



Niklas Kühl

Professor Wirtschaftsinformatik
und humanzentrische KI

Der „Meat Proxy“ 
... und die Rolle von KI-
Erklärungen

01 Wie viel Vertrauen in KI ist *richtig*?

02 Wer kann KI-Empfehlungen noch *beurteilen*?

03 Was macht eine KI-Erklärung *nützlich*?

Dieses ~~Meme~~ Bild wollen wir heute füllen!



Die Simpsons: Staffel 7, Episode 7



Quelle: 20th Century Fox, The Simpsons

2026: Aus der Karikatur ist eine Rechtspflicht geworden.

Art. 14 KI-VO Menschliche Aufsicht

- (1) Hochrisiko-KI-Systeme werden so konzipiert und entwickelt, dass sie während der Dauer ihrer Verwendung – auch mit geeigneten Instrumenten einer Mensch-Maschine-Schnittstelle – von natürlichen Personen wirksam beaufsichtigt werden können.
- (2) Die menschliche Aufsicht dient der Verhinderung oder Minimierung der Risiken für Gesundheit, Sicherheit oder Grundrechte, die entstehen können, wenn ein Hochrisiko-KI-System im Einklang mit seiner Zweckbestimmung oder im Rahmen einer vernünftigerweise vorhersehbaren Fehlanwendung verwendet wird, insbesondere wenn solche Risiken trotz der Einhaltung anderer Anforderungen dieses Abschnitts fortbestehen.
- (3) Die Aufsichtsmaßnahmen müssen den Risiken, dem Grad der Autonomie und dem Kontext der Nutzung des Hochrisiko-KI-Systems angemessen sein und werden durch eine oder beide der folgenden Arten von Vorkehrungen gewährleistet:
 - a) Vorkehrungen, die vor dem Inverkehrbringen oder der Inbetriebnahme vom Anbieter bestimmt und, sofern technisch machbar, in das Hochrisiko-KI-System eingebaut werden;
 - b) Vorkehrungen, die vor dem Inverkehrbringen oder der Inbetriebnahme des Hochrisiko-KI-Systems vom Anbieter bestimmt werden und dazu geeignet sind, vom Betreiber umgesetzt zu werden.

Artikel 14 EU AI Act verlangt für Hochrisiko-Systeme wirksame Aufsicht durch natürliche Personen.

Die Frage ist also nicht mehr...

Sitzt ein Mensch („irgendwo / irgendwie“) dabei?

SONDERN:

Ist seine Aufsicht wirksam?

Quelle: EU KI Verordnung

Zwei Fragen an Sie.

1

Wer von Ihnen setzt ein KI-System produktiv ein?

2

Wer von Ihnen hat für dieses System eine Person benannt, die die Befugnis und die Information hat, zu widersprechen?

→ Um diese Lücke soll es heute gehen.

Das Problem der effektiven menschlichen Aufsicht ist kein akademisches.

Künstliche Intelligenz

KI erfindet jede dritte Antwort

Stand: 27.10.2025 • 14:46 Uhr

Eine neue Studie der Europäischen Rundfunkunion zeigt alarmierende Fehlerquoten bei populären KI-Chatbots. Die Systeme erfinden regelmäßig Informationen und geben sogar falsche Quellen an.

[1]



Vertrauenslücke

Der effektiven menschlichen Aufsicht fehlt die Evidenz: KI-Systeme irren sich - ob Menschen die Fehler erkennen, ist ungeprüft.

[4]



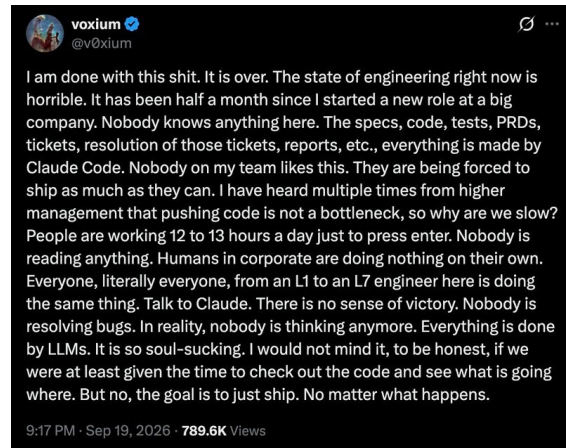
[2]



Gestaltungslücke

Effektive menschliche Aufsicht fordert der AI Act - praxistaugliche Standards und Prüfverfahren dafür fehlen.

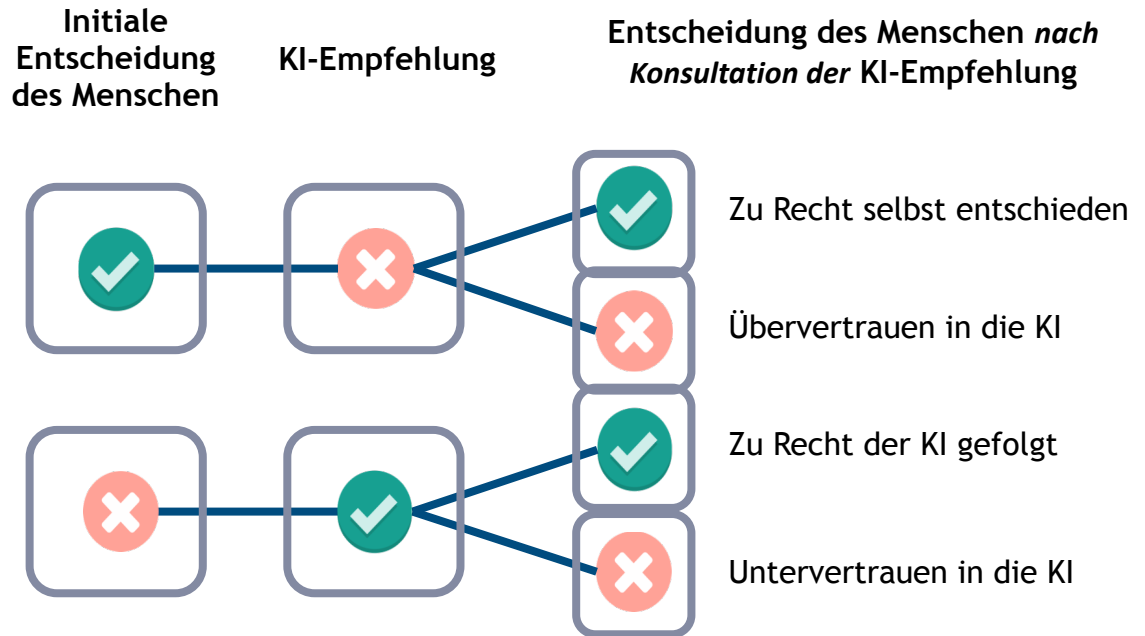
[5]



