VISCHER

ChatGPT & Co.

Datenschutz beim Einsatz von generativer KI

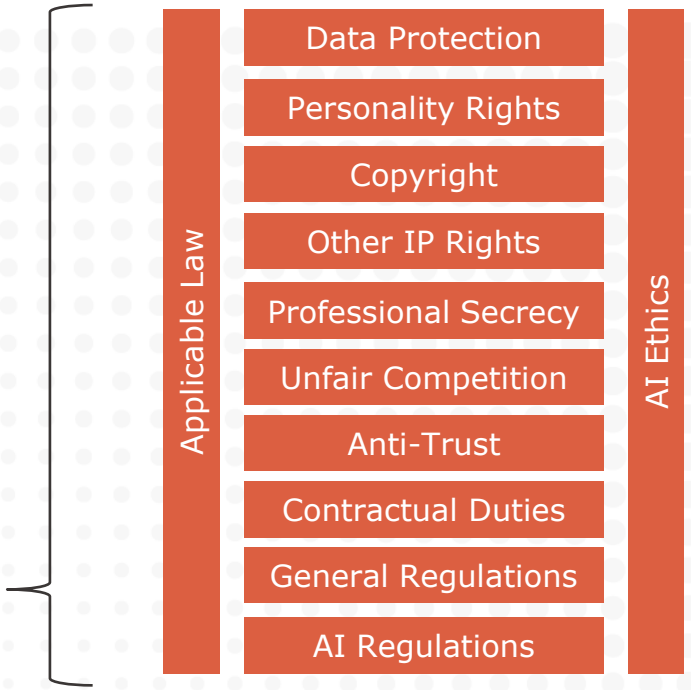
David Rosenthal, Partner, VISCHER AG
5. Oktober 2023

Generative KI …

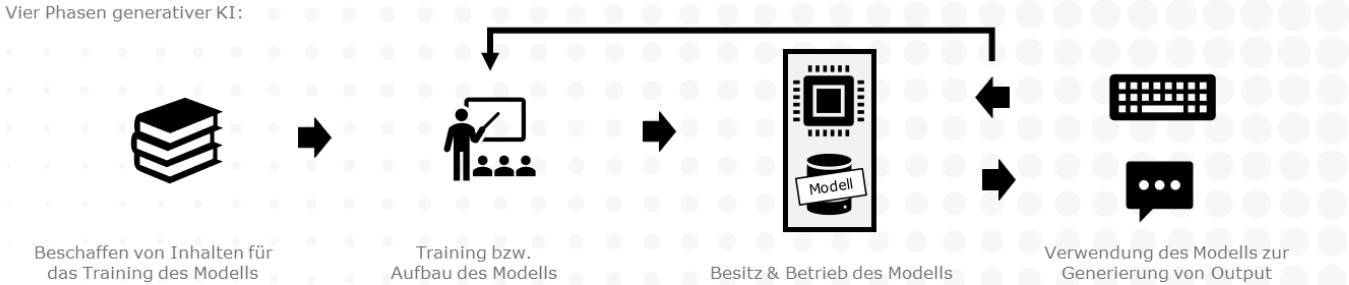
- KI-Systeme, die Texte, Bilder oder andere Inhalte mithilfe generativer Modelle **erzeugen**
  - ChatGPT, Bing, Bard, Dall-E, Midjourney etc.
- Kann das rechtlich in zulässiger Weise genutzt werden?

  Ja!
- Stellen sich viele rechtliche Fragen?

  Ja!



