

VISCHER

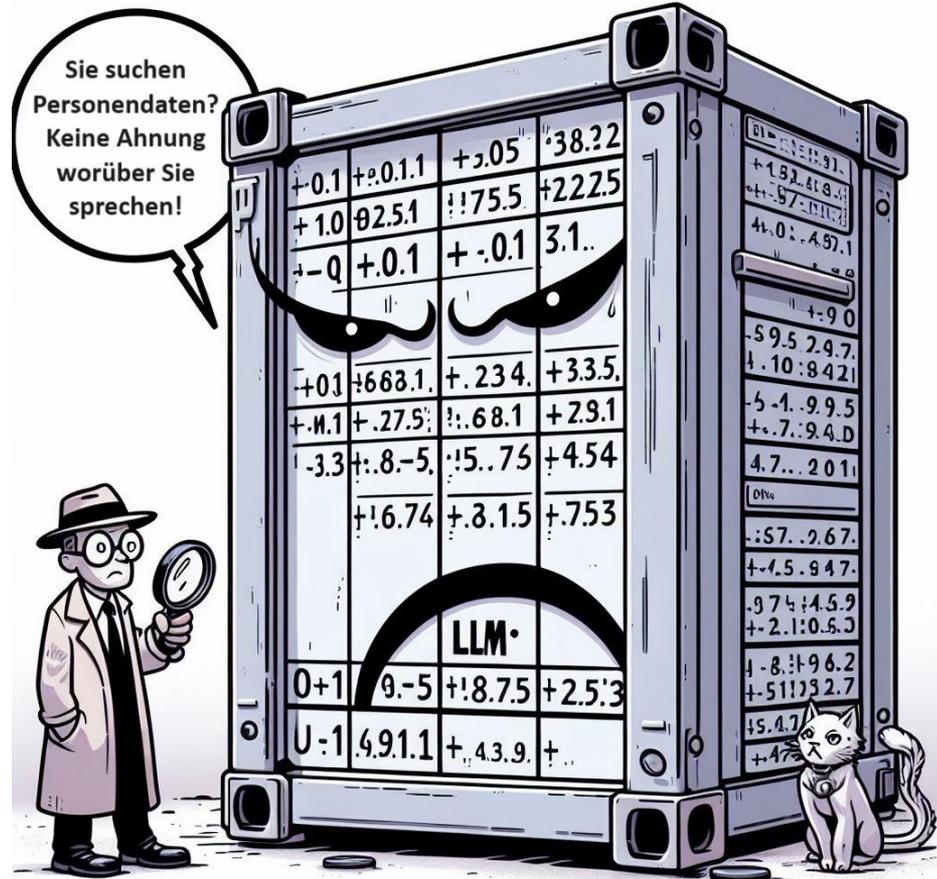
Künstliche Intelligenz.

Problemstellungen im Bereich Daten- und
Geheimnisschutz

David Rosenthal, Partner, VISCHER AG
1. Oktober 2024

Was steckt
an Daten in
Sprachmodellen
überhaupt drin?

Werden Inhalte
aus dem Training
oder gar der
Benutzung
"memorisiert"?



Was gehen Aufsichtsbehörden mit KI um?

- 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? Nein!

The image shows a document from the Hamburg Data Protection Authority (Der Hamburgische Beauftragte für die Datenschutzangelegenheiten). At the top, there is a red logo of a castle and the text 'Der Hamburgische Beauftragte für die Datenschutzangelegenheiten'. Below this, the text reads: '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.' Below this, there are three numbered points: '2. Mangelnde 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.' and '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.' At the bottom, there is a footnote: '* Gemeint sind hierbei allein die Modelle als wichtiger, aber nicht alleiniger Bestandteil eines KI-Systems (z. B. eines LLM-basierten Chatbots).' and contact information: 'www.datenschutz-hamburg.de', 'E-Mail: mail@datenschutz.hamburg.de', 'Ludwig-Erhard-Straße 22 · D-20459 Hamburg · Tel.: 040 · 4 28 54 - 40 40 · Fax: 040 · 4 28 54 - 40 00', and 'Unsere öffentlichen PDF-Dokumente sind im Internet verfügbar (Printserver: 0800 3 798 383, 0823 4030; E-Mail: 0800 8484 8377 3707)'.

