraft hasst den Entwickler, weil sie das Zimmer immer schmutzig hinterlässt.

Jonas Troles & Ute Schmid (WMT 2021). Extending Challenge Sets to Uncover Gender Bias in Machine Translation – Impact of Stereotypical Verbs and Adjectives

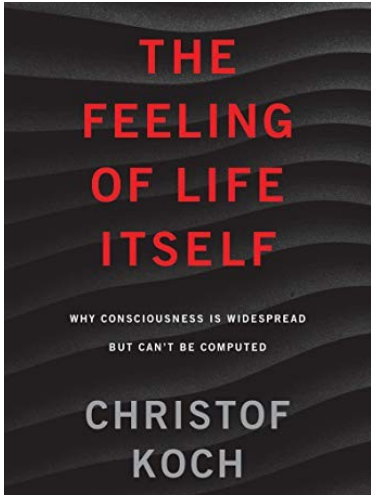
# Weak/Strong/General AI

- Most AI systems are restricted to one very specific domain (**narrow,** not **general AI**)
  - *A system which is good at classifying animals cannot classify traffic signs or skin cancer*
- Inadmissible anthropomorphization (also due to terms such as intelligence, reasoning, perception, decision making, learning)
  - Term `intelligence´ as used in everyday life is typically associated with an above average level of cognitive abilities (excellent chess player, PhD in physics vs. recognizing a cat or loading a dish washer)
- Most AI researchers are concerned with algorithms which can solve specific problems and are not making claims about or are concerned with the question whether AI systems can have intelligence similar to human intelligence – AI simulates and is not equivalent to human cognitive processes (**weak**, not **strong AI**)
- **AGI** (artificial general intelligence) as special branch of AI research striving for approaches which are similar to human intelligence

# Artificial General Intelligence?

- AI systems can simulate many aspects of human reasoning, problem solving, and learning
- ➔ But: General AI requires consciousness in a broader sense, including
  - meta cognition
  - self awareness
  - intentionality
  - qualia

*Why consciousness is widespread but can't be computed*




"I said in Dorian Gray that the great sins of the world take place in the brain: but it is in the brain that everything takes place. We know now that we do not see with the eyes or hear with the ears. They are really channels for the transmission, adequate or inadequate, of sense impressions. It is in the brain that the poppy is red, that the apple is odorous, that the skylark sings."

— Oscar Wilde, De Profundis

# Sampling Biases

e.g. gender bias
Amazon Recruiting Tool 2015
Rating applicants for software developer jobs

e.g. ethnic bias
Google Photos

PetaPixel News Reviews Guides Learn Equipment Glossary Newsl

**Google's Photos App is Still Unable to Find Gorillas**

MAY 22, 2023 PESALA BANDARA

Overcoming Racial Bias In AI Systems And Startlingly Even In AI Self-Driving Cars

Racial bias in a medical algorithm favors white patients over sicker black patients

AI expert calls for end to UK use of 'racially biased' algorithms

AI Bias Could Put Women's Lives At Risk - A Challenge For Regulators

**Gender bias in AI: building fairer algorithms**

**Bias in AI: A problem recognized but still unresolved**

Amazon, Apple, Google, IBM, and Microsoft worse at transcribing black people's voices than white people's with AI voice recognition, study finds

**Millions of black people affected by racial bias in health-care algorithms**

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

**When It Comes to Gorillas, Google Photos Remains Blind**

Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

*The Week in Tech: Algorithmic Bias Is Bad. Uncovering It Is Good.*

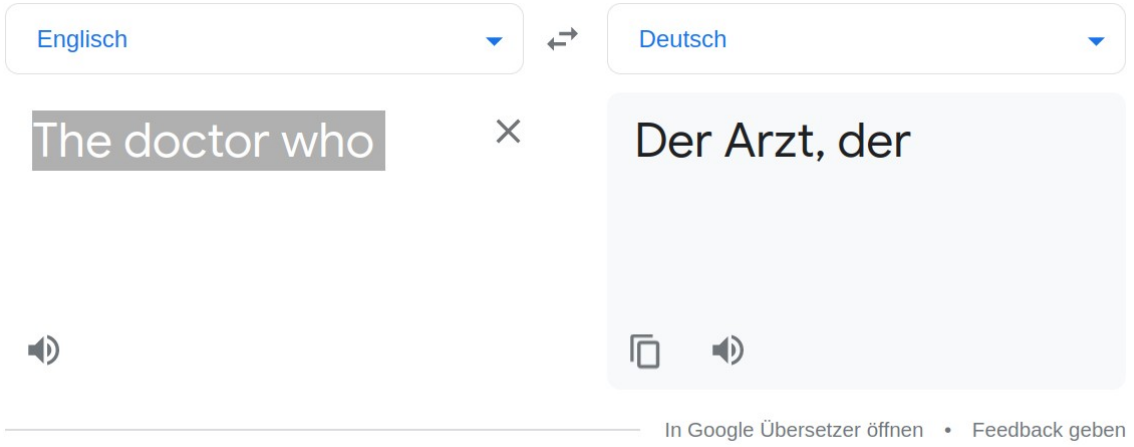
Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

