Sicherheit in der Ära der künstlichen Intelligenz


NETFLIX
STB PWR
1.@ 2abc 3def
4ghi 5jkl 6mno
7pqrs 8tuv 9wxyz
LIST 0
QUICK ACCESS
AD
P
RECENTS
GUIDE
amazon
LIVE ZOOM
FOCUS

Moonfall

Foto: Leonine

DER SPIEGEL
C 7007 CX
Nr. 16
32. Jahrgang · DM 3,–
17. April 1978

Die Computer-Revolution

Fortschritt macht arbeitslos

DER SPIEGEL
Nr. 36 / 3.9.2016
Deutschland €4,90

Sie sind entlassen!

Wie uns Computer und Roboter die Arbeit wegnehmen – und welche Berufe morgen noch sicher sind

Sportarzt Müller-Wohlfahrt
Heilende Hände, aber Probleme mit den Finanzen

Eine ukrainische Karriere
Erst Gefängnis, dann Präsidentenpalast?

Steinzeit-Drama
Warum Neandertaler zu Kannibalen wurden

www.spiegel.de

DER SPIEGEL
Nr. 10
4.3.2023

GELD
Versicherungen
Welche Sie wirklich brauchen

FLEISCH-ERSATZ
Der Mann, der uns vegan machen will

MEDIZIN
Rettende Baby-OP im Mutterleib

KÜNSTLICHE INTELLIGENZ

Die neue Weltmacht

Wie **ChatGPT und Co.** unser Leben verändern


Stand 2023

Use Case

Use Case

Use Case

Use Case

Use Case

Use Case

Fundament

Daten


Big Data

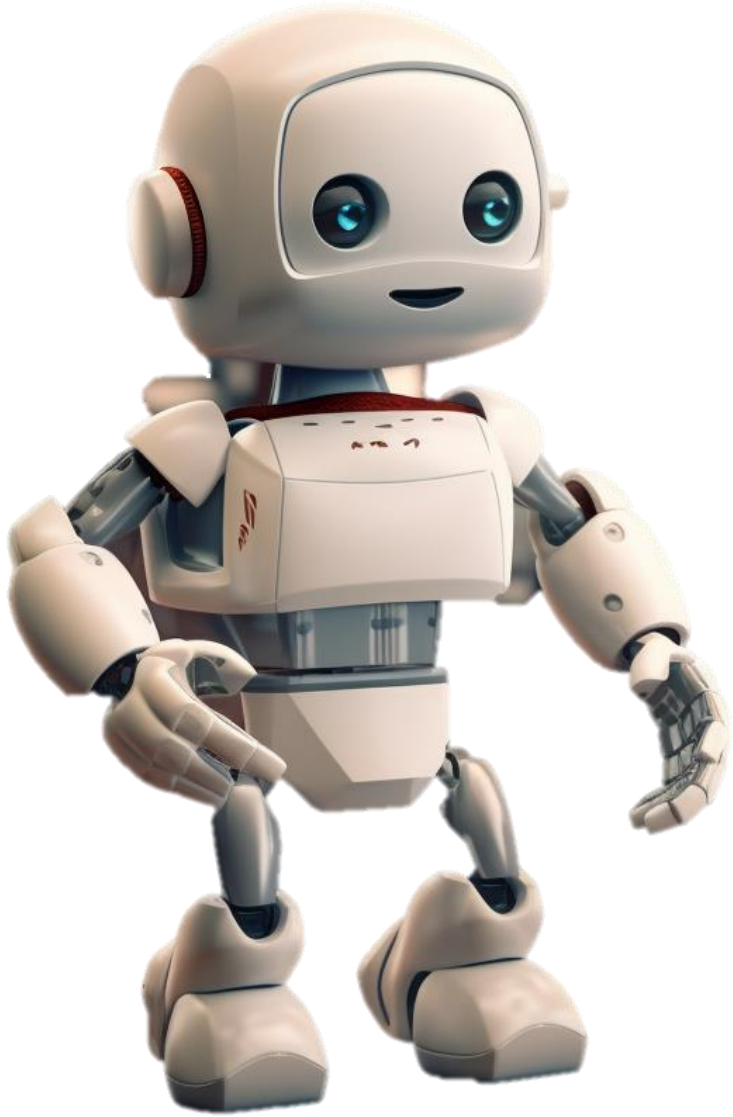
Definition

Künstliche Intelligenz

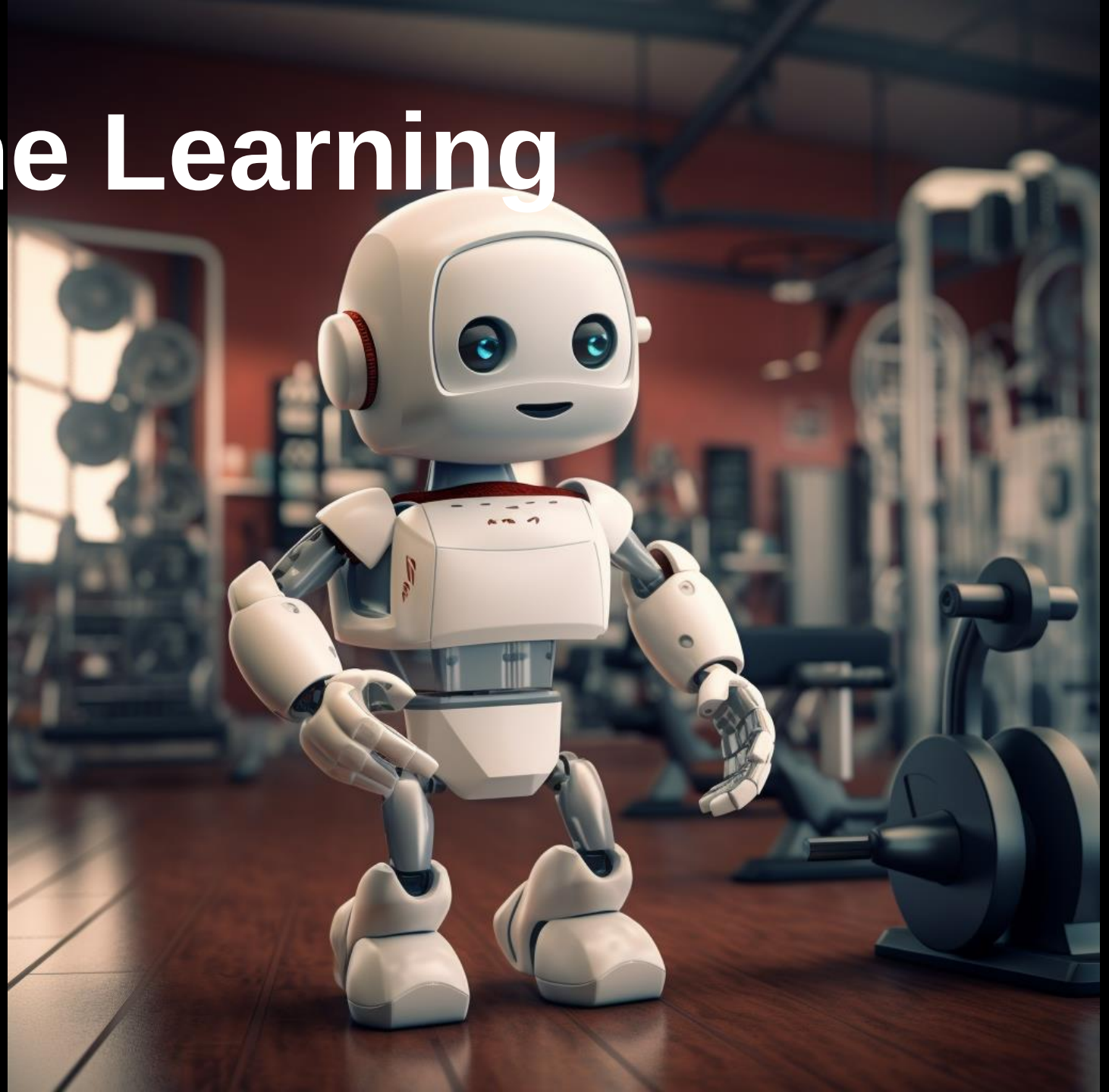
Machine Learning

Deep Learning

Reinforcement Learning

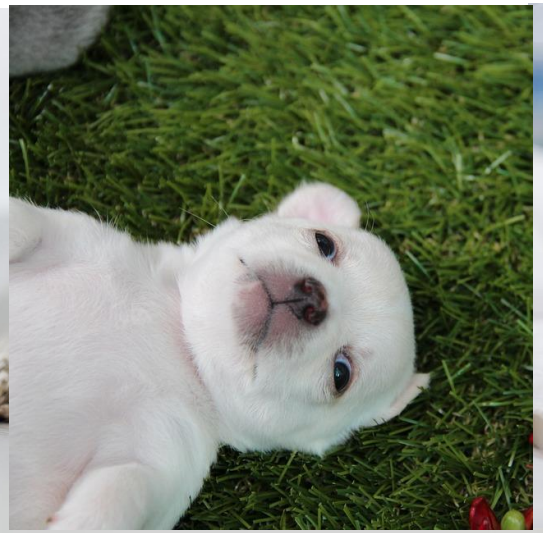
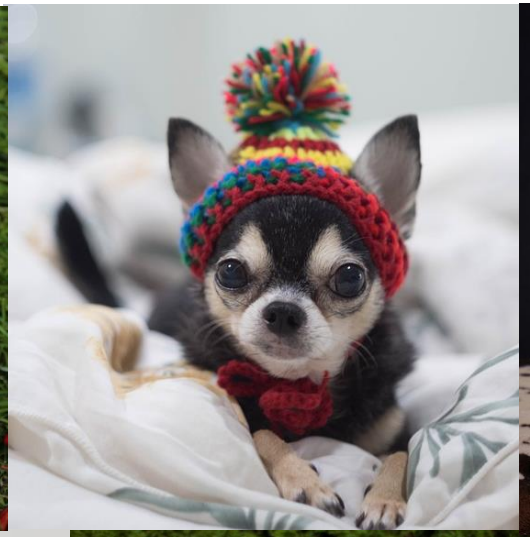


