— Künstliche Intelligenz —
Funktion+Verantwortung

VAS Arbeitsagogik · 2024-11-15 · Marcel Waldvogel · DNIP.ch

KI: Funktion und Verantwortung

Geschichte

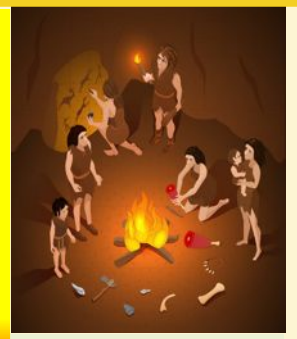
Funktion

Verantwortung

Ausblick

Entstehung der KI

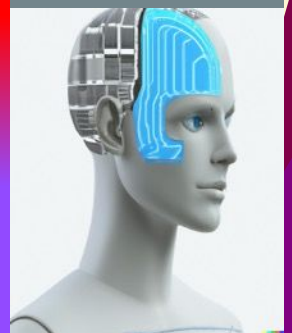
vergrössern



erweitern

ergänzen

ersetzen?



Expertensysteme

Regeln (wenn/dann)

Eliza (1966)

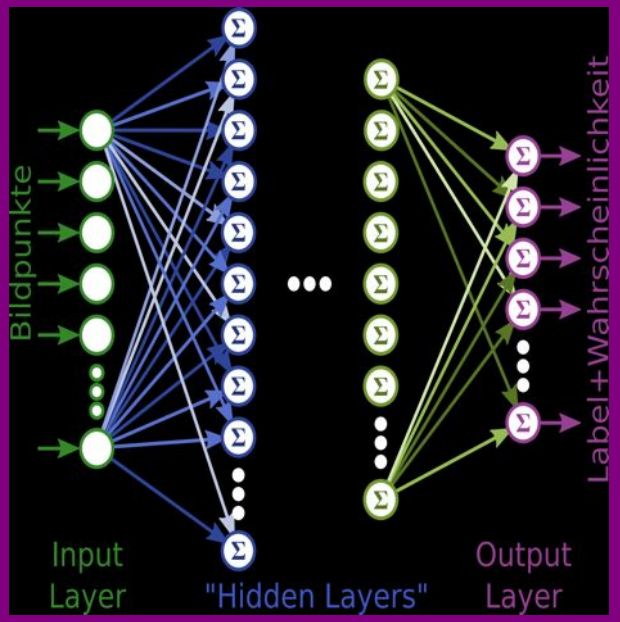
LISP, Prolog

Machine Learning

Neuronale Netze

Wahrscheinlichkeit

Datenmengen



KI: Funktion und Verantwortung

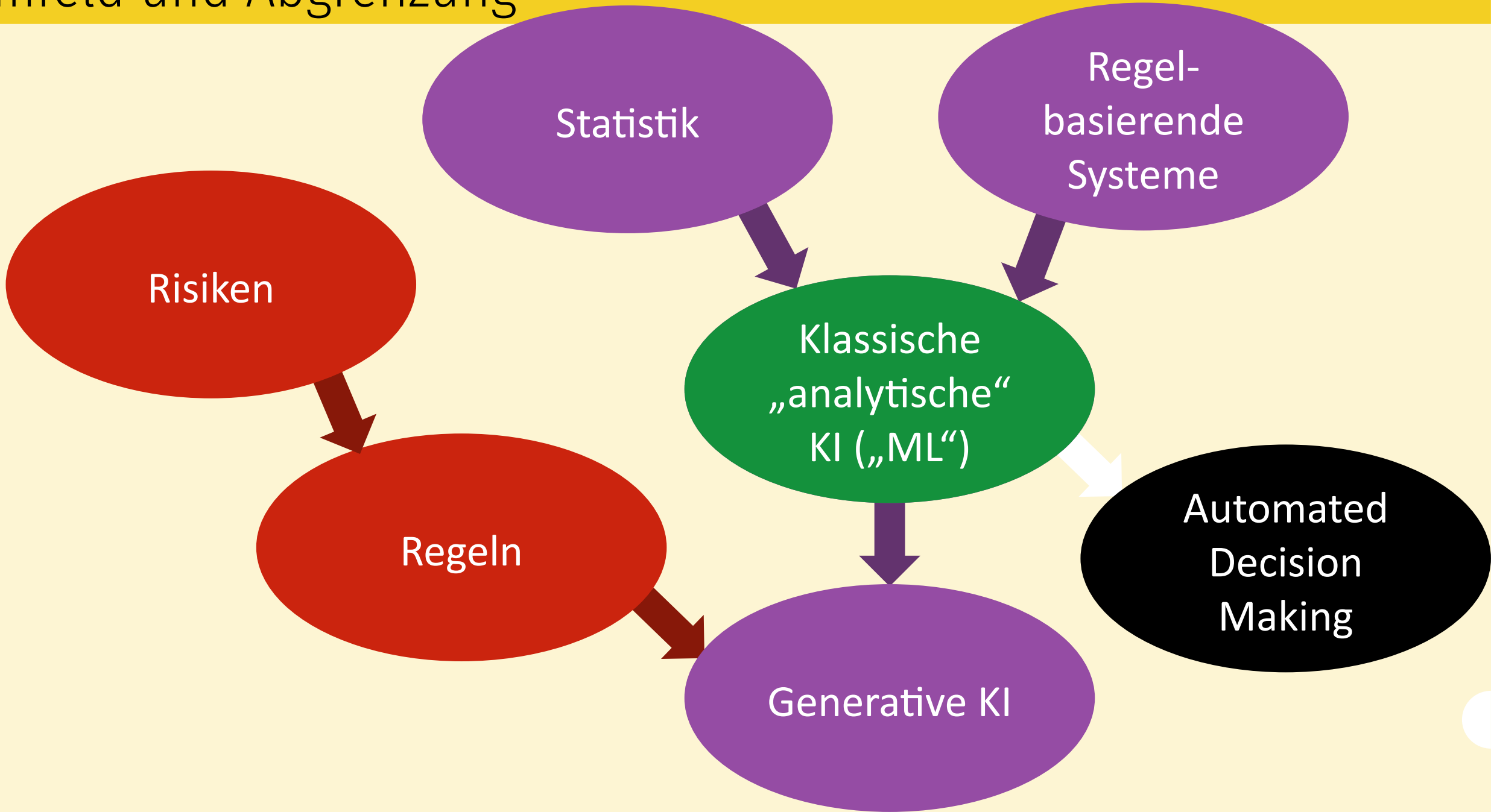
Geschichte

Funktion

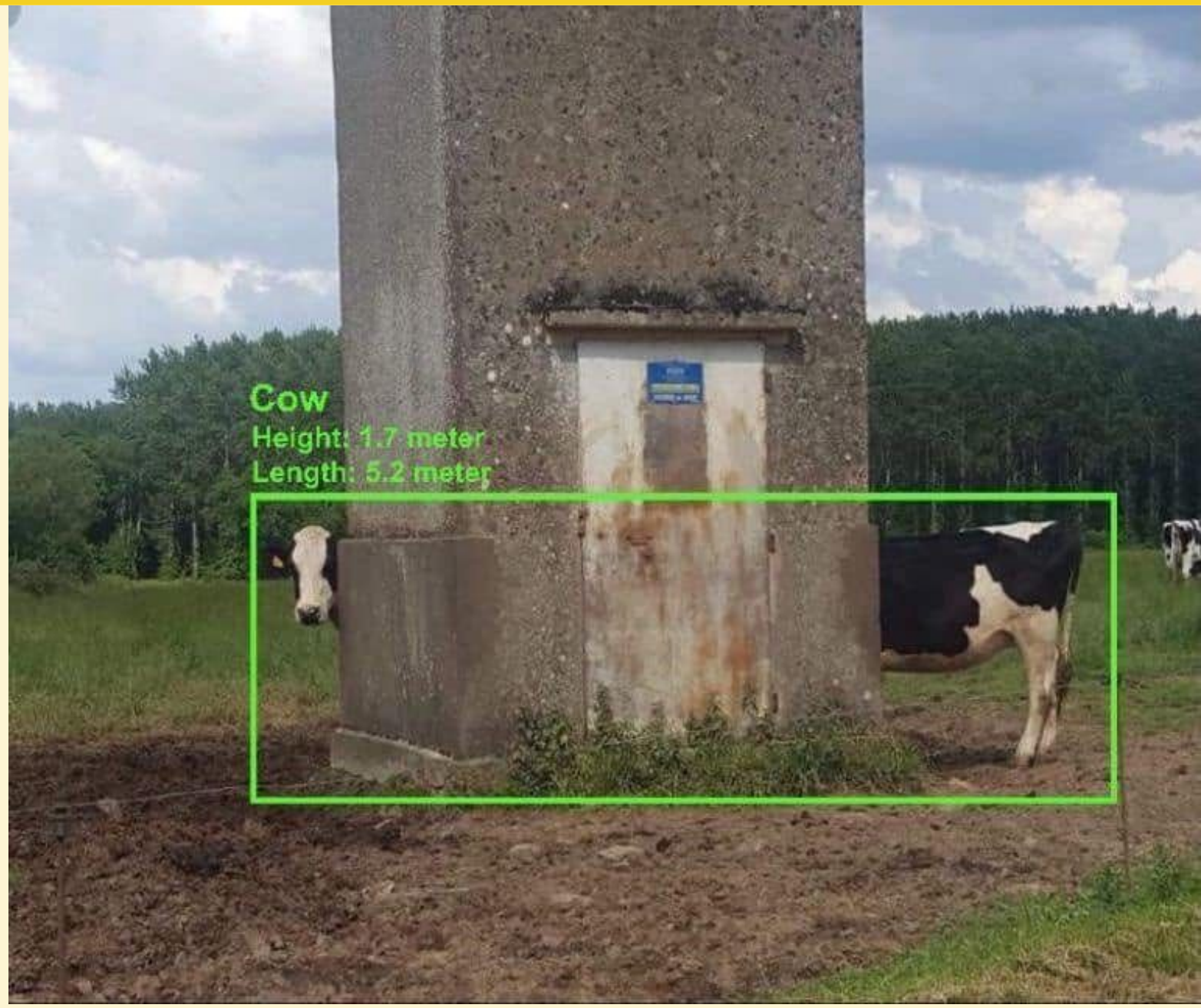
Verantwortung

Ausblick

Umfeld und Abgrenzung

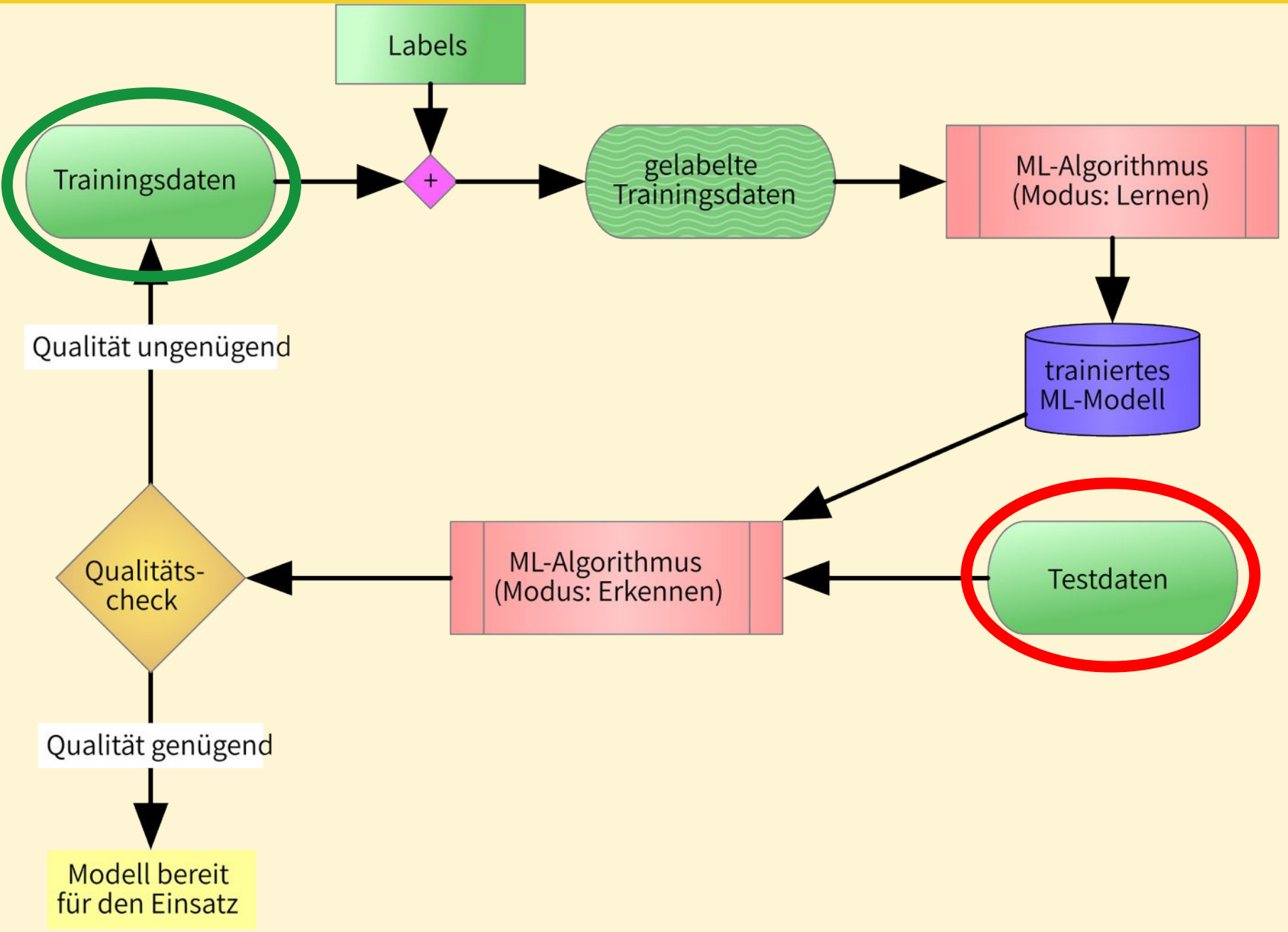