Datenschutz-Themen

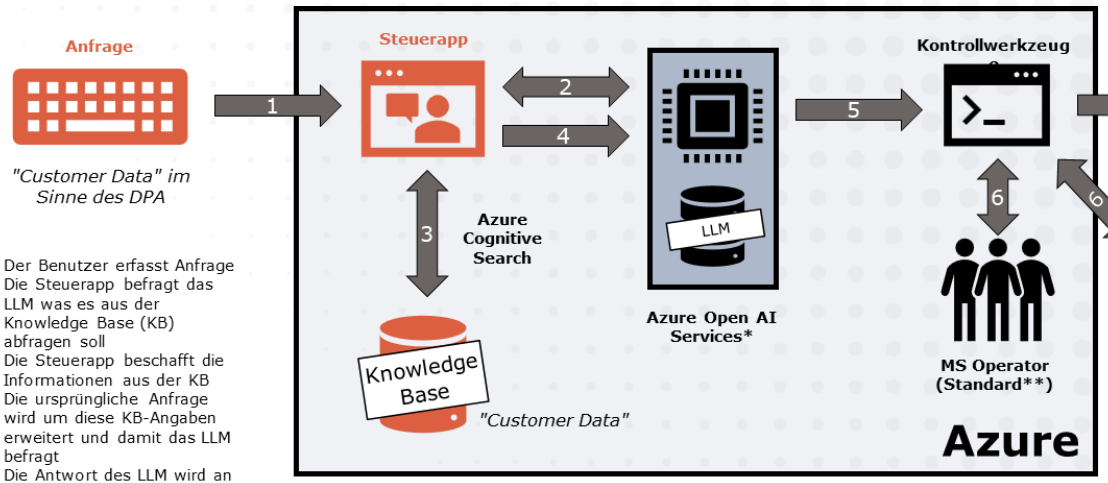


Zehn Fragen für die Praxis

- **Wo** wird der Input ("Prompts") hingesandt und bearbeitet?
- Besteht mit dem Provider ein **Auftragsbearbeitungsvertrag**?
- Ist die **Datensicherheit** gewährleistet?
- **Welchen** Input dürfen Mitarbeitende machen?
- Wird der Input für das **Provider-Training** des Modells benutzt?
- Wird der Output ("Completions") vom Provider **geprüft**?
- Wie gehen wir mit **fehlerhaftem/ungewolltem Output** um?
- Muss und kann ein **"Data Leakage"** vermieden werden?
- Haben wir eine Pflicht auf die KI-Verwendung **hinzuweisen**?
- Welches Konzept besteht für **Betroffenenrechte**?

Retrieval-Augmented Generation

Beispiel Microsoft mit OpenAI



1. Der Benutzer erfasst Anfrage
2. Die Steuerapp befragt das LLM was es aus der Knowledge Base (KB) abfragen soll
3. Die Steuerapp beschafft die Informationen aus der KB
4. Die ursprüngliche Anfrage wird um diese KB-Angaben erweitert und damit das LLM befragt
5. Die Antwort des LLM wird an Apps von Microsoft zur Prüfung der Einhaltung der Vorgaben gemeldet
6. Bei Bedarf wird die Antwort manuell überprüft
7. Die Antwort wird ausgespielt

Von/für Kunde

Von Microsoft

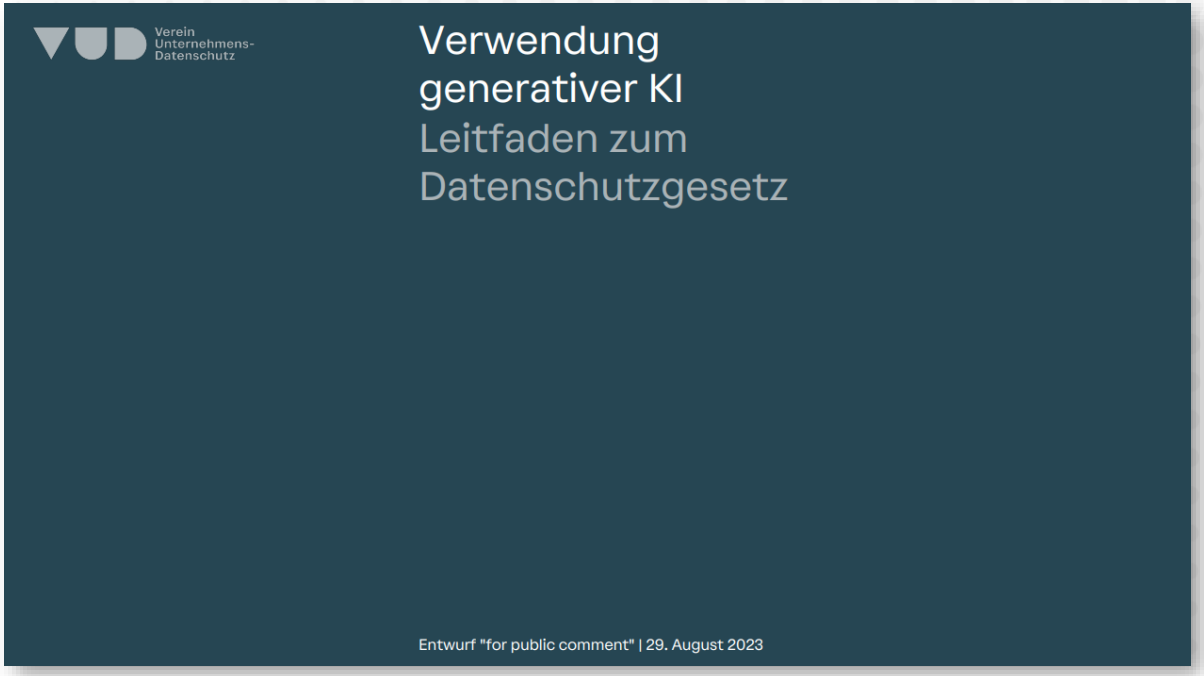
→ Einsatz von Azure im Rahmen der bestehenden Vertragsbedingungen