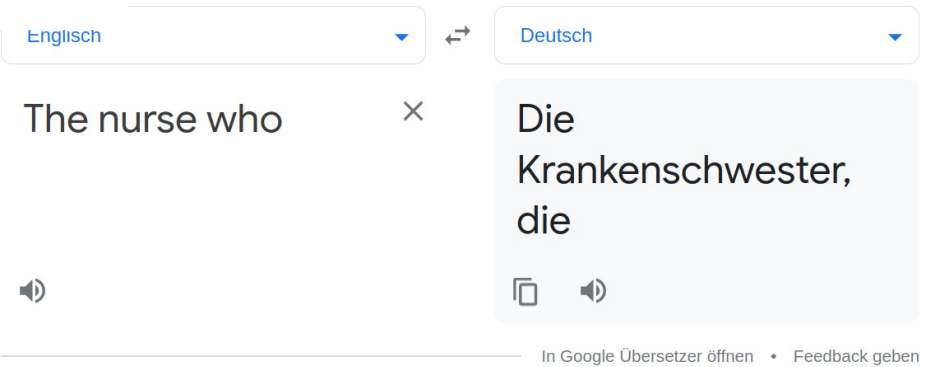
Artificial Intelligence has a gender bias problem – just ask Siri

**The Best Algorithms Struggle to Recognize Black Faces Equally**

US government tests find even top-performing facial recognition systems misidentify blacks at rates five to 10 times higher than they do whites.

# Bias in Machine Translation




Jonas Troles & Ute Schmid (WMT 2021). Extending Challenge Sets to Uncover Gender Bias in Machine Translation – Impact of Stereotypical Verbs and Adjectives

# Limited Robustness


NEWS CAREERS COMMENTARY JOURNALS Science

News Home All News ScienceInsider News Features DONAT

HOME > NEWS > ALL NEWS > A TURTLE—OR A RIFLE? HACKERS EASILY FOOL AIS INTO SEEING THE WRONG THING

NEWS | TECHNOLOGY

**A turtle—or a rifle? Hackers easily fool AIs into seeing the wrong thing**

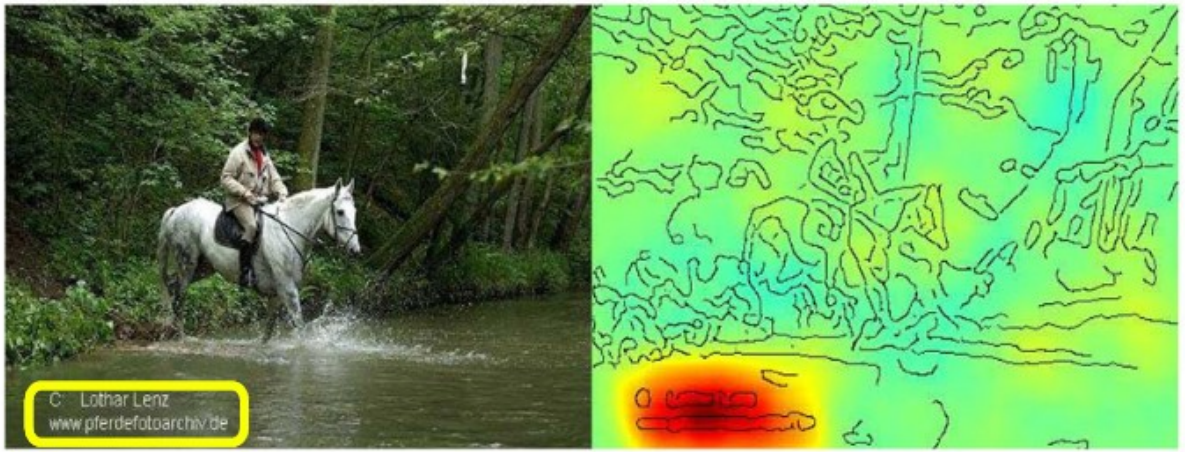
Adversarial attacks highlight lack of security in machine learning algorithms

