KI regulieren?
Mögliche Ansätze für die Schweiz

SCIENCE ET POLITIQUE

à table !

akademien der wissenschaften schweiz

Programm

- **Ist die KI von morgen zuverlässiger? Woran aktuell geforscht wird**
  Philippe Cudré-Mauroux, Professor für Informatik an der Universität Freiburg
- **Was bedeutet der «AI Act» der EU für die Schweiz?**
  Nadja Braun Binder, Professorin für Öffentliches Recht an der Universität Basel
- **Chancen von KI kontrolliert nutzen – wie kann dies gelingen?**
  Thomas Burri, Professor für Europarecht und Völkerrecht an der Universität St. Gallen
- **Diskussion**

UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Ist die KI von morgen zuverlässiger? Woran aktuell geforscht wird

Prof. Dr. Philippe Cudré-Mauroux

eXascale Infolab
Université de Fribourg

eXascale Infolab

Science et politique à table
03.12.2024

Was ist moderne KI?

Künstliche Intelligenz (KI)

Theorien und Techniken zur Herstellung von Maschinen, welche die menschliche Intelligenz simulieren können

Maschinelles Lernen (ML)

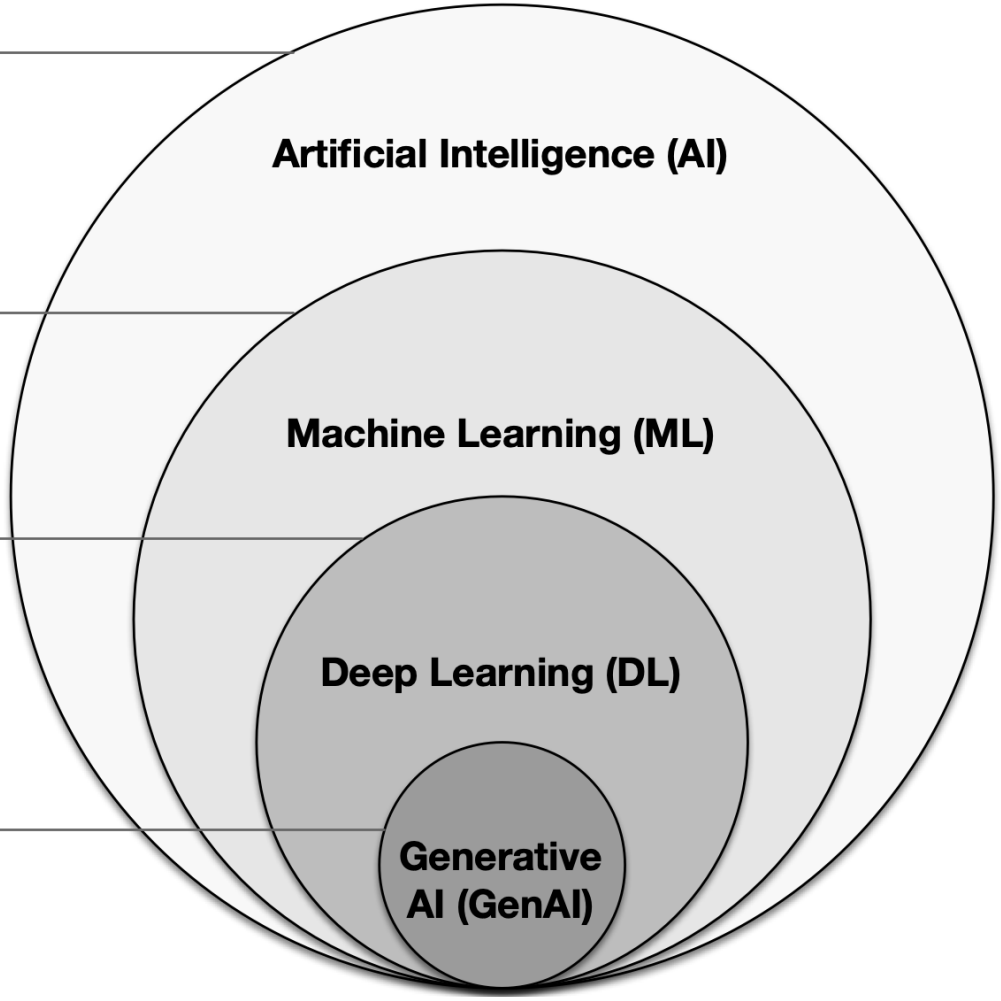
Statistische Methoden, die Modelle anhand von Daten trainieren

Deep Learning (DL)

Maschinelles Lernen mit mehrschichtigen künstlichen neuronalen Netzen, um komplexe Modelle aus Megadaten zu lernen