Maschinelles Lernen: Bilderkennung

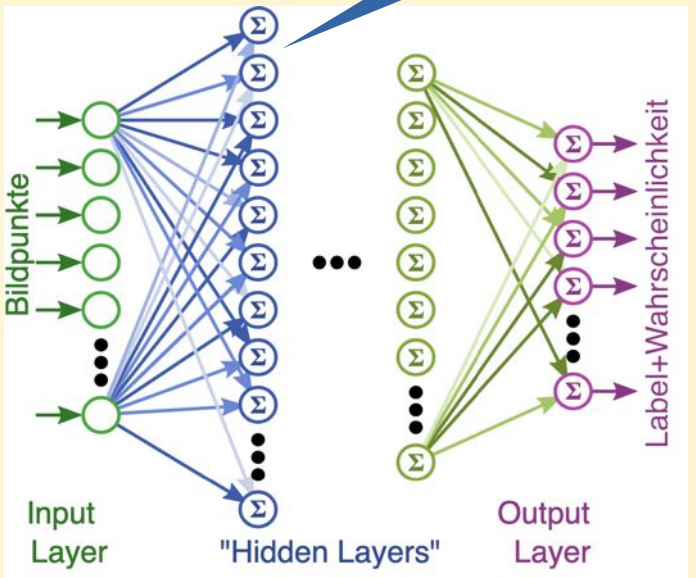


Bildquelle: https://twitter.com/bercut2000/status/1558461830565175299

ML-Training: Labelling, Training, Testing

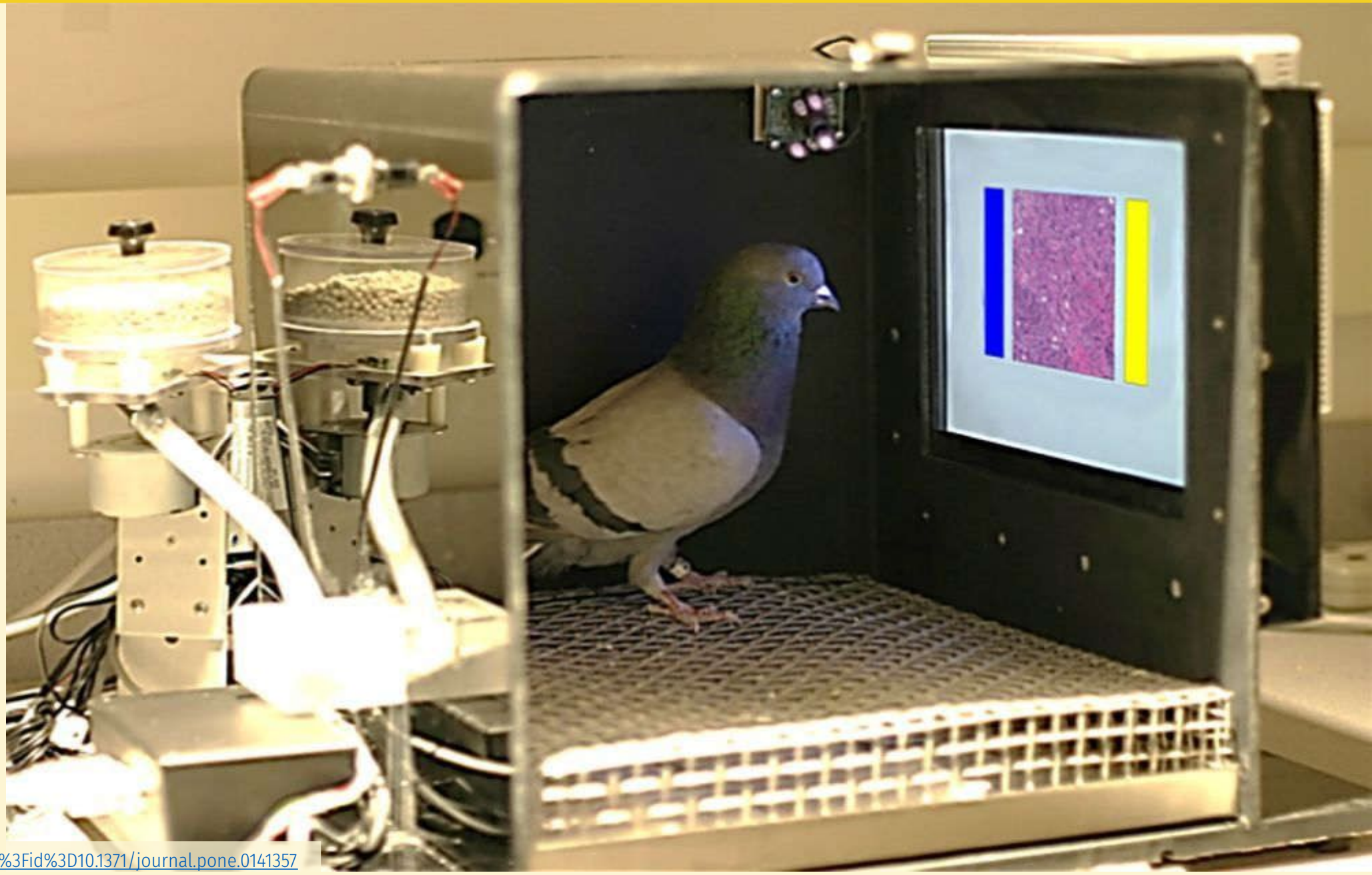


„Justiertes Mischpult“



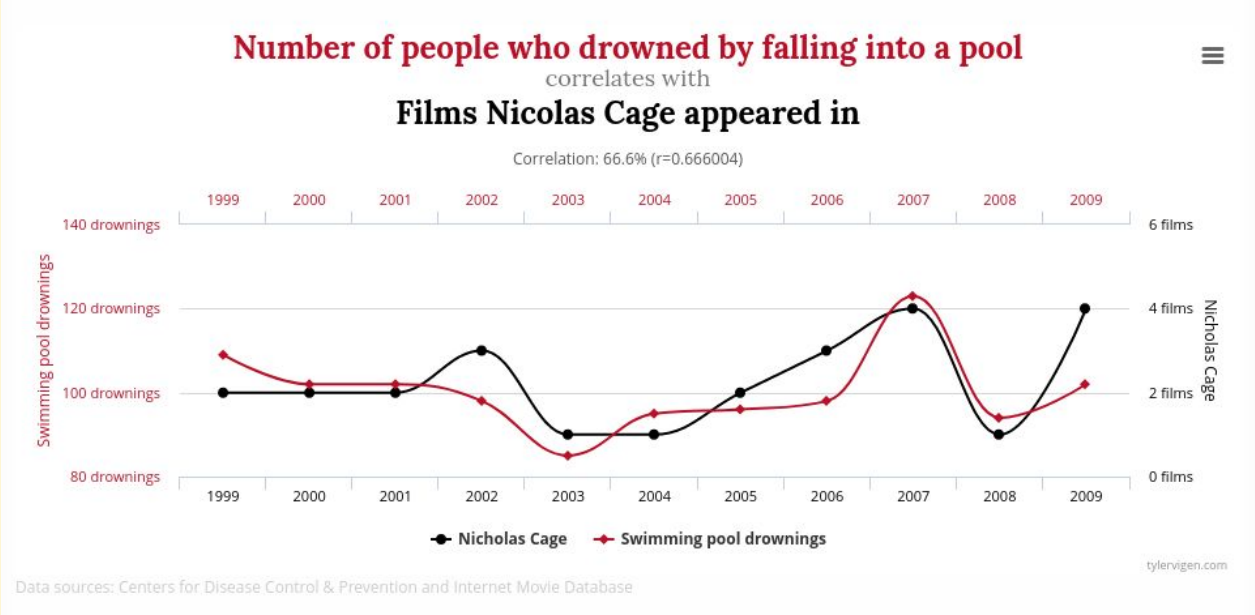
Grafik aus Machine Learning — Künstliche Faultier-Intelligenz

