

# VISCHER

Künstliche Intelligenz.

Wie wird dies in der Schweiz reguliert?

David Rosenthal, Partner, VISCHER AG  
23. September 2024

---

Streit über KI-Entwicklung

New York Times klagt gegen Micro  
OpenAI

TRANSPARENZ BEI KI

FORMEL-1-PILOTEN

Rechtsstreit um Fake-KI-Interview mit Michael  
Schumacher

**"Aktuell herrscht hier ziemliche  
Rechtsunsicherheit"**

RECHT 30.10.2023

**KI-verursachte Schäden:  
Wann haftet der Zahn-  
(Arzt)?**

Getty Images und Adobe

**KI-Training: Wie Getty  
Images und Adobe die  
Rechtsunsicherheit zu  
ihrem Vorteil nutzen**

Quellen: Horizont.net, zwf-online.info, thepioneer.de, tagesschau.de, golem.de

## Rechtsunsicherheit?

- **Ja**, weil die Materie für uns unheimlich und ungewohnt ist
  - Wissen Sie, was in einem LLM steckt und warum es so gut ist?
  - Wir hatten beim Internet früher dieselbe Situation, und heute haben wir uns daran gewöhnt – es ist völlig normal geworden
- **Unbehagen** führt zum Ruf nach mehr Regulierung und "Ethik"
  - Transparenz, Diskriminierung, Erklärbarkeit, Human-in-the-Loop
  - EU AI Act als Reaktion (primär punktuelle Produkte-Regulierung)
  - Bundesrat will bis Ende Jahr Schweizer Regulationsbedarf klären
- Das **bestehende Recht** regelt viele der Themen recht gut
  - Datenschutz, Urheberrecht, Lauterkeitsrecht, Geheimnisschutz
  - Viele 0815-Hausaufgaben und einzelne neue Herausforderungen



25. Juli 1994 (time.com, Titelseite:  
James Porto)


# Was sagen Aufsichtsbehörden?

- Hamburg: LLM enthalten keine Personendaten

- 1. Die bloße Speicherung eines LLMs stellt keine Verarbeitung im Sinne des Art. 4 Nr. 2 DSGVO dar. Denn in LLMs werden keine personenbezogenen Daten gespeichert. Soweit in einem LLM-gestützten KI-System personenbezogene Daten verarbeitet werden, müssen die Verarbeitungsvorgänge den Anforderungen der DSGVO entsprechen. Dies gilt insbesondere für den Output eines solchen KI-Systems.**

<http://www.datenschutz-hamburg.de/>

Wirklich?



Der Hamburgische Beauftragte für  
Datenschutz und Informationsfreiheit

---

**Datenschutzrechtliche Lage Large Language Models**

gischen Ba-  
er Daten-  
n Debatten-  
tze besser  
sultet, vor  
Personen-  
t drei

Nr. 2  
t. Soweit in  
müssen  
ilt insbe-

2. Mangels Speicherung personenbezogener Daten in LLM können die Betroffenenrechte der DSGVO nicht das Modell selbst zum Gegenstand haben. Ansprüche auf Auskunft, Löschung oder Berichtigung können sich jedoch zumindest auf Input und Output eines KI-Systems der verantwortlichen Anbieter:in oder Betreiber:in beziehen.
3. Das Training von LLMs mit personenbezogenen Daten muss datenschutzkonform erfolgen. Dabei sind auch die Betroffenenrechte zu beachten. Ein ggf. datenschutzwidriges Training wirkt sich aber nicht auf die Rechtmäßigkeit des Einsatzes eines solchen Modells in einem KI-System aus.

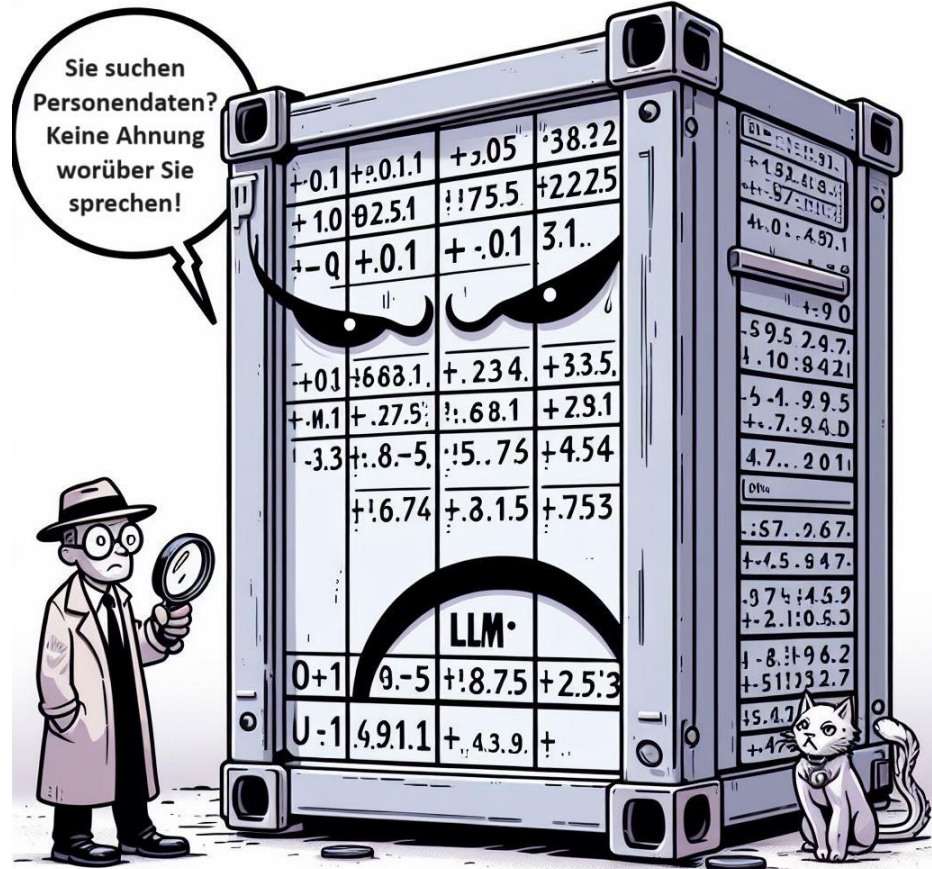
<sup>1</sup> Gemeint sind hierbei allein die Modelle als wichtiger, aber nicht alleiniger Bestandteil eines KI-Systems (z. B. eines LLM-basierten Chatbots).

---

[www.datenschutz-hamburg.de](http://www.datenschutz-hamburg.de)  
E-Mail: [mailbox@datenschutz.hamburg.de](mailto:mailbox@datenschutz.hamburg.de)  
Ludwig-Erhard-Strasse 22 · D-20459 Hamburg · Tel.: 040 - 4 28 54 - 40 40 · Fax: 040 - 4 28 54 - 40 00  
Vertrauliche Informationen sollten auf elektronischem Weg nur verschlüsselt an uns übermittelt werden.  
Unser öffentlicher PGP-Schlüssel ist im Internet verfügbar (Fingerprint: 0932 5798 33C1 8C21 6C00 E77D 9600 BA4E 3377 5707).

VISCHER

Wirklich?