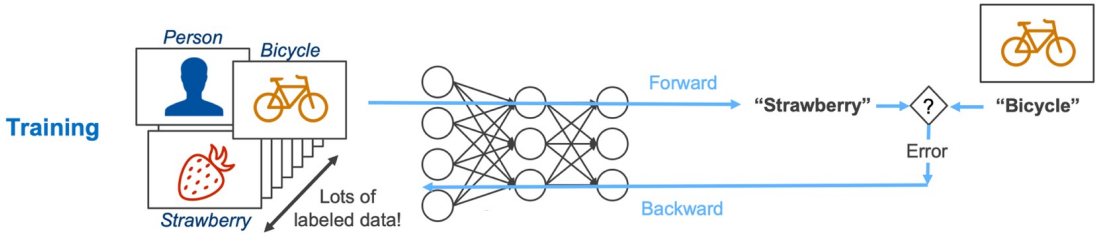
Generative KI

Teilmenge des Deep Learning, dessen Ziel es ist, neue Daten zu generieren

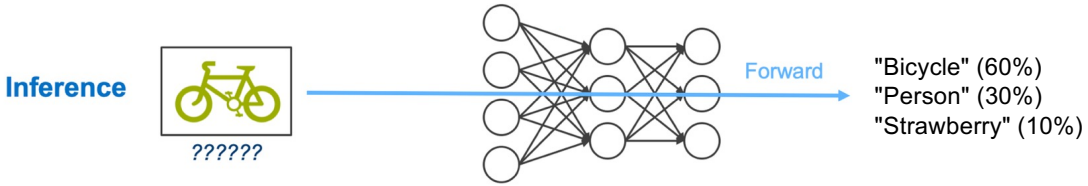


Training & Inferenz

Künstliche neuronale Netze werden im ersten Schritt an Millionen von Beispielen **trainiert**.



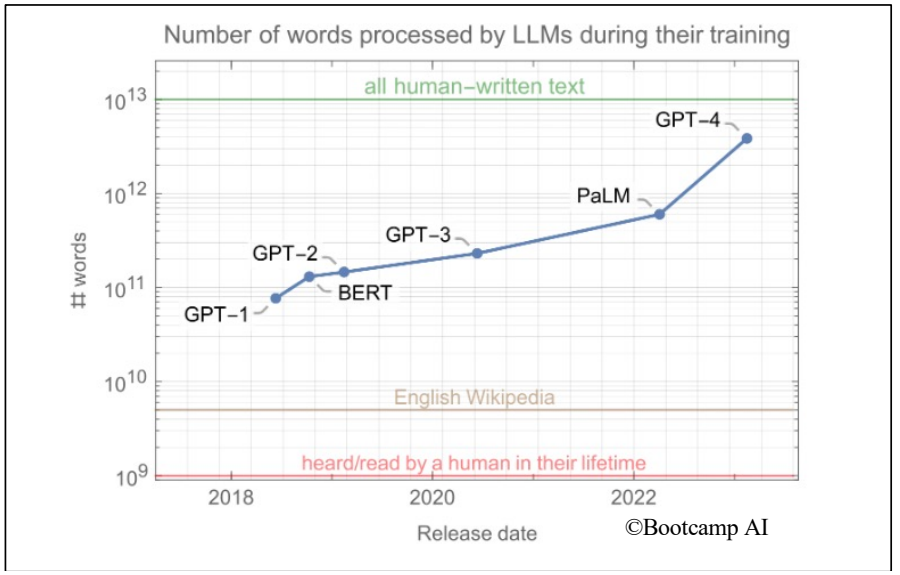
Sobald sie trainiert sind, werden sie verwendet, um Vorhersagen (**Inferenzen**) über neue Daten zu treffen.