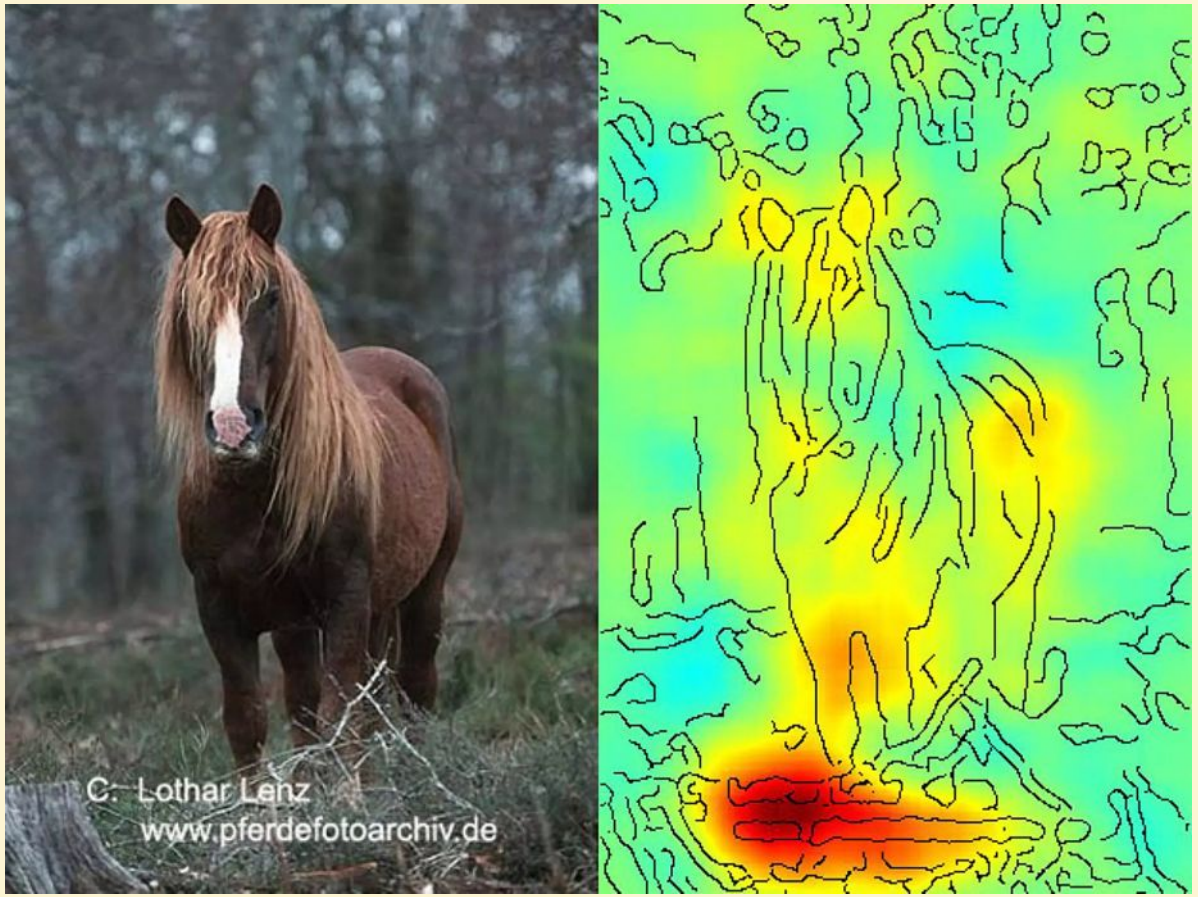
ML-Training



https://journals.plos.org/plosone/article%3Fid%3D10.1371/journal.pone.0141357

Glückliche Zufälle: «Correlation ≠ Causation»





Pferdemodell: Fraunhofer HHI via https://www.heise.de/-4665982; Korrelationen: https://tylervigen.com/spurious-correlations; Erklärungen: Machine Learning — Künstliche Faultier-Intelligenz