19 JUL 2018 · BY MATTHEW HUTSON

https://www.etsmtl.ca/en/news/brittleness-of-deep-learning-models

# Explainable AI (XAI)

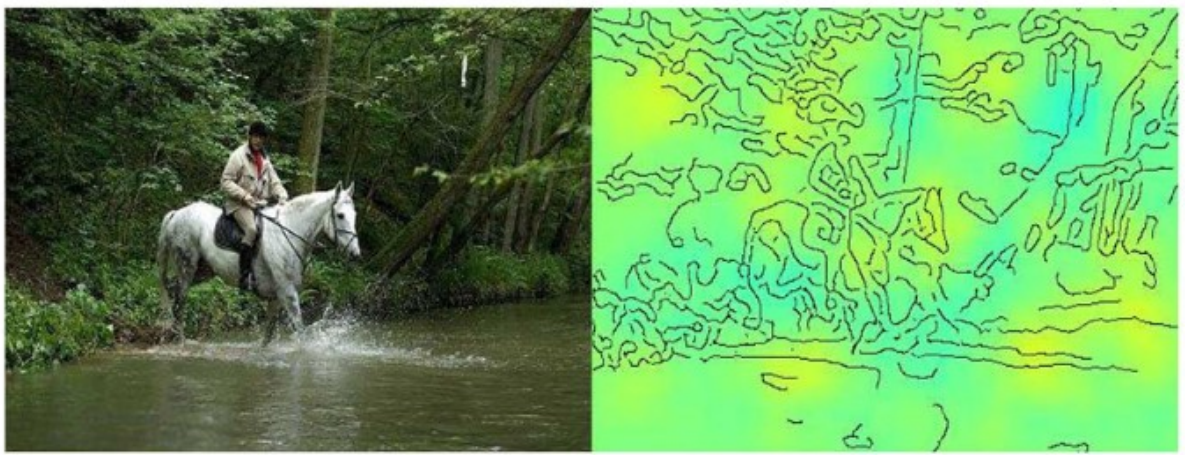
Horse-picture from Pascal VOC data set




Source tag present

↓

Classified as horse



No source tag present

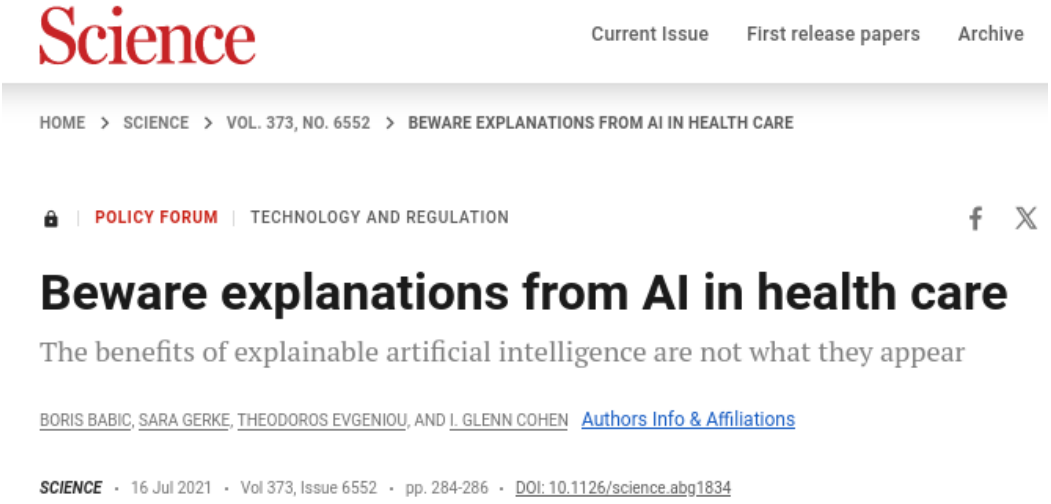
↓

Not classified as horse

Lapuschkin, Sebastian, et al. "Unmasking Clever Hans predictors and assessing what machines really learn." Nature communications 10.1 (2019): 1096.

# Tailored Explanations

- Explainability is not useful per se
  - ➔ Explain to whom and for what information need


Science

Current Issue First release papers Archive

HOME > SCIENCE > VOL. 373, NO. 6552 > BEWARE EXPLANATIONS FROM AI IN HEALTH CARE

POLICY FORUM | TECHNOLOGY AND REGULATION

**Beware explanations from AI in health care**

The benefits of explainable artificial intelligence are not what they appear

BORIS BABIC, SARA GERKE, THEODOROS EVGENIOU, AND I. GLENN COHEN Authors Info & Affiliations

