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

FORMEL-1-PILOTEN

Rechtsstreit um Fake-KI-Interview mit Michael Schumacher

TRANSPARENZ BEI KI

"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, zwp-online.info, thepioneer.de, tagesschau.de, golem.de

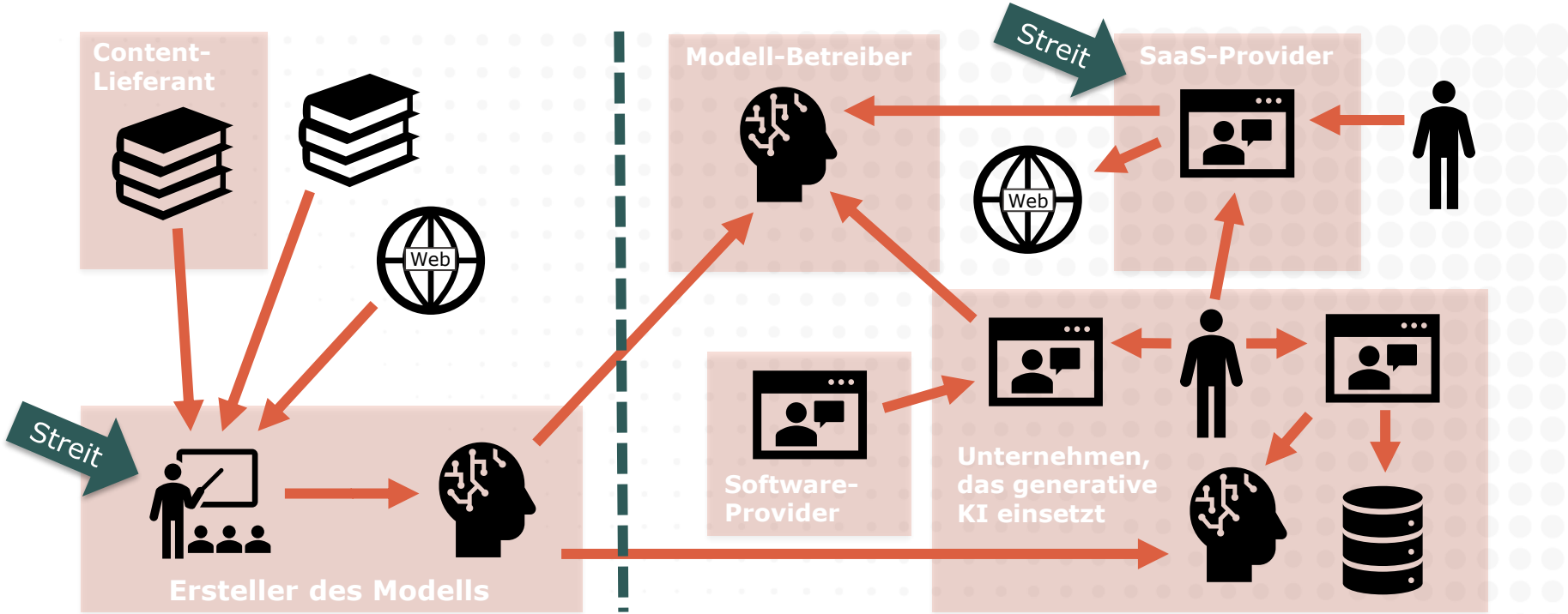
Rechtsunsicherheit?

- **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 Regulierungsbedarf klären
- Das **bestehende Recht** regelt viele der Themen recht gut
  - Datenschutz, Urheberrecht, Lauterkeitsrecht, Geheimnisschutz
  - Viele 0815-Hausaufgaben und einzelne neue Herausforderungen

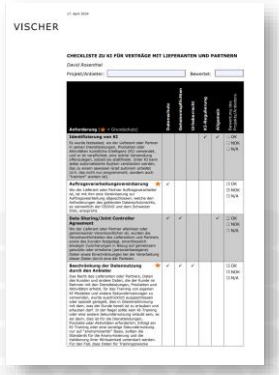
25. Juli 1994 (time.com, Titelseite: James Porto)