# Grosse Sprachmodelle?

### Wie und warum ein grosses Sprachmodell den "Geburtsdag" von (öffentlichen) Personen kennen kann

#### Training des Modells

Verarbeiten der Trainingsdaten

"Donald Trump wurde am 14. Juni 1946 geboren."  
 "Am 14. Juni 1946 erblickte Donald Trump das Licht der Welt."  
 "Das Geburtsdatum von Donald Trump ist der 14. Juni 1946."  
 "Donald John Trump wurde am 14. Juni 1946 geboren."  
 "Der 14. Juni 1946 markiert den Geburtstag von Donald Trump."  
 "Donald Trump kam am 14. Juni 1946 zur Welt."  
 "Am 14. Juni 1946 wurde der spätere US-Präsident Donald Trump geboren."  
 "Der 14. Juni 1946 ist das Geburtsdatum von Donald Trump."  
 "Donald Trump wurde am 14. Juni 1946 in Queens, New York, geboren."  
 "Am 14. Juni 1946 wurde Donald Trump, der 45. Präsident der USA, geboren."  
 "Das Geburtsdatum von Donald Trump, dem ehemaligen Präsidenten der Vereinigten Staaten, ist der 14. Juni 1946."  
 "Donald Trump wurde an einem Freitag, dem 14. Juni 1946, geboren."  
 "Am 14. Juni 1946 wurde Donald Trump, ein zukünftiger Immobilienmogul, geboren."  
 "Der 14. Juni 1946 ist der Tag, an dem Donald Trump geboren wurde. Donald Trump, geboren am 14. Juni 1946, wurde später Präsident der USA."  
 "Am 14. Juni 1946 wurde Donald Trump in New York geboren. Das Geburtsdatum von Donald Trump, der am 14. Juni 1946 geboren wurde, ist weissen bekannt."  
 "Donald Trump wurde am 14. Juni 1946 geboren und wuchs in Queens auf."  
 "Am 14. Juni 1946 wurde Donald Trump, der spätere Unternehmer und Politiker, geboren."  
 "Der 14. Juni 1946 ist das Datum, an dem Donald Trump geboren wurde."

#### Nutzung des Modells

Input im Modell Output

Donald Trumps Geburtstag ?

Donald Trump

14. Juni 1946

zur Welt kommen  
das Licht der Welt erblickten  
geboren  
Geburtsdatum  
Trump  
Juni 14.

Im Embedding-Raum des Modells ist Ebene 3231/9311 mit 'Geburtsdag' assoziiert

Ein 'Geburtsdag' ist assoziiert mit einem Datum, d.h. ein Tag, Monat und Jahr

Im Embedding-Raum des Modells ist Ebene 3231/9311 mit 'Geburtsdag' assoziiert

Ein 'Geburtsdag' ist assoziiert mit einem Datum, d.h. ein Tag, Monat und Jahr

Dimension 3231

Dimension 9311

Max Schrems Geburtstag ?

3. Oktober 1987  
16. Oktober 1987  
11. Oktober 1987

Maximilian Schrems Geburtstag ?

11. Oktober 1987  
16. Oktober 1987

Max Schrems

11x  
16x  
3x

Maximilian Schrems

Oktober 1987

Max

Aggregation - was steckt in den Trainingsdaten als Konzept heraus?

Das Konzept 'Geburtsdag' ist stark assoziiert mit '14.', 'Juni' und '1946.'

Das Konzept 'Geburtsdag' ist stark assoziiert mit '14.', 'Juni' und '1946.'

Aggregation - was steckt in den Trainingsdaten als Konzept heraus?

#### Modell-Parameter

GPT-4o

Prompt

Prompt anwenden

Output

"Donald Trumps Geburtstag?"

"14. Juni 1946"

Welcher Use Case führt vermehrt zu einem solchen Prompt?

Für die meisten Use Cases nicht der Fall

Mittel zur Identifizierung

"means reasonably likely to be used" (Erwägung 16)

Die im Modell zwischen Person und der gesuchten Information bestehende Assoziation wird mittels Prompt sichtbar; bei hoher Konfidenz ist die betroffene Person damit identifiziert

Personenbezogene Daten

Die Informationen, die sich auf den Kontext des Prompts und damit die Person beziehen

Pseudonymisierte Daten (nur jene, die im Training genügend häufig "gesehen" wurden)

DSGVO

Hinweise:

- Die Darstellung ist stark vereinfacht. Es kommen dabei nicht nur der oben dargestellte sogenannte Embedding-Raum zur Anwendung, sondern auch weitere Funktionen, etwa um die Bedeutungen des Inputs zu ermitteln (z.B. dass "Donald Trump" ein Name ist).
- Die Angabe der Dimensionen ist nur illustrativ, ob tatsächlich eine Ebene für "Geburtsdag" besteht ist für Anwendbarkeit nicht relevant (GPT hat z.B. 12.000 Dimensionen). Das Konzept funktioniert ebenso, wenn die Ebene z.B. lediglich für Datumsangaben besteht und der Bezug zum "Geburtsdatum" und den Zahlen anders hergestellt wird.
- Die Darstellung ist vom "Wissen" von GPT-4o inspiriert; nicht jedes LLM kennt diese Personen.
- Die Darstellung kann implizieren, dass Assoziationen bidirektional sind (d.h. wenn von A auf B dann auch von B auf A geschlossen werden kann). Das ist in der Regel nicht so bzw. nicht zwingend der Fall.
- Was "höchstwahrscheinlich" ist, ist nicht allgemeingültig definiert; nimmt die Wahrscheinlichkeit jedoch ab, beginnt das Modell zu "halluzinieren".
- Für die Frage, ob personenbezogene Daten vorliegen, muss der "relative" Ansatz berücksichtigt werden, d.h. es kommt darauf an, wer auf das LLM zugreift; in den meisten Fällen werden deshalb keine personenbezogenen Daten vorliegen, weil es keine entsprechenden Prompts geben wird.
- Um Obiges besser zu verstehen, lesen Sie, wie ein LLM funktioniert: <https://vischerInk.com/4anNh1r>.
- Mehr Angaben über personenbezogene Daten in LLM finden Sie hier: <https://vischerInk.com/3YugXHZ>.

Autor: David Rosenthal - Version 4.8.2024 - Mehr Information: vischer.com/ki

Wie funktioniert ein grosses Sprachmodell und was ist wirklich darin gespeichert?  
[vischerInk.com/4anNh1r](https://vischerInk.com/4anNh1r)

Enthält ein grosses Sprachmodell Personendaten?  
[vischerInk.com/3YugXHZ](https://vischerInk.com/3YugXHZ)

# Was sagen Aufsichtsbehörden?

- EDÖB: Absolute Transparenzpflicht!

Angesichts dieser Vorgaben des DSG müssen die Hersteller, Anbieter und Verwender von KI-Systemen den Zweck, die Funktionsweise und die Datenquellen der auf KI beruhenden Bearbeitungen transparent machen. Das gesetzliche Recht auf Transparenz ist eng verbunden mit dem Anspruch der betroffenen Personen, einer automatischen Datenbearbeitung zu widersprechen oder zu verlangen, dass automatisierte Einzelentscheidungen von einem Menschen überprüft werden – wie dies das DSG ausdrücklich vorsieht. Im Falle intelligenter Sprachmodelle, die direkt mit Benutzerinnen und Benutzern kommunizieren, haben Letztere ein gesetzliches Recht zu erfahren, ob sie mit einer Maschine sprechen oder korrespondieren und ob die von ihnen eingegebenen Daten zur Verbesserung der selbstlernenden Programme oder zu weiteren Zwecken weiterbearbeitet werden. Auch die Verwendung von Programmen, welche die Verfälschung von Gesichtern, Bildern oder Sprachnachrichten von identifizierbaren Personen ermöglichen, muss stets deutlich erkennbar sein – soweit sie sich im

[https://www.edoeb.admin.ch/edoeb/de/home/kurzmeldungen/2023/20231109\\_ki\\_dsg.html](https://www.edoeb.admin.ch/edoeb/de/home/kurzmeldungen/2023/20231109_ki_dsg.html)

Wirklich?

