

VISCHER

KI-Regulierung.

Welche materiellen Vorgaben für Schweizer Unternehmen wirklich wichtig sind

David Rosenthal, Partner, VISCHER AG
20. Juni 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?

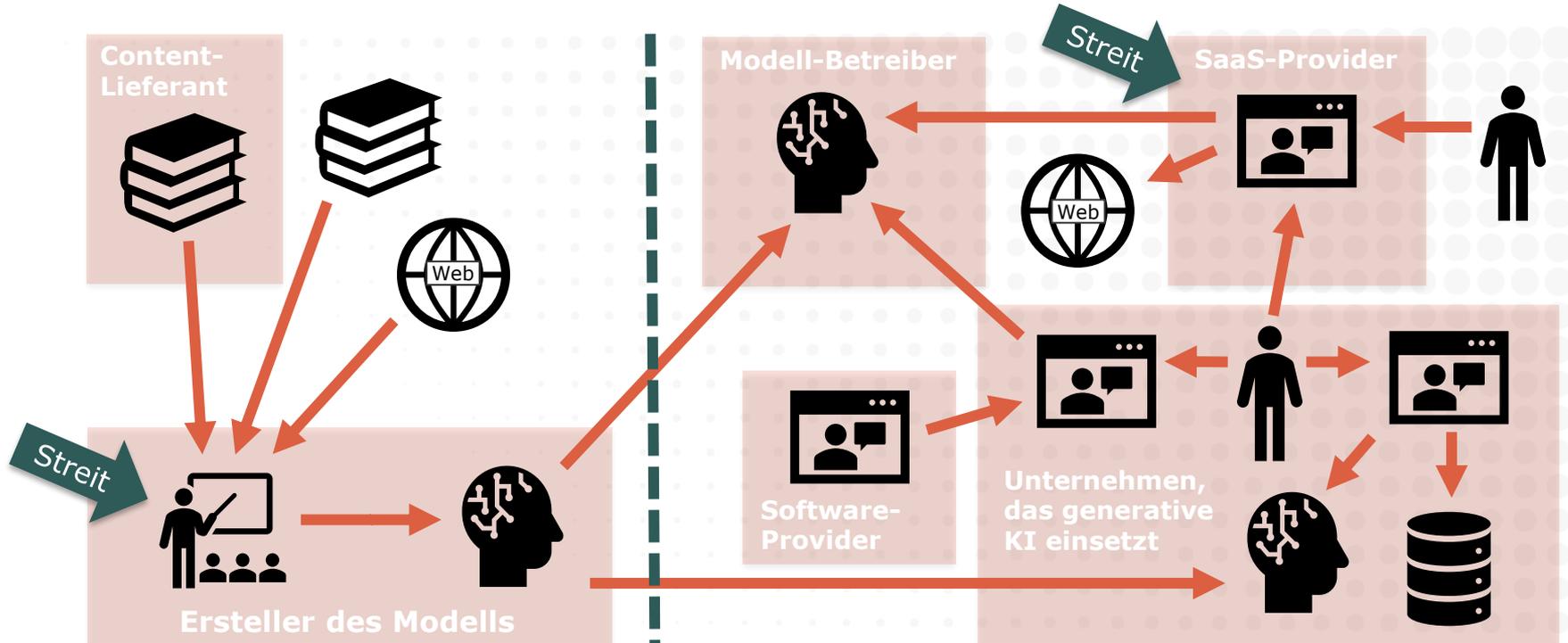
- **Ja**, weil die Materie für uns unheimlich und ungewohnt ist
 - Wissen Sie, was in einem LLM steckt und warum es so gut ist?
 - Wir hatten beim Internet früher dieselbe Situation, und heute haben wir uns daran gewöhnt – es ist völlig normal geworden
- **Unbehagen** führt zum Ruf nach mehr Regulierung 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

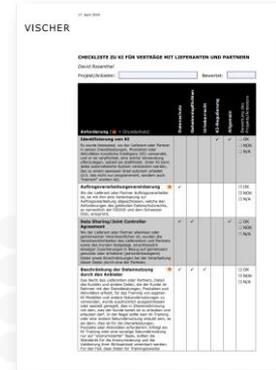


25. Juli 1994 (time.com, Titelseite:
James Porto)

"KI" schon vielerorts im Einsatz: OCR, biometrische Authentifizierung, Stauwarner, Übersetzung, Malware-Erkennung, Sprachsteuerung etc.