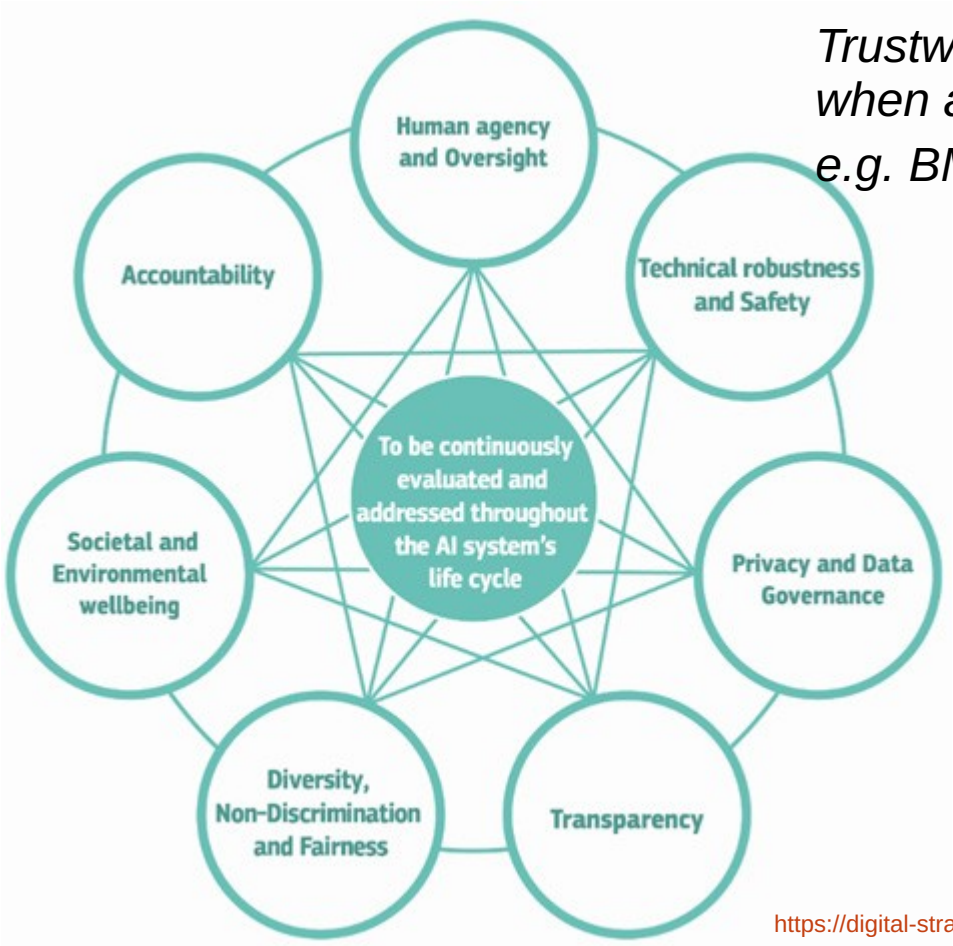
SCIENCE · 16 Jul 2021 · Vol 373, Issue 6552 · pp. 284-286 · DOI: 10.1126/science.abg1834

https://www.sciencemediacenter.de/angebote/explainable-ai-in-der-medizin-21098

- **For model developers:** overfitting, biases
- **For domain experts:** comprehensibility of AI decision making, calibrated (not naive) trust, explain to revise
- **For end users:** transparency of data-based decision algorithms (insurance, health-apps)

Atzmueller, M., Fürnkranz, J., Kliegr, T., & Schmid, U. (2024). Explainable and interpretable machine learning and data mining. Data Mining and Knowledge Discovery, 38(5), 2571-2595.

# Requirements for Trustworthy AI




*Trustworthiness is of high relevance when adopting AI tools in work settings*

*e.g. BMFTR project Ethyde at bidt*

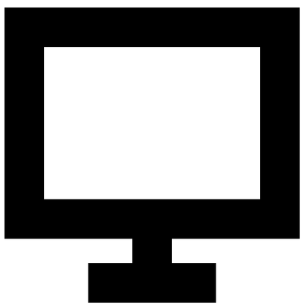

European AI Act
High-level Expert Group

- Knowledge informed (contributes to robustness)
- Explainable/Comprehensible (contributes to transparency)
- Revisable, „better together“ (contributes to human agency and oversight)

Schmid, U. (2024). Trustworthy Artificial Intelligence: Comprehensible, Transparent and Correctable. In In: Werthner, H., et al. Introduction to Digital Humanism. Springer, Cham. https://doi.org/10.1007/978-3-031-45304-5_10