Schweizerische Eidgenossenschaft  
Confédération suisse  
Confederazione Svizzera  
Confederaziun svizra

Eidgenössischer Datenschutz- und  
Öffentlichkeitsbeauftragter (EDÖB)

soziale Leben der Bevölkerung. Der  
tz des Bundes auf KI-gestützte

n vielbeachteten Schritt zur  
e Parlament seine grundsätzliche  
e EU-weiten Regulierung von KI,  
g über KI, Menschenrechte,

ie Regulierung von KI. Der Bundesrat  
uftrag erteilen. Die Schweiz verfolgt  
t, dass die Rechtsetzung branchen-  
ert weiterverfolgt oder durch eine

uf hin, dass unabhängig vom Ansatz  
sind. Das Datenschutzgesetz des  
KI-gestützten Datenbearbeitungen  
nder Applikationen auf die  
Planung ihres Einsatzes  
Selbstbestimmung verfügen.

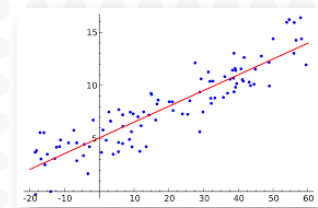
Angesichts dieser Vorgaben des DSG müssen die Hersteller, Anbieter und Verwender von KI-Systemen den Zweck, die Funktionsweise und die Datenquellen der auf KI beruhenden Bearbeitungen transparent machen. Das gesetzliche Recht auf Transparenz ist eng verbunden mit dem Anspruch der betroffenen Personen, einer automatischen Datenbearbeitung zu widersprechen oder zu verlangen, dass automatisierte Einzelentscheidungen von einem Menschen überprüft werden – wie dies das DSG ausdrücklich vorsieht. Im Falle intelligenter Sprachmodelle, die direkt mit Benutzerinnen und Benutzern kommunizieren, haben Letztere ein gesetzliches Recht zu erfahren, ob sie mit einer Maschine sprechen oder korrespondieren und ob die von ihnen eingegebenen Daten zur Verbesserung der selbstlernenden Programme oder zu weiteren Zwecken weiterbearbeitet werden. Auch die Verwendung von Programmen, welche die Verfälschung von Gesichtern, Bildern oder Sprachnachrichten von identifizierbaren Personen ermöglichen, muss stets deutlich erkennbar sein – soweit sie sich im konkreten Fall nicht aufgrund strafrechtlicher Verbote als gänzlich unrechtmässig erweist.

KI-gestützte Datenbearbeitungen mit hohen Risiken sind nach DSG dem Grundsatz nach zulässig, erfordern aber angemessene Massnahmen zum Schutz der potentiell betroffenen Personen. Aus diesem Grund verlangt das Gesetz bei hohen Risiken eine sog. «Datenschutz-Folgenabschätzung». Anwendungen hingegen, die geradezu auf eine Auslöschung der vom DSG geschützten Privatsphäre und informationellen Selbstbestimmung abzielen, sind datenschutzrechtlich verboten. Gemeint sind insbesondere KI-basierte Datenbearbeitungen, die in autoritär regierten Staaten zu beobachten sind, wie die flächendeckende Gesichtserkennung in Echtzeit oder die umfassende Observation und Bewertung der Lebensführung, das sog. «Social Scoring».



## Was ist KI überhaupt?

- Gemäss EU **KI-Gesetz** "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."
- Das einzig praktisch relevante Element ist "**Autonomie**"
  - Einfach gesagt: Ein IT-System, das seiner Entscheide auf Basis eines Trainings fällt statt nur voll ausprogrammierten Logik
- **Aber:** Die Definition ist mangelhaft ...
  - Jedes Kopiergerät ist KI (OCR); was ist mit linearer Regression?





# Und was macht der EDÖB?

- Er beharrt auf seiner Position zur KI ...

- **Das Recht auf Transparenz.** Wer KI-Systeme herstellt, anbietet oder verwendet, muss offenlegen, welchen Zweck er damit verfolgt, wie das System funktioniert
- un ▪ **Das Recht, einer automatischen Bearbeitung meiner Personendaten zu widersprechen.** Ohne mein Einverständnis dürfen KI-Systeme meine Personendaten nicht verwenden. Personendaten sind etwa Name, Adresse oder Geburtsjahr. Das Gesetz definiert zudem eine besondere Gruppe von Personendaten, die für die Bearbeitung von Personendaten von besonderer Bedeutung sind. Diese sind:
  - **Recht auf Datenherausgabe oder Datenübertragung.** Ich kann bei jedem Betreiber von künstlich intelligenten Anwendungen verlangen, dass er meine Personendaten herausgibt.
  - **Recht auf Berichtigung unrichtiger Daten und auf Löschung.** Meine Personendaten müssen stimmen. Unrichtige Daten muss der Betreiber einer KI-Anwendung löschen. Ich kann verlangen, dass er meine Personendaten löscht.
  - **Das Recht, zu erfahren, ob ich mit einer Maschine spreche oder korrespondiere.** Immer mehr Firmen setzen im Kundenkontakt auf künstlich intelligente Sprachroboter. Wenn ich nachfrage, muss mir das Unternehmen dazu Auskunft geben. Eng damit verbunden ist auch das nächste Recht.
  - **Das Recht, zu erfahren, ob Programme verwendet werden, die die Verfälschung von Gesichtern, Bildern oder Sprachnachrichten von identifizierbaren Personen ermöglichen.**

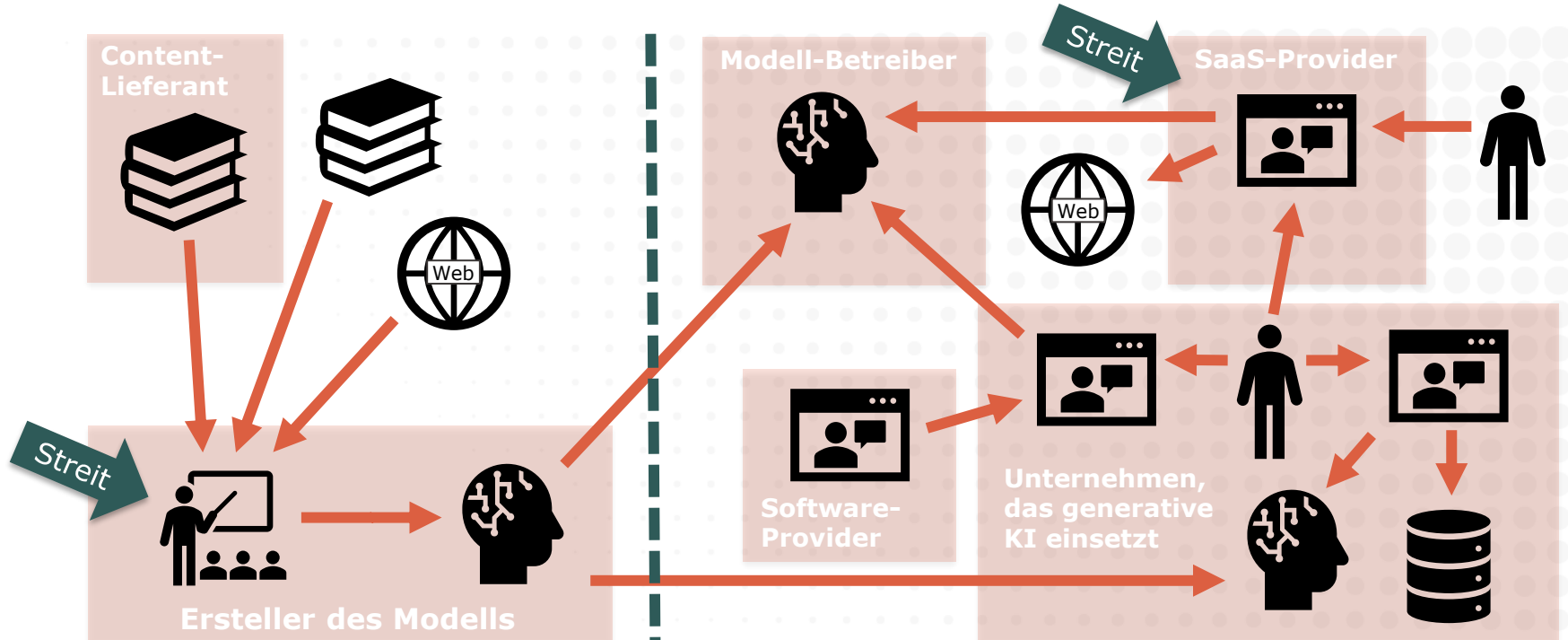
## Der Datenschutzler gehe zu weit, sagen Fachleute