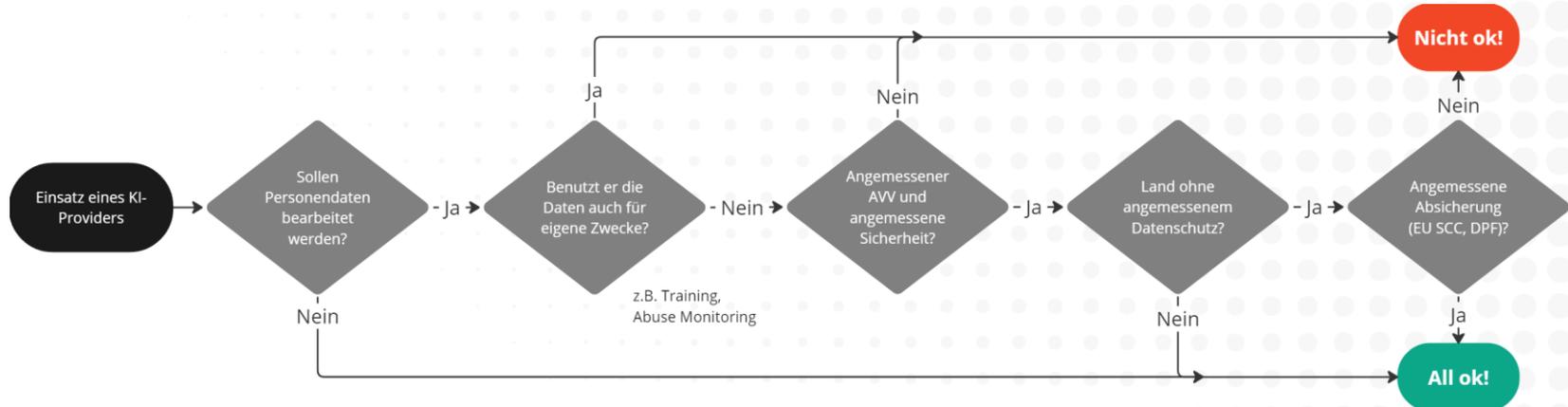
Verantwortlichkeit für generative KI





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?



vischerInk.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

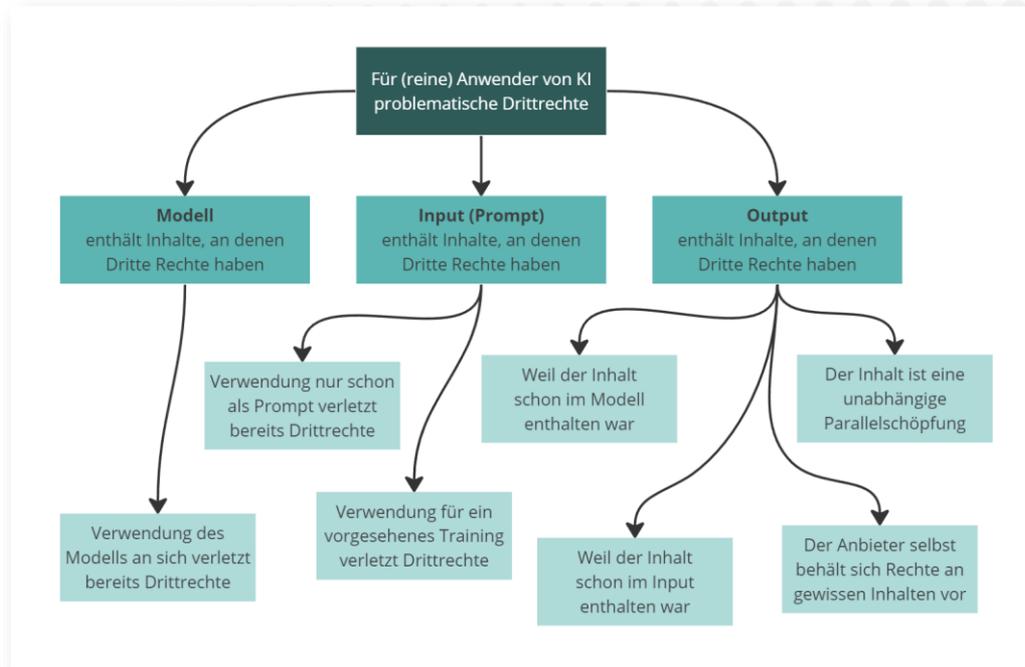
- **Vorgaben** beim Einsatz insb. **ausländischer Provider**
 - Einhaltung der Geheimhaltung seitens des Providers, auch wenn dieser im Ausland ist (vertragliche Verpflichtung)
 - Angemessene Informationssicherheit, keine Zweckentfremdung
 - Kein Grund zur Annahme, dass es via Provider zu ausländischem 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



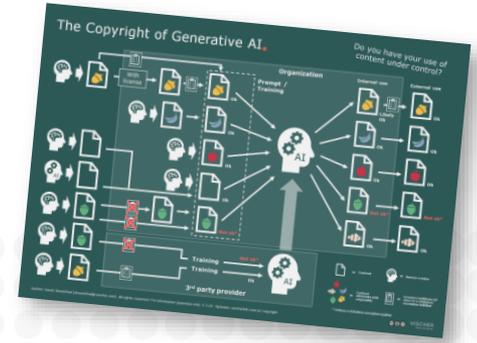
vischerlnk.com/4ck2J0L
vischerlnk.com/4bGEARE

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

Urheberrecht



Mehr: vischerlnk.com/3vYAPqG



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
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").	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.	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.	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).	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.	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.

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

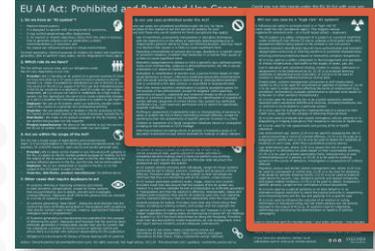
vischerink.com/30PTpaA

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?

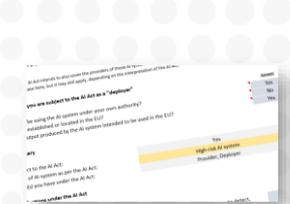


Und der AI Act?

- 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



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) den Anweisungen des Anbieters zu folgen, (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, und (ix) Anfragen betroffener Personen bezüglich solcher Entscheidungen zu befolgen.

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

Quelle: vischerlnk.com/gaira

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?

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



VISCHER
1892 1000 1000

Blog-Beitrag und weitere Infos:

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

Hier muss jedes Unternehmen seinen Weg finden

vischerlink.com/ki-compliance-kurz

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?**

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