ML/Analytische KI

Correlation ≠ Causation:
Überzeugend, kein Verständnis

- Nur Korrelation, evt. ohne ursächlichen Zusammenhang
- Hohe Unsicherheit (Fehlerquote) bei mangelnden/mangelhaften Trainingsdaten
- Unabsichtliche „Fehlschlüsse“
- Mutwillige Fehlklassifikationen: „Adversarial Images/Data“

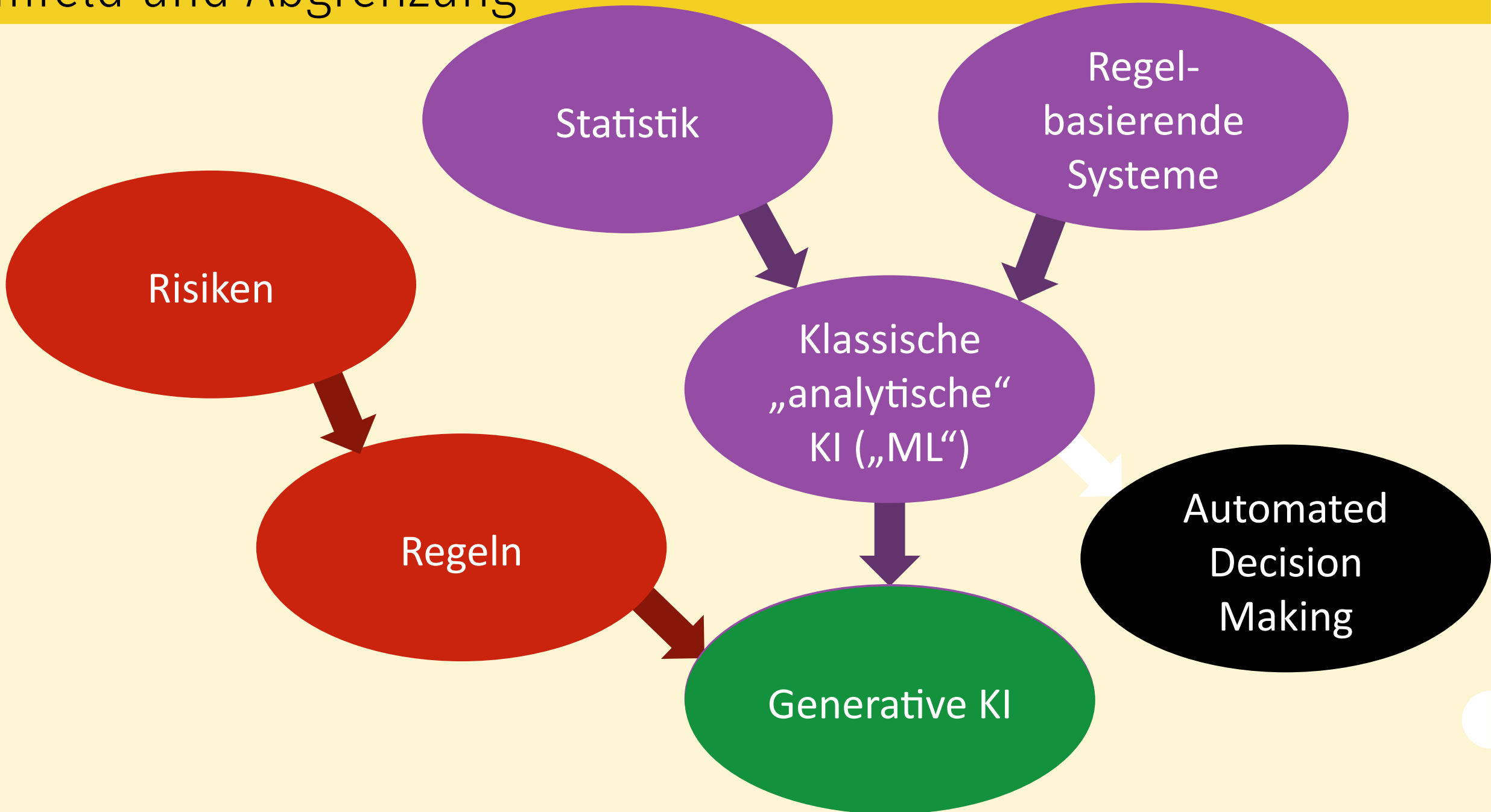
Schwierig nachzuvollziehen

- Testing; Regression Tests

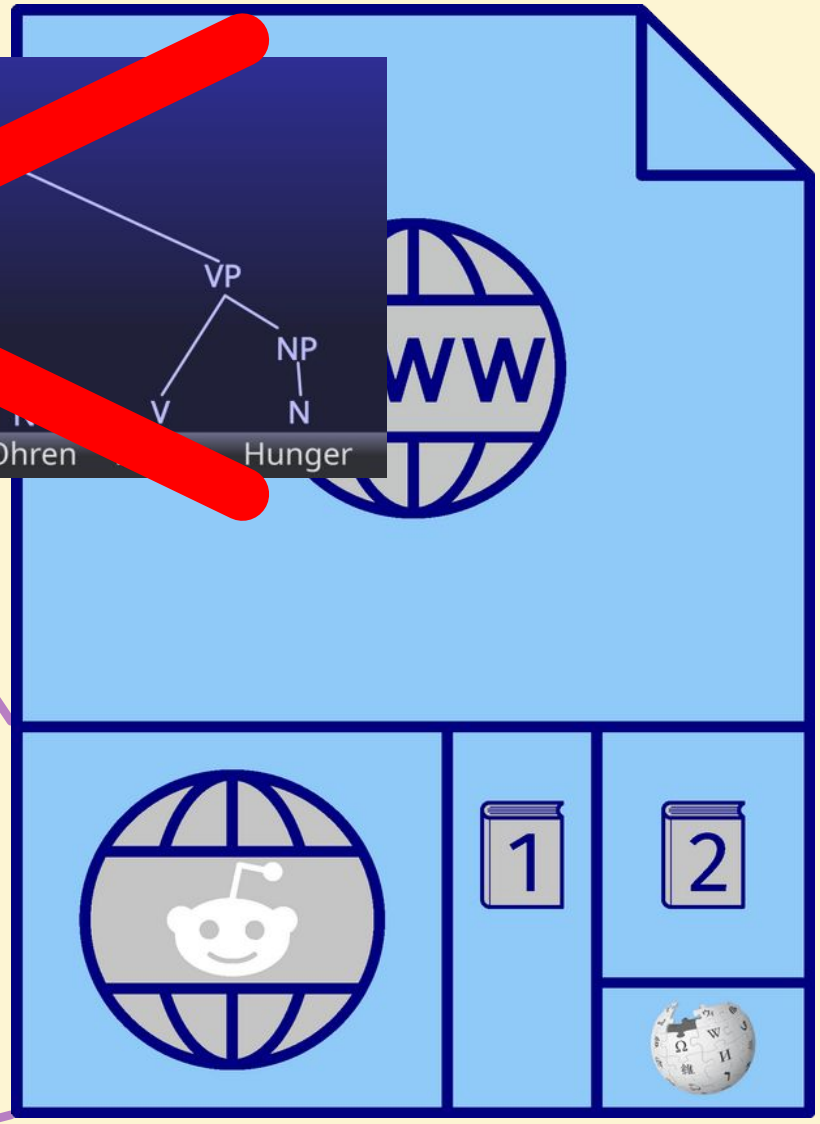
Verführerisch schnell erste Erfolge;
viel Aufwand für Verbesserungen