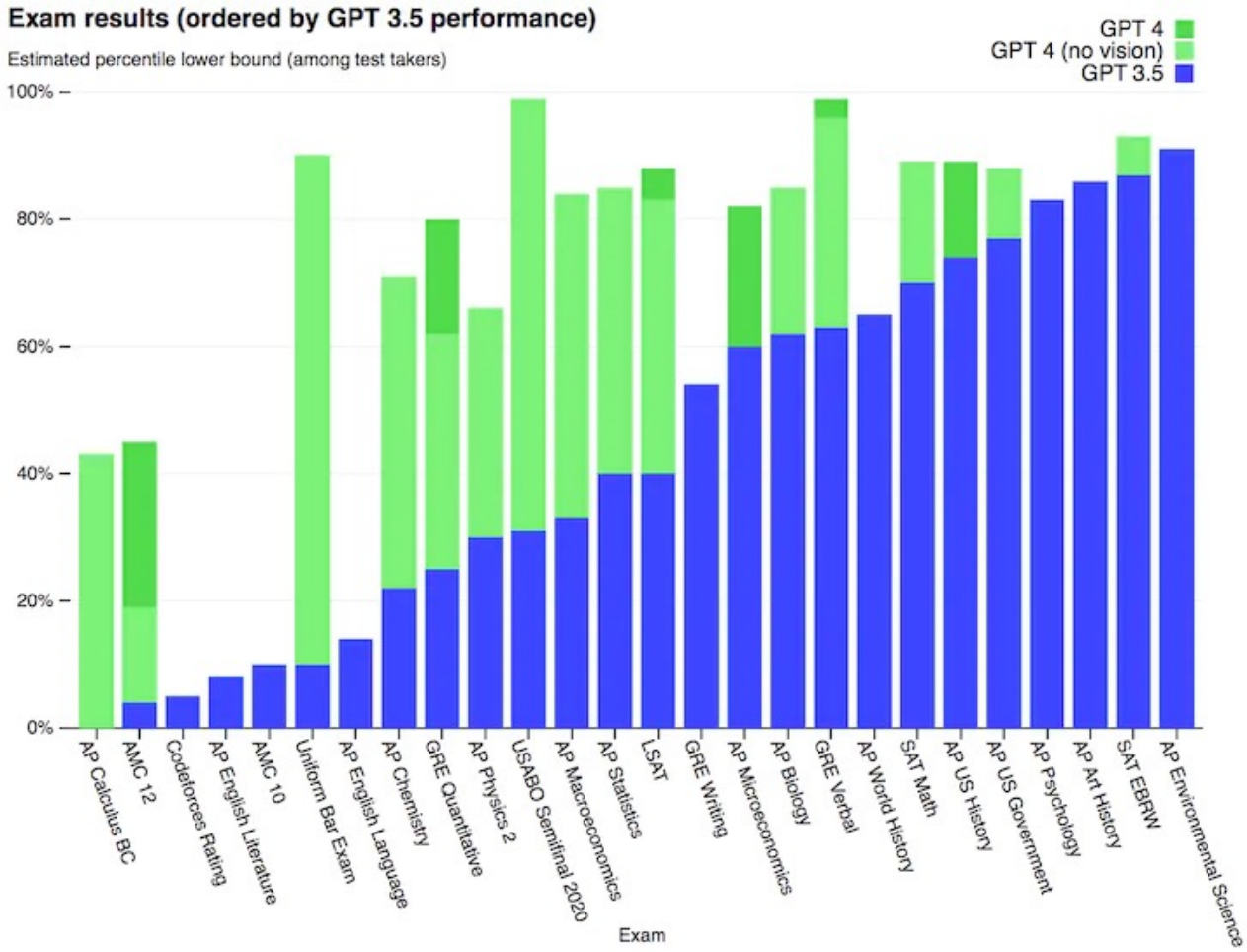
Foundation & Large Language Models

Die leistungsstärksten neueren Modelle (wie GPT-4)

- haben Hunderte von Millionen **Parametern**
- sind an riesigen **Datensätzen** trainiert
- anhand von Zehntausenden **Graphikprozessoren**
- für Kosten in Höhe von zweistelligen oder sogar dreistelligen **Millionenbeträgen**.



… mit spektakulären Ergebnissen



... und leider unzuverlässigen Ergebnissen

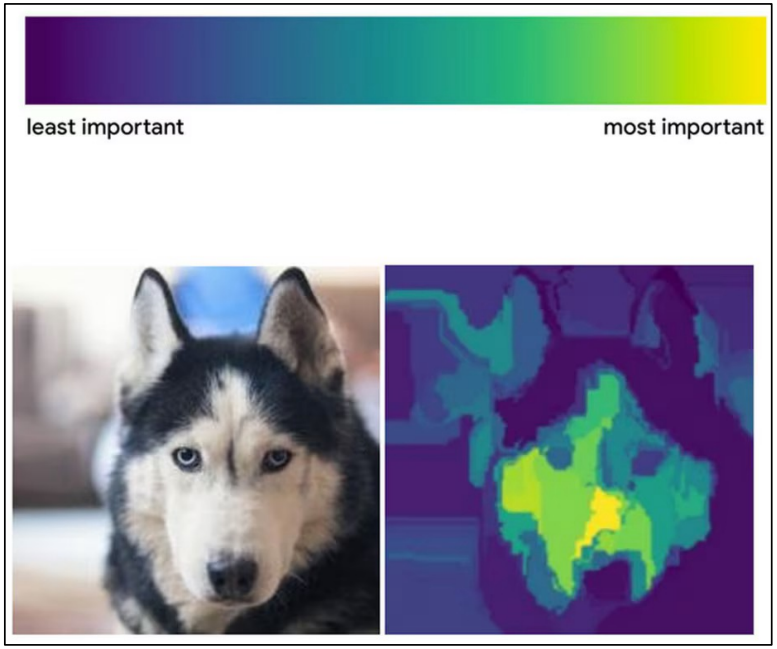
- **Unsicherheit**: Modelle liefern immer mehrere wahrscheinliche Ergebnisse
- **Overfitting**: Modelle produzieren zufällige Ergebnisse für neue Fälle (die während des Trainings nicht gesehen wurden)
- **Bias**: Modelle spiegeln die Verzerrung ihrer Daten wider
- **Halluzinationen**: GenAI-Modelle erzeugen neue Inhalte, die nicht unbedingt den Fakten entsprechen

Zuverlässige KI - möglich oder nicht?