[1] <https://www.tagesschau.de/wissen/technologie/kuenstliche-intelligenz-fakten-100.html> [2] <https://background.tagesspiegel.de/digitalisierung-und-ki/briefing/stop-the-clock-forderung-sinnvoll-oder-unkoordinierter-rueckzieher> [3] <https://www.theguardian.com/society/2026/jun/09/doctors-nhs-could-be-sued-mistakes-ai-tools-medical-protection-society-report> [4] Jussupow, E., Benbasat, I., & Heinzl, A. (2024). An integrative perspective on algorithm aversion and appreciation in decision-making. *MIS Quarterly*, 48(4), 1575-1590. <https://doi.org/10.25300/MISQ/2024/18512> [5] Lu, Q., Zhu, L., Xu, X., Whittle, J., Zowghi, D., & Jacquet, A. (2024). Responsible AI pattern catalogue: A collection of best practices for AI governance and engineering. *ACM Computing Surveys*, 56(7), Article 173. <https://doi.org/10.1145/3626734.16> Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *Journal of Strategic Information Systems*, 34(2), 101885. <https://doi.org/10.1016/j.isis.2024.101885>

Wodurch zeichnet sich eine effektive menschliche Aufsicht denn aus?

- **Beispiel:** Aufsicht über ein Atomkraftwerk „Reaktor hochfahren“.
- **Annahme:** Es gibt nur richtige oder falsche Entscheidungen.
- **Annahme:** Wir wissen, was richtig und was falsch ist.
- **Legende:**
 - ✓ = richtig
 - ✗ = falsch



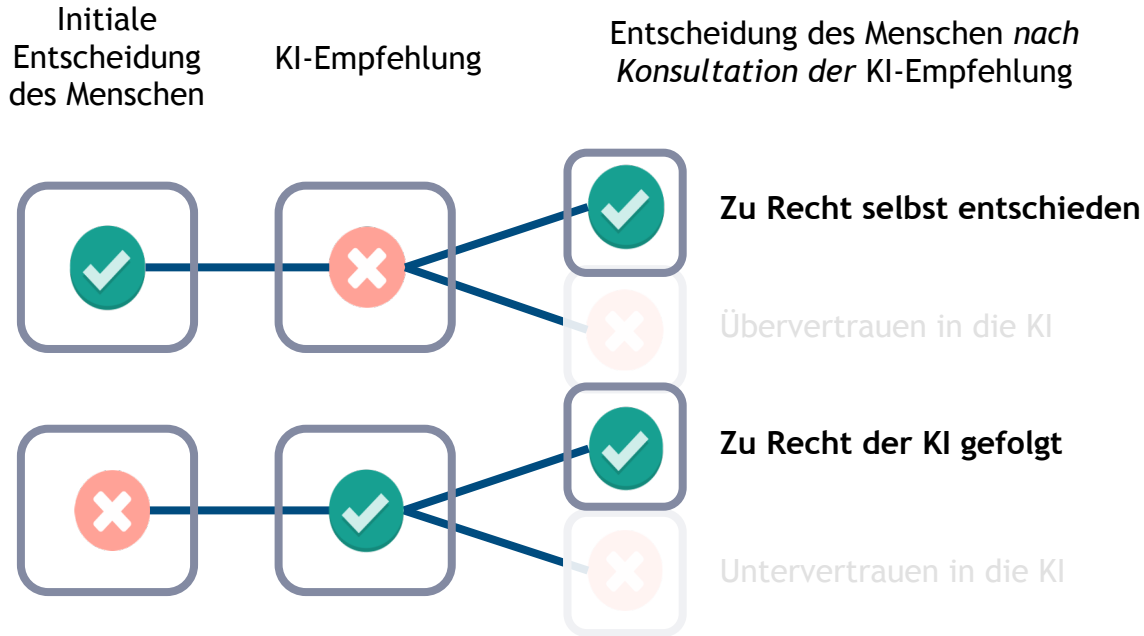
Quelle: Schemmer, M., Kühl, N., Benz, C., Bartos, A., & Satzger, G. (2023, March). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In Proceedings of the 28th international conference on intelligent user interfaces (pp. 410-422).

Was haben wir gelernt?

Kalibriertes Vertrauen sollte
immer das Ziel sein, nicht
Vertrauen pauschal.



Aber wie komme ich dahin?



Wenn das Ziel *kalibriertes Vertrauen* ist (und nicht *pauschales*) – wie kann ich dies erreichen?

- Schulung der Mitarbeitenden
- Unsicherheitskommunikation der KI
- Erklärungen & Transparenz der KI
- „Cognitive Forcing“
- ...

Die naheliegende Antwort zum Erreichen von kalibriertem Vertrauen lautet: Erklärungen bzw. Explainable AI (XAI).

Denn: Wenn das System begründet, kann der Mensch prüfen. Genau darauf setzen Regulierung und Produktdesign derzeit flächendeckend.