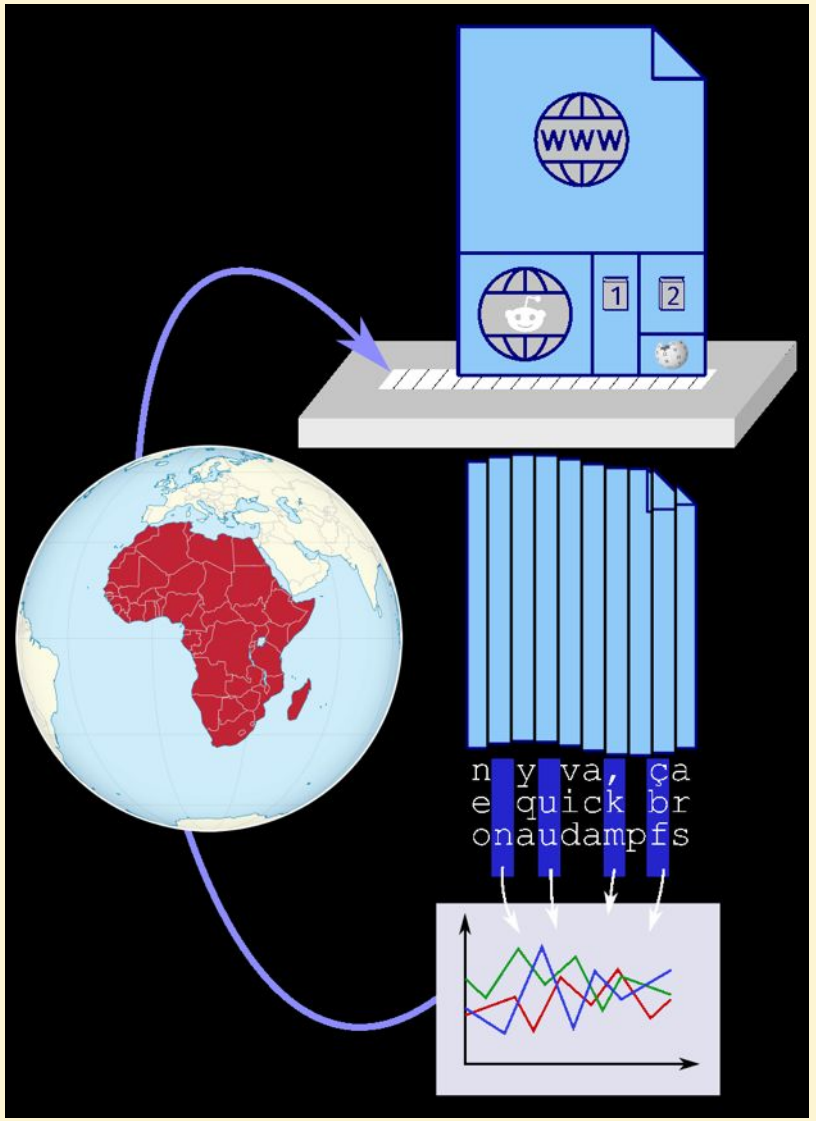
Stoppschild: Eykholt et al. (2018); Erklärungen, u.a. zu Adversarial Images: Machine Learning — Künstliche Faultier-Intelligenz

Umfeld und Abgrenzung



Datenquellen: ChatGPT und andere LLMs





Daten: https://arxiv.org/abs/2005.14165 und https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/; Pizza Story, Reddit Post „My cheese slices off the pizza too easily“

GPT-basierte LLMs (ChatGPT, Gemini, Llama, …)

LLMs sind Textvervollständigungen, welche Fragen anhand von Mustern vervollständigen, die aus geschredderten Texten aus dem Internet extrahiert wurden.

Die Ausgabe wird aufgrund von Häufigkeiten erzeugt, wie wahrscheinlich eine Zeichensequenz («Token») auf eine andere Zeichensequenz folgt; moduliert von diesen extrahierten Mustern.

Jeder Output ist zufällig. Vorausschauende Planung gibt es nicht.

Mit FAQs werden sie auf Beantwortung getrimmt.

Noch mehr Statistik

```
The top in a world by susphoring grace.

LUCIO:
We muse hath resistes him so sovere: son't his other wrough
stands of coverent sh'd: he has here, and stand it
and poor exceeder or a Henry's last, stay
not in faith, forewell's base of graves, thanks, happy comparel,
warmentfully: may as face by the courst, that strangth
errise hath breathed. Hastings come to Valenting.

HERMIONE:
Well have been bolly poor late
Is the lords.

ABELLA:
Let's found: I will kind him;
I do braw'sy him business wherein far his face.

LUCENTIO:
He is last afford: make him diseably to London,
Take him great Hastings, boldness in his natic keeps,
```

Trainingsdaten

```
First Citizen:
Before we proceed any further,

All:
Speak, speak.

First Citizen:
You are all resolved rather to

All:
Resolved. resolved.

First Citizen:
First, you know Caius Marcius i
```

Zeichenstatistik

t → h, i, e, …
i → n, s, t, …

```
POPSousthe?
LA:

VI
LLINotom.
S: we, ar,
Sewed t--qurds ce hand the al OSTistir:
IORICinid thas r eas th here anthe t prose
F ticleasuolur d
T:
pesomo owst pugino d

ARif w k ithore th, Roue s ped tha okifok,
Ar.
```