Was gehen Aufsichtsbehörden mit KI um?

- 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

Wirklich? Nein!



Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

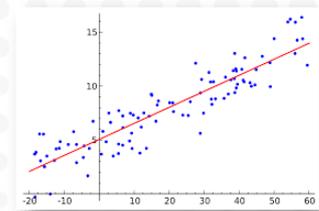
Eidgenössischer Datenschutz- und
Öffentlichkeitsbeauftragter (EDÖB)

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.

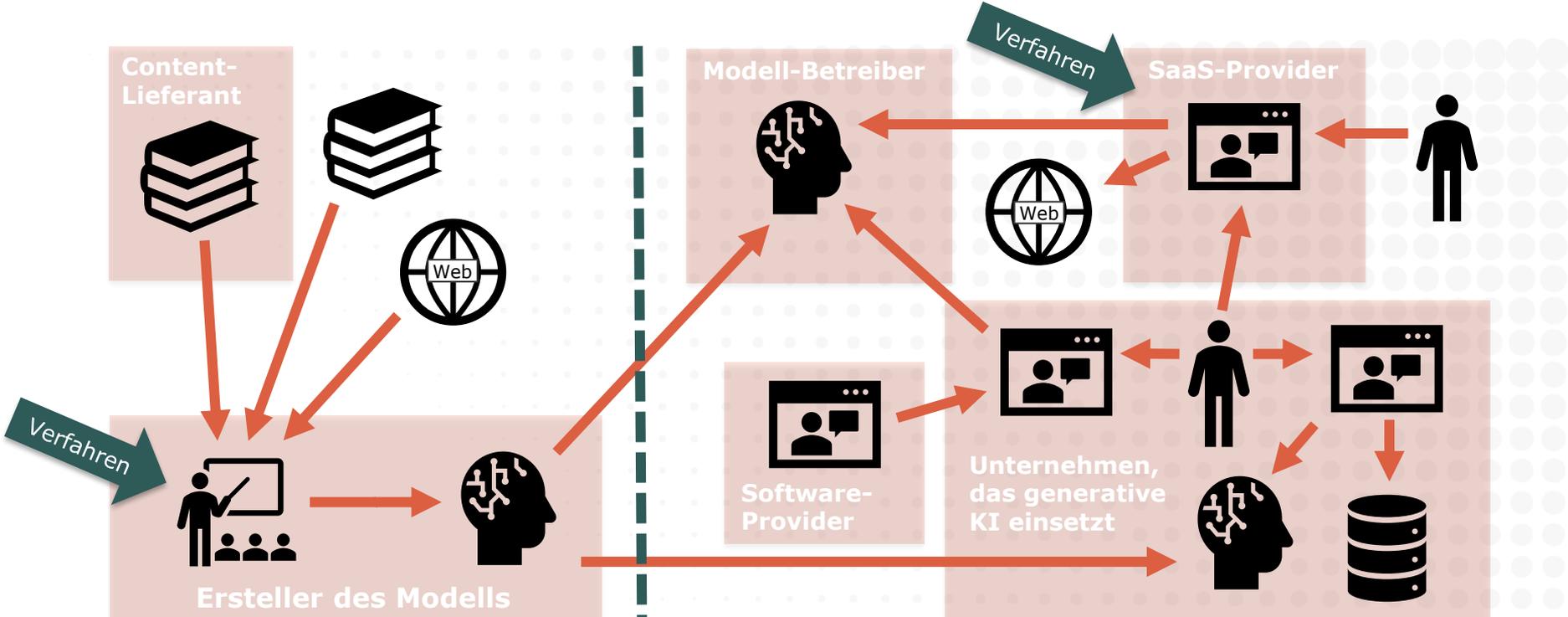
KI-gestützte Datenbearbeitungen mit hohen Risiken sind nach DSG dem Grundsatz nach zulässig, erfordern aber angemessene Massnahmen zum Schutz der potentiell betroffenen Personen. Aus diesem Grund verlangt das Gesetz bei hohen Risiken eine sog. «Datenschutz-Folgenabschätzung». Anwendungen hingegen, die geradezu auf eine Aushöhlung 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 seine Entscheide auf Basis eines Trainings fällt statt einer voll ausprogrammierten Logik
- **Aber:** Die Definition ist mangelhaft ...
 - Jedes Kopiergerät ist KI (OCR); was ist mit linearer Regression?