Forbes

FORBES > INNOVATION > CONSUMER TECH

Explainable AI Is Trending And Here's Why

Article | [Open access](#) | Published: 28 December 2024

Explainable AI improves task performance in human–AI collaboration

[Julian Senoner](#), [Simon Schallmoser](#), [Bernhard Kratzwald](#), [Stefan Feuerriegel](#) & [Torbjørn Netland](#) 



AI can see clearly now: Why transparency leads to ethical and fair AI systems

Forbes

FORBES > INNOVATION > ENTERPRISE TECH

Building Trust In AI: The Case For Transparency

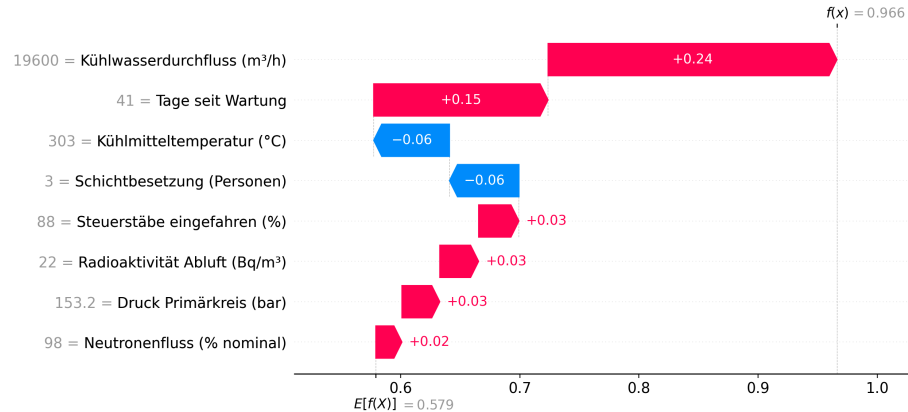
Vainio-Pekka, H., Agbese, M. O. O., Jantunen, M., Vakkuri, V., Mikkonen, T., Rousi, R., & Abrahamsson, P. (2023). The role of explainable AI in the research field of AI ethics. *ACM Transactions on Interactive Intelligent Systems*, 13(4), 1-39. Forbes: <https://www.forbes.com/sites/jenniferhicks/2022/07/28/explainable-ai-is-trending-and-heres-why/> Silicon Angle: <https://siliconangle.com/2024/03/31/ai-can-see-clearly-now-transparency-leads-ethical-fair-ai-systems/> Senoner, J., Schallmoser, S., Kratzwald, B., Feuerriegel, S., & Netland, T. (2024). Explainable AI improves task performance in human-AI collaboration. *Scientific Reports*, 14(1), 31150. Forbes: <https://www.forbes.com/sites/bernardmarr/2024/05/03/building-trust-in-ai-the-case-for-transparency/>

Was sind Erklärungen? Zwei Beispiele aus der Praxis.



„Reaktor hochfahren!“

Warum empfiehlt das Modell, den Reaktor hochzufahren?

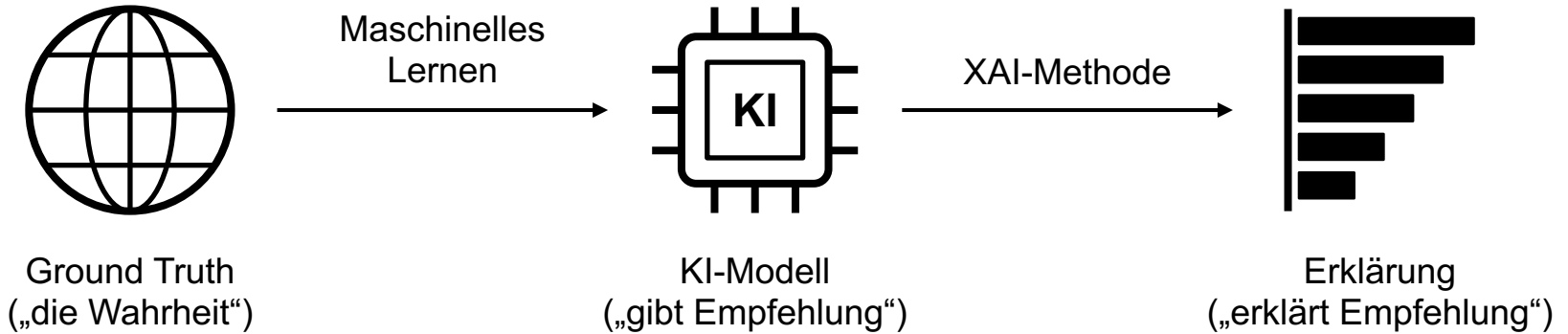


**Wichtigkeit der
„Features“**

„Erklärung: hochfahren vor allem wegen des hohen Kühlwasserdurchflusses und kürzlicher Wartung. Erhöhte Kühlmitteltemperatur und nur 3-Personen-Schichtbesetzung sprechen dagegen, wiegen aber weniger.“

