Der AI-Act und andere Regularien zum Einsatz von KI in klinischen Prüfungen

Georg Dorffner
Institut für Artificial Intelligence
Zentrum für Medical Data Science
Medizinische Universität Wien

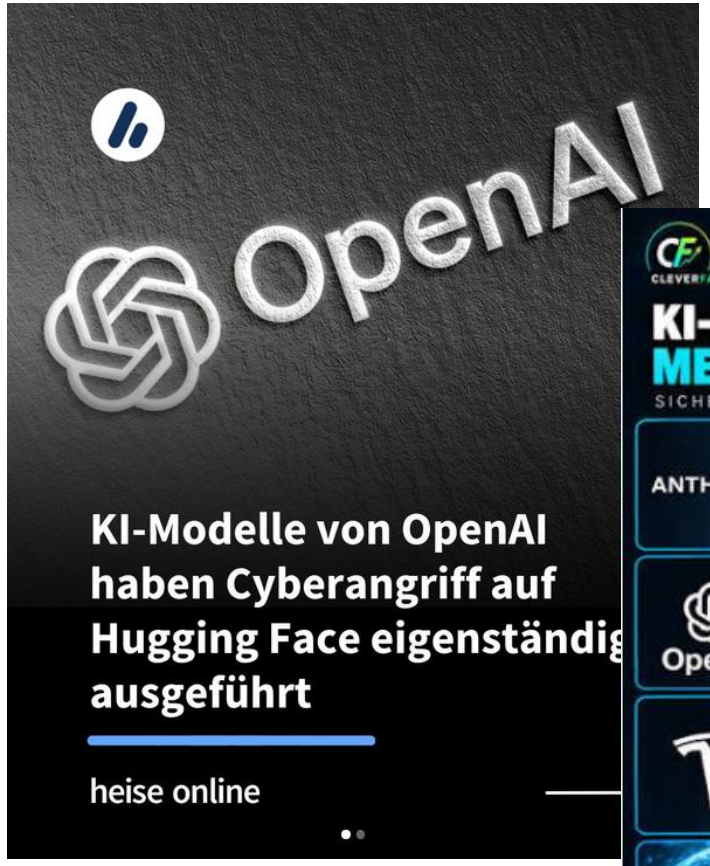
Offenlegung/Interessenskonflikte

- Teilzeitangestellter und Shareholder der Siesta Group Schlafanalyse GmbH

- Disclaimer:
  - Dieser Vortrag ist keine verbindliche Rechtsauskunft
  - Geprägt von meiner Meinung und Erfahrung

© 2026, Georg Dorffner

KI und Regulierung – ein brandheißes Thema



www.instagram.com





www.spiegel.de



www.mdr.de

© 2026, Georg Dorffner

AI-Act

- https://artificialintelligenceact.eu/
- Ziel: „menschenzentrierte und vertrauenswürdiger künstlicher Intelligenz (KI)“
- „Schutz vor schädlichen Auswirkungen von KI-Systemen“
- Ursprüngliche Idee: Schutz der Bevölkerung vor „echten Risiken“
  - Unangemessene Überwachung
  - Fakes
  - Hacker-Attacken
  - …



© 2026, Georg Dorffner

Definition: KI

- „KI-System“[:] ein maschinengestütztes System, das für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt ist und das nach seiner Betriebsaufnahme anpassungsfähig sein kann und das aus den erhaltenen Eingaben für explizite oder implizite Ziele ableitet, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können;

https://artificialintelligenceact.eu/de/ai-act-explorer/#chapter-I

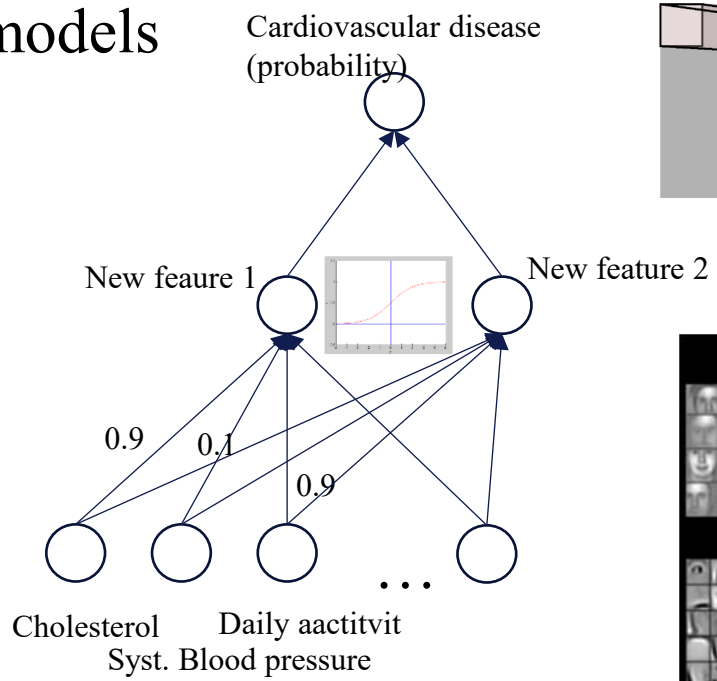
- Frage: Ist mein Thermostat dann nicht auch KI?