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

Verantwortlichkeit für generative KI

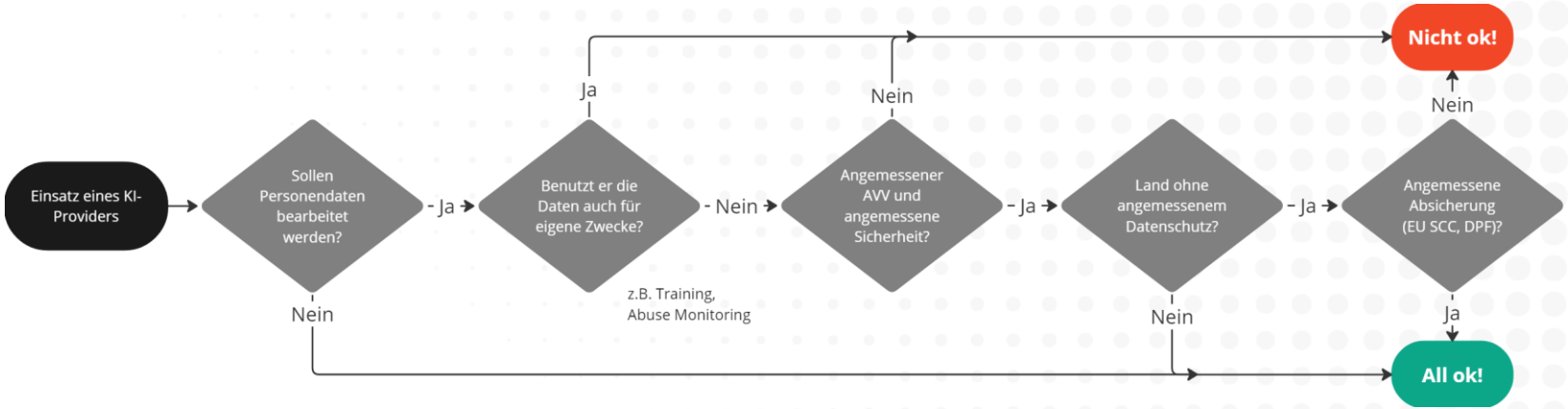


vischerlnk.com/ki-provider-check



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

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

Berufs- und Amtsgeheimnis

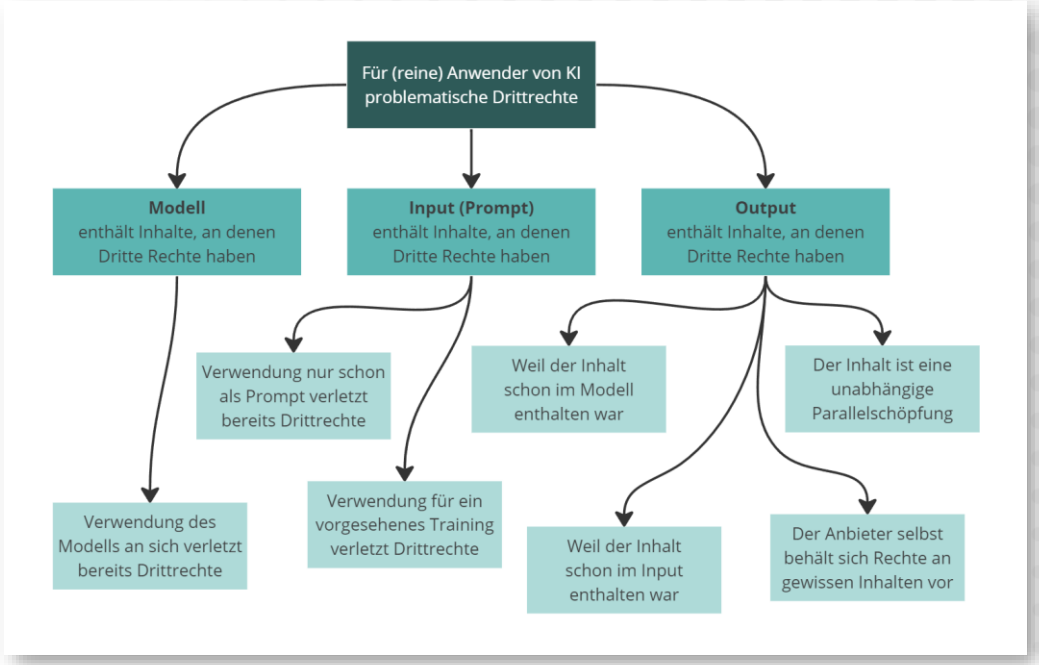
Prüfung Lawful Access Risiko mit "Methode Rosenthal"
vischerlnk.com/flara
vischerlnk.com/flarafaq

- **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

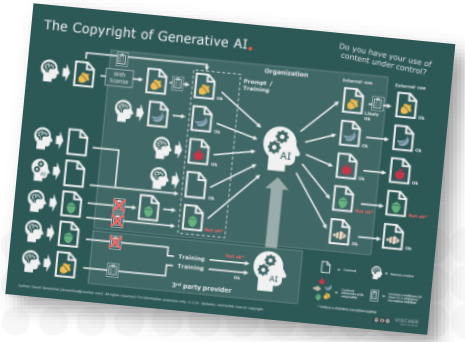
vischerlnk.com/4ck2J0L
vischerlnk.com/4bGEARe

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

Werk "zweiter Hand"? Nein!