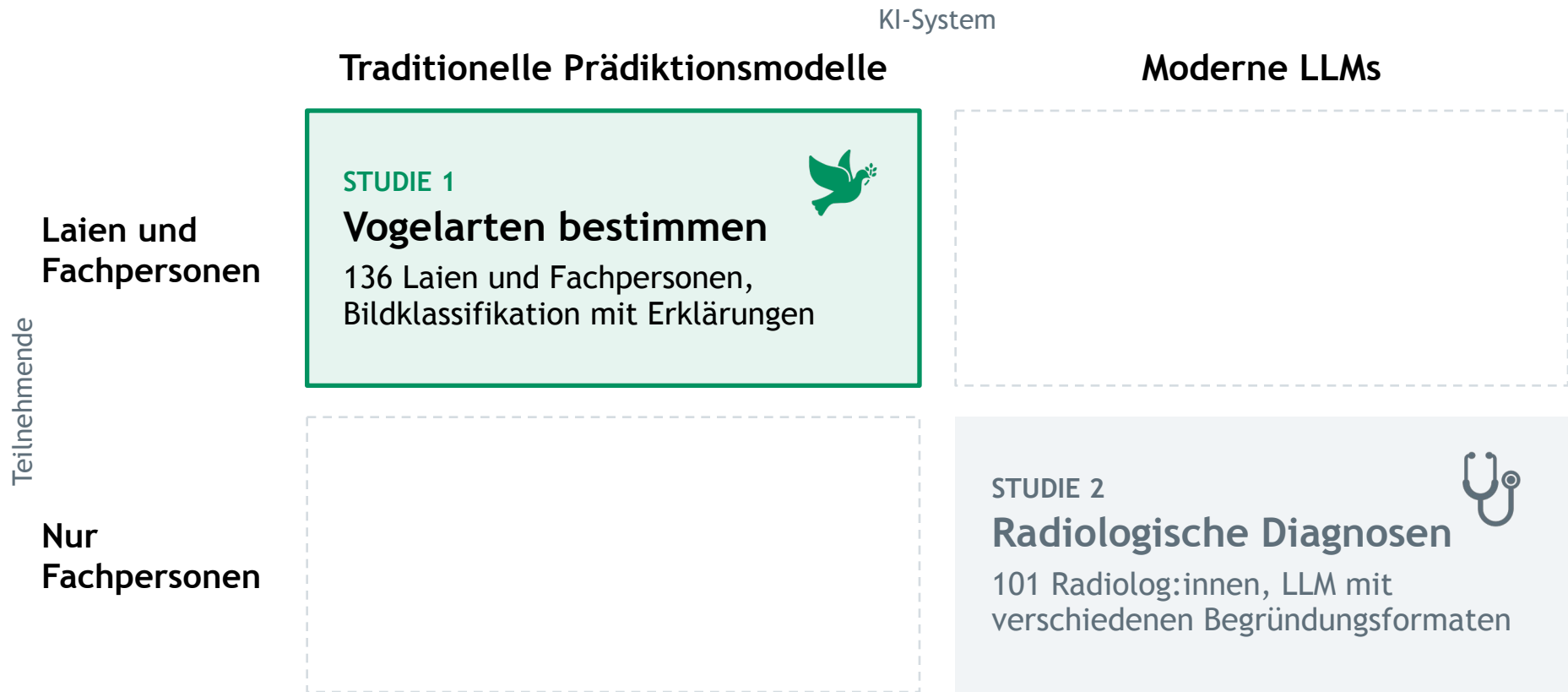
**Natürlichsprachliche
Erklärungen**

Wie kommen denn Erklärungen zustande?

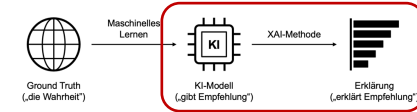


Ok, cool. Aber vertrauen Menschen nun kalibrierter, wenn wir ihnen KI-Empfehlungen erklären?

Übersicht der Studien



Was bedeutet das in der Praxis?



„KI erklärt KI“:



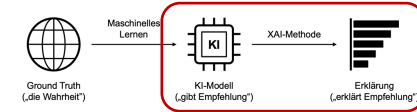
„Das ist ein Vogel mit
gelbem Bauch, schwarzen
Streifen auf der Brust und
einem grauen Kopf“

Vorhersage der KI:
Magnolienwaldsänger

Richtige Vorhersage

Quelle: Spitzer, P., Morrison, K., Turri, V., Feng, M., Perer, A., & Kühl, N. (2025). Imperfections of XAI: Phenomena influencing AI-assisted decision-making. ACM Transactions on Interactive Intelligent Systems, 15(3), 1-40.

...und das stimmt immer?



Vorhersage der KI:
Pappelwaldsänger

Richtige Vorhersage



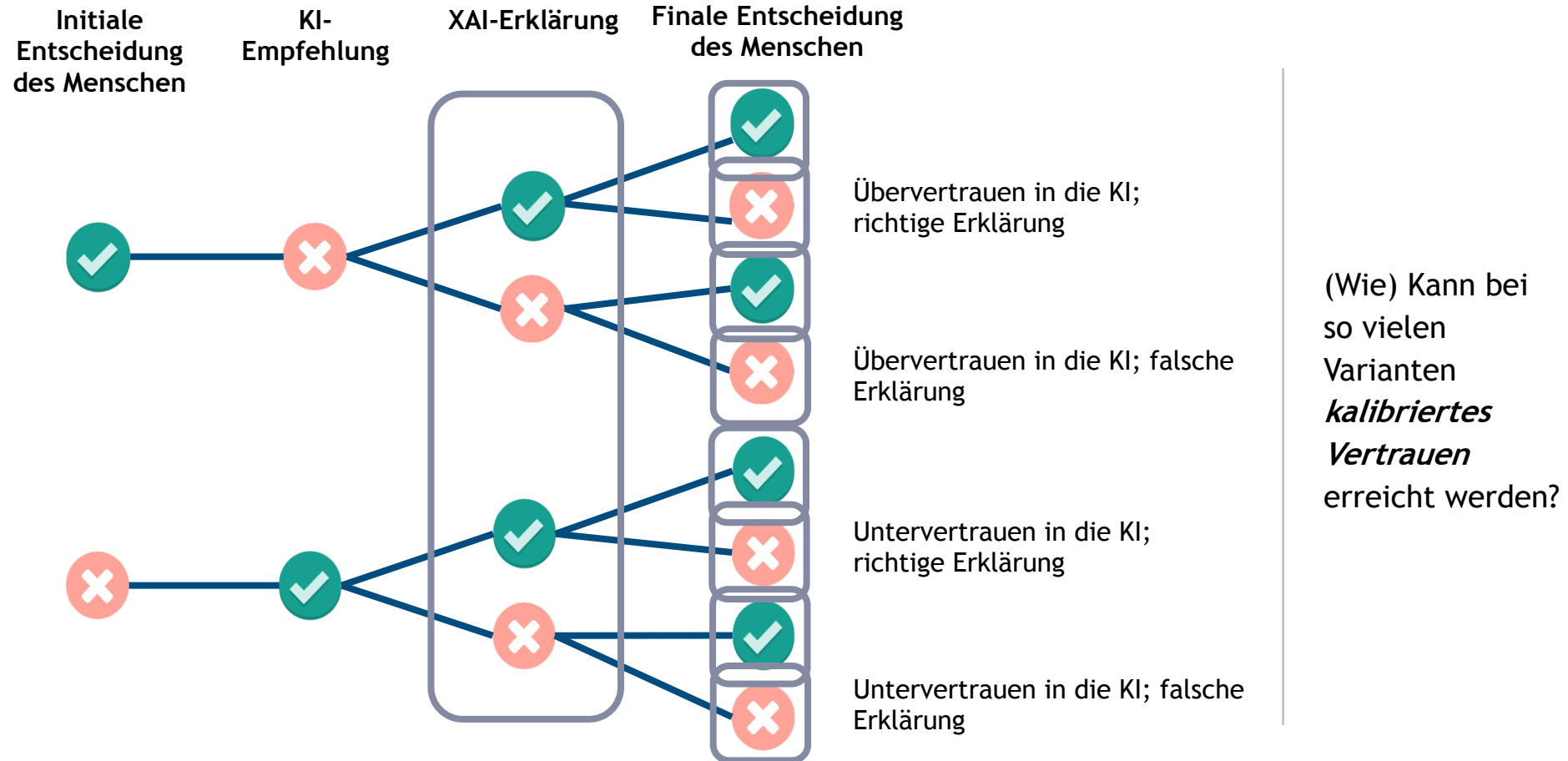
„Das ist ein kleiner Vogel mit
weißem Bauch und einem gelben
Kopf“