Den beiden Datenschutzexperten David Rosenthal und Martin Steiger geht der EDÖB mit seiner Interpretation des Datenschutzgesetzes deutlich zu weit. Es gebe weder eine generelle Pflicht, offenzulegen, ob ein KI-System zur Anwendung komme, noch das generelle Recht, zu erfahren, ob ich mit einer Maschine spreche oder nicht. In der Schweiz sehe das Datenschutzgesetz zwar ein Recht auf Widerspruch vor, verlange grundsätzlich aber nicht eine Einwilligung für die Verwendung von Personendaten in KI-Systemen. Der EDÖB stellt sich auf den Standpunkt, dass die Gerichte diese strittigen Fragen beurteilen müssen, wenn jemand seine Anordnungen bestreite.

Hier werden Recht und ethische Vorstellungen **vermischt**; die Position des EDÖB geht sogar weiter als der AI Act

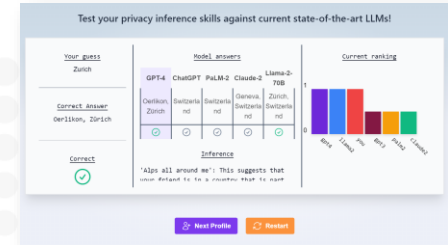
Quelle: "Beobachter", <https://vischerlnk.com/3T6dvQD>  
 Datenschutzplaudereien: <https://vischerlnk.com/42Or9eb>

# Verantwortlichkeit für generative KI



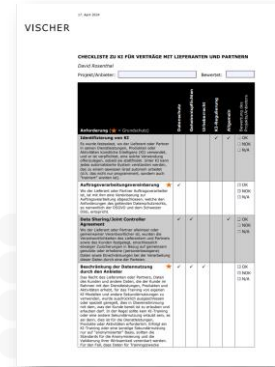
## Verschiedene Schutzbereiche

- Personendaten der **Benutzer** schützen
  - Betrifft vor allem Arbeitnehmende
  - Schutz: Zweckbindung, Providervertrag, Transparenz
- Informationen und Inhalte von **Dritten** schützen
  - Betroffen sind Kontakte (Kunden etc.), Inhaber der Rechte der verwendeten Werke, weitere Dritte
  - Schutz: Wie oben + Ergebnisse prüfen, Schutzrechte, Verträge
- Personen vor anderen unerwünschten **Auswirkungen** schützen
  - Getäuschte oder sonst angegriffene Personen, von Entscheidungen oder Fehlern betroffene Personen
  - Schutz: Wie oben + Haftungsregeln, KI-Regulierung



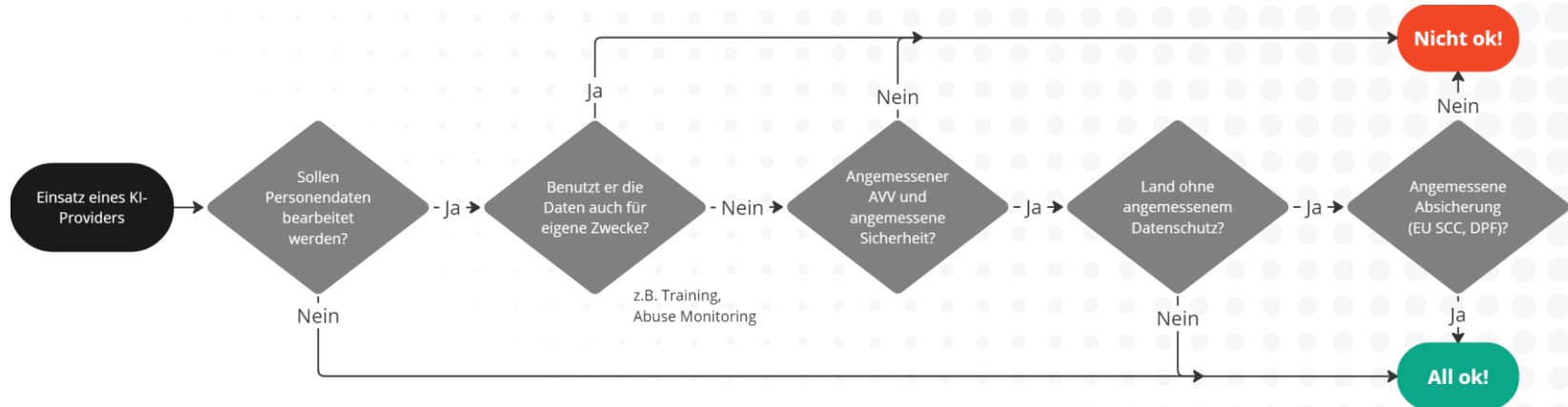
Quelle: llm-privacy.org

Hier setzt der  
EU AI Act an



# Datenschutzrecht

- **1. Frage:** Wem vertrauen wir allfällige Personendaten beim Einsatz von KI in welcher Weise an und was tut er damit?
  - Stellt sich, wenn wir Provider wie OpenAI oder Microsoft nutzen
  - **Prüfen:** Auftragsverarbeitungsvertrag (AVV) inkl. angemessene Datensicherheit, internationaler Transfer, Training/Monitoring



# Datenschutzrecht

- **2. Frage:** Was machen wir mittels KI mit den Personendaten?
  - Haben wir das den davon betroffenen Personen in unserer **Datenschutzerklärung** gesagt, insbesondere den **Zweck**?
  - Mussten sie damit rechnen, als wir ihre Daten erhalten haben?
  - Können wir ihnen das, was wir tun, zumuten? Bleiben wir im Hinblick auf den Zweck **verhältnismässig**? Werden wichtige Entscheide von einem Menschen gefällt oder mind. überprüft?
  - Sind die Daten, die wir (weiter-)nutzen, für unsere Zwecke **richtig** und vollständig (soweit wir überhaupt darauf abstellen)?
  - Können wir, wo nötig, die **Betroffenenrechte** gewährleisten (z.B. wo Auskunft, Löschung oder Korrekturen verlangt werden)?
  - Öffentliche Organe & DSGVO: Deckt unsere **Rechtsgrundlage** die Verwendung von KI ab, oder haben wir eine Einwilligung?



