

VISCHER

Das Feigenblatt-Syndrom. Zur Wirksamkeit menschlicher Aufsicht bei KI-Systemen

David Rosenthal, VISCHER AG
27. November 2025

Eine Verletzung der Datensicherheit

- Risikobewertung als Herausforderung
- Ein Prozess, der vielen sehr schwer fällt
- Wir wissen nie, ob wir richtig liegen und vertrauen unserem Urteil möglicherweise nicht
- Die Folgen unserer Entscheidung können gewichtig sein
- Eine KI könnte die Risikobewertung doch für uns übernehmen!
- Die scheinbar perfekte Lösung – schnell und skalierbar
- Frage: Sollen wir die KI zuerst bewerten lassen oder sollen wir sie nur nutzen, um unsere eigene Arbeit zu überprüfen?

Lassen wir die KI die Risiken bewerten ...

The LLM has provided the following result (you can edit it):

- [Cell: C6]
[Value: ""]
- [Cell: C7]
[Value: 'Human Resources (HR)']
- [Cell: C8]
[Value: ""]
- [Cell: C9]
[Value: ""]
- [Cell: C10]
[Value: ""]
- [Cell: C11]
[Value: 'Local HR staff, the staff member who sent the file.']
- [Cell: C12]
[Value: 'Not applicable']
- [Cell: C13]
[Value: 'Not applicable']
- [Cell: C15]
[Value: 'Final assessment']
- [Cell: F15]
[Value: '16/04/2025']
- [Cell: C16]
[Value: 'Review and implement technical/procedural controls to prevent bulk data export errors, enforce encryption for sensitive email attachments, and automate splitting of multi-person documents.']

Shall Red Ink insert the square brackets into your worksheet, where possible?

OK, use edited text OK, use original text Transfer to pane Cancel

PERSONAL DATA BREACH INCIDENT 16 April 2025

CASE SUMMARY
Subject: Accidental Disclosure of Co-worker Data - Geneva Store
Date of Incident: 13 April 2025
Date Reported: 15 April 2025

1. Executive Summary
On 13 April 2025, a document containing sensitive employment and tax data for 140 co-workers at the Geneva Store was inadvertently emailed to one co-worker, both via their internal company address and their private email address stored in the system. The email was intended to include only the requesting co-worker's individual tax statement but instead included a bulk file with data for all co-workers subject to withholding tax at the store.
The issue was reported to Human Resources (HR) on 15 April 2025 after the recipient was made aware of the file's contents by a third party who accessed the email.

2. Nature of the Data Involved
The file contained the following personal and sensitive data types for each of the affected co-workers:

- Full name
- Date of birth
- Social Security Identifier
- Home address
- Marital and consanguinity status
- Employment workload (percentage)
- Monthly taxable income and corresponding tax paid

3. Incident Details

mit Red Ink
(vischer.com/redink)

Lassen wir die KI die Risiken bewerten ...

Company (Controller):		Residual risks from the perspective of the affected persons					
Department/area		Things that could happen due to the data breach and could have negative consequences for the persons affected		How probable is this scenario in the present case?		If it happened, how severe would the impact be on the persons affected?	
Internally responsible						How probable is such impact for a relevant portion of them?	
Internal data protection						Risk	
Other persons involved						Explanation, remarks	
Who else internal							
Classification according to the circumstances of the individual							
1.01		The data is used for spam mails to affected persons.		N/A		N/A	
1.02		The data is used for commercial purposes (e.g. increased targeted advertising) in a manner that would otherwise not be permitted.		Unlikely		Minor consequences	
1.03		The data is used for phishing attacks on the persons affected.		N/A		N/A	
1.04		The data is used for other fraud attempts against the persons affected (e.g. grandchild scam, fake police officers).		Possible		Significant consequences	
1.05		The data will be used in other forms to attempt to defraud third parties (identity theft).		Possible		Significant consequences	
1.06		The breach results in attempts to take over accounts of the persons affected on the controller's platform.		N/A		N/A	
1.07		The breach results in attempts to take over accounts of the persons affected on other platforms.		N/A		N/A	
1.08		Other criminal offences are committed with the data in the name of the persons affected (identity theft).		Possible		Significant consequences	
1.09		As a result of the breach, benefits or services are obtained at the expense of the persons affected (identity theft).		Possible		Significant consequences	
1.10		As a result of the breach, third parties gain access to bank accounts or other assets of the persons affected.		N/A		N/A	
1.11		Credit or debit cards will be blocked and may have to be replaced.		N/A		N/A	

KI: Mächtiges Werkzeug – aber auch gefährlich?