- Es ist technisch schwierig (wenn nicht gar unmöglich), diese Probleme zu beseitigen
- Jedoch entwickelt die akademische Forschung seit einigen Jahren vielversprechende Methoden, um sie **zu erkennen**, **zu verringern** oder **auszugleichen**

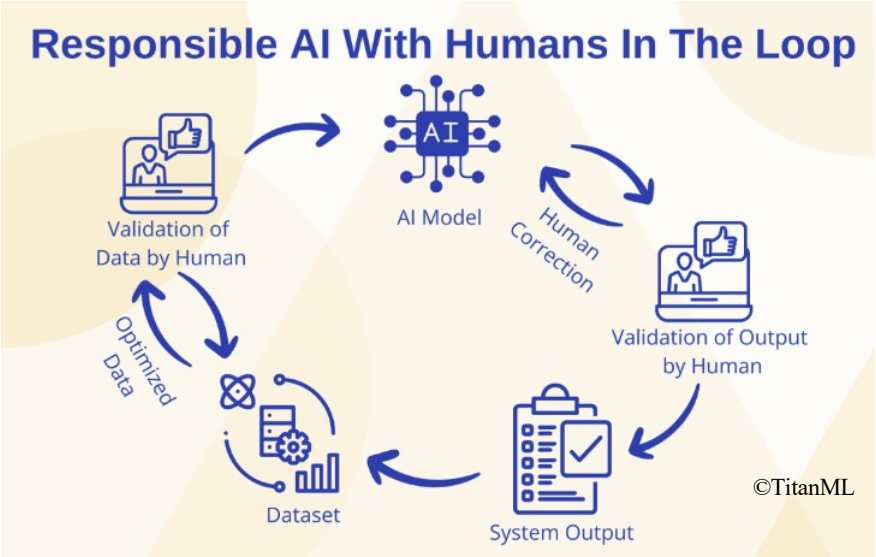
eXplainable AI (xAI)

- Die Methoden der erklärbaren KI werden verwendet, um ein KI-Modell, seine erwarteten Auswirkungen und seine möglichen Verzerrungen zu beschreiben



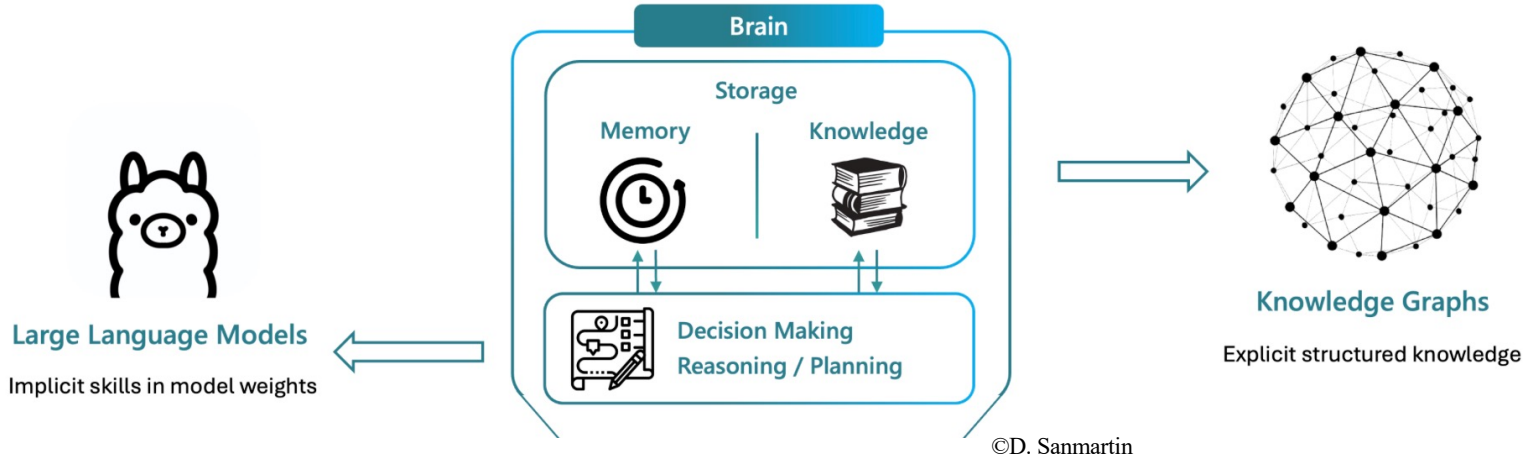
Human-in-the-Loop AI

- Integriert systematisch und kontinuierlich menschliches Feedback, um die Genauigkeit, Zuverlässigkeit und Anpassungsfähigkeit der Modelle zu verbessern