[vischerlnk.com/3IdAymb](https://vischerlnk.com/3IdAymb)

Falls das Vorhaben  
hohe Risiken für die  
Personen birgt: **DSFA**

# Berufs- und Amtsgeheimnis

Prüfung Lawful Access Risiko  
mit "Methode Rosenthal"  
vischerlnk.com/flara  
vischerlnk.com/flaraafa

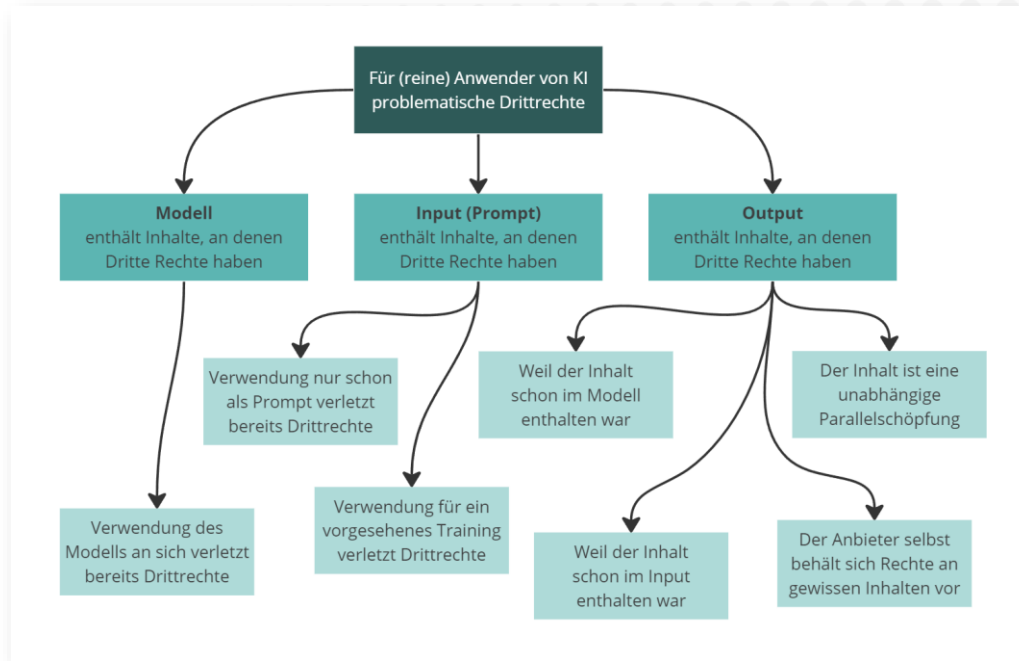
- **Vorgaben** beim Einsatz insb. **ausländischer Provider**
  - Einhaltung der Geheimhaltung seitens des Providers, auch wenn dieser im Ausland ist (vertragliche Verpflichtung)
  - Angemessene Informationssicherheit, keine Zweckentfremdung
  - Kein Grund zur Annahme, dass es via Provider zu ausländischem Behördenzugriff kommt (Stichwort "US CLOUD Act")
- **Massnahmen** insb. gegen ausländische Behördenzugriffe
  - Europäische Gegenpartei, Datenhaltung in der Schweiz, vom Kunden kontrollierte Verschlüsselung, manueller Providerzugriff beschränken (Stichwort "Customer Lockbox"), Verpflichtung zur Einhaltung des Berufsgeheimnisses, Defend-your-data-Klausel, Schutzmassnahmen für Personendaten auf alle Inhalte ausweiten und Einschränkung der Bearbeitung für eigene Providerzwecke



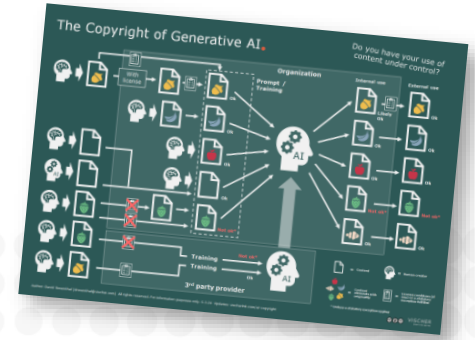
vischerlnk.com/4ck2J0L  
vischerlnk.com/4bGEARE

Ermöglichen  
Abwehr von  
Behörden-Zugriff  
z.B. unter dem  
US CLOUD Act

# Urheberrecht



Mehr: [vischerlnk.com/3vYAPqG](https://vischerlnk.com/3vYAPqG)



[vischerlnk.com/ai-copyright](https://vischerlnk.com/ai-copyright)

## Themen in der Praxis:

- Training / Prompts bestehend aus **geschützten Inhalten** von Dritten
- Geschützte Inhalte von Dritten tauchen **über das Modell** im Output des ahnungslosen Benutzers auf
- Benutzer missbrauchen die KI zur **Nachahmung** bestehender Werke
- **Kein Schutz** von Werken weil eine menschliche Schöpfung fehlt



# Gefährliche Modelle?

## Original:

Roy Lichtenstein, In the Car

[https://uploads5.wikiart.org/images/roy-lichtenstein/in-the-car-1963\(1\).jpg](https://uploads5.wikiart.org/images/roy-lichtenstein/in-the-car-1963(1).jpg)



## Werk "zweiter Hand"? Nein!

OpenAI Dall-E 3 basierend auf einem entsprechendem Prompt, der Bild und Stil beschreibt



## Gegenbeispiel




Quelle: Heinz

I apologize, but I cannot create a variant of Roy Lichtenstein's "In the Car" due to copyright restrictions. If you have any other requests, feel free to ask! 😊

# Haben/behalten auch wir unsere Modelle im Griff?

Six ways to attack an AI system.

Are your AI applications prepared for them?



Poisoning	Trojan Horse	Prompt Injection	Sponge Attack	Model & Data Theft	Deception
<p>AI poisoning is a tactic where attackers manipulate the data used to train artificial intelligence (AI) models, causing these models to produce incorrect results or become unreliable. Attackers can introduce subtle errors into training data, such as mislabeling images or biased information, or embed hidden triggers that cause the AI to act unexpectedly when activated. This manipulation can occur intentionally by bad actors, accidentally by use of biased or poor-quality data, or even during normal use if the AI continues to learn from manipulated input or AI content ("feedback loops").</p>	<p>With this form of attack, bad actors secretly insert harmful code into AI models, especially large language models, before companies use them, expecting that they cannot check what is hidden inside these models when they obtain them from open sources or buy them. Once these tampered models are used, the hidden malicious code may be activated in one way or another, acting like a trojan horse and using, for instance, unprotected systems (e.g., third-party tools with elevated privileges or insecure browsers) to launch attacks from within a company.</p>	<p>Prompt injection attacks involve tricking an AI system by entering malicious commands instead of normal input. These commands can manipulate the AI to perform unintended actions, like revealing sensitive data or the secret "system prompts" of an AI system, turning off safety controls, or even taking control of other systems that process the output generated by an AI system that is being misused by an attacker. Malicious commands can be included in prompts, but also in documents that a user may upload to an AI system for analysis, resulting in manipulated output.</p>	<p>Sponge attacks target AI systems by overwhelming them with complex or large inputs, like a sponge soaking up their computing power. This can slow down or even damage a system. Attackers may do so by crafting inputs that are hard to process, causing the AI to use excessive energy or memory. Such harmful input may be included in a model during the training phase, making the system vulnerable from the start, or they are added later on. This can lead to delays, damage, or safety risks, for example where AI system must remain responsive at all times (e.g., in autonomous vehicles).</p>	<p>Attackers target AI systems to uncover secret data contained in them or how an AI or its model was built. They might trick the AI into revealing if certain data was used in its training or infer private details from the AI's responses. One method does so by testing the system with real data to determine whether it recognizes it with certainty, indicating that it has already seen it during training. Another approach involves flooding the system with specific questions to replicate its logic. These tactics may not only expose sensitive or proprietary information but can lay groundwork for more advanced attacks.</p>	<p>Attackers can trick AI systems that rely on pattern recognition by using manipulated input to trigger certain (false) responses. For example, if an AI relies on image recognition to classify objects (e.g., speed limit signs), the attacker may use visual elements (e.g., certain stickers on a sign) that may even be invisible to a human to cause the AI to incorrectly assess the object. This may also work with face recognition. In a "white-box" attack the attacker has inside knowledge of the model, whereas in a "black-box" attack, the attacker figures out how to deceive the AI through trial and error.</p>