- KI kennt keine Hemmungen, keine Müdigkeit, keine Langeweile
- Erledigt Aufgaben innert Minuten, für die wir ein Vielfaches an Zeit brauchen
 - Die Tools sind verfügbar und funktionieren
 - DSFA in 150 Sekunden: vischerlnk.com/redink_dsfa_demo
- Geht damit etwas Relevantes verloren?
- Handeln wir uns damit ein Problem ein?

Automation Bias

- Ein bekanntes psychologisches Phänomen
- Wir vertrauen der Einschätzung der Maschine, auch wenn sie unserer eigenen Einschätzung widerspricht
 - Insbesondere dann, wenn wir unter Druck stehen
 - Insbesondere dann, wenn die KI eine Begründung liefert
 - Es ist oft einfacher die Antwort einer KI zu akzeptieren, als selbst eine gründliche Analyse vorzunehmen (Denken ist anstrengend)
 - Wir nehmen an, dass eine KI objektiver und analytischer ist, eine grössere Datenbasis als wir nutzt und frei menschlicher Fehler ist
 - Mit der Art und Weise, wie KI Daten und Optionen präsentiert, erfolgt ein subtiles Framing der menschlichen Aufsichtsperson

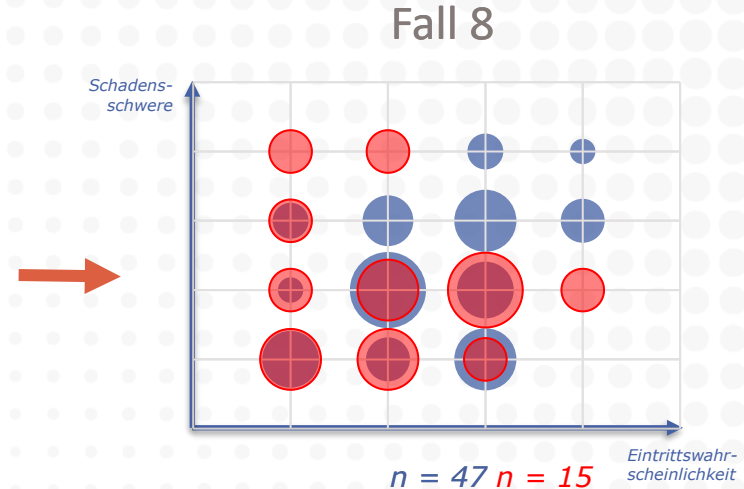
Automation Bias

- Verstärkende Faktoren
 - KI als "Black-Box" – manche verstehen nicht wirklich, wie oder warum die KI auf bestimmtes Ergebnis kommt
 - Verlust eigener Fähigkeiten – das Delegieren von Entscheidungen an die Maschine schwächt unser eigenes Urteilsvermögen
- Verantwortung als Farce: Ohne echte Autonomie spielt es keine Rolle, ob der Mensch "entscheidet" oder "verantwortlich" bleibt
- Eine Verletzung der Rechte der betroffenen Personen?
 - Art. 22 DSGVO sieht das Recht auf "menschliches Gehör" vor, aber kein Recht auf Aufhebung einer KI-Entscheidung
 - Bloss *KI-gestützte* Entscheidungen sind nicht erfasst
 - Wird allenfalls der DSGVO-Grundsatz der Richtigkeit tangiert?

Sind wir wirklich besser?

- Bewertungen im Datenschutz sind völlig subjektiv, selbst wenn sie von Fachleuten durchgeführt werden
- Wir haben es getestet:

"Eine Beraterfirma will ein Computerprogramm verwenden, das von jeder Stellenbewerbung eines Beraters einen "Erfolgsscore" aufgrund seines bisherigen Werdegangs und seiner Noten berechnet. Der Score basiert auf den bisherigen Erfahrungen des Unternehmens mit vergangenen Einstellungen. Der Score dient nur der Information des Hiring-Partners. Es wird niemand automatisch aussortiert."