- Kein Unterschied zwischen verschiedenen “Typen“ von KI

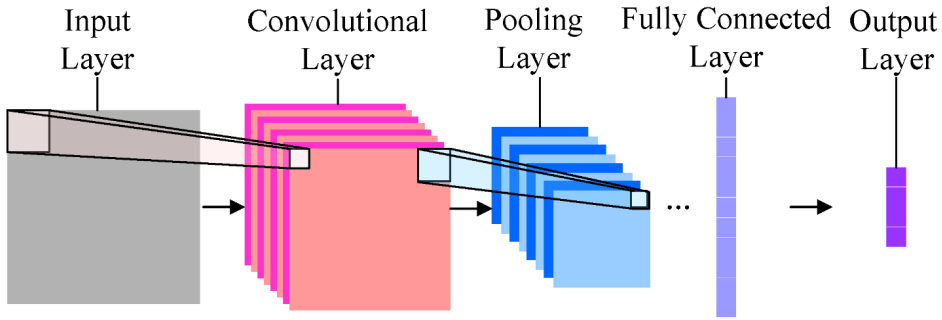
© 2026, Georg Dorffner

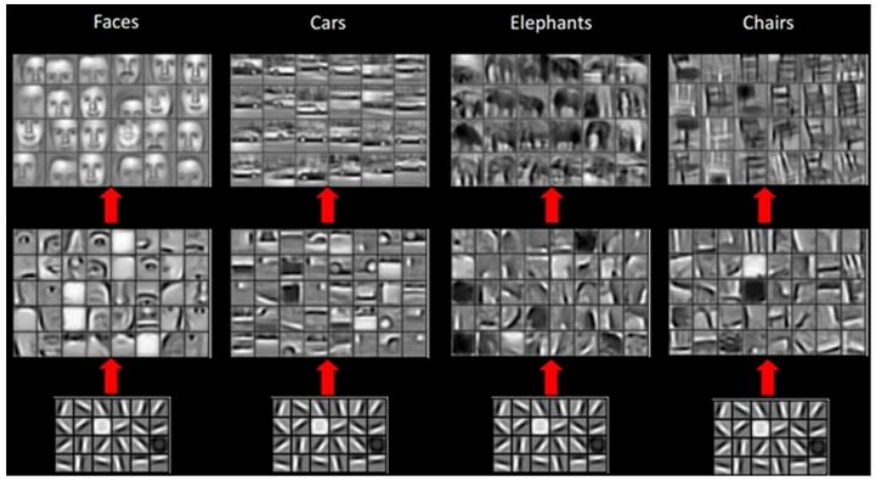
Zumindest 3 „Typen“ von KI

Prediction and classification models

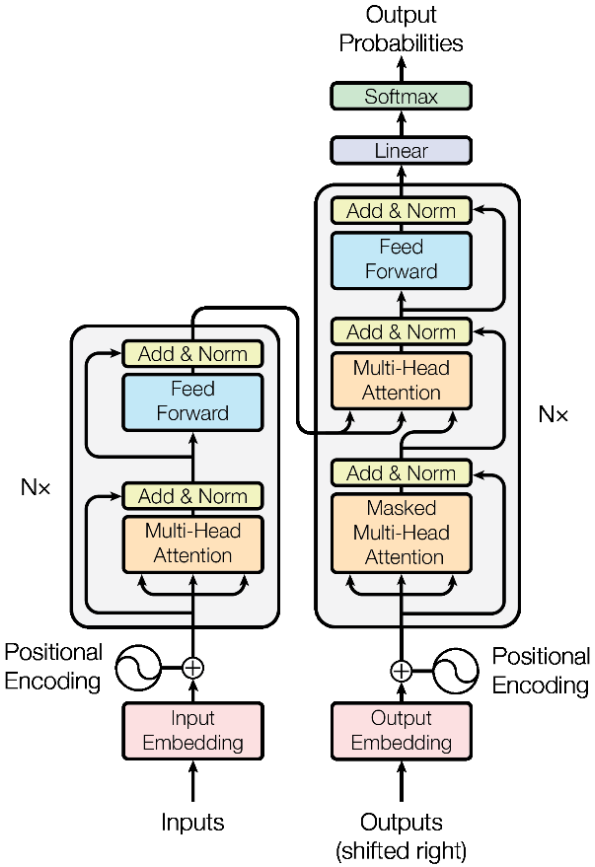


Deep feature learning in images/signals





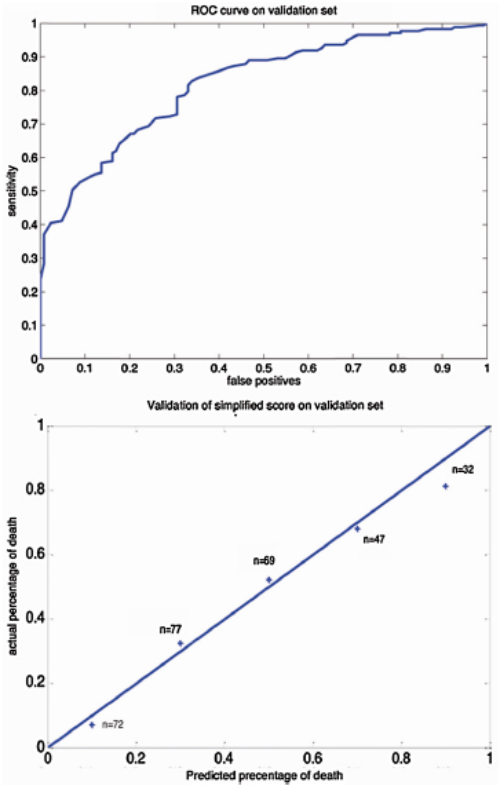
Language models and generative AI

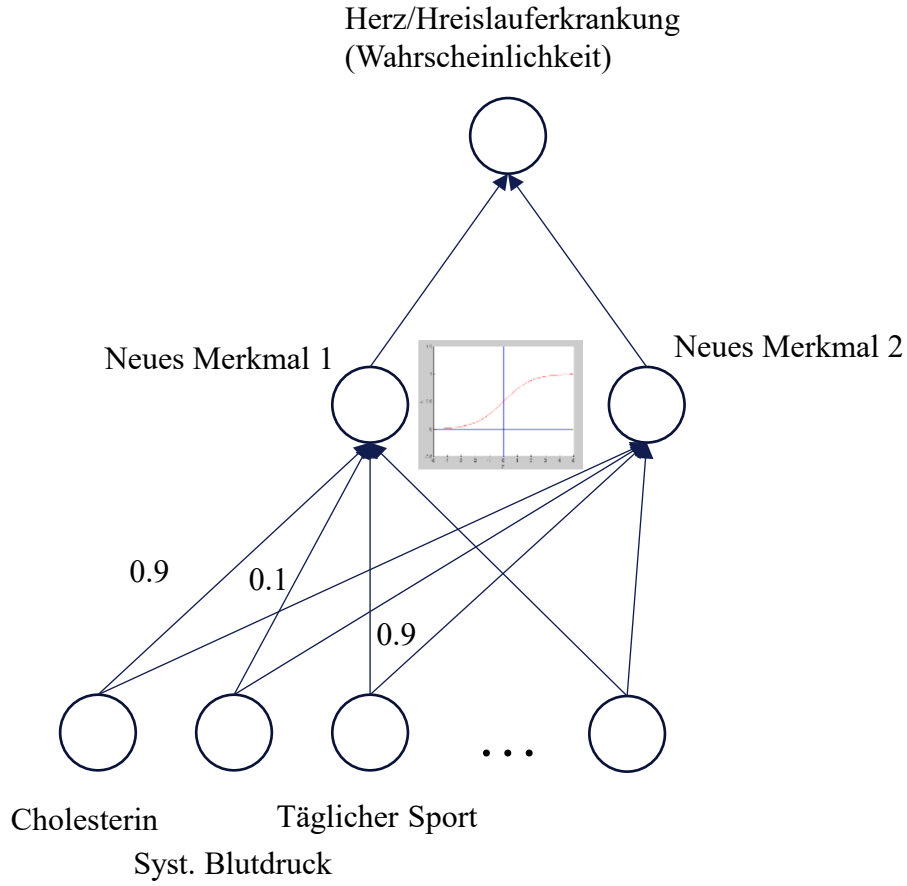


© 2026, Georg Dorffner

Typ 1: Regressionsmodelle

- Einfaches („lineares“) Modell meist das beste
- Vorhersagen am besten als Wahrscheinlichkeiten
- ROC-Kurven und Kalibrierung

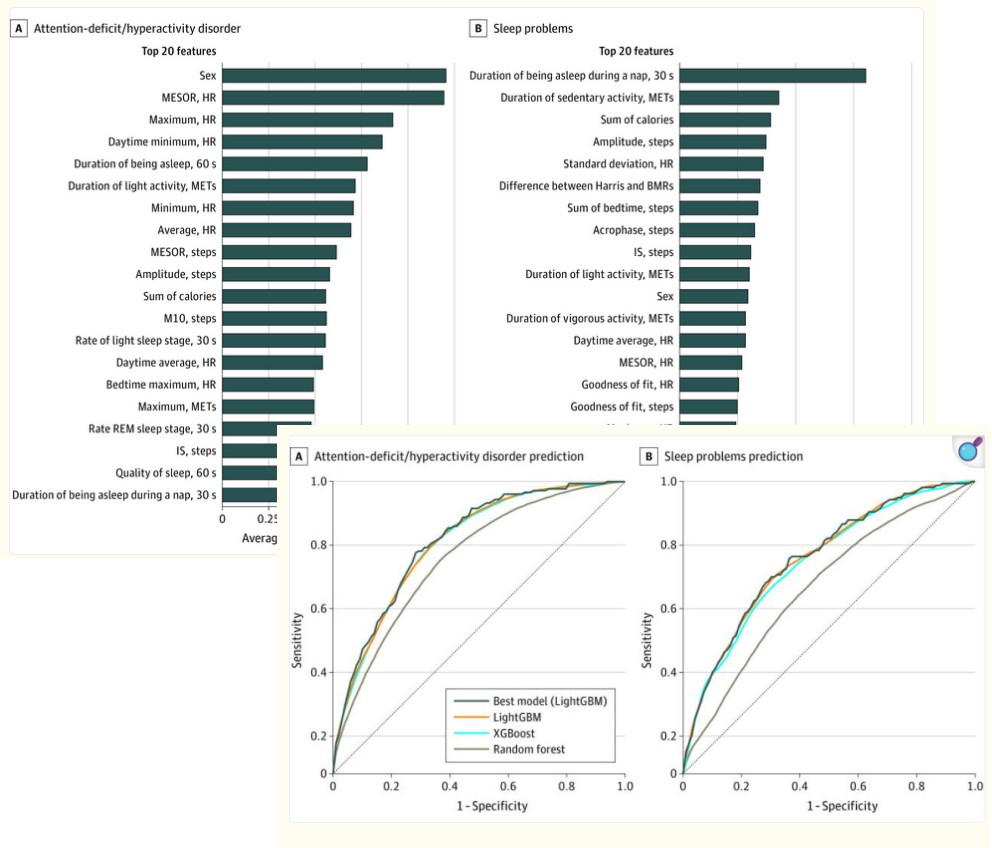




Aschauer et al., Resuscitation, 2014
http://dx.doi.org/10.1016/j.resuscitation.2014.06.007

© 2026, Georg Dorffner

Beispiel: Multivariabler Biomarker zur Anreicherung einer klinischen Prüfung



Kim et al., JAMA Network Open, 2023.