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

# Joint Decision Making & Problem Solving

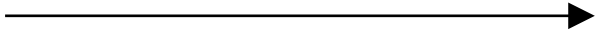


Class of Tissue is P3

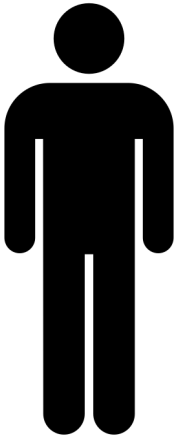
Component nok

Reduce fertilizer by 20%

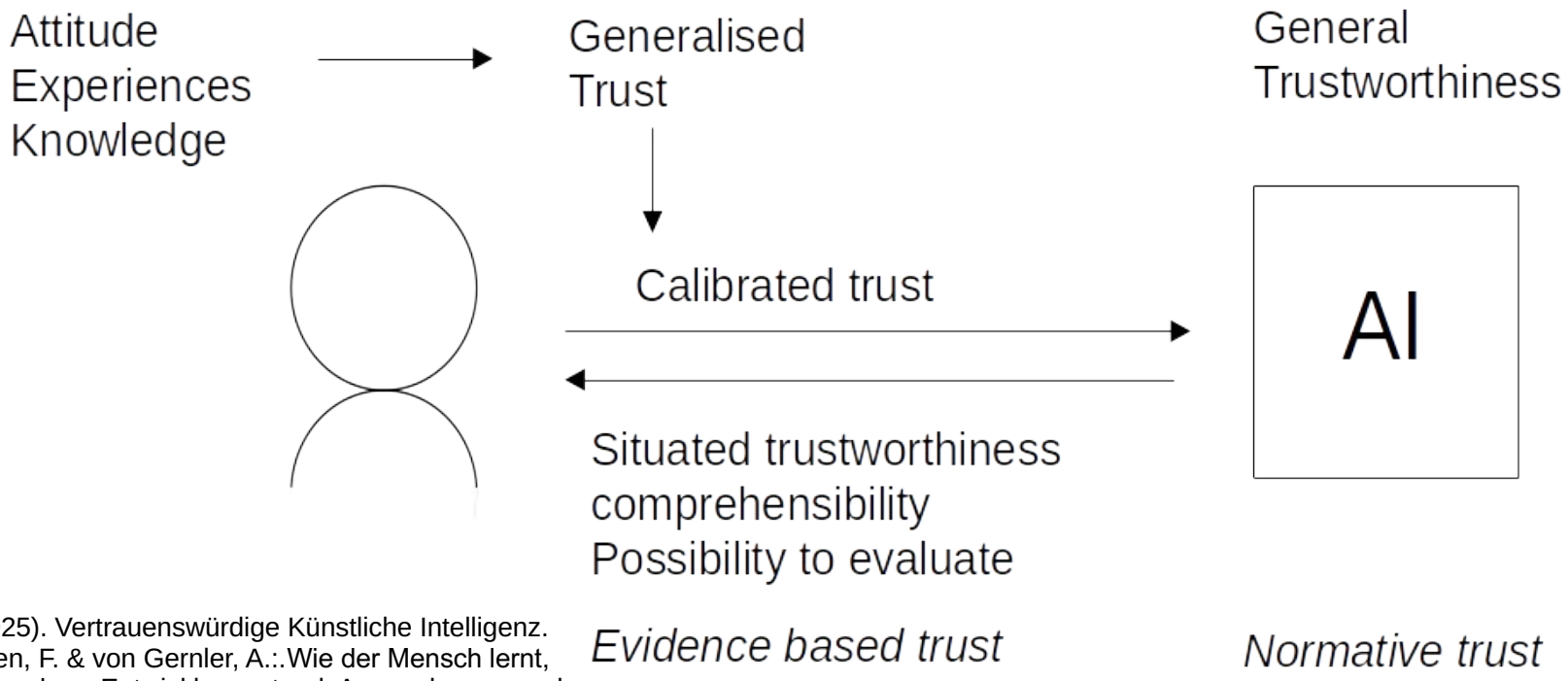
Give loan



```
statistics.stdev([...])
```



# Human Trust and Trustworthiness




Schmid, U. (2025). Vertrauenswürdige Künstliche Intelligenz. In Schmiedchen, F. & von Gernler, A.:.Wie der Mensch lernt, die KI zu beherrschen: Entwicklungsstand, Anwendungen und Steuerungserfordernisse der Künstlichen Intelligenz. Springer.

Article | Open access | Published: 28 October 2024

# When combinations of humans and AI are useful: A systematic review and meta-analysis

Michelle Vaccaro, Abdullah Almaatouq & Thomas Malone

*Nature Human Behaviour* **8**, 2293–2303 (2024) | Cite this article

- On average among recent experiments, human–AI systems did not exhibit synergy: the human–AI groups performed worse than either the human alone or the AI alone
- Most (>95%) of the human–AI systems in our dataset involved humans making the final decisions after receiving input from AI algorithms. In these cases, one potential explanation of our result is that, when the humans are better than the algorithms overall, they are also better at deciding in which cases to trust their own opinions and in which to rely more on the algorithm's opinions.

# Requirements for Successful Human-AI Teams

**Problems:**

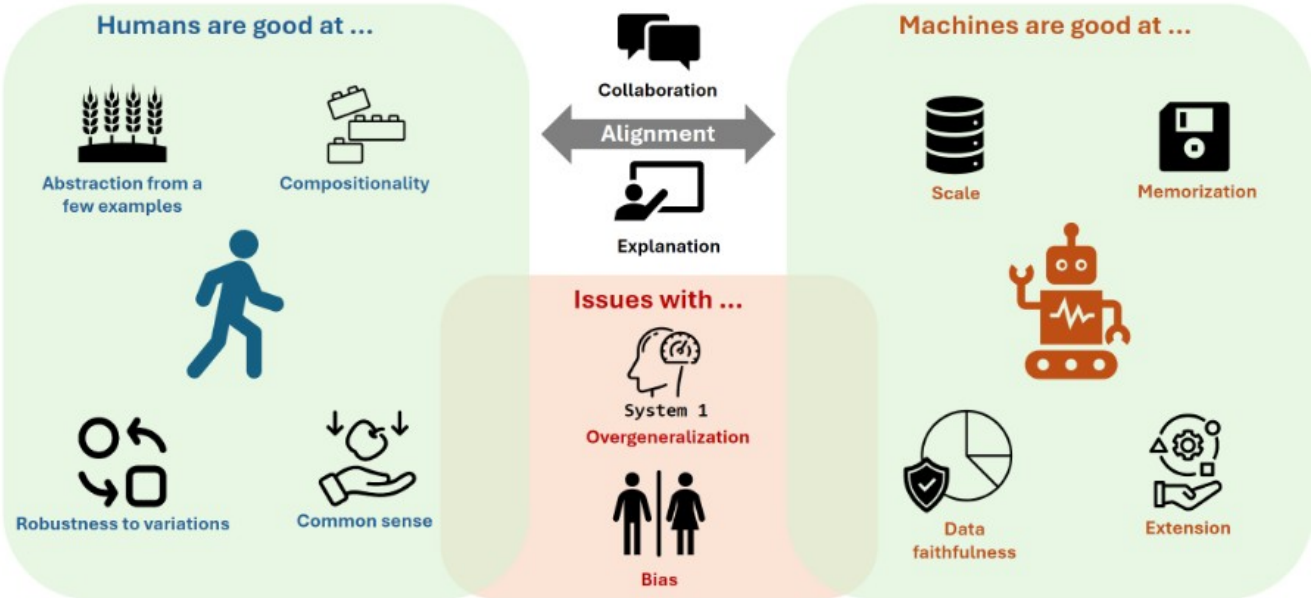
- ✗ Overtrust (automation bias) or undertrust → need of support of trust calibration
- ✗ Domain competence is relevant → reduce danger of deskilling by over-delegation to AI systems