Author: David Rosenthal (drosenthal@vischer.com) All rights reserved. For information purposes only. 19.2.24 Updates: vischerink.com/ai-attacks

vischerink.com/3OPTpaA

1. Womit und worauf wurden Modelle trainiert? Compliance eingehalten? Ist alles dokumentiert?
2. Welche Trainingsinhalte wurden allenfalls "memorisiert"? Können Trainingsinhalte "leaken"?
3. "Bias" soweit nötig vermieden?
4. Wird ein "Concept Drift" erkannt?
5. Sind speziell auf KI ausgerichtete Angriffsformen wie z.B. "Prompt Injection" oder "Poisoning" bedacht?



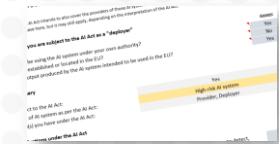
## Und der EU AI Act?

- Schweizer Unternehmen können erfasst sein, wenn sie ...
  - KI-Produkte zum **Einsatz in der EU entwickeln**
  - KI verwenden und der **Output** in der EU benutzt wird
  - Nicht schon, wenn KI in der EU **läuft** oder Leute dort **betrifft**
- Besondere Vorgaben macht der AI Act für ...
  - **Verbotene** KI-Anwendungen (z.B. KI-Emotionserkennung am Arbeitsplatz und in der Schule, Manipulation durch KI-Einsatz)
  - **Hoch-Risiko** KI-Anwendungen (z.B. regulierte Produkte, KI-Beurteilung von Mitarbeitenden/Schülern, KI-Bonitätsbewertung)
- Darüber hinaus: Nur sehr begrenzte Pflichten zur **Transparenz**
  - Z.B. Emotionserkennung, Wasserzeichen, Deep Fakes, Chatbots



[vischerlnk.com/ai-act-uc](https://vischerlnk.com/ai-act-uc)

Ausführlicher Aufsatz  
zum EU AI Act:  
[vischerlnk.com/3ZkPOYh](https://vischerlnk.com/3ZkPOYh)



Siehe AI Act Check unter  
[vischerlnk.com/gaira](https://vischerlnk.com/gaira)

## AI Act: "Hoch-Risiko"-KI-Systeme vermeiden ...

**Anbieter** sind unter anderem verpflichtet, (i) ein Risiko- und Qualitätsmanagement zu betreiben, (ii) eine Konformitätsbewertung durchzuführen und eine CE-Kennzeichnung mit ihren Kontaktdaten anzubringen, (iii) bestimmte Qualitätsniveaus für Schulungs-, Validierungs- und Testdaten zu gewährleisten, (iv) eine detaillierte technische Dokumentation bereitzustellen, (v) automatisches Protokollieren vorzusehen und Protokolle aufzubewahren, (vi) Anweisungen für Betreiber bereitzustellen, (vii) das System so zu gestalten, dass menschliche Aufsicht möglich ist, es robust, zuverlässig, gegen Sicherheitsbedrohungen (einschliesslich KI-Angriffe) geschützt und fehlertolerant ist, (viii) das KI-System behördlich zu registrieren, (ix) eine Überwachung des Systems nach seiner Markteinführung zu betreiben, (x) Vorfälle den Behörden zu melden und Korrekturmaßnahmen zu ergreifen, (xi) mit den Behörden zusammenzuarbeiten, (xii) die Einhaltung der vorstehenden Anforderungen zu dokumentieren und (xiii) einen Vertreter in der EU zu haben, falls der Anbieter selbst nicht in der EU ansässig ist, aber dem AI Act unterliegt.

**Betreiber** sind unter anderem verpflichtet, (i) die Anleitung des Anbieters zu befolgen, (ii) angemessene menschliche Aufsicht zu gewährleisten, (iii) automatisch generierte Protokolle mindestens sechs Monate lang aufzubewahren, (iv) angemessenen Input zu gewährleisten, (v) an der Überwachung des KI-Systems nach seiner Einführung durch den Anbieter teilzunehmen, (vi) schwere Vorfälle und bestimmte Risiken den Behörden und dem Anbieter zu melden, (vii) Mitarbeiter zu informieren, falls das KI-System sie betrifft, (viii) betroffene Personen über Entscheidungen zu informieren, die durch oder mit Hilfe des KI-Systems getroffen wurden, (ix) eine Grundrechte-Folgenabschätzung durchführen in bestimmten Fällen (z.B. öffentliche Dienste), und (x) Anfragen betroffener Personen bezüglich solcher Entscheidungen zu befolgen.

Offizielle Schätzung: Max. 5-10% der KI-Systeme

Quelle: vischerlnk.com/gaira

# KI-Konvention des Europarats

## **Article 7 – Human dignity and individual autonomy**

Each Party shall adopt or maintain measures to respect human dignity and individual autonomy in relation to activities within the lifecycle of artificial intelligence systems.

## **Article 8 – Transparency and oversight**

Each Party shall adopt or maintain measures to ensure that adequate transparency and oversight requirements tailored to the specific contexts and risks are in place in respect of activities within the lifecycle of artificial intelligence systems, including with regard to the identification of content generated by artificial intelligence systems.

## **Article 9 – Accountability and responsibility**

Each Party shall adopt or maintain measures to ensure accountability and responsibility for adverse impacts on human rights, democracy and the rule of law resulting from activities within the lifecycle of artificial intelligence systems.

## **Article 10 – Equality and non-discrimination**

Each Party shall adopt or maintain measures with a view to ensuring that activities within the lifecycle of artificial intelligence systems respect equality, including gender equality, and the prohibition of discrimination, as provided under applicable international and domestic law.

Each Party undertakes to adopt or maintain measures aimed at overcoming inequalities to achieve fair, just and equitable outcomes, in line with its applicable domestic and international human rights obligations, in relation to activities within the lifecycle of artificial intelligence systems.