Infos: https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy

* Das LLM speichert keine Customer Data und benutzt sie nicht für sein Training

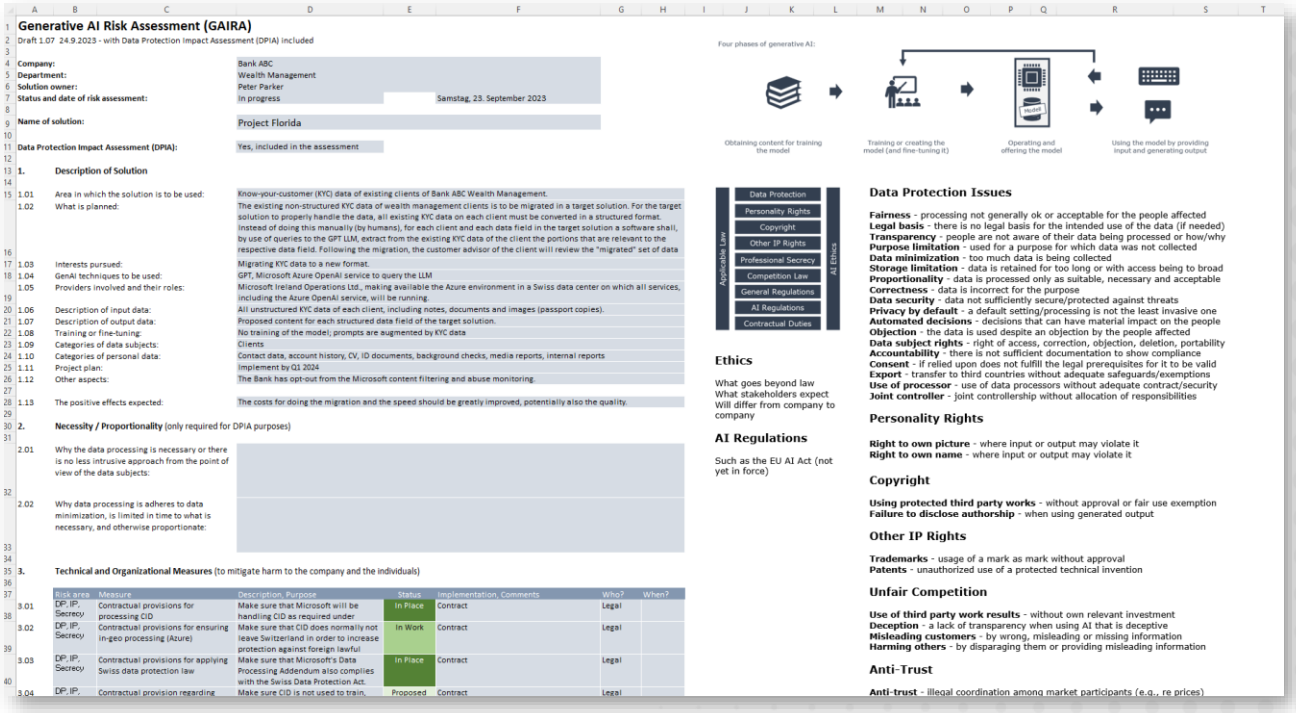
** Der Kunde kann unter Umständen die Prüfung der Antwort durch Microsoft abwählen (und sie mit den Werkzeugen von Microsoft ggf. selbst vornehmen)

Weitere Informationen



www.vud.ch

Risikobeurteilung von generativer KI



Generative AI Risk Assessment (GAIRA)

Draft 1.07 24.9.2023 - with Data Protection Impact Assessment (DPIA) included

Company:	Bank ABC
Department:	Wealth Management
Solution owner:	Peter Parker
Status and date of risk assessment:	In progress — Samstag, 23. September 2023
Name of solution:	Project Florida
Data Protection Impact Assessment (DPIA):	Yes, included in the assessment