**Trust calibration requires:**

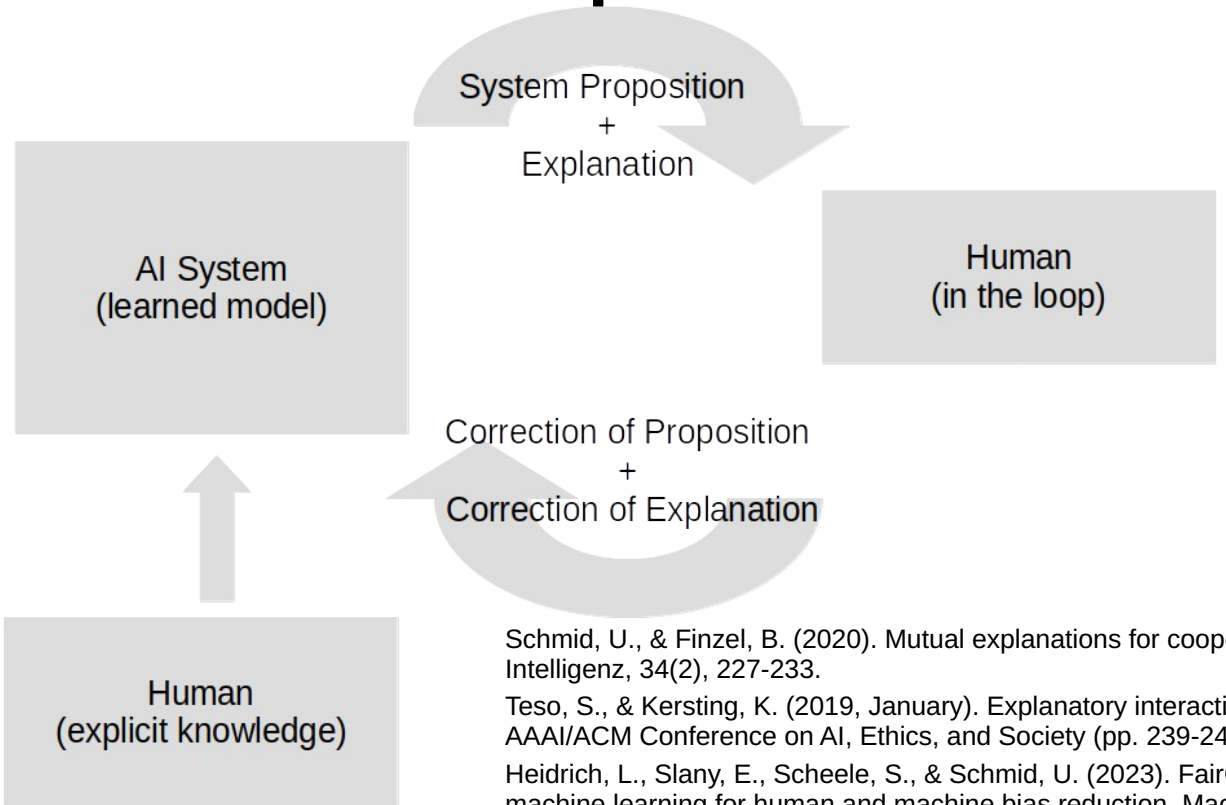
- ✔ Domain competence
- ✔ AI Literacy
- ✔ Understanding of why the system computed a specific output (XAI)
- ✔ Cognitive involvement in the task: explanatory interactive machine learning
- ✔ Alignment of human cognitive processes and AI system