Verschiedene Stakeholder (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

Es gelten die allgemeinen Regeln und sie führen grundsätzlich zu angemessenen Ergebnissen

Datenschutz

Geheimnisschutz,
Urheberrecht und
Lauterkeitsrecht

EU AI Act, andere
KI-Regulierung

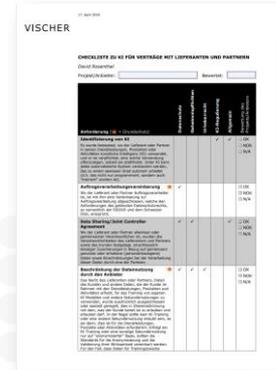
Datenschutzrecht

- **1. 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?



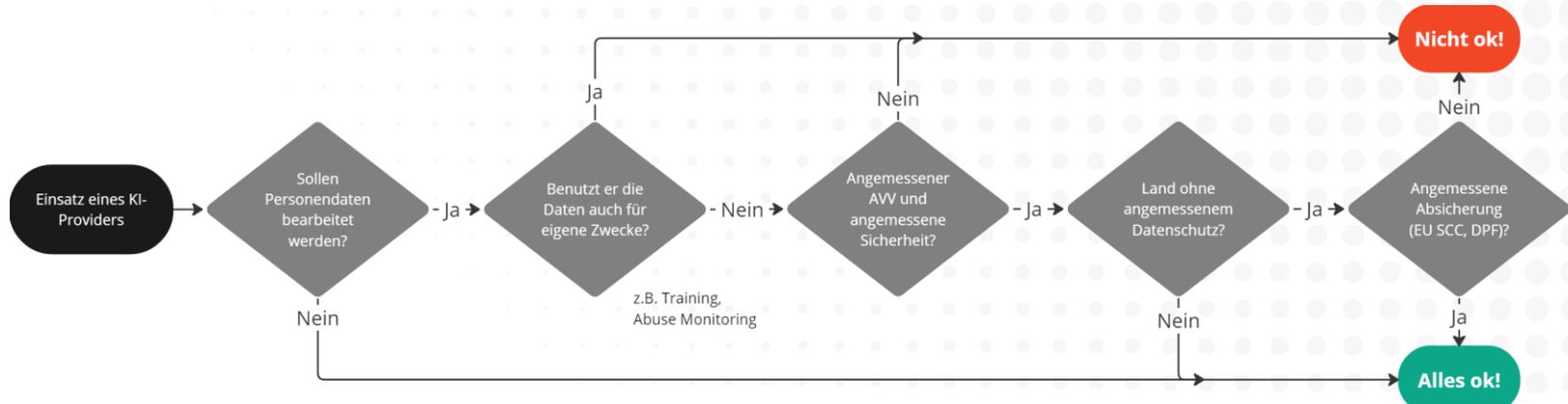
vischerlnk.com/3IdAymb

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



Datenschutzrecht

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



Berufs- und Amtsgeheimnis

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

- **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/3Xfz16e

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

KI-Einsatz in der Kanzlei

- **Vier Möglichkeiten**

- Allzweck-Dienste wie ChatGPT oder Copilot
- Spezial-KI-Lösungen wie Herlock.ai, CoCounsel
- KI-Modell eines Cloud-Providers für eigene Anwendung
- Selbst betriebenes KI-Modell für eigene Anwendung

- **Praktische Herausforderungen**

- Vielen Provider fehlt die nötige Maturität; viel "heisse Luft"
- Fehlende Transparenz und Stabilität der Angebote
- AVV wird zwar angeboten, aber Berufsgeheimnisklauseln selten
- Speicherung/Bearbeitung in der Schweiz, kein Operator Access
- Opt-out des menschlichen "Abuse Monitoring" durch Provider



VISCHER GPT
Kleiner Zettel

Alle Informationen sind dem Rechtsanwaltsbüro...