Neuro-Symbolic AI

- Kombiniert neuronale und symbolische KI-Architekturen (die z.B. logische Regeln oder Wissensgraphen verwenden), um eine robustere KI zu erstellen



Schlussfolgerung

- Moderne KI-Modelle sind **extrem breit gefächert** und **komplex**, sehr **leistungsfähig**, aber **unzuverlässig**
- Diese Unzuverlässigkeit ist den verwendeten Trainingsmethoden **immanent** und wird nicht verschwinden
- Jedoch entwickelt die akademische Forschung vielversprechende Techniken, um diese Probleme zu **erkennen**, **zu mindern** oder **auszugleichen**

Vielen Dank für Ihre Aufmerksamkeit!

https://exascale.info

Universität Basel

Was bedeutet der «AI Act» der EU für die Schweiz?

Prof. Dr. Nadja Braun Binder

03.12.2024

ePIAF
Electronic Public Institutions and Administrations Research Forum

1. Ausgangslage & Rahmenbedingungen

Aktuelles
Europäisches Parlament

Ursula von der Leyen stellt dem Plenum ihre Leitlinien vor

Pressemitteilung PLENARTAGUNG 16-07-2019 - 14:26

A Europe fit for the digital age

«**In my first 100 days in office, I will put forward legislation for a coordinated European approach on the human and ethical implications of Artificial Intelligence**. This should also look at how we can use big data for innovations that create wealth for our societies and our businesses.»

https://www.europarl.europa.eu/news/de/press-room/20190711IPR56823/ursula-von-der-leyen-stellt-dem-plenum-ihre-leitlinien-vor

1. Ausgangslage & Rahmenbedingungen

Elemente der rechtliche Erfassung von KI in der EU:

- **Zugang zu hochwertigen Daten** als wesentlicher Faktor von leistungsstarken, robusten KI-Systemen
  - Data Act
  - Data Governance Act
- **Vertrauen in KI**
  - Europäischer Rechtsrahmen für KI
  - Rahmen für die zivilrechtliche Haftung
  - Überarbeitung sektoraler Sicherheitsvorschriften
- **Bestehender** Rechtsrahmen
  - z.B. Datenschutzgrundverordnung (DSGVO)

1. Ausgangslage & Rahmenbedingungen

Ziele der KI-Verordnung (→ Art. 1 KI-VO)

- Verbesserung der Funktionsweise des europäischen **Binnenmarktes**
- Förderung von **menschenzentrierter** und **vertrauenswürdiger** KI
- Hohes Schutzniveau für **Gesundheit**, **Sicherheit** und **Grundrechte** – inkl. Demokratie, Rechtsstaatlichkeit & Umweltschutz – gegenüber schädlichen Auswirkungen (→ Produktsicherheit)

2. Zentrale Bestimmungen der KI-Verordnung

Definition KI-System

Art. 3 Begriffsbestimmungen

«Für die Zwecke dieser Verordnung bezeichnet der Ausdruck

1. „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;»

2. Zentrale Bestimmungen der KI-Verordnung

Persönlicher Anwendungsbereich der KI-Verordnung (→ Art. 2)

- Anbieter, Betreiber, Einführer, Händler, Produkthersteller (die KI-Systeme zusammen mit ihrem Produkt unter ihrem eigenen Namen oder ihrer Handelsmarke in Verkehr bringen oder in Betrieb nehmen)
  - mit Sitz in der Union oder wenn sie sich in der Union befinden
  - Anbieter und Betreiber, die ihren Sitz in einem **Drittland** haben oder sich in einem Drittland befinden, **wenn die vom KI-System hervorgebrachte Ausgabe in der Union verwendet wird**
- Betroffene Personen, die sich in der Union befinden

2. Zentrale Bestimmungen der KI-Verordnung

Kategorisierung von KI-Systemen

- KI mit spezifischem Verwendungszweck (single-purpose AI)

→ Einteilung nach Risiko ihrer **Anwendung**:

  - **Verbotene** Praktiken (Art. 5)
  - **Hochrisiko-KI-Systeme** (Art. 6 ff.)
  - KI-Systeme mit **begrenztem** Risiko: Transparenzpflichten (Art. 50)
  - KI-Systeme mit **minimalem** Risiko: keine Vorgaben

2. Zentrale Bestimmungen der KI-Verordnung

Kategorisierung von KI-Systemen

- KI mit allgemeinem Verwendungszweck (general purpose AI, GPAI) & Basismodelle

→ Einstufung nach **Leistung & Reichweite** des Basismodells:

  - GPAI: Zusätzliche Transparenzpflichten
  - GPAI mit **systemischem Risiko** (Art. 51): Zusätzliche Transparenzpflichten & weitere Pflichten (Überwachung, Modellbewertung, Angriffstests)

3. Beispiel aus einer Bestimmung aus der KI-Verordnung

Art. 10 KI-VO: Daten und Daten-Governance

[1] (...)