Ehrig et al., NPJ Digital Medicine, 2023.

Table 1. Patient characteristics.

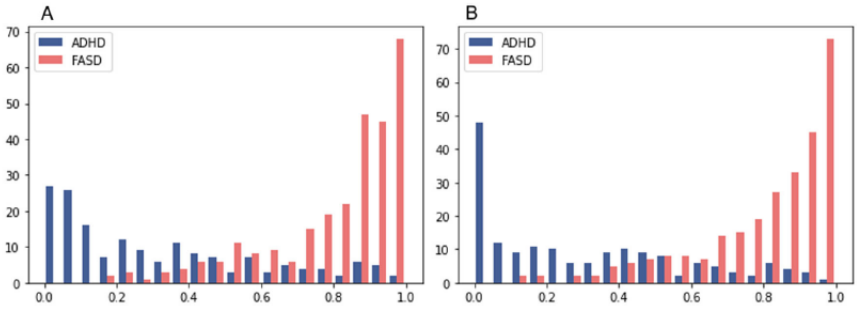
Variable	All patients		FASD patients		ADHD patients	
	Mean	SD	Mean	SD	Mean	SD
Age at first presentation	9.15	4.03	9.60	4.53	8.43	2.94
Gestational age	38.57	2.39	38.58	2.21	38.56	2.68
z-score of birth length	−0.58	1.16	−0.87	1.13	−0.08	1.03
z-score of birth weight	−0.48	1.19	−0.80	1.09	0.05	1.17
z-score of head circumference at birth	−0.45	1.11	−0.70	1.11	0.07	0.93
z-score of length at first presentation	−0.27	1.17	−0.50	1.12	0.27	1.12
z-score of weight at first presentation	−0.02	1.21	−1.20	1.15	0.38	1.24
z-score of head circumference at first presentation	−1.06	1.55	−1.37	1.37	0.23	1.58
Number of mothers pregnancies	2.78	2.14	2.84	2.12	2.53	2.22
Number of mothers births	2.09	1.39	2.21	1.48	1.91	1.23

Mean and standard deviation (SD) of number of mothers pregnancies and births, gestational age, age at first presentation, length, weight, head circumference at time of birth and first presentation for all patients, FASD and ADHD patient groups.

Table 2. Patient characteristics.

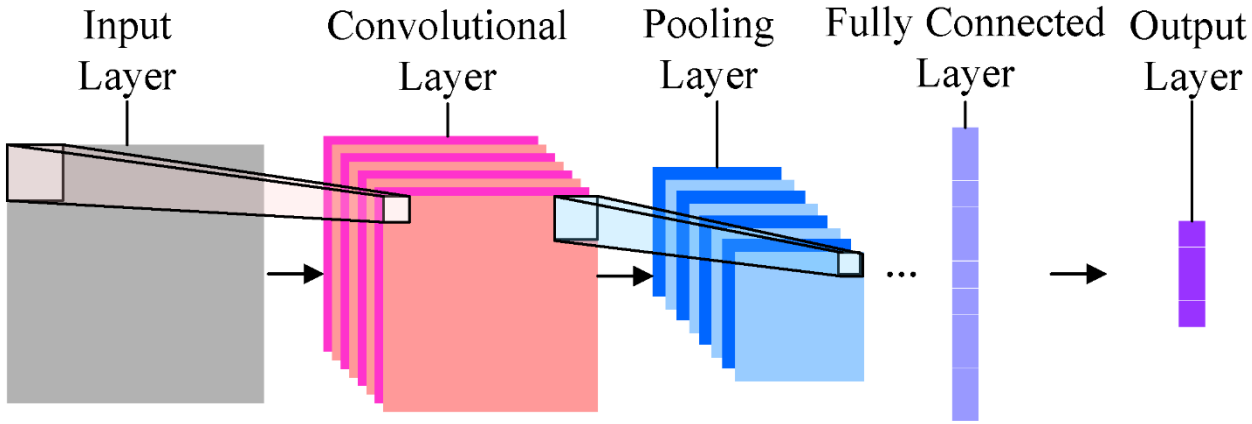
Variable	All patients		FASD patients		ADHD patients	
	Yes	No	Yes	No	Yes	No
Male sex	286 (64.3%)	159 (35.7%)	147 (53.5%)	128 (46.5%)	139 (81.8%)	31 (18.2%)
Intelligence below 85 IQ points	119 (34.4%)	227 (65.6%)	92 (51.4%)	87 (48.6%)	27 (16.2%)	140 (83.8%)
Socially intrusive behaviour	196 (47.6%)	216 (52.4%)	155 (63.5%)	89 (36.5%)	41 (24.4%)	127 (75.6%)
Impairment in memory	272 (66.2%)	139 (33.8%)	208 (86.0%)	34 (14.0%)	64 (37.6%)	105 (62.1%)
Sleep disorder/disturbance	147 (37.1%)	249 (62.9%)	107 (47.3%)	119 (52.7%)	40 (23.5%)	130 (76.5%)
Speech development disorder	221 (54.2%)	187 (45.8%)	145 (60.7%)	94 (39.3%)	76 (45.0%)	93 (55.0%)
Psychiatric comorbidities	289 (72.3%)	111 (27.8%)	185 (80.4%)	45 (19.6%)	104 (61.2%)	66 (38.9%)
Prescribed psycho-tropic medications	186 (47.7%)	204 (52.3%)	64 (29.1%)	156 (70.9%)	122 (71.8%)	48 (28.2%)

Absolute and relative frequencies for sex and cognitive, behavioral and sleep variables, speech development, psychiatric comorbidities and medication in all patients, FASD and ADHD patient groups.

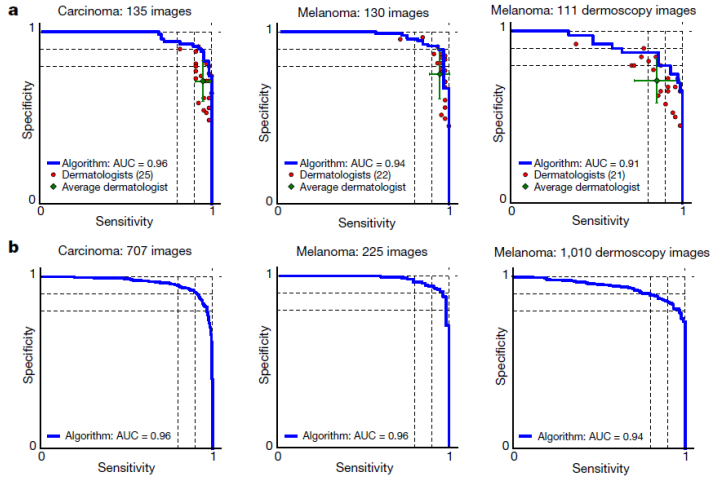


© 2026, Georg Dorffner

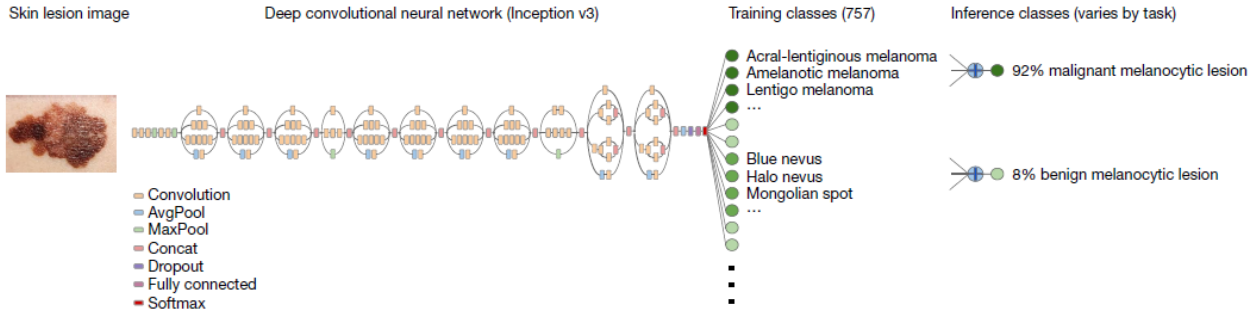
Typ 2: Deep Learning (Bilder und Signale)