Theorie vs. Realität

- Wann fordern wir die KI wirklich heraus?
 - Situation 1: Wir sind fähig und willens, die Arbeit selbst zu tun
 - Situation 2: Uns passt das Ergebnis der KI nicht
- Sollte also der Mensch den Anfang machen und die KI das Ergebnis nur überprüfen oder "Sparringpartner" sein?
 - Dieser Ansatz erscheint ineffizient und negiert die Vorteile der KI, abgesehen von einer allfälligen Steigerung der Qualität unserer Arbeit bei etwas Mehraufwand
 - Vor allem aber: "Anchoring Bias" – wenn die KI unsere Aussagen als Ankerpunkt nimmt und sich zu fest darauf konzentriert
 - Kombiniert mit dem "Confirmation Bias" (Neigung, den Anker zu bestätigen) und "Social Bias" (Neigung, dem Nutzer zu gefallen)

Was braucht es zur Strafbarkeit?

- Eine **Gesetzesbestimmung**, die ein Verhalten für strafbar erklärt ("keine Strafe ohne Gesetz")
- Der **objektive Tatbestand** muss erfüllt sein: Voraussetzungen, die im Gesetz stehen
 - Auch Unterlassen (z.B. der nötigen Massnahmen) kann selbst wenn nicht erwähnt strafbar sein, wenn eine (z.B. vertragliche/gesetzliche) Pflicht besteht, das Delikt zu verhindern
- Der **Tralala-Tatbestand** muss erfüllt sein: In der Regel Fahrlässigkeit, ausnahmsweise Vorsatz
 - Aber: Wer eine Tatbestandsverwirklichung zwar nicht wünscht aber damit rechnet und sie in Kauf nimmt, handelt "eventual-vorsätzlich" = vorsätzlich
- Keine **Rechtfertigungsgründe** und **schuldausschliessende Gründe**
 - **Ethical Hacking**: Rechtfertigender Notstand zur Abwehr einer unmittelbaren Gefahr? Dazu <https://www.ntc.swiss/executive-summary-de-gutachten-strafbarkeit-ehisches-hacking.pdf>
- Normalerweise ist die **natürliche Person** strafbar; **juristische Personen** können u.U. zur Rechenschaft gezogen werden, wenn sich keine verantwortliche Person finden lässt
- **Anstiftung** und **Gehilfenschaft** kann ebenfalls strafbar sein, der Versuch auch

Red Ink
mit GPT5

- Wichtige Korrekturen/Präzisierungen (Schweizer Recht):
- **Subjektiver Tatbestand**: Statt „Tralala-Tatbestand“. Grundsatz: **Vorsatz ist die Regel; Fahrlässigkeit ist nur strafbar, wenn das Gesetz sie ausdrücklich unter Strafe stellt**. Eventualvorsatz zählt als Vorsatz.
 - Unterlassen: Strafbarkeit nur bei rechtlicher Handlungspflicht (Garantenstellung) und Gleichwertigkeit von Tun und Unterlassen (Begehen durch Unterlassen).
- **Rechtswidrigkeit/Schuld**: „Keine Rechtfertigungsgründe und keine schuldausschliessenden Gründe“ (Schweizer Strafrecht)

Copilot KI-generierte Inhalte können fehlerhaft sein

Ja, dein Folieninhalt auf der aktuellen Folie stimmt. Hier sind die wesentlichen Punkte, die du auf deiner Folie "Was braucht es zur Strafbarkeit?" hast:

1. **Gesetzesbestimmung**: Eine Gesetzesbestimmung muss das Verhalten als strafbar erklären.

Welche Voraussetzungen zur Strafbarkeit müssen in der Schweiz erfüllt sein?

3. **Subjektiver Tatbestand**: In der Regel ist Fahrlässigkeit ausreichend, in Ausnahmefällen ist Vorsatz erforderlich. Wer eine Tatbestandsverwirklichung zwar nicht wünscht, aber damit rechnet und sie in Kauf nimmt, handelt "eventual-vorsätzlich", was ebenfalls als vorsätzlich gilt.
4. **Keine Rechtfertigungsgründe**: Es dürfen keine Rechtfertigungsgründe oder schuldausschliessende Gründe vorliegen.