Machine Learning

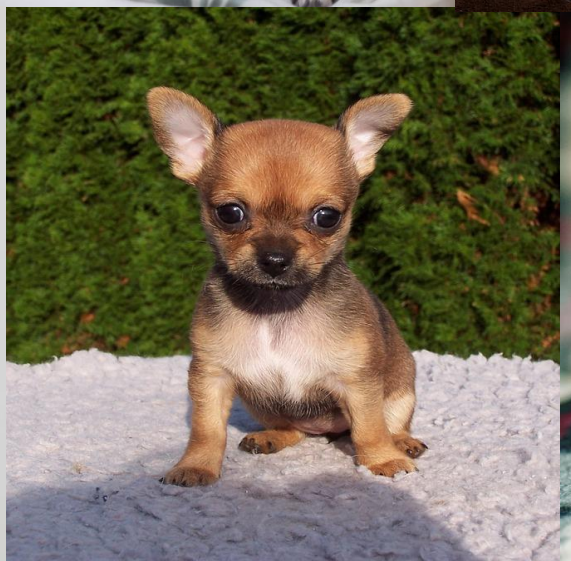


/fh///
st.pölten

Training






Training





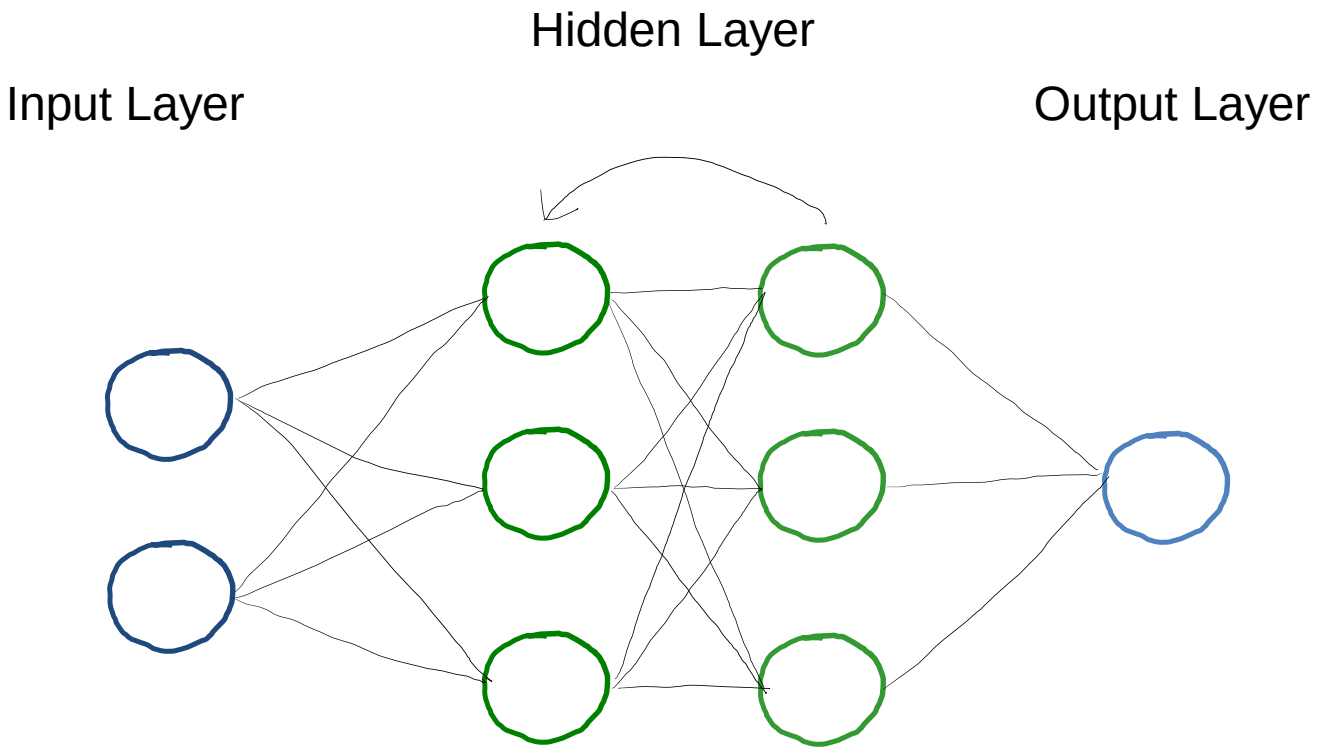


Problem

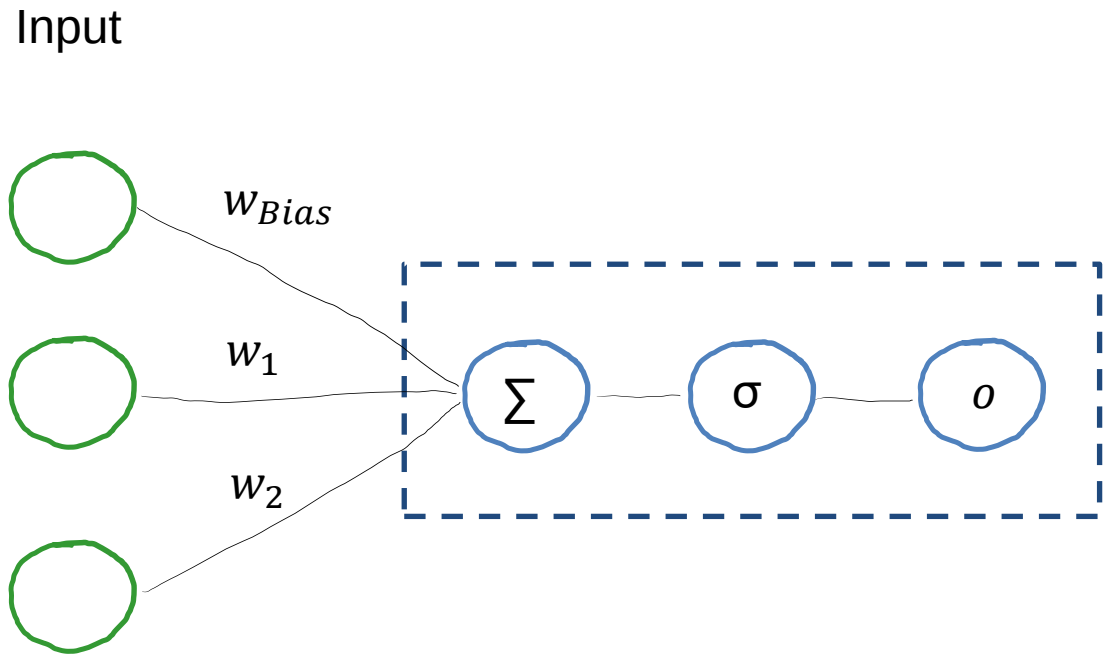
fh st.pölten

@teenybiscuit

Deep Learning



Deep Learning



Reinforcement Learning

Aktion

Neue Zustand

Belohnung / Bestrafung










FUCK

ChatGPT

ChatGPT

Examples	Capabilities	Limitations
"Explain quantum computing in simple terms" →	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?" →	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?" →	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021