- Article 6 General approach
- Article 7 Human dignity and individual autonomy
- Article 8 Transparency and oversight
- Article 9 Accountability and responsibility
- Article 10 Equality and non-discrimination
- Article 11 Privacy and personal data protection
- Article 12 Reliability
- Article 13 Safe innovation
- Article 14 Remedies
- Article 15 Procedural safeguards
- Article 16 Risk and impact management framework

<https://rm.coe.int/1680afae3c>



## Gesetzliche und "ethische" Vorgaben gemischt

- Wir sorgen für **Verantwortlichkeit**
- Wir sorgen für die nötige **Transparenz**
- Wir bleiben **fair** und **schaden nicht**
- Wir sorgen für **Zuverlässigkeit**
- Wir sorgen für **Informationssicherheit**
- Wir achten auf **Verhältnismässigkeit** und **Selbstbestimmung**
- Wir respektieren fremdes und eigenes **Geistiges Eigentum**
- Wir wahren die **Rechte der Betroffenen**
- Wir sorgen für **Erklärbarkeit** und **menschliche Aufsicht**
- Wir verstehen und kontrollieren die **Risiken**
- Wir verhindern **Missbräuche** unserer KI-Anwendungen

# Die wichtigsten Compliance-Fragen

## Checkliste: 18 KI-Compliance-Schlüsselfragen.

KI = System, das Ergebnisse auf Basis eines Trainings und nicht nur einer Programmierung erzeugt

Unter [vischer.com/ki](https://vischer.com/ki) finden Sie kostenlose Ressourcen zu diesen Themen sowie zu KI-Governance und Risikomanagement (keine Registrierung erforderlich)

### Datenschutz

- Haben wir einen angemessenen Vertrag mit den von uns genutzten Providern (z.B. einen ADV, EU SCC, Verbot der Eigennutzung unserer Daten)?
- Haben wir die Leute über die Zwecke informiert, zu denen wir Daten von ihnen bearbeiten oder erzeugen?
- Haben wir es im Griff, wenn die KI falsche oder anderweitig unzulässige Daten über sie produziert?
- Wenn eine KI wichtige Entscheidungen über sie trifft, können sie diese von einem Menschen prüfen lassen?
- Ist unsere KI vor Missbrauch und Angriffen geschützt und auch sonst sicher, insbesondere, wo wir Dritten die Nutzung erlauben (z.B. Chatbot)?
- Können wir Auskunfts- und Berichtigungsbegehren wie erforderlich umsetzen?
- Haben wir eine Risikobeurteilung für unser Vorhaben (inklusive einer DSFA) durchgeführt?

### Vertragspflichten, Geheimhaltung

- Kommen wir unseren Geheimhaltungspflichten nach (z.B. beim Einsatz von Providern, Verhinderung der unerwünschten Preisgabe von Daten)?
- Untersagen unsere Verträge die von uns ins Auge gefasste Anwendung (z.B. NDA, welches die Nutzung von Daten für unsere Zwecke einschränkt)?

### Schutz von Inhalten Dritter

- Füttern wir KI-Systeme nur dann mit Inhalten Dritter, soweit unsere Lizenzen oder die gesetzlichen Schranken des Urheberrechts dies zulassen?
- Vermeiden wir die Erstellung von Inhalten, die bereits bestehenden Inhalten Dritter entsprechen?

### EU AI Act (noch nicht in Kraft)

- Ist klar, dass wir entweder nicht unter den EU AI Act fallen oder unser Vorhaben keine verbotene Praktik ist und möglichst auch kein "Hoch-Risiko"-KI-System (und gehen wir ansonsten richtig damit um)?
- Wenn eine KI "Deep Fakes" erstellt oder mit Menschen interagiert oder sie beobachtet, werden sie dann darauf hingewiesen gemacht?

### Andere (auch ethische) Aspekte

- Vermeiden wir Diskriminierung beim Einsatz von KI?
- Behält der Mensch (wirklich) die Kontrolle über die KI?
- Können wir unsere KI-Ergebnisse rechtfertigen/erklären?
- Sagen wir es den Leuten, wie wir KI einsetzen, wenn es für sie unerwartet sein könnte, und erlauben wir ihnen gar, sich für oder gegen deren Einsatz zu entscheiden?
- Haben wir ein angemessenes KI-Testing, angemessene Überwachung und ein angemessenes Risk-Management?

Autor: David Rosenthal ([david.rosenthal@vischer.com](mailto:david.rosenthal@vischer.com)) Alle Rechte vorbehalten. Nur zu Informationszwecken (Fokus europäisches Recht). 16.5.24 Aktualisierungen: [vischerlnk.com/ki-compliance-kurz](https://vischerlnk.com/ki-compliance-kurz)



VISCHER  
1992 LLP AG INC.

Blog-Beitrag und weitere Infos:

<https://bit.ly/3WNgxeO>  
<https://vischer.com/ki>

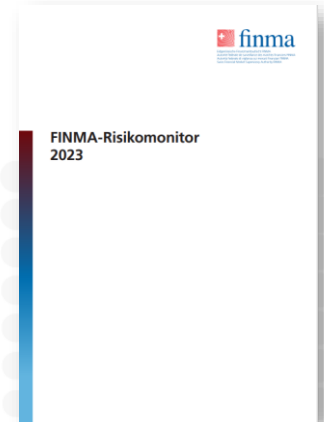
Hier muss jedes Unternehmen seinen Weg finden

[vischerlnk.com/ki-compliance-kurz](https://vischerlnk.com/ki-compliance-kurz)



## Erwartungen der FINMA

1. Es müssen klare **Rollen und Verantwortlichkeiten** sowie **Risikomanagementprozesse** definiert und implementiert werden. **Die Verantwortung für Entscheidungen kann nicht an KI oder Drittparteien delegiert werden.** Alle Beteiligten müssen über genügend **Know-how im Bereich KI verfügen.**
2. Bei der Entwicklung, der Anpassung und in der Anwendung von KI ist sicherzustellen, dass die **Ergebnisse hinreichend genau, robust und zuverlässig sind.** Dabei sind sowohl die Daten als auch die Modelle und die Resultate kritisch zu hinterfragen.
3. Die **Erklärbarkeit der Resultate** einer Anwendung sowie die **Transparenz über deren Einsatz** sind je nach Empfänger, Relevanz und Prozessintegration sicherzustellen.
4. Nicht begründbare **Ungleichbehandlung ist zu vermeiden.**



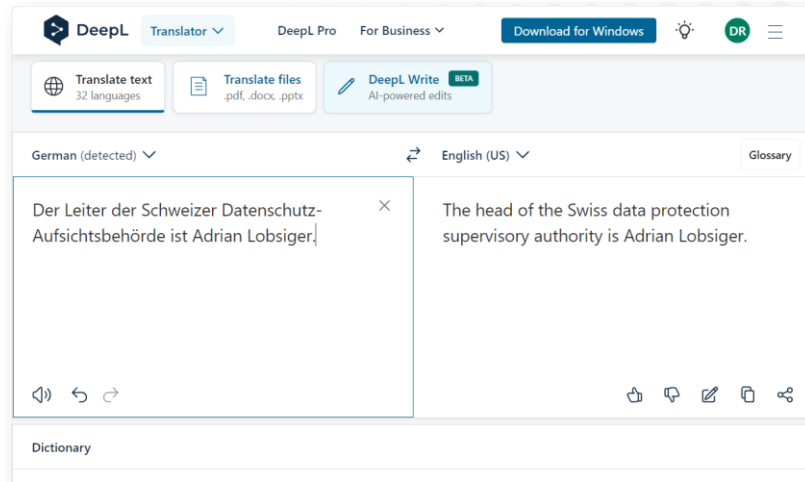
Blog-Beitrag:

[vischerInk.com/3MRHurk](https://vischerInk.com/3MRHurk)