[2] Für Trainings-, Validierungs- und Testdatensätze gelten Daten-Governance- und Datenverwaltungsverfahren, die für die Zweckbestimmung des Hochrisiko-KI-Systems **geeignet** sind. (…)

[3] Die Trainings-, Validierungs- und Testdatensätze müssen im Hinblick auf die Zweckbestimmung **relevant**, hinreichend **repräsentativ** und so weit wie möglich **fehlerfrei** und **vollständig** sein. (…)

[4] Die Datensätze müssen, soweit dies für die Zweckbestimmung erforderlich ist, die entsprechenden Merkmale oder Elemente berücksichtigen, die für die besonderen geografischen, kontextuellen, verhaltensbezogenen oder funktionalen Rahmenbedingungen, unter denen das Hochrisiko-KI-System bestimmungsgemäss verwendet werden soll, **typisch** sind.

(…)

4. CH: Autonomer Nachvollzug AI Act?

- EU hat einen anderen Fokus (**Binnenmarktharmonisierung**) als die Schweiz
- Normen des AI Act sind sehr **vage** – eine Konkretisierung wäre noch notwendig
- Anerkennung von **Konformitätsbewertungsverfahren** alles andere als gesichert
- Gefahr von **Doppelungen** mit bereits bestehenden, sektorspezifischen Regelungen für Produktsicherheit
- Konkretisierung in technischen Normen, die **de facto** auch in der Schweiz relevant sein werden
- Regulierung stärkt die Marktposition von «**Big Tech**»

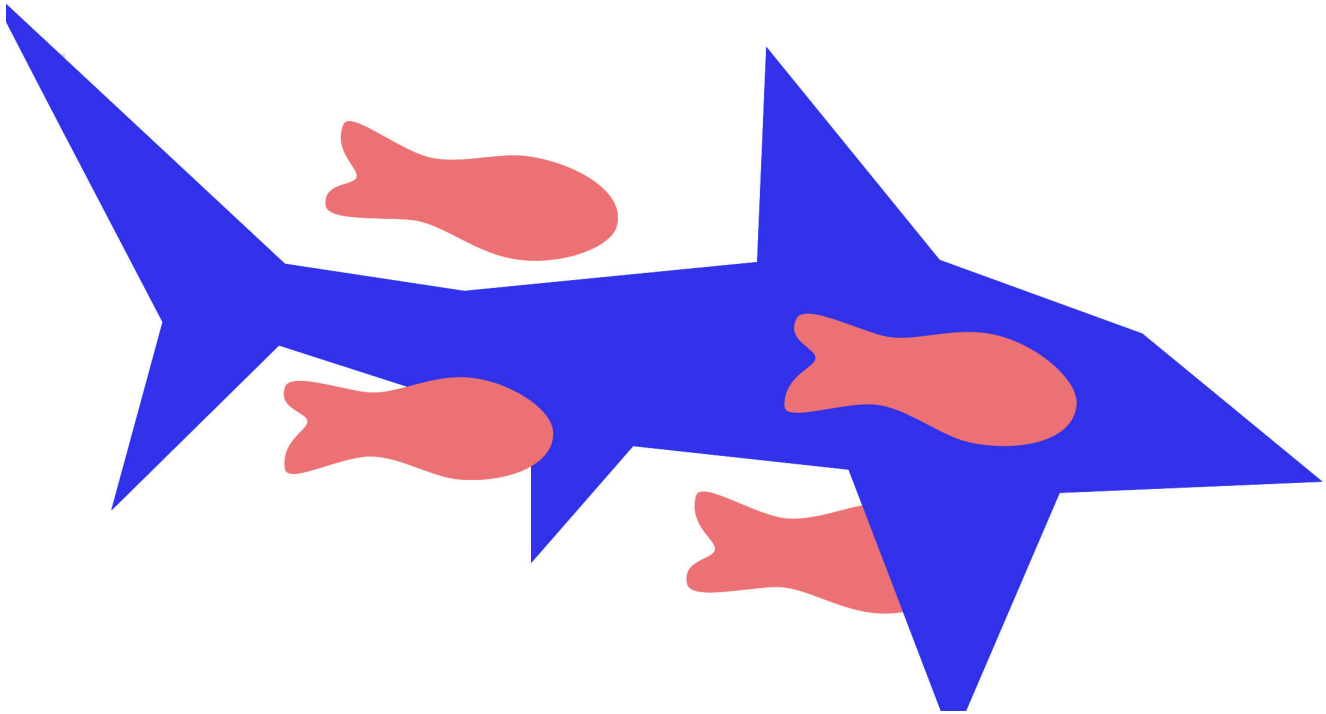
Universität Basel

Vielen Dank
für Ihre Aufmerksamkeit.

ePIAF
Electronic Public Institutions and Administrations Research Forum

Chancen von KI kontrolliert nutzen – wie kann das gelingen?