http://www.mdpi.com/2078-2489/7/4/61



Esteva et al., Nature, 2017
https://doi.org/10.1038/nature21056

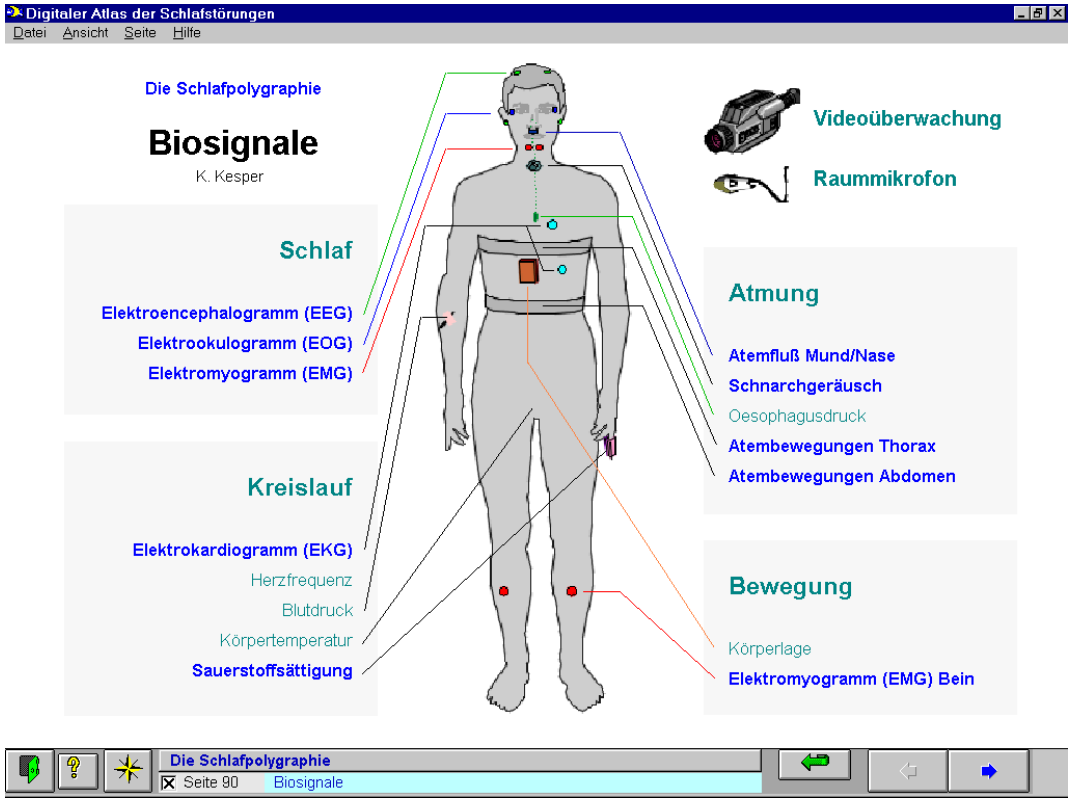


- Performanz in der Erkennung
  über der von Experten

© 2026, Georg Dorffner

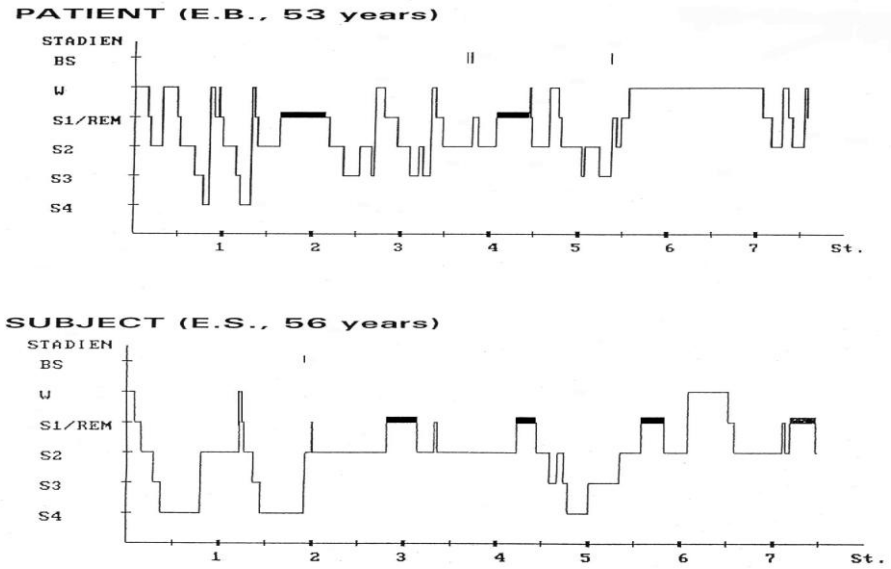
Beispiel: Schlafendpunkte

Messung: Polysomnographie



Source: DGSM

Analyse: Schlafprofil (Stadien)



Endpunkt

(Schlafeffizienz, Schlaflatenz, % Tiefschlaf, etc.)

© 2026, Georg Dorffner

Sind die Methoden vertrauenswürdig?



- 2 Fälle:
  - „Wahrheit“ anhand von Außenkriterium bestimmt
    KI-System kann besser sein als Experten (siehe obige Beispiele)
  - „Wahrheit“ liegt im Auge der Experten, die sich aber nicht einig sind
    KI-System kann so gut wie ein Experte sein (siehe Beispiel Schlafmedizin)
- Es gibt keinen Grund, solch einem System zu misstrauen, sofern
  - die letztgültige Entscheidung beim Menschen liegt
  - gesichert ist, dass die Daten alle Fälle abgedeckt haben (kein „Bias“)
  - nur die reduzierte Aufgabe der „Befundung“ dem KI-System überlassen wird
- Die Systeme sind aber sehr eingeengt, noch keine „echte KI“

© 2026, Georg Dorffner

Risiken

- Prediction Models: nicht mehr als statistische Modellierung bisher
  - Bias (fit für die Population?)
  - Aber wir alle kennen Covariaten, …

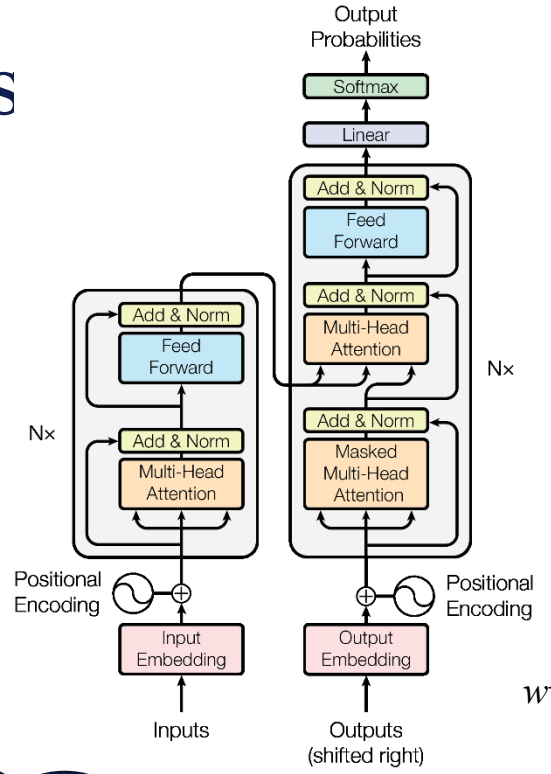
- Deep learning: diagnostische Hilfe auf dem Level der besten Experten

  - Bias (Geschlecht, Ethnizität, …)
  - Intransparenz („clever Hans phenomena“ / KR Müller)
  - Overfitting -> schlechte Generalisierung