# Aligning Generalisation Between Humans and Machines

Filip Ilievski[1*], Barbara Hammer[2], Frank van Harmelen[1], Benjamin Paassen[2], Sascha Saralajew[3], Ute Schmid[4], Michael Biehl[5], Marianna Bolognesi[6], Xin Luna Dong[7], Kiril Gashteovski[3,8], Pascal Hitzler[9], Giuseppe Marra[10], Pasquale Minervini[11,12], Martin Mundt[13,14], Axel-Cyrille Ngonga Ngomo[15], Alessandro Oltramari[16], Gabriella Pasi[17], Zeynep G. Saribatur[18], Luciano Serafini[19], John Shawe-Taylor[20], Vered Shwartz[21,22], Gabriella Skitalinskaya[23], Clemens Stachl[24,25], Gido M. van de Ven[10], Thomas Villmann[26,27]



# Explain to Revise




- Learned models are never 100% correct
- In highly specialized domains, getting a sufficient amount of high quality expert labeled data is very expensive
- Decision making needs to be in human control
- → **Explanatory interactive ML**


Schmid, U., & Finzel, B. (2020). Mutual explanations for cooperative decision making in medicine. KI-Künstliche Intelligenz, 34(2), 227-233.

Teso, S., & Kersting, K. (2019, January). Explanatory interactive machine learning. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 239-245). CAIPI

Heidrich, L., Slany, E., Scheele, S., & Schmid, U. (2023). FairCaipi: a combination of explanatory interactive and fair machine learning for human and machine bias reduction. Machine learning and knowledge extraction, 5(4), 1519-1538.

Finzel, B., Knoblach, J., Thaler, A., & Schmid, U. (2024). Near hit and near miss example explanations for model revision in binary image classification. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 260-271). Springer

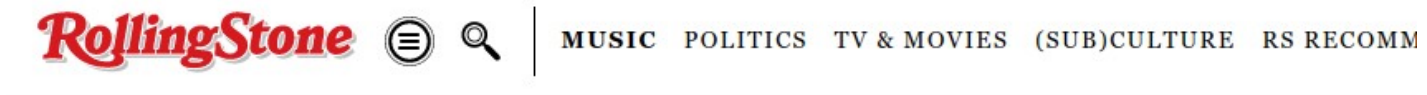
# Efficiency-Competency Dilemma

- Discriminative and generative AI both are based on information/knowledge given by humans
  - ground truth labeling of training data
  - human-in-the-loop reinforcement learning, chain of thought prompting, retrieval augmented generation (RAG)
- What if the “generation genAI” does no longer acquire necessary skills?
- Deskilling by over-delegation

# Problems with Over-delegation

- Deskilling includes
  - Loss of crucial abilities (e.g. structured writing, programming)
  - Together with loss of quality criteria necessary for evaluating own as well as given content
- GenAI models mostly generate “average” solutions – originality will be lost
- We want to use AI tools to support solving of real world problems: AI agents talking with AI agents lead to a loss of grounding in reality

# ChatGPT Reception – Synthetic friendliness, `Californication‘




UNCANNY VALLEY

## Nick Cave Slams AI Attempts at Nick Cave Songs

Fans tasked controversial AI bot ChatGPT to write songs in the musician's trademark style, and he was not amused

BY CHARISMA MADARANG

JANUARY 16, 2023

He continued, “Mark, thanks for the song, but with all the love and respect in the world, this song is bullshit, a grotesque mockery of what it is to be human.”

# Ways out of the Dilemma

AI Literacy

*Learning about AI*

New Human-AI-Interfaces

*Learning with AI*



for Learning Despite AI

# AI in Education

## AI as Topic

- Basic concepts
- Machine learning
- Generative AI
- Knowledge-based methods
- Competent use of AI tools

## AI as Tool

**Intelligent Tutorsystems** (ITS)

Individual support for specialized domains

**Learning Analytics**

Management and Diagnostics

**Generative AI**

- Educational materials
- Test generation
- Support for grading
- Give ideas

# Learning with AI

- Digital tools have been ignored for a long time in education (schools as well as universities) – Wikipedia, spell checking and grammer checking, machine translation
- ChatGPT has finally triggered discussion about use of AI (by students and teachers)
- Forbidding use is not helpful (who cared when home work has been done by other persons?)
- Necessary discussion about which competencies are relevant in each discipline/subject, which competencies become less relevant, which new competencies are necessary
- Danger of simple edu-tech solutions: nudging leads to loss of intrinsic motivation, multiple-choice tests hinder aquistion of deeper understanding and transferable problem solving competencies