1. Description of Solution

1.01	Area in which the solution is to be used:	Know-your-customer (KYC) data of existing clients of Bank ABC Wealth Management.
1.02	What is planned:	The existing non-structured KYC data of wealth management clients is to be migrated in a target solution. For the target solution to properly handle the data, all existing KYC data on each client must be converted in a structured format. Instead of doing this manually (by humans), for each client and each data field in the target solution a software shall, by use of queries to the GPT LLM, extract from the existing KYC data of the client the portions that are relevant to the respective data field. Following the migration, the customer advisor of the client will review the "migrated" set of data
1.03	Interests pursued:	Migrating KYC data to a new format.
1.04	GenAI techniques to be used:	GPT, Microsoft Azure OpenAI service to query the LLM
1.05	Providers involved and their roles:	Microsoft Ireland Operations Ltd., making available the Azure environment in a Swiss data center on which all services, including the Azure OpenAI service, will be running.
1.06	Description of input data:	All unstructured KYC data of each client, including notes, documents and images (passport copies).
1.07	Description of output data:	Proposed content for each structured data field of the target solution.
1.08	Training or fine-tuning:	No training of the model; prompts are augmented by KYC data
1.09	Categories of data subjects:	Clients
1.10	Categories of personal data:	Contact data, account history, CV, ID documents, background checks, media reports, internal reports
1.11	Project plan:	Implement by Q1 2024
1.12	Other aspects:	The Bank has opt-out from the Microsoft content filtering and abuse monitoring.
1.13	The positive effects expected:	The costs for doing the migration and the speed should be greatly improved, potentially also the quality.

2. Necessity / Proportionality (only required for DPIA purposes)

2.01	Why the data processing is necessary or there is no less intrusive approach from the point of view of the data subjects:	
2.02	Why data processing is adheres to data minimization, is limited in time to what is necessary, and otherwise proportionate:	

3. Technical and Organizational Measures (to mitigate harm to the company and the individuals)

	Risk area	Measure	Description, Purpose	Status	Implementation, Comments	Who?	When?
3.01	DP, IP, Secrecy	Contractual provisions for processing CID	Make sure that Microsoft will be handling CID as required under	In Place	Contract	Legal	
3.02	DP, IP, Secrecy	Contractual provisions for ensuring in-geo processing (Azure)	Make sure that CID does normally not leave Switzerland in order to increase protection against foreign lawful	In Work	Contract	Legal	
3.03	DP, IP, Secrecy	Contractual provisions for applying Swiss data protection law	Make sure that Microsoft's Data Processing Addendum also complies with the Swiss Data Protection Act.	In Place	Contract	Legal	
3.04	DP, IP,	Contractual provision regarding	Make sure CID is not used to train,	Proposed	Contract	Legal	

Four phases of generative AI:

Applicable Law: Data Protection; Personality Rights; Copyright; Other IP Rights; Professional Secrecy; Competition Law; General Regulations; AI Regulations; Contractual Duties

AI Ethics

Ethics

What goes beyond law
What stakeholders expect
Will differ from company to company

AI Regulations

Such as the EU AI Act (not yet in force)

Data Protection Issues

Fairness - processing not generally ok or acceptable for the people affected
Legal basis - there is no legal basis for the intended use of the data (if needed)
Transparency - people are not aware of their data being processed or how/why
Purpose limitation - used for a purpose for which data was not collected
Data minimization - too much data is being collected
Storage limitation - data is retained for too long or with access being to broad
Proportionality - data is processed only as suitable, necessary and acceptable
Correctness - data is incorrect for the purpose
Data security - data not sufficiently secure/protected against threats
Privacy by default - a default setting/processing is not the least invasive one
Automated decisions - decisions that can have material impact on the people
Objection - the data is used despite an objection by the people affected
Data subject rights - right of access, correction, objection, deletion, portability
Accountability - there is not sufficient documentation to show compliance
Consent - if relied upon does not fulfill the legal prerequisites for it to be valid
Export - transfer to third countries without adequate safeguards/exemptions
Use of processor - use of data processors without adequate contract/security
Joint controller - joint controllership without allocation of responsibilities

Personality Rights

Right to own picture - where input or output may violate it
Right to own name - where input or output may violate it

Copyright

Using protected third party works - without approval or fair use exemption
Failure to disclose authorship - when using generated output

Other IP Rights

Trademarks - usage of a mark as mark without approval
Patents - unauthorized use of a protected technical invention

Unfair Competition

Use of third party work results - without own relevant investment
Deception - a lack of transparency when using AI that is deceptive
Misleading customers - by wrong, misleading or missing information
Harming others - by disparaging them or providing misleading information

Anti-Trust

Anti-trust - illegal coordination among market participants (e.g., re prices)

Mit dem Einsatz generativer KI kommen weitere Ebenen hinzu

Generative AI Risk Assessment (GAIRA)

Risikobeurteilung *durch* generative KI

Datenschutz-Folgenabschätzung (DSFA)
Version 25.9.2023 for public comment - Private CH-DSG/DSGV