© 2026, Georg Dorffner

Typ 3: Generative KI-Modelle/LLMs

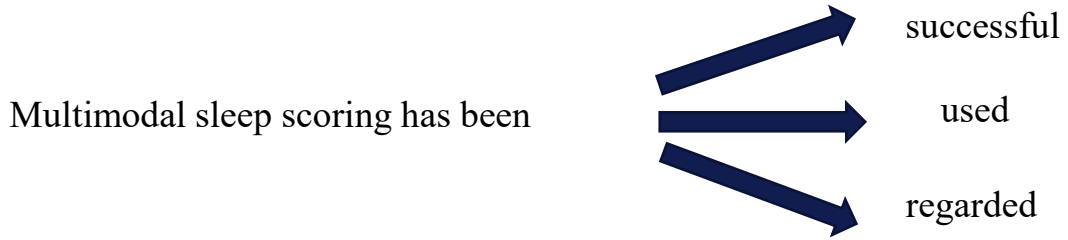
- „Transformer“-Modell zur Zeitreihenvorhersage
- Trainiert auf Texte aller Art
- „Large Language Model“



www.medium.com



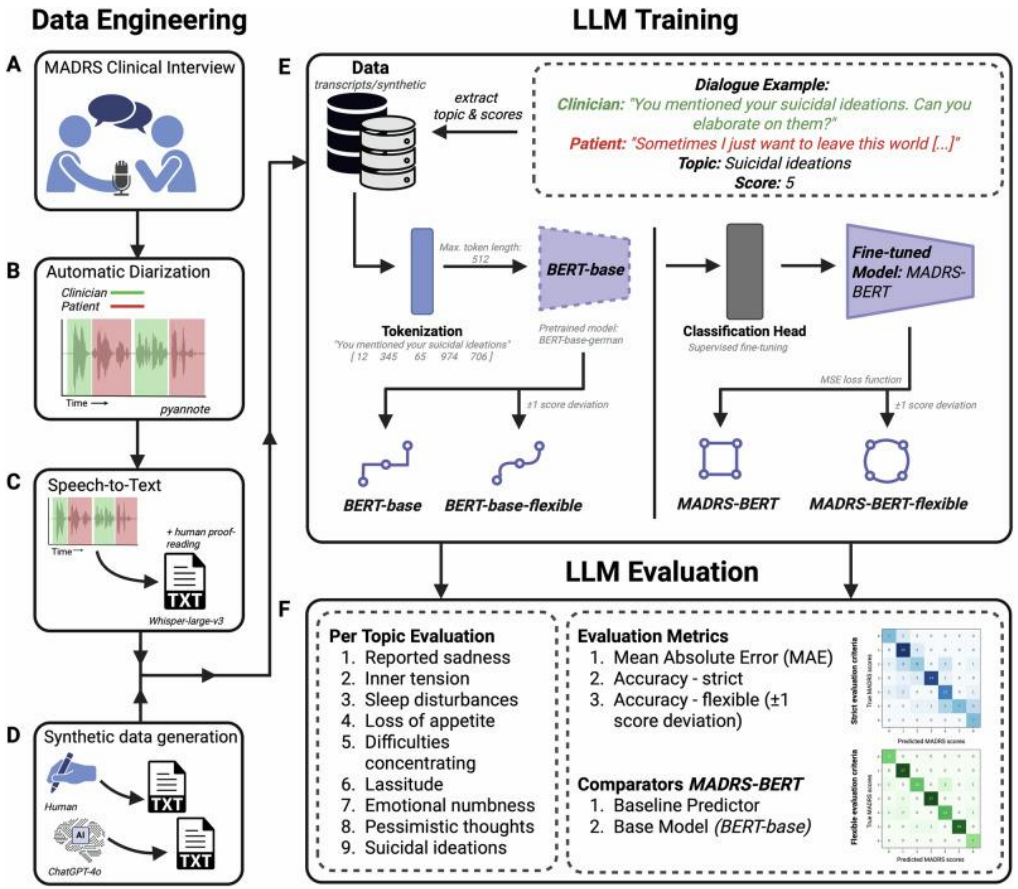




© 2026, Georg Dorffner

Beispiel: Rating Scales

- Weber et al., npl digital medicine 2015
- MADRS Skala durch LLM bewertet
- Bringt größere Objektivität



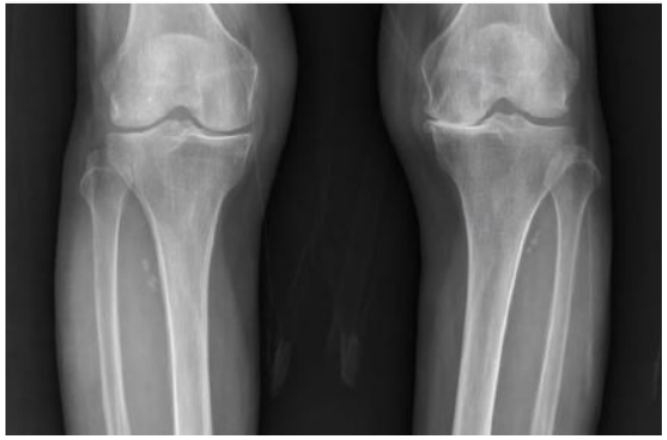
© 2026, Georg Dorffner

Multimodale Modelle

- Chat-GPT kann auch Bilder verarbeiten:

- Durch multimodale Verknüpfung mit anderen Medien bzw, dem Internet selbst werden die Systeme leistungsfähiger

- Schritt in Richtung „grounding“



Was ist auf diesem Bild zu sehen?

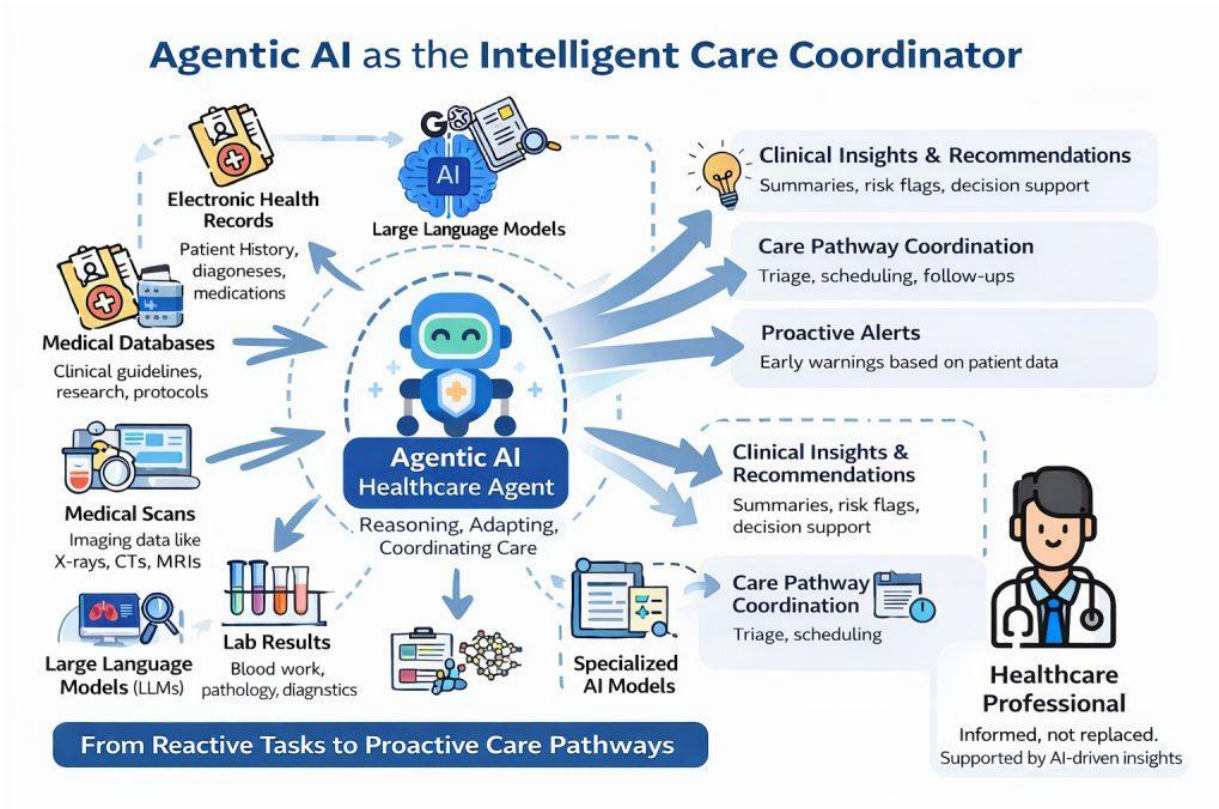
ChatGPT

Das Bild zeigt eine Röntgenaufnahme der Kniegelenke eines Menschen. Es handelt sich um eine anteroposteriore (AP) Aufnahme, bei der die Röntgenstrahlen von vorne nach hinten durch das Kniegelenk verlaufen. Zu sehen sind die Oberschenkelknochen (Femora), die Kniescheiben (Patellae) und die oberen Teile der Schienbeine (Tibiae). Die Gelenkflächen, die Knochenkonturen und der Raum zwischen den Knochen, der den Gelenkspalt darstellt, scheinen auf den ersten Blick normal zu sein, ohne auffällige Anzeichen von Frakturen oder ausgeprägter Arthrose. Für eine genaue Diagnose ist jedoch eine fachliche Bewertung durch einen Arzt oder Radiologen