ChatGPT Feb 13 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

GPT-3.5

Generative Pre-trained

Transformer 3.5

Language Model

Das Wetter in Österreich wird _?_.

$P_\theta(? = schön|h) = 0.15$

$P_\theta(? = bewölkt|h) = 0.2$

$P_\theta(? = regnerisch|h) = 0.4$

$P_\theta(? = hungrig|h) = 0.02$

...

/fh///
st.pölten

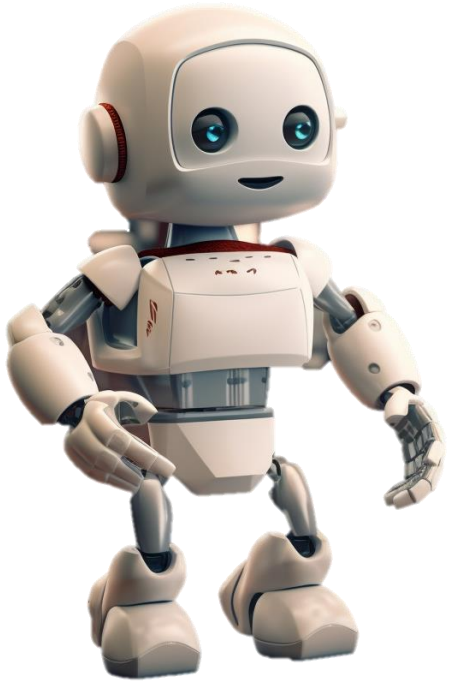
Transformer

+

Reinforcement Learning from Human Feedback

ChatGPT

Interaktionen werden möglich!

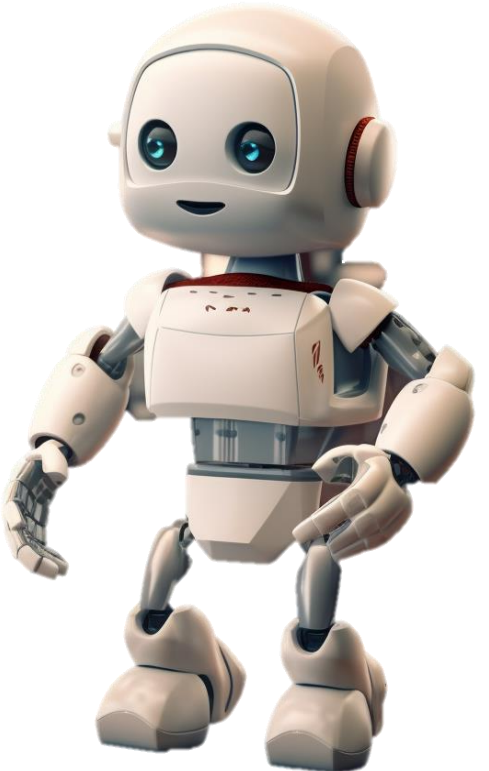


ChatGPT ist ein

Sprachmodell und

arbeitet mit

Wahrscheinlichkeiten.




Es fühlt sich so real an…

Ziel ist auch, dass ChatGPT „**menschelt**“.

Halluzinationen

Digital Life

Anwalt reicht ChatGPT-formulierte Klage mit erfundenen Fällen ein

28.05.2023

Gericht in New York deckt auf, dass Anklage mit gefälschten Fällen und Zitaten gespickt ist. Anwalt schiebt Schuld auf ChatGPT.

Sensible Informationen

Bias

presenting data visualizations to board of directors - @marlies (fast)

Threats

Alternativen

Human (in the Loop)

Expertise

Herzlichen Dank für die Aufmerksamkeit!

Bilder: Pixabay und Midjourney