Methode erklärbarer KI:
Natürlichsprachliche Erklärung

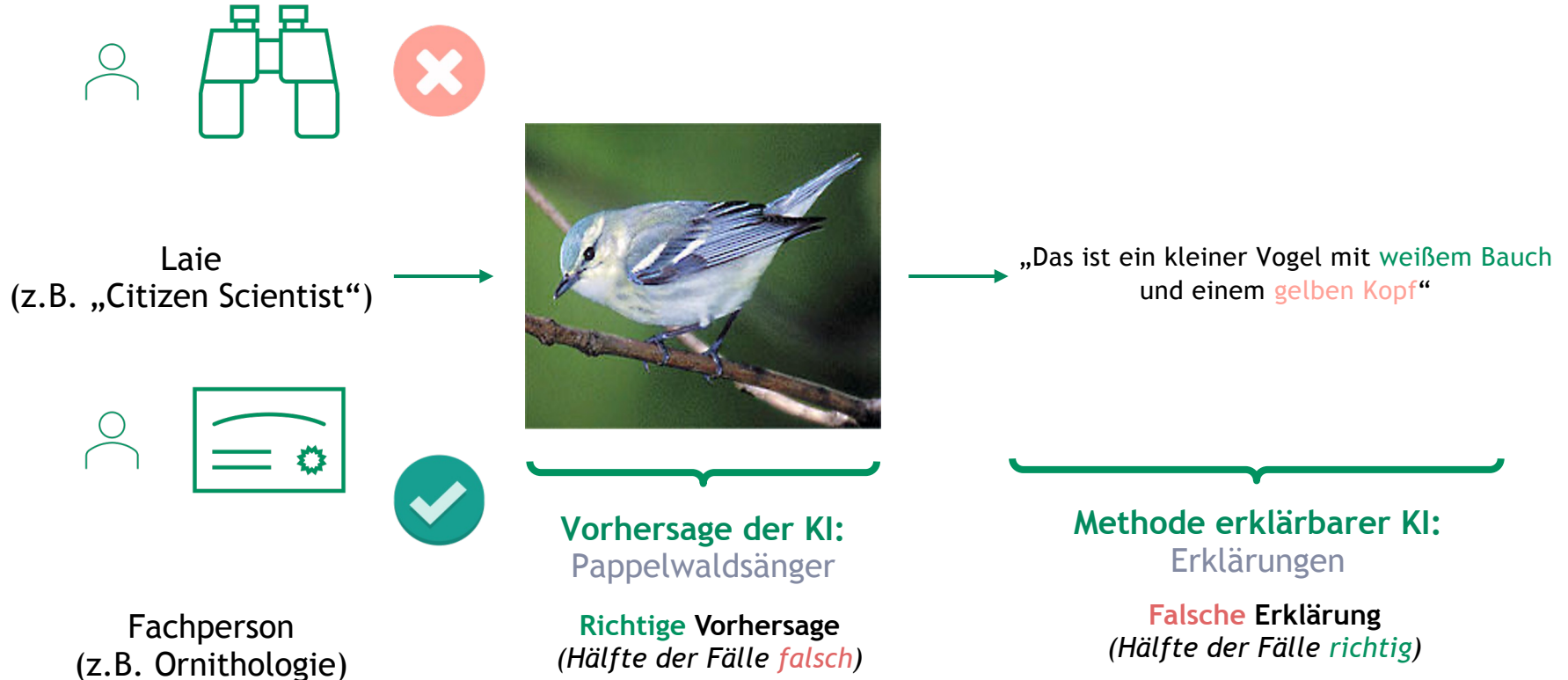
Falsche Erklärung

Quelle: Spitzer, P., Morrison, K., Turri, V., Feng, M., Perer, A., & Kühl, N. (2025). Imperfections of XAI: Phenomena influencing AI-assisted decision-making. ACM Transactions on Interactive Intelligent Systems, 15(3), 1-40.

Was bedeutet das für unsere Konzeptualisierung von kalibriertem Vertrauen?



Studie 1: 136 Menschen bestimmen Vogelarten – erst allein, dann mit KI.



Was kam raus? Laien folgen – und überstimmen praktisch nie.



War die KI richtig, folgten Laien ihr in fast **9 von 10 Fällen***; war sie falsch, blieben sie fast nie bei ihrer eigenen richtigen Antwort (**0-6 %**).



Die Trefferquote der Laien stieg von **6-8 % (ohne KI)** auf **43-46 % (mit KI)** – blieb aber unter der KI allein (**50 %**): anwesend, aber ohne eigenen Beitrag.



Fachpersonen schlugen die KI um **5 bis 7 Prozentpunkte**: Teamgewinn (Mensch+KI>KI allein) entstand **nur dort, wo Expertise vorhanden war**.