Prof. Dr. Thomas Burri, Universität St. Gallen

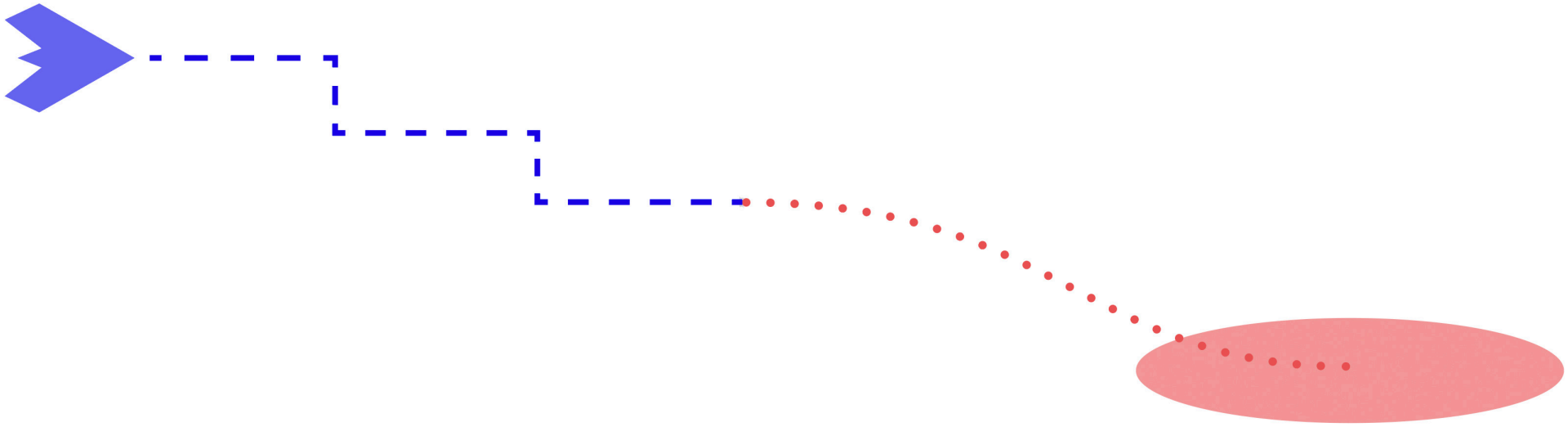


Thomas Burri

Menschliche Kontrolle und menschliche Aufsicht

Heisse und kalte Künstliche Intelligenz

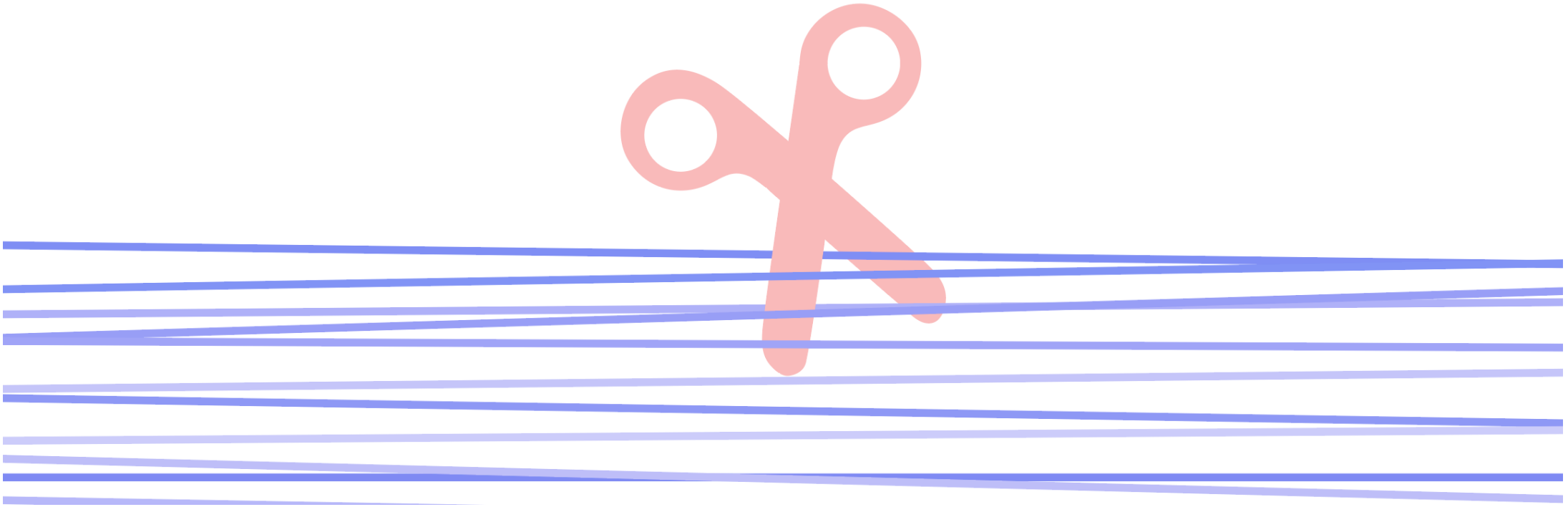
Landung mit Hilfe von Künstlicher Intelligenz