Längere statistische Zusammenhänge

```
And the Ror
Thow and is and thrad thom obe to tarver-and that hauss ar hapie us hat tot?
Wedtlad ane aw crupeak,
Do'n om onour
Yowns, tof it he cove lend lincats if ees, hain lat Het drovets, and to poman is wables
knamopetell lownomthy wod moth keeoal--so wher eiicks to thour rive cees,
We
An so mower; toure kad nocrupt for to igis! my to thy ale ontat af Pried my of.
```

Detailliertere Beschreibung als Text: „Wie funktioniert eigentlich ChatGPT“; Screenshots aus Andrej Karpathys „Wir bauen uns NanoGPT“-Video

Generative KI: Risiken kennen

Risiken der zugrunde liegenden Bausteine:

- Correlation ≠ Causation:
  Überzeugend, kein Verständnis
- Schwierig nachzuvollziehen
- Verführerisch schnelle erste Erfolge;
  viel Aufwand für Verbesserungen

Neue Risiken:

- Sehr überzeugend, „menschlich“
  - Preisgabe von Daten
- Externe Anbieter (oft weitgehende Rechte)
  - Nutzung, Weitergabe von Daten
- Unklare Trainingsdaten (und Rechte)
  - Evt. veraltet, einseitig, voreingenommen
  - Evt. Verletzung Rechte Dritter
  - Fragwürdige Ethik
- Unklare Resultate (und Rechte)
  - Evt. Verletzung Rechte Dritter
  - Evt. nicht schützenswert
  - Fehlerhafte, unwahre Informationen
    (z.B. Satire als echt)
  - Fehlerhafter, unsicherer Programmcode

KI: Funktion und Verantwortung

Geschichte

Funktion

Verantwortung

Ausblick

Generative KI: Risiken kennen

Trends:

- KI in jedem Produkt, jeder Dienstleistung
- Von hilfreich über Buzzword bis abweisend
- Gegentrend als Qualitätsmerkmal

Chancen:

- Innovatives Image
- Unterstützung für Mitarbeitende
- Selbstbedienung für Dritte ⚠️
- Einfachere, effizientere Prozesse

KI erst, nachdem der Geschäfts-prozess maximal vereinfacht wurde (neue Komplexität!)

KI-spezifische Risiken:

- Datenschutz: AGB
- Zuverlässigkeit: Halluzination/Fabulieren
- (Schutzrechte)

Nicht KI-spezifische Risiken:

- Nutzung von Cloud-Diensten privat (BYOC, BYOAI; Übersetzungsdienst, Online-Zeichenprogramm, Cloud-Speicher, Umfragetools, Datenanalyse, …)
  - Technisch nicht zu verhindern (Privathandy)
- IT-Dienste: Datensicherheit

KI: Funktion und Verantwortung

Geschichte

Funktion

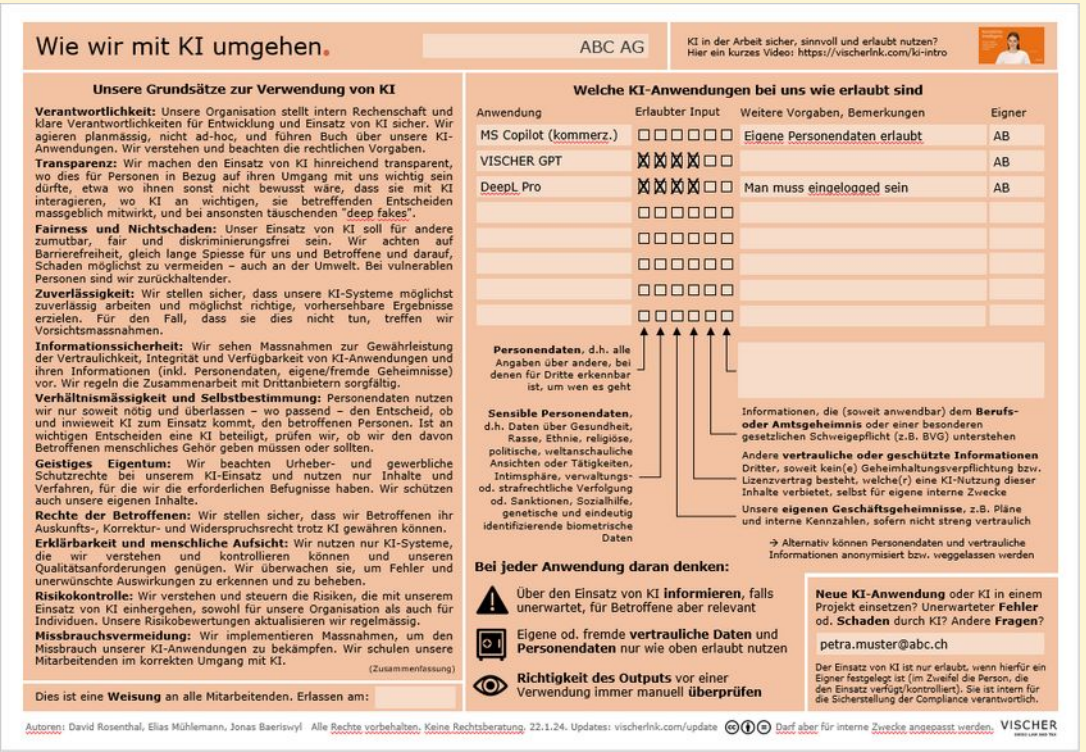
Verantwortung

Ausblick

Generative KI: Wie regeln?

Für die Organisation:

- Prinzipien erstellen für den Umgang mit IT-Diensten (nicht nur KI!)
  - Auch (gerade!) kostenlose Dienste: Datenschutz, Datensicherheit evaluieren
  - Informieren, überzeugen
  - «Gut genug» zur Verfügung stellen, Best Practices



Wie wir mit KI umgehen. ABC AG

KI in der Arbeit sicher, sinnvoll und erlaubt nutzen? Hier ein kurzes Video: https://vischerlnk.com/ki-intro

Unsere Grundsätze zur Verwendung von KI

Verantwortlichkeit: Unsere Organisation stellt intern Rechenschaft und klare Verantwortlichkeiten für Entwicklung und Einsatz von KI sicher. Wir agieren planmässig, nicht ad-hoc, und führen Buch über unsere KI-Anwendungen. Wir verstehen und beachten die rechtlichen Vorgaben.

Transparenz: Wir machen den Einsatz von KI hinreichend transparent, wo dies für Personen in Bezug auf ihren Umgang mit uns wichtig sein dürfte, etwa wo ihnen sonst nicht bewusst wäre, dass sie mit KI interagieren, wo KI an wichtigen, sie betreffenden Entscheiden massgeblich mitwirkt, und bei ansonsten täuschenden "deep fakes".