Hinweis: Eine Anleitung zum Ausfüllen dieser DSFA und zur KI-gestützten Ausfüllhilfe (optional, nur in der Version des Excels mit Makros) findet sich am Ende dieses Arbeitsblatt

Unternehmen (Verantwortlicher): Musterfirma AG

Abteilung:

Verantwortlich intern:

Status der DSFA:

Name des Vorhabens:

Aktivität gemäss Bearbeitu

1. Beschreibung de

1.01 In welchem Bereic bzw. welcher Gesc

1.02 Was vorgesehen i

1.03 Welche Interesser

1.04 Welche Mittel unc

1.05 Welche Dritten an

1.06 Welche Daten bea

1.07 Wessen Daten bei

1.08 Wo überall Daten

1.09 Wann die Datenb

1.10 Weitere Besonder

2. Erforderlichkeit /

2.01 Warum die Daten Sicht der betroffer eingreifende Vorg

2.02 Warum die Datenbearbeitung datensparsam, zeitlich auf das nötige begrenzt und auch sonst verhältnismässig ist:

Zugriffen geschützt werden.
Die Datenbearbeitung ist datensparsam, da nur der Stimmabdruck, die ID der Person und Tonaufnahmen gespeichert werden, die für die Identitätsverifizierung notwendig sind. Die Datenbearbeitung ist zeitlich begrenzt, da der Stimmabdruck bei jedem Anruf neu erstellt und nicht länger als nötig gespeichert wird. Die Datenbearbeitung ist verhältnismässig, da sie zur Sicherheit der Anrufer im Call-Center beiträgt und die einzigen Daten bearbeitet werden, die dafür erforderlich sind.

4. Risiken von negativen Folgen für die betroffenen Personen, die trotz der obigen Massnahmen verbleiben

10 Risiken vorschlagen (überschreibt bisherige Werte)

Hinweis: Falls die ermittelten Risiken als zu hoch erscheinen oder sich zeigt, dass es noch weitere Massnahmen zur Minimierung gibt, sollten diese oben unter Ziff. 3 eingetragen werden und bei der Risikobeurteilung hier berücksichtigt werden.

	Mögliche unerwünschte negative Folgen Weiteres Risiko vorschlagen*	Was wir dagegen tun Massnahmen vorschlagen* / Aus obigen formulieren*	Wie wir das Restrisiko einschätzen Risikobeurteilung vorschlagen*	Mögliche Folgen für die Person	Eintrittswahrscheinlichkeit (alles in allem)	Risiko (1-16)
4.01	Personendaten des Vorhabens gelangen wegen eines Fehlers oder absichtlich an **unbefugte Dritte**. Diese missbrauchen sie zum Schaden der betroffenen Personen.	- Berechtigungskonzept: Da wir nur autorisierten Personen Zugriff auf die Personendaten geben, wird das Risiko von unbefugtem Zugriff und Missbrauch reduziert. - Schulung: Durch Schulungen stellen wir sicher, dass die Mitarbeitenden die Lösung korrekt und sicher nutzen, was das Risiko von Fehlern und Missbrauch verringert. - Zugriffskontrolle: Durch die Beschränkung des Zugriffs auf autorisierte Personen können wir Missbräuche und unbefugte Nutzung der Personendaten verhindern. - Verschlüsselung "at rest": Die Verschlüsselung der Personendaten in unserem System schützt vor unautorisiertem Zugriff, falls jemand physischen Zugriff auf die Speichermedien erhält. - Datenlöschungsfunktionen: Durch die Möglichkeit, nicht mehr benötigte Personendaten zu löschen oder zu anonymisieren, minimieren wir das Risiko eines unbefugten Zugriffs auf diese Daten.	Das konkrete Restrisiko für die betroffene Person besteht darin, dass ihre Personendaten aufgrund eines Fehlers oder absichtlich an unbefugte Dritte gelangen könnten. Diese könnten die Daten dann zum Schaden der betroffenen Person nutzen, beispielsweise für Identitätsdiebstahl oder Missbrauch in sozialen Medien. Die Wahrscheinlichkeit dieses Szenarios ist jedoch insgesamt gering, da strenge Sicherheitsmassnahmen wie Zugriffskontrollen und Verschlüsselung implementiert wurden.	Substanziell	Tief	Mittel (6)
4.02	Personendaten des Vorhabens gelangen wegen eines Fehlers oder absichtlich an eine **unbefugte interne Person**.					

vud.ch/dsfa

VISCHER

Danke für Ihre Aufmerksamkeit!

Fragen: drosenthal@vischer.com

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