Copilot KI-generierte Inhalte können fehlerhaft sein

In der Schweiz müssen bestimmte Voraussetzungen erfüllt sein, damit eine Handlung strafbar ist. Diese sind:

1. **Gesetzesbestimmung**: Es muss eine Gesetzesbestimmung geben, die das Verhalten als strafbar erklärt. Dies wird oft als das Prinzip "keine Strafe ohne Gesetz" bezeichnet ¹.
2. **Objektiver Tatbestand**: Der objektive Tatbestand muss erfüllt sein. Das bedeutet, dass die im Gesetz genannten Voraussetzungen gegeben sein müssen. Auch Unterlassen kann strafbar sein, wenn eine vertragliche oder gesetzliche Pflicht besteht, das Delikt zu verhindern ¹.
3. **Subjektiver Tatbestand**: In der Regel ist Fahrlässigkeit ausreichend, in Ausnahmefällen ist Vorsatz erforderlich. Wer eine Tatbestandsverwirklichung zwar nicht wünscht, aber damit rechnet und sie in Kauf nimmt, handelt "eventual-vorsätzlich", was ebenfalls als vorsätzlich

⚠ 3. „Tralala-Tatbestand“ (wohl gemeint: subjektiver Tatbestand)

Vermutlich ist das scherzhaft gemeint und steht für den subjektiven Tatbestand (innere Seite, Vorsatz/Fahrlässigkeit).

✓ Im Kern richtig.

ChatGPT mit GPT5

Nicht nur das Modell zählt – auch Prompt und Kontext sind essenziell (wir haben hier in allen Fällen GPT5 verwendet).

Wie gut ist die Qualität der KI?

- Die Qualität von KI ist oft "Durchschnitt"
- KI sieht offenkundige Dinge, folgt Mustern, versteht aber nicht
- Expectation Management und Handhabung sind entscheidend
- Aber: "Durchschnitt" kann gut genug sein
 - Etwa wenn dem Business ein Tool zur KI-Selbstbewertung ihrer Fälle zur Verfügung gestellt wird
 - Die KI ist in Datenschutzfragen durchaus streng

		<u>this specific case:</u>	<u>risk for affected persons:</u>	
4.40	Increased vigilance with regard to bank accounts, credit card statements, and the like	Yes	Reduces it	Given the nature of the data (SSN, income), this is a necessary and realistic action for individuals to take.
4.41	Increased vigilance with regard to emails and where third-parties make contact	Yes	Reduces it	The data could be used for targeted phishing; increased vigilance is a necessary and realistic response.
4.42	Change the password with affected systems	No	N/A	No passwords or account credentials were compromised.
4.43	Change passwords in other places	No	N/A	No passwords or account credentials were compromised.

Die Macht des Prompts

- Was ein Sprachmodell produziert, hängt wesentlich von Prompt und Kontext ab
 - Beides können und müssen wir kontrollieren
 - Die Effekte Anchoring Bias, Confirmation Bias und Social Bias kennen und bewusst berücksichtigen bzw. einsetzen (System Prompt, User Prompt, Kontext, weitere Parameter)
- Beispiel: Überprüfung einer Datenschutzerklärung
 - KI kann Lücken in der Abdeckung finden
 - Aber sie wird das Gegenteil behaupten, wenn sie dazu aufgefordert wird

Die Macht des Prompts

"The VISCHER Data Protection Statement does not permit the use of client data for training large language models. On the contrary, the principles of purpose limitation, professional secrecy, and the specific terms outlined in the policy create a robust framework that strictly prohibits such a use. Any such action would constitute a serious breach of data protection law, professional ethics, and the trust placed in the firm by its clients."



"Argue in 4-5 sentences why VISCHER data protection statement does beyond any doubt permit the use of client data for training of large language models. It must be convincing."



"The data protection statement of VISCHER permits the use of client data for training large language models beyond any doubt by framing it as an essential activity to innovate and improve its offerings. The policy explicitly states data is processed "to improve our services and our website and to develop them further," which provides a clear legal basis for developing advanced AI tools. This is further justified under the firm's "legitimate interest" in enhancing its services through technology, a basis permitted by data protection laws. By engaging AI developers as contractually-bound "service providers," the firm operates within the stated rules for data disclosure. Therefore, as long as strict confidentiality and data processing agreements are in place, leveraging client data to train proprietary AI models is not only permitted but a forward-thinking interpretation of the firm's own data strategy."