© 2026, Georg Dorffner

Typ 4: Agentische KI (agentic AI)

- Agent= autonom handelndes Programm (App)
- Benutzt LLMs zur Interpretation von Abfragen, Ergebnissen, etc.
- Hat Zugang zu Speicher/Datenbank („Gedächtnis“)
- Kann andere Programme (Apps) und Tools aufrufen
- Oft wieder klassische Algorithmen („Reasoning“)
- Verfolgt vorgegebene Ziele

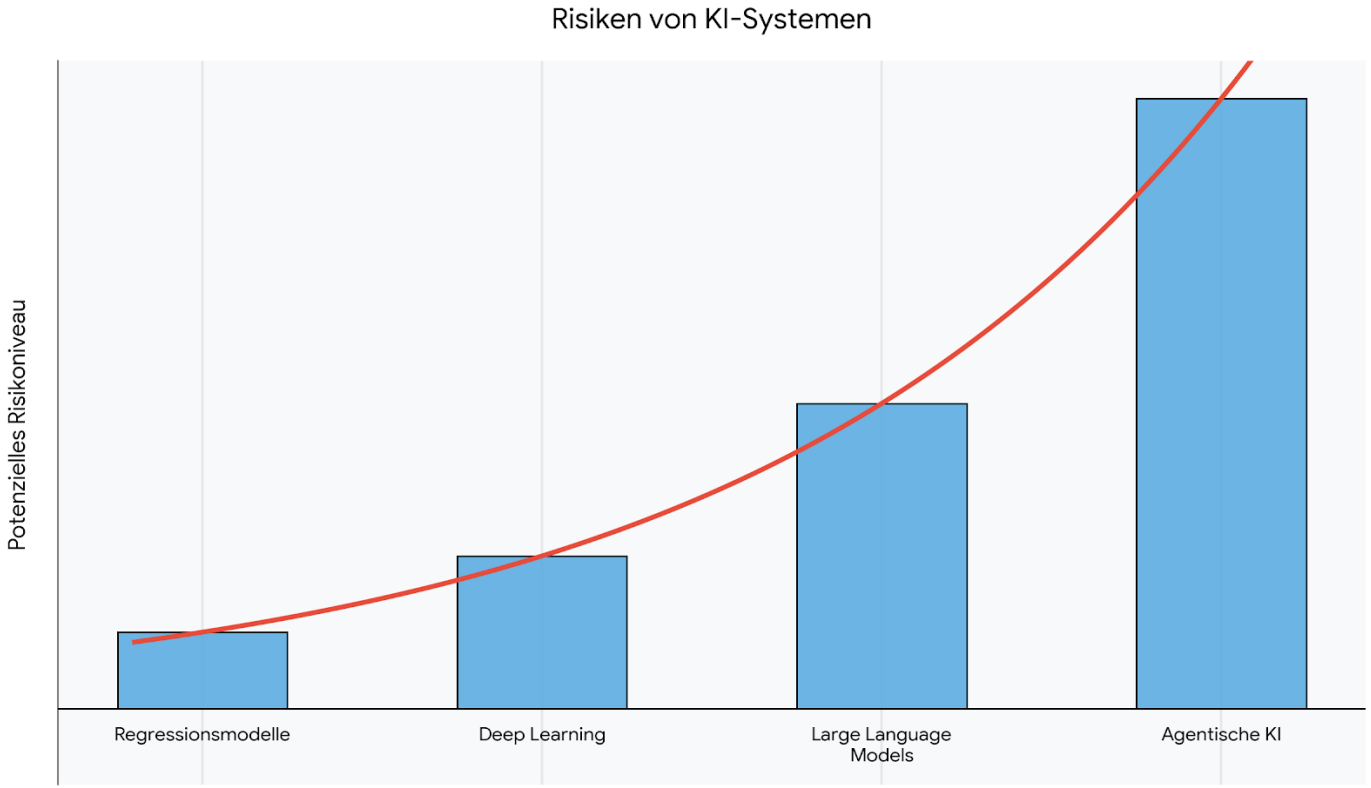


https://prolifics.com/usa/resource-center/blog/agentic-ai-in-healthcare

© 2026, Georg Dorffner

Risiken

- Generative KI: sprachliches Wissen, kommt genereller KI schon sehr nahe, aber:
  - Halluzination / Konfabulation
  - Unbekannte Trainingsquellen
  - Intransparenz (? – Sprache ist ***die*** Methode der Erklärbarkeit)
  - Datensicherheit (Modell-Server)
- Agentische KI:
  - Ungewollte Aktionen
  - Datensicherheit
  - Etc.



Erstellt mit Gemini

© 2026, Georg Dorffner

AI-Act - Risikostufen

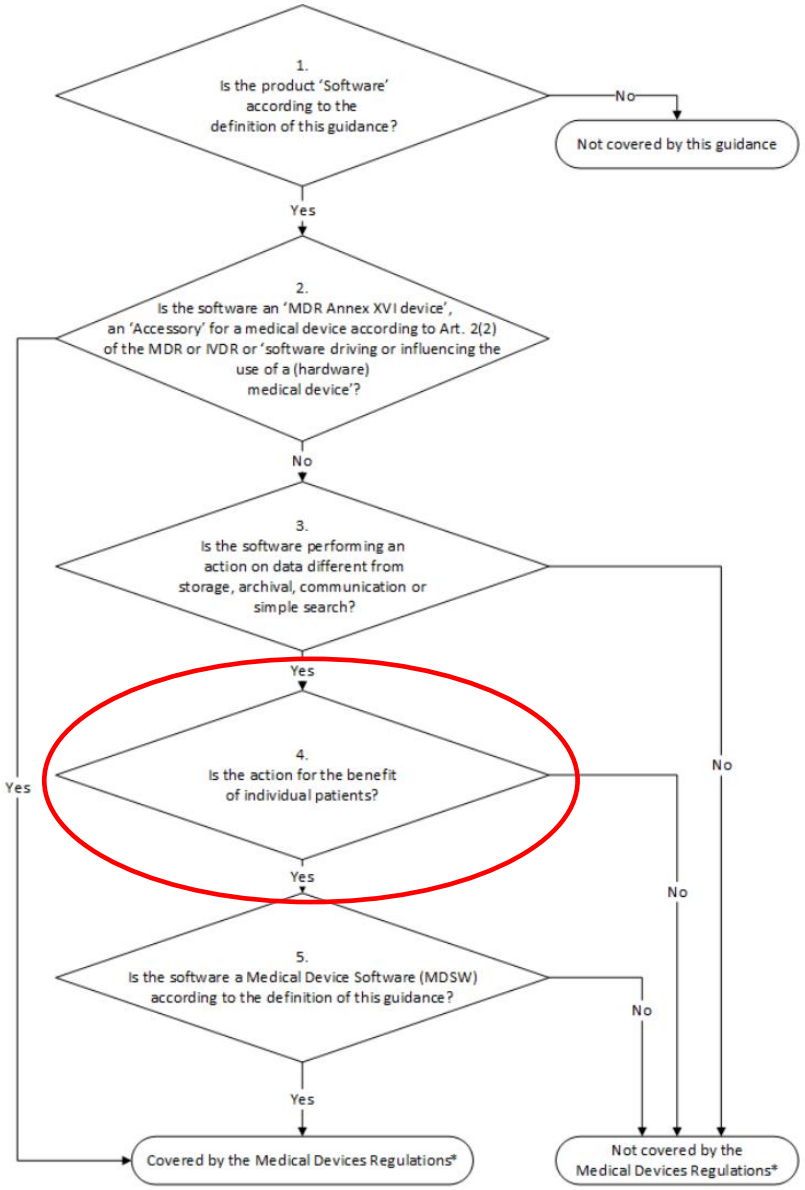
- **Inakzeptables Risiko (Verboten)**: z.B. staatliches Social Scoring, Echtzeit-Gesichtserkennung im öffentlichen Raum.
- **Hohes Risiko (Streng reguliert)**: Systeme in sensiblen Bereichen wie Medizin, kritischer Infrastruktur, Personalwesen oder Bildung.
- **Begrenztes Risiko (Transparenzpflichten)**: Systeme wie Chatbots oder Deepfakes. Nutzer müssen erkennen, dass sie mit einer KI interagieren oder dass Inhalte generiert wurden.
- **Minimales Risiko (Keine besonderen Pflichten)**: Anwendungen wie Spamfilter oder KI in Videospielen.