Falsche Erklärungen trafen Laien am härtesten: Bei Beispielbildern lag ihre Trefferquote mit KI bei **49 %** statt **37 %**, bei Fachpersonen nur um rund 3 Punkte.

*3 von 4 relevanten Bedingungen

Was haben wir gelernt?

Kalibriertes Vertrauen sollte
immer das Ziel sein, nicht
Vertrauen pauschal.

Kalibriertes Vertrauen setzt
genau die Expertise voraus, die
durch KI ersetzt werden soll.

Und selbst die schützt nicht.



Expertise schützt nicht vollständig – und wird knapper.



Expertise hilft, aber schützt nicht

Auch Fachleute wurden getäuscht – sogar von korrekten Erklärungen zu falschem Rat.

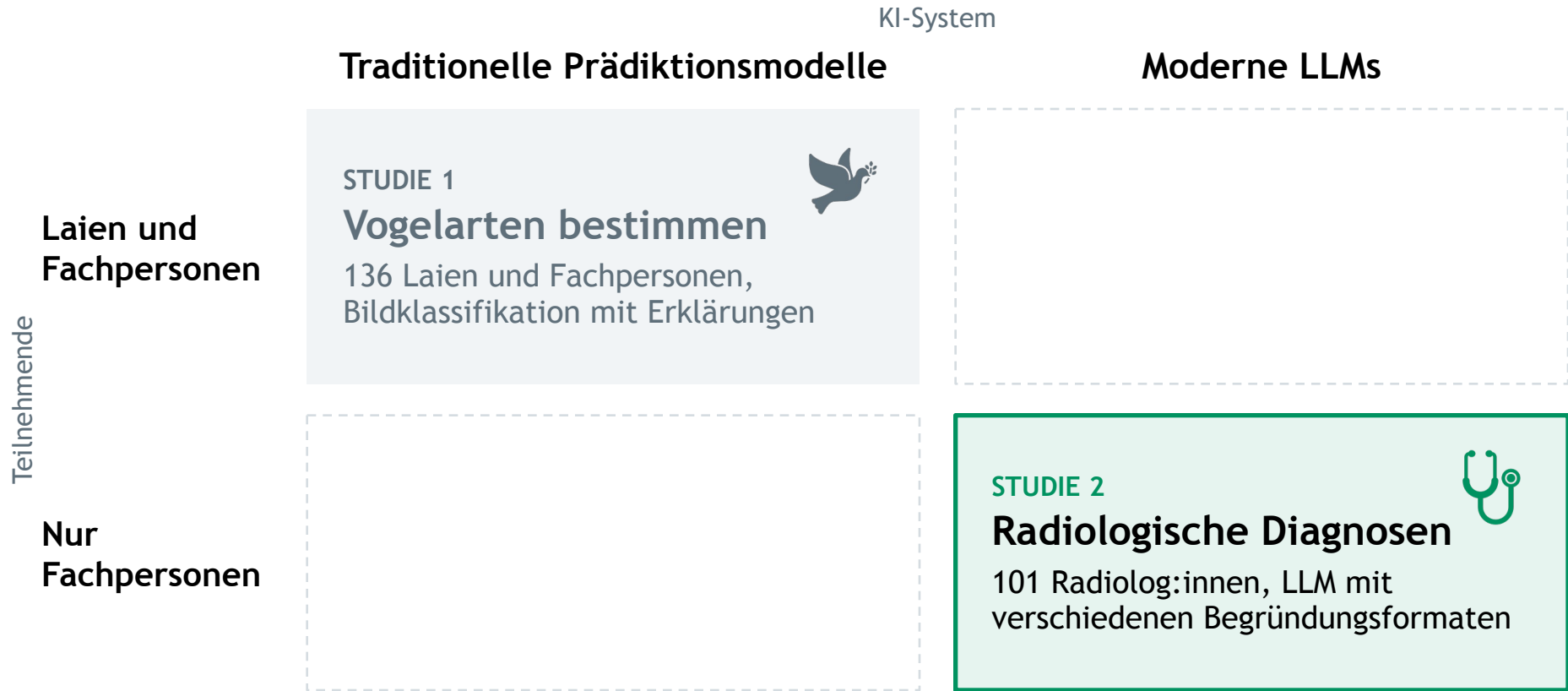


Expertise wird knapper

Wer von Anfang an mit KI arbeitet, baut das eigenständige Urteil womöglich nie auf.

Die Frage ist nicht nur, *wer* prüft (z.B. Experte vs. Laie), sondern *womit* (was hilft denn wirklich bei der Beantwortung?).

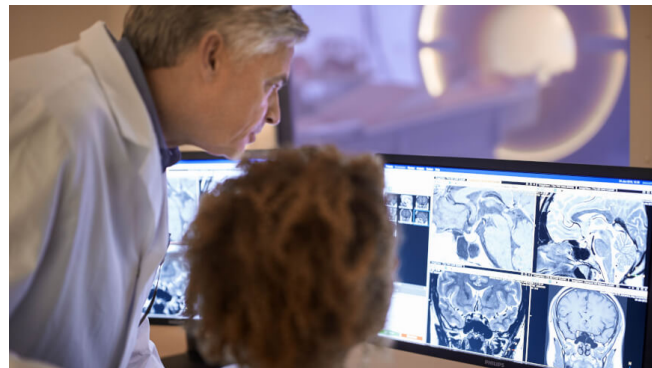
Übersicht der Studien



Studie 2: Vom Labor in die Klinik.

101 Radiolog:innen, 2.020 Bewertungen, echte Patientenfälle mit CT- und MRT-Bildern.

Variiert wurde nur eines: das Format der Erklärung von Large Language Models.



Vier randomisiert zugewiesene Bedingungen:

Kontrolle

Diagnose ohne KI-Unterstützung.



Standard

Diagnose ohne Begründung.



AI

Differentialdiagnose

Liste möglicher Diagnosen mit Erläuterung.



AI

Chain-of-Thought

Begründung Schritt für Schritt.