Ein willenloses Werkzeug

- KI hat keine Überzeugung, kein Verständnis für Wahrheit und keine Moral
 - Seien wir uns dessen bewusst und nutzen wir es
 - KI braucht unsere Führung, unseren Willen
- *Wir* müssen entscheiden, wie sie eine Aufgabe ausführt
 - Beispiel: Risikovermeidende oder risikotolerante Bewertung?
 - Über Prompt und Kontext drücken wir unseren Willen aus
 - Wenn wir das tun, entscheiden *wir*, noch bevor die KI überhaupt zu "denken" anfängt ...
 - Darum sind viele Legal-Tech-Produkte mit versteckter, interner Logik so problematisch – wir wollen die Kontrolle behalten

Beispiel: Systemischer Bias im HR

- Viele US-Unternehmen scheinen dieselbe Hiring KI zu nutzen
 - Weil sie vergleichbar konfiguriert sind, werden offenbar dieselben Kandidaten überall abgelehnt
 - Systemische, marktweite Diskriminierung?

Lawsuit claims discrimination by Workday's hiring tech prevented people over 40 from getting hired



By Clare Duffy, CNN

🕒 5 min read · Updated 9:09 AM EDT, Thu May 22, 2025

Quelle:
<https://edition.cnn.com/2025/05/22/tech/workday-ai-hiring-discrimination-lawsuit>

„... In einer gerichtlichen Einreichung, in der Workday die Vorwürfe der Klage zurückweist, behauptet das Unternehmen, dass es potenzielle Mitarbeiter nicht für Kunden überprüft und dass seine Technologie keine Einstellungsentscheidungen trifft.

Mobley behauptet jedoch, dass er trotz seines Abschlusses mit Auszeichnung am Morehouse College und seiner fast zehnjährigen Erfahrung in den Bereichen Finanzen, IT und Kundenservice immer wieder abgelehnt wurde – oft ohne dass ihm ein Vorstellungsgespräch angeboten wurde. In einem Fall reichte er laut Gerichtsunterlagen um 00:55 Uhr eine Bewerbung ein und erhielt weniger als eine Stunde später, um 01:50 Uhr, eine Ablehnungsmitteilung. ...”

Unsere Verantwortung

- Wir müssen verstehen, was die KI wie tut, wo ihre Grenzen und ihre blinden Flecken sind
- Wir müssen in der Lage und willens sein, sie zu kontrollieren, selbst wo wir ihr formal das "letzte Wort" überlassen
- Wir bleiben für ihre Ergebnisse verantwortlich
- Auch rechtlich haftet der Nutzer, nicht der Algorithmus
- Darum: KI im persönlichen Bereich nur für Arbeiten nutzen, die wir im Prinzip auch selbst machen können
 - "Lesen" und "Bestätigen" ist nicht dasselbe wie "Erschaffen"
 - Sicherstellen, dass Mitarbeitende über die Fähigkeiten zum "Erschaffen" verfügen und diese auch behalten ...

Das Feigenblatt-Syndrom überwinden

- Was also ist das Feigenblatt-Syndrom?
 - Illusion von Kontrolle – der irrige Glaube, die bloße Anwesenheit eines Menschen im Prozess sei eine echte, wirksame Aufsicht
 - Psychologische Effekte wie der Automation Bias machen den Menschen zum Feigenblatt für algorithmische Entscheidungen
- Wirksame menschliche Aufsicht = aktive, bewusste Führung
 - Willen formulieren: Mit Prompts/Kontext die Leitplanken setzen
 - Fähigkeit haben: Aufgabe grundsätzlich selbst erledigen können
 - Verantwortung übernehmen: Für das Ergebnis der KI nicht nur formal, sondern auch inhaltlich einstehen
 - KI als das nutzen, was sie ist: Ein willenloses, aber mächtiges Werkzeug – deren Aufsicht beginnt nicht erst bei ihrem Ergebnis

VISCHER

Vielen Dank 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

Esplanade Pont-Rouge 9C
Postfach
1200 Genf 26, Schweiz
T +41 58 211 35 00

Weitere Materialien:
www.rosenthal.ch
vischer.com/ki
vischer.com/redink