Thomas Burri

8 Denkanstösse

1. Aufsichtspflichten über KI anwendungsbezogen konkretisieren
2. Menschlicher Eingriff in heisse KI kein Allheilmittel
3. Keine Erklärbarkeit moderner KI annehmen



Thomas Burri

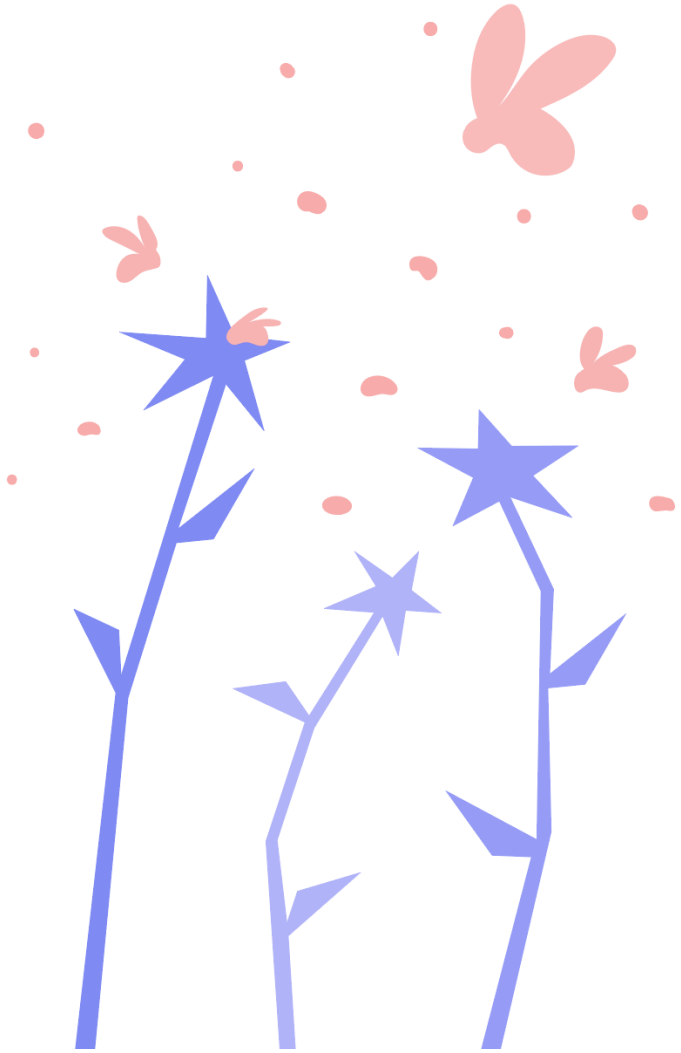
4. Umfassende KI-Regulierung wirkt sich stark auf Innovation aus

5. Bestehendes Recht erfasst KI zu erheblichen Teilen

6. Soziale Auswirkungen v. KI separat von Betriebsrisiken angehen

7. Vielversprechendster Ansatz: Allg. Grundprinzipien mit spezialgesetzl. Anpassungen kombinieren

8. EU-Verordnung zu KI vorerst nicht übernehmen



Danke!

Prof. Dr. Thomas Burri, Universität St. Gallen

SNF Projekt Nr. 407740_187494
ArmaSuisse W+T Verträge 8003540020, 8003538711, 8003535283

Weitere Fachleute (neben Referent:innen)

Alice Delorme Benites,
Institut für Übersetzen und Dolmetschen, ZHAW

Roger Abächerli,
Departement für Gesundheitswissenschaften und Technologie der ETH Zürich; SATW

Emanuela Keller,
Professorin und leitende Ärztin am Universitätsspital Zürich

Markus Christen, Digital Society Initiative der Universität Zürich, TA-Swiss

akademien der wissenschaften schweiz

KI regulieren?
Mögliche Ansätze für die Schweiz

SCIENCE ET POLITIQUE

à table !

akademien der wissenschaften schweiz