AI

Quelle: Spitzer, P., Hendriks, D., Rudolph, J., Schlaeger, S., Ricke, J., Kühl, N., ... & Feuerriegel, S. (2026). The effect of medical explanations from large language models on diagnostic accuracy in radiology. *npj Digital Medicine*, 9(1), 333.

Studie 2: Vom Labor in die Klinik.

Standard

Diagnose ohne Begründung.

Standard output

The MRI of the orbit shows a significant mass in the right eye, which, along with the patient's history of eye inflammation and blindness, suggests an intraocular tumor.

The presence of numerous hepatic masses, lymphadenopathy, and vertebral sclerotic osseous lesions on abdominal and thoracic imaging indicates metastatic disease.

The findings are consistent with a diagnosis of metastatic cancer, with the primary site likely being the right eye, possibly from a choroidal melanoma.



Differentialdiagnose

Liste möglicher Diagnosen mit Erläuterung.

Differential diagnosis

List: Metastatic carcinoma, Orbital lymphoma, Orbital pseudotumor (Idiopathic Orbital Inflammatory Disease), Orbital sarcoma, Sarcoidosis

- **Metastatic carcinoma:** Best Fit: Metastatic carcinoma is most likely given the presence of numerous hepatic masses, abdominal and thoracic lymphadenopathy, and vertebral sclerotic osseous disease which are common sites for metastases. The mass in the orbit could also be a metastasis.
- **Orbital lymphoma:** Orbital lymphoma is possible given the orbital mass; however, it is less likely because it typically would not present with extensive systemic metastases like the hepatic and vertebral involvement seen here.
- **Orbital pseudotumor:** Orbital pseudotumor usually presents with inflammation and pain in the eye, but it does not explain the hepatic masses, lymphadenopathy, and vertebral disease, making it less likely.
- **Orbital sarcoma:** While orbital sarcoma can present with a mass in the orbit, it is less common and less likely to present with such widespread systemic disease (hepatic masses and vertebral involvement) compared to metastatic carcinoma.
- **Sarcoidosis:** Least Likely: Sarcoidosis can present with lymphadenopathy and orbital involvement, but the presence of numerous hepatic masses and the pattern of bone involvement are atypical, making it the least likely diagnosis.



Chain-of-Thought

Begründung Schritt für Schritt.

Chain-of-thought explanation

The patient is likely suffering from metastatic cancer, specifically **metastatic ocular melanoma**. This diagnosis is based on several key findings:

- **Clinical Presentation:** A 59-year-old woman presented with a 4-day history of inflammation and pain in the right eye, where she had been blind for several years. The chronic blindness and recent symptoms suggest a long-standing ocular issue that has now become symptomatic.
- **Imaging Findings:** Magnetic resonance imaging revealed a right orbital mass, indicating a significant lesion in the affected eye.
- **Systemic Involvement:** Abdominal and thoracic imaging showed numerous hepatic masses, abdominal and thoracic lymphadenopathy, and vertebral sclerotic osseous disease, indicating widespread metastatic disease.
- **Ocular History:** The fact that the patient had been blind in the eye for several years suggests a long-standing primary ocular condition, such as melanoma, which can metastasize to other organs.
- **Pathological Confirmation:** The right eye was enucleated for palliative relief and to obtain tissue for diagnosis. Enucleation is often performed in cases of severe ocular melanoma to control symptoms and obtain a definitive diagnosis.
- **Symptoms Correlation:** The combination of an orbital mass, widespread metastatic lesions, and the patient's ocular history strongly suggests metastatic ocular melanoma.

In conclusion, the combination of clinical presentation, imaging findings, and systemic involvement strongly supports the diagnosis of metastatic ocular melanoma.



Nicht ob erklärt wird, sondern wie – das Format entscheidet.

Anteil richtiger Diagnosen je Gruppe (in Klammern: Differenz zur Kontrollgruppe)

Kontrolle

ca. 62 %

ohne KI-Unterstützung



Standard

ca. 67 %

(+5,0 pp, kein klarer Effekt)



Differentialdiagnose

ca. 64 %

(+2,5 pp, kein klarer Effekt)



Chain-of-Thought

ca. 74 %

(+12,2 pp, klarer Effekt)



n = 101 Radiolog:innen, 2.020 Bewertungen.

Nur ein Format verbessert die Diagnosen messbar.

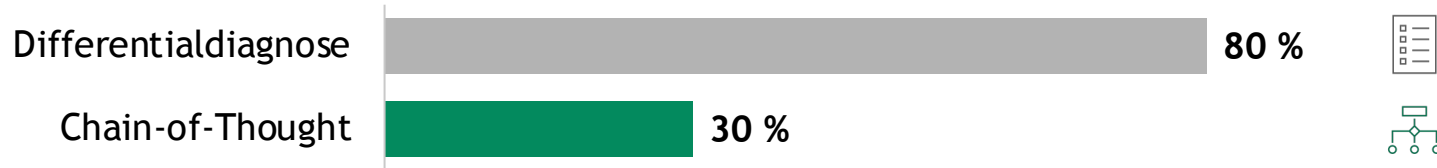
Die Schritt-für-Schritt-Begründung: +12,2 pp gegenüber der Kontrollgruppe.

Quelle: Spitzer, P., Hendriks, D., Rudolph, J., Schlaeger, S., Ricke, J., Kühl, N., ... & Feuerriegel, S. (2026). The effect of medical explanations from large language models on diagnostic accuracy in radiology. *npj Digital Medicine*, 9(1), 333.

Der Mechanismus: Prüfbarkeit statt Überzeugungskraft.

Folgebereitschaft bei **falscher** KI-Empfehlung

Anteil der Fälle mit Übernahme der falschen Diagnose · falsche Empfehlung, falsche Begründung



War die KI falsch, folgten ihr 80 %, trotz falscher Begründung.
Das gilt für die Differentialdiagnose; bei Chain-of-Thought waren es 30 %.

Quelle: Spitzer, P., Hendriks, D., Rudolph, J., Schlaeger, S., Ricke, J., Kühl, N., ... & Feuerriegel, S. (2026). The effect of medical explanations from large language models on diagnostic accuracy in radiology. *npj Digital Medicine*, 9(1), 333.