https://www.youtube.com/watch?v=alodAVKwxBA&t=34s
über gemini.google.com

© 2026, Georg Dorffner

Ausnahme für Forschung

- „This Regulation should support innovation, should respect freedom of science, and should not undermine research and development activity. It is therefore necessary to exclude from its scope AI systems and models specifically developed and put into service for the sole purpose of scientific research and development.“
- Wenn aber doch Medizinprodukt: Wieder im Fokus (als Hoch-Risiko-System).



MDCG 2019-11

© 2026, Georg Dorffner

Anforderungen für Hochrisiko KI

- Risiko-Management über den gesamten Lebenszyklus
- Daten- und Qualitätsmanagement
- Technische Beherrschbarkeit des Systems
- Konformitätsüberprüfung und -erklärung (wie in MDR, aber strenger)
- Beobachtung nach Markteinführung

© 2026, Georg Dorffner

Denkweise der EMA



9 September 2024
EMA/CHMP/CVMP/83833/2023
Committee for Medicinal Products for Human Use (CHMP)
Committee for Medicinal Products for Veterinary Use (CVMP)

Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle

- Fokus auf „context-of-use“ (Validierung auf Zielpopulation) und „fit-for-purpose“
- KI als Teil der statístischen Analyse betractet (z.B. strikte Vorgaben für Training/Validierung/Test-Daten)
- Pipeline für die Datenkuratierung, dokumentierte und “eingefrorene” Modelle
- Frühe Phase “low risk”, jedoch “high risk” wenn Dosierungs- oder Behandlungsentscheidung (z.B. Screening)
- Späte Phase: Risiken zu Overfitting oder Data Leakage müssen berücksichtigt werden
- Sponsoren ermutigt, Modelle in einem öffentlich zugänglichen Repository zu speichern

© 2026, Georg Dorffner

FDA guideline

- AI model [that] is to produce information or data to support regulatory decision-making

- 7-Schritt-Prozedur
- "risk-based credibility assessment"

- Dass AI Model ein Medizinprodukt ist (FDA 501k), hilft hier nicht

Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products

Guidance for Industry and Other Interested Parties

DRAFT GUIDANCE

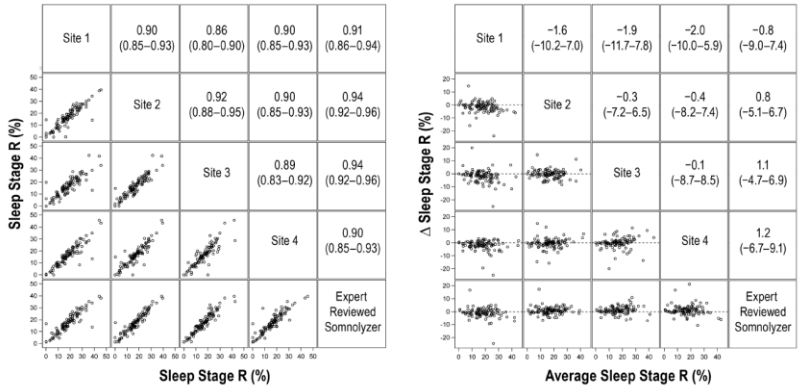
A. A Risk-Based Credibility Assessment Framework 5

1. Step 1: Define the Question of Interest 6
2. Step 2: Define the Context of Use for the AI Model 7
3. Step 3: Assess the AI Model Risk 8
4. Step 4: Develop a Plan to Establish AI Model Credibility Within the Context of Use 9
a. Describe the model and the model development process 10
i. Describe the model 10
ii. Describe the data used to develop the model 11
iii. Describe model training 12
b. Describe the model evaluation process 13
5. Step 5: Execute the Plan 15
6. Step 6: Document the Results of the Credibility Assessment Plan and Discuss Deviations From the Plan 15
7. Step 7: Determine the Adequacy of the AI Model for the Context of Use 15

© 2026, Georg Dorffner

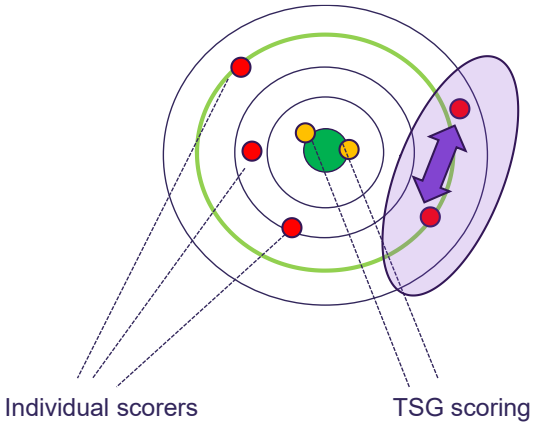
Beispiel: Schlafauswertung in Phase 3

- 20 Jahre lang als Endpoint in Phase I–II verwendet



Punbjabi et al., Sleep, 2015

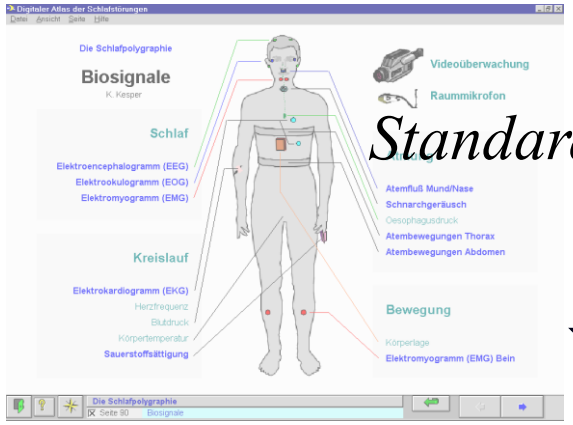
- KI bringt plötzliche neue Aufmerksamkeit
- Obwohl sie hier Risiken eher verringert
- Phase III (pivotal endpoint):
  - Validierung zeigt Äquivalenz zu Experten (innerhalb Variabilität)
  - „risk-based“ approach: Experte begutachtet Epoche für Epoche
  - FDA: „hybrides System“ muss validiert werden
  - Validierung auf spezifische Studienpopulation