Fairness und Nichtschaden: Unser Einsatz von KI soll für andere zumutbar, fair und diskriminierungsfrei sein. Wir achten auf Barrierefreiheit, gleich lange Spiesse für uns und Betroffene und darauf, Schaden möglichst zu vermeiden – auch an der Umwelt. Bei vulnerablen Personen sind wir zurückhaltender.

Zuverlässigkeit: Wir stellen sicher, dass unsere KI-Systeme möglichst zuverlässig arbeiten und möglichst richtige, vorhersehbare Ergebnisse erzielen. Für den Fall, dass sie dies nicht tun, treffen wir Vorsichtsmassnahmen.

Informationssicherheit: Wir sehen Massnahmen zur Gewährleistung der Vertraulichkeit, Integrität und Verfügbarkeit von KI-Anwendungen und ihren Informationen (inkl. Personendaten, eigene/fremde Geheimnisse) vor. Wir regeln die Zusammenarbeit mit Drittanbietern sorgfältig.

Verhältnismässigkeit und Selbstbestimmung: Personendaten nutzen wir nur soweit nötig und überlassen – wo passend – den Entscheid, ob und inwieweit KI zum Einsatz kommt, den betroffenen Personen. Ist an wichtigen Entscheiden eine KI beteiligt, prüfen wir, ob wir den davon Betroffenen menschliches Gehör geben müssen oder sollten.

Geistiges Eigentum: Wir beachten Urheber- und gewerbliche Schutzrechte bei unserem KI-Einsatz und nutzen nur Inhalte und Verfahren, für die wir die erforderlichen Befugnisse haben. Wir schützen auch unsere eigenen Inhalte.

Rechte der Betroffenen: Wir stellen sicher, dass wir Betroffenen ihr Auskunfts-, Korrektur- und Widerspruchsrecht trotz KI gewähren können.

Erklärbarkeit und menschliche Aufsicht: Wir nutzen nur KI-Systeme, die wir verstehen und kontrollieren können und unseren Qualitätsanforderungen genügen. Wir überwachen sie, um Fehler und unerwünschte Auswirkungen zu erkennen und zu beheben.

Risikokontrolle: Wir verstehen und steuern die Risiken, die mit unserem Einsatz von KI einhergehen, sowohl für unsere Organisation als auch für Individuen. Unsere Risikobewertungen aktualisieren wir regelmässig.

Missbrauchsvermeidung: Wir implementieren Massnahmen, um den Missbrauch unserer KI-Anwendungen zu bekämpfen. Wir schulen unsere Mitarbeitenden im korrekten Umgang mit KI.

(Zusammenfassung)

Dies ist eine **Weisung** an alle Mitarbeitenden. Erlassen am:

Welche KI-Anwendungen bei uns wie erlaubt sind

Anwendung	Erlaubter Input	Weitere Vorgaben, Bemerkungen	Eigner
MS Copilot (kommerz.)	☐☐☐☐☐☐	Eigene Personendaten erlaubt	AB
VISCHER GPT	☒☒☒☒☐☐		AB
DeepL Pro	☒☒☒☒☐☐	Man muss eingelogged sein	AB
	☐☐☐☐☐☐		
	☐☐☐☐☐☐		
	☐☐☐☐☐☐		
	☐☐☐☐☐☐		
	☐☐☐☐☐☐		

Personendaten, d.h. alle Angaben über andere, bei denen für Dritte erkennbar ist, um wen es geht

Sensible Personendaten, d.h. Daten über Gesundheit, Rasse, Ethnie, religiöse, politische, weltanschauliche Ansichten oder Tätigkeiten, Intimsphäre, verwaltungs- od. strafrechtliche Verfolgung od. Sanktionen, Sozialhilfe, genetische und eindeutig identifizierende biometrische Daten

Informationen, die (soweit anwendbar) dem **Berufs- oder Amtsgeheimnis** oder einer besonderen gesetzlichen Schweigepflicht (z.B. BVG) unterstehen

Andere **vertrauliche oder geschützte Informationen** Dritter, soweit kein(e) Geheimhaltungsverpflichtung bzw. Lizenzvertrag besteht, welche(r) eine KI-Nutzung dieser Inhalte verbietet, selbst für eigene interne Zwecke

Unsere **eigenen Geschäftsgeheimnisse**, z.B. Pläne und interne Kennzahlen, sofern nicht streng vertraulich

→ Alternativ können Personendaten und vertrauliche Informationen anonymisiert bzw. weggelassen werden

Bei jeder Anwendung daran denken:

Über den Einsatz von KI **informieren**, falls unerwartet, für Betroffene aber relevant

Eigene od. fremde **vertrauliche Daten** und **Personendaten** nur wie oben erlaubt nutzen

Richtigkeit des Outputs vor einer Verwendung immer manuell **überprüfen**

Neue KI-Anwendung oder KI in einem Projekt einsetzen? Unerwarteter **Fehler** od. **Schaden** durch KI? Andere **Fragen**?

petra.muster@abc.ch

Der Einsatz von KI ist nur erlaubt, wenn hierfür ein Eigner festgelegt ist (im Zweifel die Person, die den Einsatz verfügt/kontrolliert). Sie ist intern für die Sicherstellung der Compliance verantwortlich.

Autoren: David Rosenthal, Elias Mühlemann, Jonas Baeriswyl Alle Rechte vorbehalten, Keine Rechtsberatung. 22.1.24. Updates: vischerlnk.com/update Darf aber für interne Zwecke angepasst werden. VISCHER

KI-Prinzipienblatt: David Rosenthal et al., Vischer

KI-Basisregelung für Nutzende:

- Input/Datenschutz
  Keine personenidentifizierenden, interne, vertraulichen, geheimen Daten in Prompts
  - Ausnahme: Speziell geprüfte Dienste

«80%-Regel»:
KI nur für Dinge einsetzen, welche man zu mindestens 80% selbst gut versteht.
(Überprüfung, Psychologie)

- Output/Korrektheit
  - Überprüfung vor Nutzung/Weitergabe
  - Informationspflicht
  - Ausnahme: Weitergabe an Personen, welche dies tun
- Schutzrechte

Fragen und Diskussion

Beispiel für Halluzination mangels Trainingsdaten und Wissen/Logik («world model»)

Hier ist das Bild, das eine Großstadt mit Chalets in den Alpen bei Nacht zeigt, alle Lichter ausgeschaltet, und darüber ein Helikopter mit einem Suchscheinwerfer.