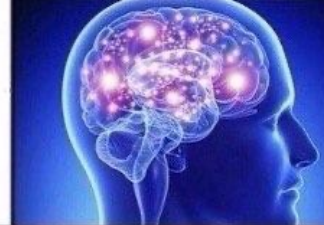
Was haben wir gelernt?

Kalibriertes Vertrauen sollte immer das Ziel sein, nicht Vertrauen pauschal.

Kalibriertes Vertrauen setzt genau die Expertise voraus, die durch KI ersetzt werden soll.

Und selbst die schützt nicht.

Nicht ob erklärt wird, zählt, sondern wie. Nur Prüfbarkeit erzeugt kalibriertes Vertrauen.



01

Wie viel Vertrauen in KI ist richtig?

02

Wer kann KI-Empfehlungen noch beurteilen?

03

Was macht eine KI-Erklärung nützlich?

Die KI allein war am genauesten, haften muss der Mensch.

Befund aus Studie 2

KI allein 80 %

Mensch + KI (Chain-of-Thought) ca. 74 %

Mensch ohne KI ca. 62 %

Anteil richtiger Diagnosen, Vignettenstudie mit 20 Fällen

Warum der Mensch trotzdem entscheidet

Vignetten sind keine Klinik.

Genauigkeit ist kein Patientenergebnis.

Haftung folgt nicht der Trefferquote.

Solange Menschen haften, muss ihre Aufsicht wirksam sein.

Ob sie das ist, zeigen drei Prüffragen.

Zwei Fragen an Sie.

1

Wer von Ihnen setzt ein KI-System produktiv ein?

2

Wer von Ihnen hat für dieses System eine Person benannt, die die Befugnis und die Information hat, zu widersprechen?

Drei Prüffragen für jedes KI-System in Ihrem Haus.

A



Einblick

Bekommt die prüfende Person den Gedankengang oder nur das Ergebnis?

B



Sachkenntnis

Hat sie die Sachkenntnis, ihn zu beurteilen? Kann sie falsche Erklärungen identifizieren?

C



Konsequenz

Hat ihr Widerspruch Konsequenzen?

Was passiert, wenn wir dies nicht durchdenken?

Die Simpsons: Staffel 7, Episode 7 - Das Finale



Quelle: 20th Century Fox, The Simpsons

Es gibt inzwischen ein Wort dafür: Meat Proxy.

STATEMENT 1 OF 6
● ● ● ● ● ●

“I’ve pasted an AI answer I didn’t read all the way through.”

Yes. 🍖 No. 🧠

Live Meat-o-meter 0 of 6 yes

STATEMENT 4 OF 6
● ● ● ● ● ●

“I put my name on summaries I didn’t write.”

Yes. 🍖 No. 🧠

← Previous statement

Live Meat-o-meter 3 of 6 yes

STATEMENT 5 OF 6
● ● ● ● ● ●

“When the AI is down, so is my output.”

Yes. 🍖 No. 🧠

← Previous statement

Live Meat-o-meter 4 of 6 yes

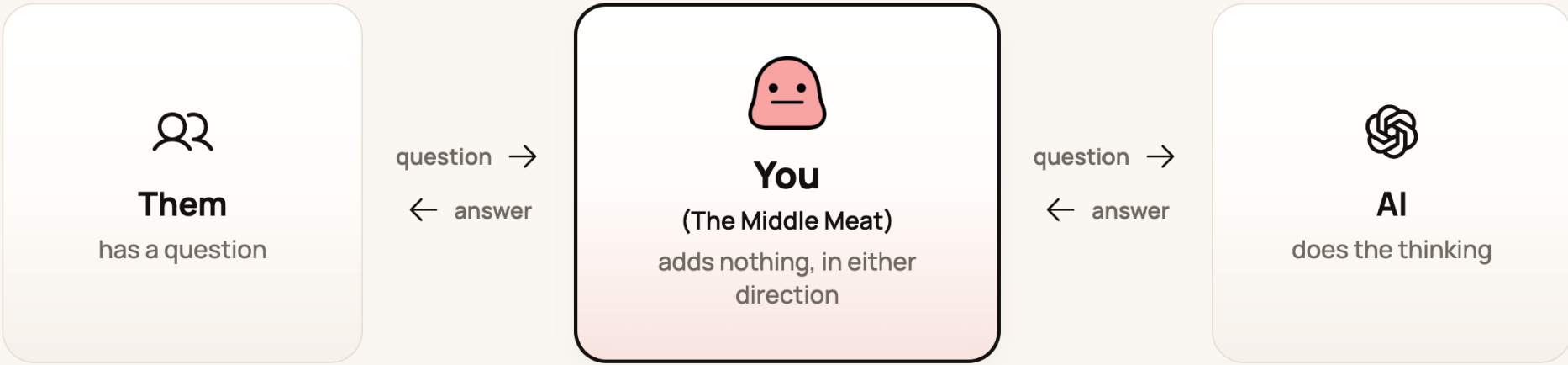
Quelle:



meatproxy.me

Early human errors at the beginning of the AI age.

Es gibt inzwischen ein Wort dafür: Meat Proxy.



There is no reason for this step to exist. Be more than a meat bridge.

Quelle:



meatproxy.me

Early human errors at the beginning of the AI age.

Wer entwirft eigentlich wirksame Aufsicht?

Wir. Und das funktioniert nur interdisziplinär – mit einer Mischung aus (Wirtschafts-)Informatik, Recht, Ethik und Philosophie. Daran glaube ich 😊

AI Responsibility Centre (ARC)

Diesen Sommer an der Universität Bayreuth gegründet, um genau diese Fragen zu beantworten.

Report an die EU-Kommission

Ich fasse gerade zusammen, was effektive menschliche Aufsicht nach Artikel 14 AI Act heißt.



Vielen Dank für die Aufmerksamkeit! Visit us @ <https://ai-responsibility-centre.eu/>