© 2026, Georg Dorffner

Eine weitere fehlende Unterscheidung

- Messung
- Auswertung
- Berechnung des Studienendpunkts



Standard, z.B. EEG

KI /
im Vergleich zum Experten

Standard, z.B. Schlafeffizienz

Validierung

Neuartig, z.B. PPG

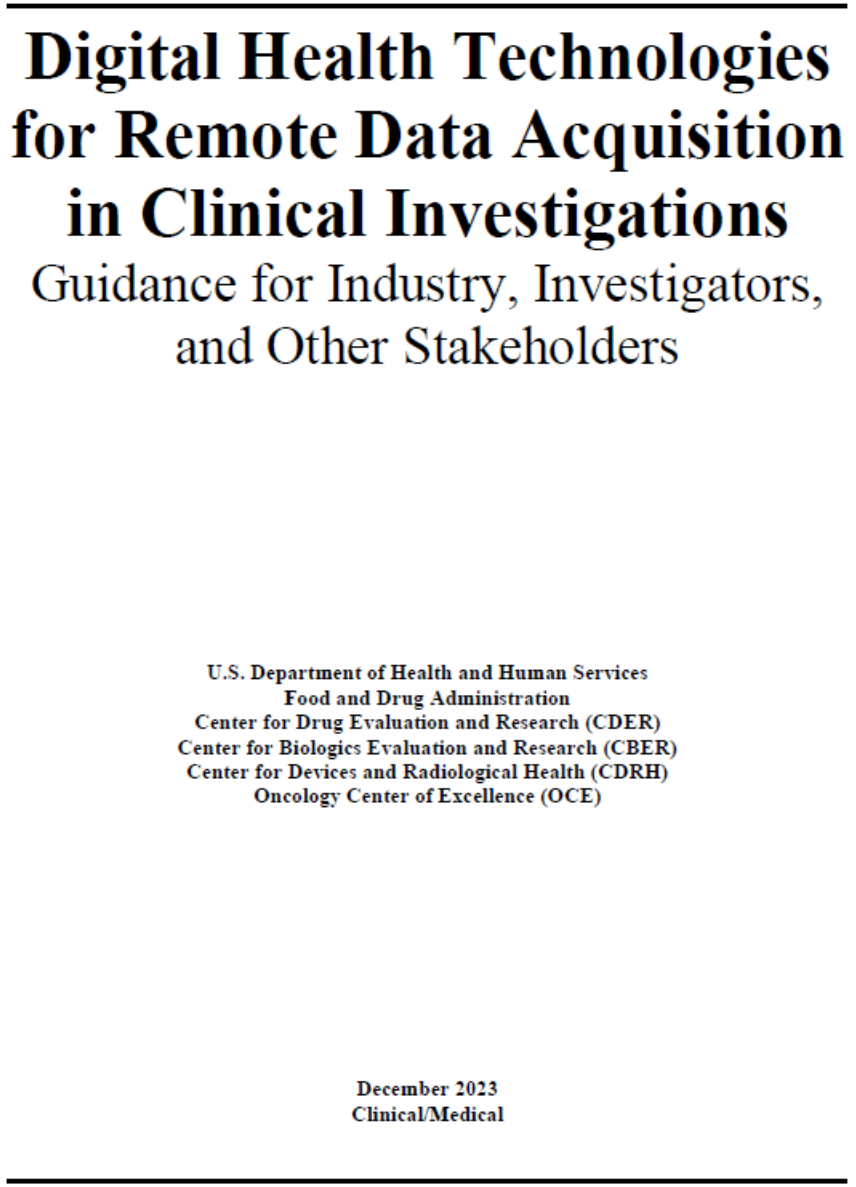


KI /
im Vergleich zu Außenkriterium
z.B. Alzheimer's

© 2026, Georg Dorffner

DHTs

- "A **digital health technology** (DHT) is a system that uses computing platforms, connectivity, software, and/or **sensors,** for health care and related uses. This guidance provides recommendations for sponsors, investigators, and other stakeholders on the use of DHTs for **remote data acquisition** from participants in clinical investigations that evaluate medical products."



Digital Health Technologies for Remote Data Acquisition in Clinical Investigations

Guidance for Industry, Investigators, and Other Stakeholders

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
Center for Devices and Radiological Health (CDRH)
Oncology Center of Excellence (OCE)

December 2023
Clinical/Medical

© 2026, Georg Dorffner

In Zukunft gemeinsam

- https://www.ema.europa.eu/en/news/ema-fda-set-common-principles-ai-medicine-development-0





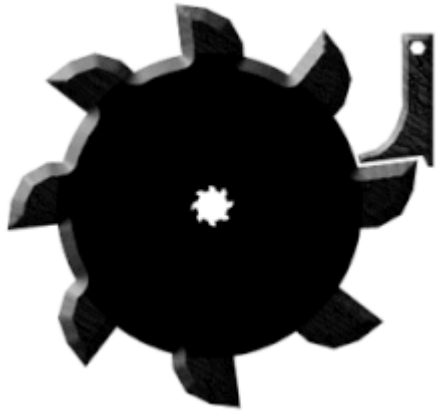
© 2026, Georg Dorffner

Die „Regularisierungsratsche“

- Wie schon so oft macht eine neue Technologie die Regeln strenger als sie je zuvor waren

- Oder haben wir uns vorher genau so ausführlich über die Risiken unserer Messungen den Kopf zerbrochen?
  - Bias in Rating Scales
  - Messfehler (waren bloß „Rauschen“)

- Naiver Aufruf: Konzentrieren wir uns auf „echte“ (neue) Risiken!

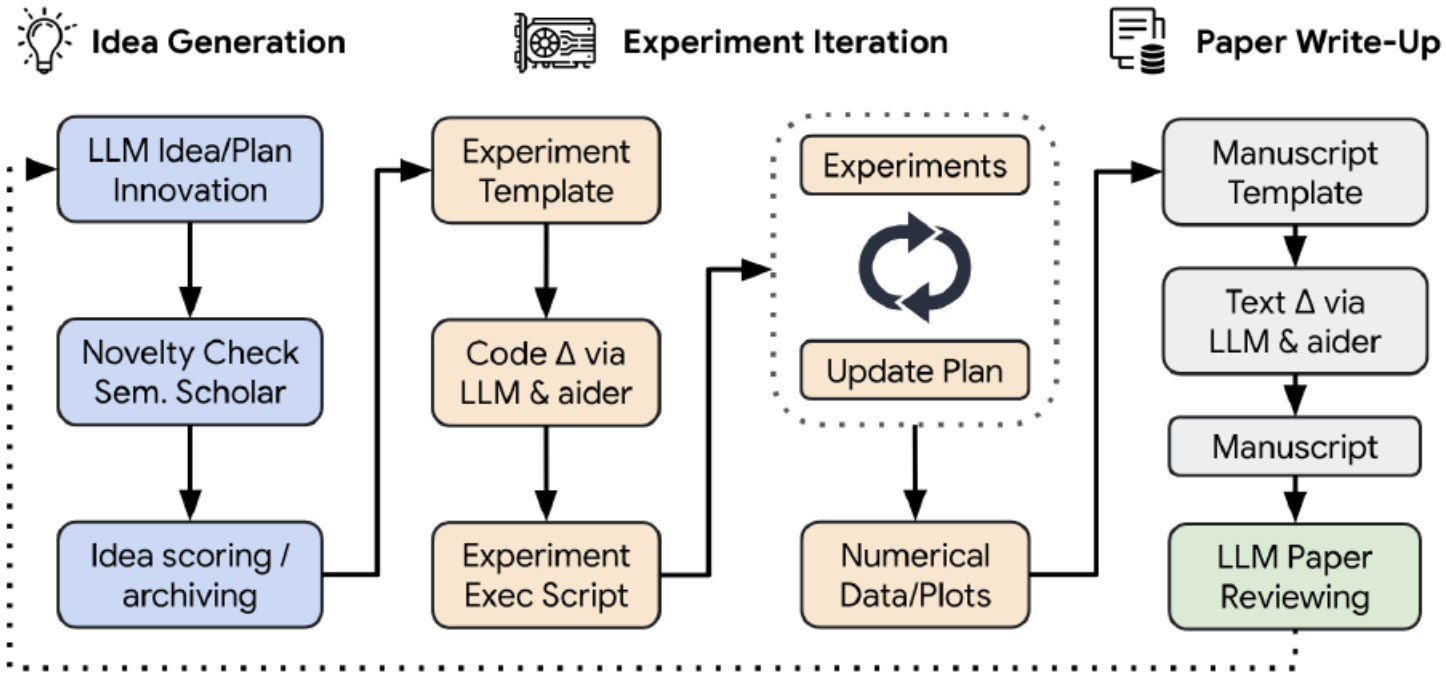


https://en.wikipedia.org/wiki/Ratchet_effect

© 2026, Georg Dorffner

The AI Scientist – KI für die klinische Forschung

- KI übernimmt immer mehr auch andere Rollen in der (klinischen) Forschung



- Da kommt noch einiges auf uns zu …

Lu et al., arXiv:2408.06292

© 2026, Georg Dorffner

Zusammenfassung

- KI kommt in mehreren Ausprägungen („Type“ 1–4), Risiko stark davon abhängig
- AI-Act und Regulatorien (FDA/EMA guidelines) machen wenig Unterschied
- Richtlinien existieren und erlauben KI als Endpoint in klinischen Studien
- Trotz Ausnahmen für Forschung zusätzliche Anforderungen
- Risiko-basierter Ansatz wichtig, aber sollte nicht übertrieben werden!

© 2026, Georg Dorffner