# Intelligent Tutorsystems

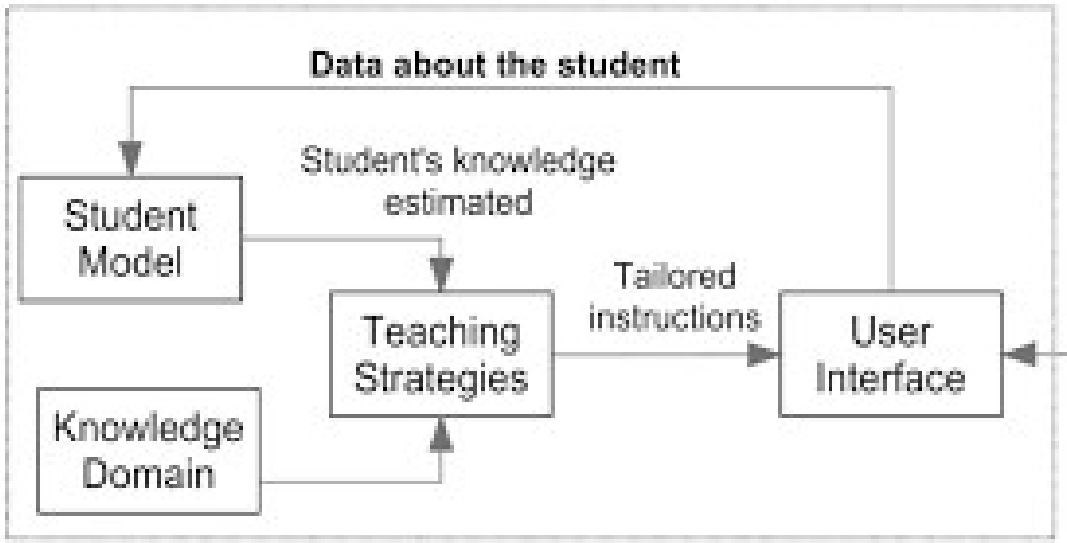


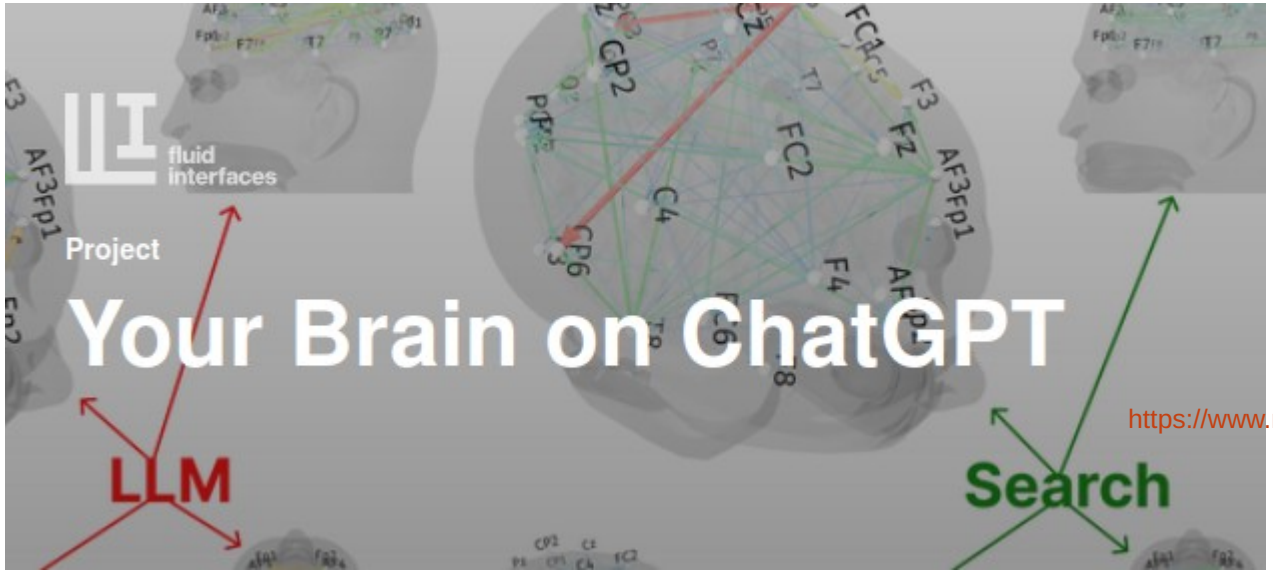

Fig. 1. Basic architecture of an ITS [7].

**Knowledge domain**
e.g. topics in physics, history, English grammar

**Student model**
Individual learning history +
Identification of misconceptions or knowledge gaps in the task context

**Teaching strategy**
Choice of topics and tasks
Targeted feedback

Butz, Hua, Maguire, 2006




cognitive cost of using an LLM in the educational context of writing an essay

https://www.media.mit.edu/projects/your-brain-on-chatgpt/overview

EEG analysis presented robust evidence that LLM, Search Engine and Brain-only groups had significantly different neural connectivity patterns, reflecting divergent cognitive strategies. Brain connectivity systematically scaled down with the amount of external support: the Brain-only group exhibited the strongest, widest-ranging networks, Search Engine group showed intermediate engagement, and LLM assistance elicited the weakest overall coupling.

The reported ownership of LLM group's essays in the interviews was low. The Search Engine group had strong ownership, but lesser than the Brain-only group. The LLM group also fell behind in their ability to quote from the essays they wrote just minutes prior.

# New Interfaces: ITS & XAI & GenAI

- Support of understanding by tailored feedback (e.g. teaching written subtraction, redox reactions, spoken dialog, essay writing)
- More resilience against desinformation (e.g. uncovering manipulative argumentative structures)
- Targeted support for human control and oversight and calibrated trust