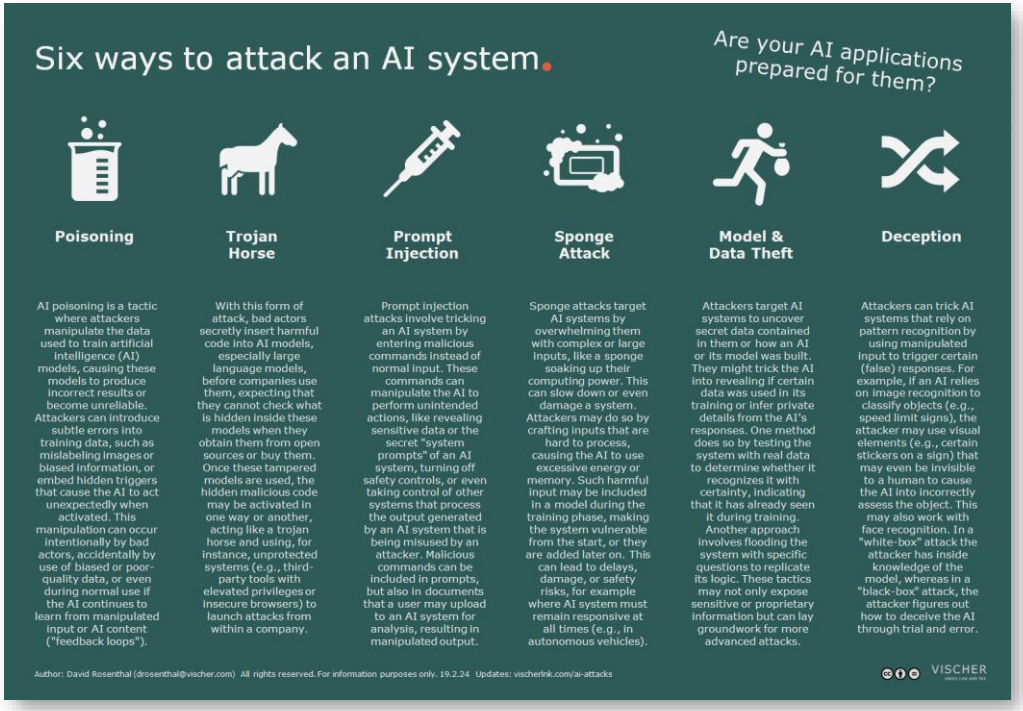
OpenAI Dall-E 3 basierend auf einem entsprechendem Prompt, der Bild und Stil beschreibt

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?

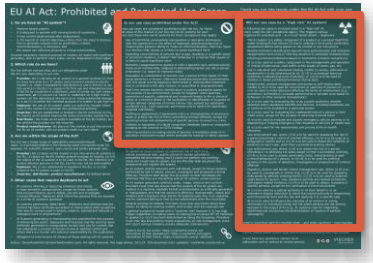


1. Womit und worauf wurden Modelle trainiert? Compliance eingehalten? Ist alles dokumentiert?
2. Welche Trainingsinhalte wurden allenfalls "memorisiert"? Können Trainingsinhalte "leaken"?
3. "Bias" soweit nötig vermieden?
4. Wird ein "Concept Drift" erkannt?
5. Sind speziell auf KI ausgerichtete Angriffsformen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

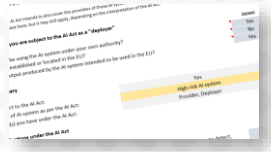
vischerlnk.com/3OPTpaA

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, (vii) das KI-System behördlich zu registrieren, (viii) eine Überwachung des Systems nach seiner Markteinführung zu betreiben, (ix) Vorfälle den Behörden zu melden und Korrekturmassnahmen zu ergreifen, (x) mit den Behörden zusammenzuarbeiten, (xi) die Einhaltung der vorstehenden Anforderungen zu dokumentieren und (xii) 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

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

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

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?

VISCHER

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

Use Case: Chatbot auf der Website

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?

- 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?

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

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

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

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

www.vischer.com