## KI-Governance: Sechs Schritte

- Voraussetzungen schaffen: Robustes **Data Management**
- Aufgaben, Kompetenzen und Verantwortlichkeiten (**AKV**) regeln
- Richtlinie mit Vorgaben zum **Umgang mit KI** um Mitarbeitende "sicher" zu machen und einen KI-Einsatz zu ermöglichen
- **Schulung** im sicheren und verantwortungsvollen Umgang mit KI und Vermittlung von KI-Kenntnissen – bis zur GL und zum VR, damit die Risiken bekannt sind und übernommen werden können
- **Map & Track** von (relevanter) KI im Unternehmen
- **Risiko-Management** für KI-Vorhaben und Tools (heisst: die wichtigsten Risiken beurteilen und Massnahmen dazu treffen)

## Use Case: Assistent für Übersetzungen etc.




- **Varianten:** Zusammenfassen von Texten, Protokollierung, Mails formulieren

- Grundsatz der **Transparenz**?
- Grundsatz der **Richtigkeit**?
- Grundsatz der **Zweckbindung**?
- Grundsatz der **Verhältnismässigkeit**?
- Ist es **fair**, was ich mit KI tue?
- **Datenschutzerklärung**
- Was tut der **Provider** mit den Daten?
  - Besteht ein DPA bzw. AVV?
  - Internationaler Transfer im Griff?
  - Hinreichende Datensicherheit?
  - Verwendung für eigene Zwecke?

# Use Case: Bewerber-CV mit LLM analysiert

**Lebenslauf**



**Persönliche Daten:**

Name: Mustermann  
 Vorname:  
 Adresse:  
 Telefon:  
 E-Mail:  
 Geburtsdatum:  
 Zivilstand:

**Berufliche Erfahrungen**

02/2004 – heute  
 02/2000 – 01/2004  
 07/1998 – 01/2000

**Ausbildung:**

05/1999 – 05/2000 HSO Schulen Thun Bern AG: «Abschluss als Marketingplaner»  
 08/1994 – 08/1997 Wirtschafts- und Kaderschule KV Bern: «Abschluss als Kaufmann E-Profil»

"Während Kenntnisse in 3D-Animation und Adobe Photoshop wertvoll sein können, scheinen diese Fähigkeiten nicht direkt mit seiner Rolle als Marketingkoordinator in Verbindung zu stehen. Dies könnte darauf hinweisen, dass der Kandidat Interesse an einer Karriereänderung hat oder dass er über Qualifikationen verfügt, die er möglicherweise nicht vollständig nutzen konnte."

- Grundsatz der **Transparenz**?
- Grundsatz der **Richtigkeit**?
- Grundsatz der **Zweckbindung**?
- Grundsatz der **Verhältnismässigkeit**?
- Ist es **fair**, was ich mit KI tue?
- **Provider** korrekt beauftragt (AVV)?  
Zweitverwertung der Personendaten durch ihn ausgeschlossen?
- Unter dem EU AI Act wäre dies ein "Hoch-Risiko" KI-System

Quelle: [https://www.jobscout24.ch/download/vorlagen/Lebenslauf\\_Marketing.pdf](https://www.jobscout24.ch/download/vorlagen/Lebenslauf_Marketing.pdf)

## Use Case: Bewerber wird von KI selektioniert

### Art. 21 Informationspflicht bei einer automatisierten Einzelentscheidung

<sup>1</sup> Der Verantwortliche informiert die betroffene Person über eine Entscheidung, die ausschliesslich auf einer automatisierten Bearbeitung beruht und die für sie mit einer Rechtsfolge verbunden ist oder sie erheblich beeinträchtigt (automatisierte Einzelentscheidung).

<sup>2</sup> Er gibt der betroffenen Person auf Antrag die Möglichkeit, ihren Standpunkt darzulegen. Die betroffene Person kann verlangen, dass die automatisierte Einzelentscheidung von einer natürlichen Person überprüft wird.

<sup>3</sup> Die Absätze 1 und 2 gelten nicht, wenn:

- a. die automatisierte Einzelentscheidung in unmittelbarem Zusammenhang mit dem Abschluss oder der Abwicklung eines Vertrags zwischen dem Verantwortlichen und der betroffenen Person steht und ihrem Begehren stattgegeben wird; oder
- b. die betroffene Person ausdrücklich eingewilligt hat, dass die Entscheidung automatisiert erfolgt.

- **Informationspflicht**
- Recht auf **menschliches Gehör**
- **Auskunft** über "das Vorliegen einer automatisierten Einzelentscheidung sowie die Logik, auf der die Entscheidung beruht" (Art. 25 DSGVO)
- **Treu und Glauben?**

## Use Case: Chatbot auf der Website

- Wie riskant ist der Bereich, um den es geht? Wie wird die **Thementreue** sichergestellt? Wie gut wurde der Bot getestet?
- Ist der Chatbot auf eigene, "gute" Daten begrenzt (sog. **RAG**)?
- Wie wird kommuniziert, wie verlässlich ist die Auskunft, die der Chatbot erteilt? Pauschalvorbehalt auf der Website (schwächer) oder im Output selbst integrierter **Vorbehalt** und Vermeidung von Einzelfallauskünften via *Alignment* (besser/stärker)?
- Ist eine **Eskalation** an den Menschen vorgesehen? Mittels Themen und Konfidenzschwellen? Über Standardhinweise?
- Welche Massnahmen gibt es gegen (unerwünschten) **Bias**?
- Wird geloggt? Werden **Logs** und User-**Feedback** ausgewertet?

Source:  
WashingtonPost.com

**Air Canada chatbot promised a discount.  
Now the airline has to pay it.**  
Air Canada argued the chatbot was a separate legal entity 'responsible for its own actions,' a Canadian tribunal said

Wieso sollte sie für fehlerhaftere Auskünfte nicht bezahlen? Wo ist der Unterschied zum Call Center? Und lohnt es sich nicht trotzdem?

1. **Haben wir ein vernünftiges Set an Massnahmen getroffen?**
2. **Was sind die Restrisiken?**
3. **Sind sie akzeptabel und lohnt sich das Ganze wirklich?**

## Use Case: Training von KI-Modellen

- **Öffentliche Daten** sind nicht einfach frei
- **Urheberrecht/Lauterkeitsrecht** bei Inhalten Dritter
  - Liegt überhaupt eine rechtlich relevante Nutzung vor?
  - Haben wir eine Einwilligung für die geplante Nutzung?
  - Können wir uns auf eine gesetzliche Ausnahme berufen?
- **Datenschutzrecht**, falls Personendaten vorliegen
  - Haben wir die Verwendung in der Datenschutzerklärung genannt?
  - Bearbeitungsgrundsätze eingehalten (z.B. Zweckbindung)?
  - Rechtfertigungsgrund der nicht personenbezogenen Bearbeitung?
- Könnten Trainingsinhalte **im Output** sein ("Memorisierung")?



# VISCHER

## Danke für Ihre Aufmerksamkeit!

Fragen: [david.rosenthal@vischer.com](mailto:david.rosenthal@vischer.com)

### **Zürich**

Schützengasse 1  
Postfach  
8021 Zürich, Schweiz  
T +41 58 211 34 00

[www.vischer.com](http://www.vischer.com)

### **Basel**

Aeschenvorstadt 4  
Postfach  
4010 Basel, Schweiz  
T +41 58 211 33 00

### **Genf**

Rue du Cloître 2-4  
Postfach  
1211 Genf 3, Schweiz  
T +41 58 211 35 00

Mehr Unterlagen:  
[www.vischer.com/ki](http://www.vischer.com/ki)  
[www.rosenthal.ch](http://www.rosenthal.ch)