Wie wir die Risiken einschätzen

Risikobewertung vorschlagen	Mögliche Folgen für die Daten	Existenzrisiko abnehmend	Risiko (1-100)
Wenn die Personendaten von unbefugten Dritten erfasst und missbraucht werden, könnte es beispielsweise zu Identitätsdiebstahl kommen, bei dem unbefugte Handlungen im Namen des Opfers durchgeführt werden, wie etwa betrügerische Kreditkartenanträge oder Zugang zu persönlichen Konten erlangen. Angesichts der geringeren Gegenmaßnahmen ist das Risiko insgesamt gering, aber sollte es dennoch auftreten, könnten die Folgen für die betroffene Person erheblich sein, vor allem Verlust des Vertrauens in die Sicherheitsmaßnahmen von Unternehmen und Institutionen.	Substanziell	Hoch	Mittel (6)

VGPT via www.rosenthal.ch
VUD DSFA via vud.ch/dsfa

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 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?

Blog-Beitrag und weitere Infos:

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

Risikobeurteilung von KI-Projekten:



vischerInk.com/gaira

vischerInk.com/ki-compliance-kurz

VISCHER

Danke für Ihre Aufmerksamkeit!

Fragen: david.rosenthal@vischer.com

Zürich

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

www.vischer.com

Basel

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

Genf

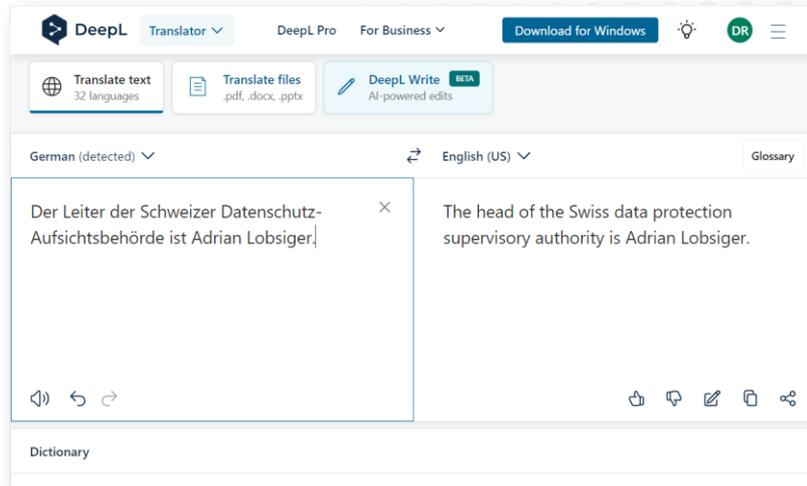
Rue du Cloître 2-4
Postfach
1211 Genf 3, Schweiz
T +41 58 211 35 00

Mehr Unterlagen:
www.vischer.com/ki
www.rosenthal.ch

VISCHER

Backup-Slides

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

Use Case: Bewerber wird von KI selektioniert

Art. 21 Informationspflicht bei einer automatisierten Einzelentscheidung

¹ 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).

² 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.

³ 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 fehlerhafter Auskünfte nicht bezahlen? Wo ist der Unterschied zum Call Center? Und lohnt es sich nicht trotzdem?

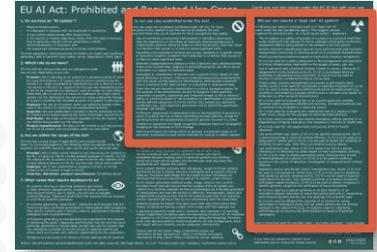
1. **Haben wir ein vernünftiges Set an Massnahmen getroffen?**
2. **Was sind die Restrisiken?**
3. **Sind sie akzeptabel und lohnt sich das Ganze wirklich?**

Use Case: Training von KI-Modellen

- **Öffentliche Daten** sind nicht einfach frei verwendbar
- **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?
- Was gilt wenn Trainingsinhalte im Modell **memorisiert** werden?

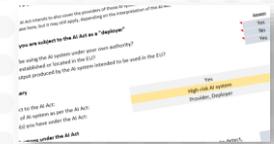
Und der EU AI Act?

- 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

Ausführlicher Aufsatz
zum EU AI Act:
vischerlnk.com/3ZkPOYh



Siehe AI Act Check unter
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

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

VISCHER
INFORMATION SECURITY

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?



vischerink.com/3OPTpaA