

# VISCHER

## The AI Revolution. How to Talk AI to Your Lawyers



David Rosenthal, Partner, VISCHER Ltd.  
October 22, 2024

---

# This is how some lawyers feel about AI ...



## This is how some lawyers feel about AI ...

### AI Act Adopted: introduction to the 458-page EU legislation

elvingerhoss.lu

### Some personal reflections on the EU AI Act: a bittersweet ending



**Kai Zenner**  
Head of Office and Digital Policy  
Adviser for MEP Axel Voss (EPP...)

<https://www.linkedin.com/pulse/some-personal-reflections-eu-ai-act-bittersweet-ending-kai-zenner-avgee/>

- Secondly, the final text is not fulfilling the law's key objective of providing legal certainty or an ecosystem of trust as the European Commission calls it. On the contrary, most definitions in Article 3 are vague, the procedures are incomplete and legally questionable (i.e. designation of systemic GPAI

Combined, those three issues could significantly raise compliance costs for providers and deployers of AI. Especially SMEs and start-ups from the EU might find it in the end too risky to develop or deploy AI ... or they are forced to draw on expensive third-party auditing and certification schemes in order to prevent heavy fines. If this

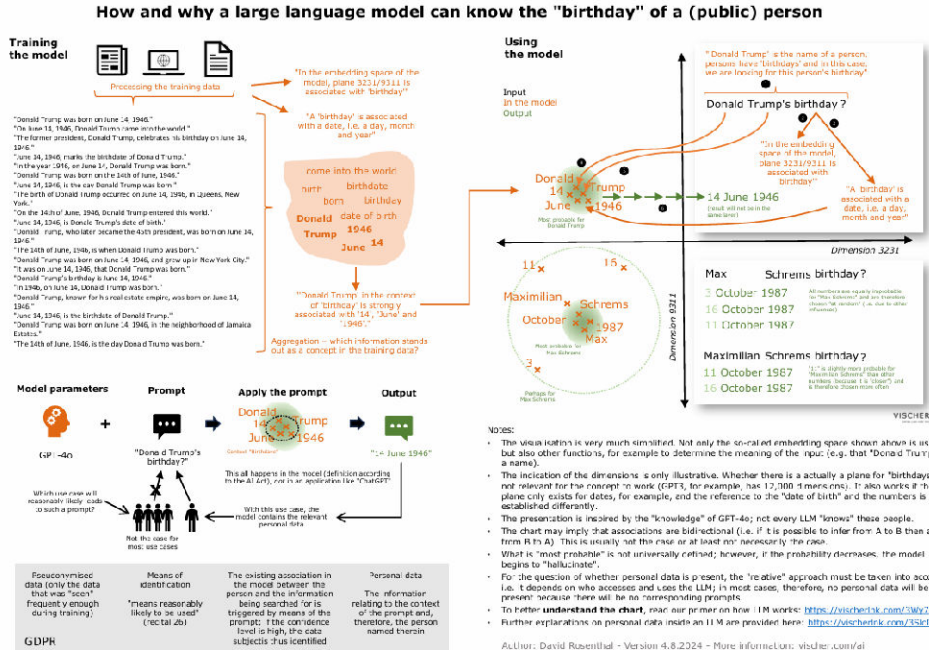
## Typical reactions from "legal" ...

- "This AI project idea of yours will have to wait."
- "We are not yet allowed to use this with personal data."
- "There are still too many unresolved legal issues."
- "Are we sure that there is no bias?"
- "We do not yet have the necessary provider contracts."
- "Are we sure that the model has not been trained illegally?"
- "Can you explain how exactly the AI comes up with this result?"
- "Can you ensure that the chatbot cannot be abused and does not make any wrong statements?"
- "The output can infringe third-party rights."

## Three things to do

- **Get aligned and bridge the gaps**
  - Educate your lawyers about AI
  - Educate yourself about the law
  - Accept that some aspects remain unclear
- **Help build a risk-based approach**
  - Establish risk categories
  - Establish risk mitigation measures
  - Establish risk assessments
- **Invest in proper governance**
  - Even if it worked well in the past without it ...

# Educate your lawyers about AI

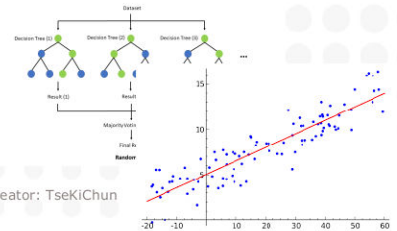


How does a large language model work and what is really stored in it?  
[vischerlnk.com/3WY7gGQ](https://vischerlnk.com/3WY7gGQ)

Does a large language model contain personal data?  
[vischerlnk.com/3SlcIum](https://vischerlnk.com/3SlcIum)

## Educate your lawyers about AI







- They have **no clear** understanding of what AI is
  - Is it a copying machine since OCR is based on a neural network?
- As per the EU **AI Act** "a machine-based system that is designed to **operate with varying levels of autonomy** and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments"
  - The only practically relevant element is "**autonomy**"
  - In simple terms: An IT system that has been **trained** on how to decide, not only using programmed logic ...
  - But to which applications in your company does this apply?




# Educate your lawyers about AI

*Are your AI applications prepared for them?*

## Six ways to attack an AI system.

					
<b>Poisoning</b>	<b>Trojan Horse</b>	<b>Prompt Injection</b>	<b>Sponge Attack</b>	<b>Model &amp; Data Theft</b>	<b>Deception</b>
<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: vischerflk.com/ai-attacks

 VISCHER  
1900 1000 1000

Have you discussed security issues with your lawyers?  
They may have a misunderstanding of the real issues in your use cases

[vischerlnk.com/3OPTpaA](https://vischerlnk.com/3OPTpaA)



# Educate yourself about the law

## Checklist: 18 Key AI Compliance Issues.

Go to [vischer.com/ai](https://vischer.com/ai) for free resources on the issues below and on AI governance & risk management (no registration required)

AI = any system that produces output on the basis of training instead of only programming

### Data Protection

- Do we have a proper contract when using a provider (e.g., a DPA, EU SCC, no own use of our data)?
- Do we tell people about the purposes for which we use their data or create data about them?
- Do we have measures in place if the AI produces wrong or otherwise improper data about them?
- When an AI makes important decisions about them, can they have it reviewed by a person?
- Is our AI protected against misuse, attacks and other security issues, in particular if we allow third parties to use it (e.g., chatbot)?
- Can we honor access and correction requests?
- Have we done a risk assessment (incl. DPIA)?

### Contractual Commitments, Secrecy

- Do we comply with our secrecy obligations (e.g., when using providers, data leakage prevention)?
- Do any of our contracts prohibit our intended use case (e.g., NDA that also restricts use of data)?

### Third-Party Content Protection

- Do we feed third-party content to AI systems only where our licenses or "fair use" rules permit it?
- Do we avoid generating content that resembles pre-existing content of third parties?

### EU AI Act (not yet in force)

- Do we make sure we are either not subject to the AI Act or what we do is not a prohibited practice and, if possible, also not a "high risk" AI system (and do we otherwise deal with it properly)?
- Where an AI creates deep fakes or interacts with or watches people, are they made aware of this?

### Other (also ethical) Aspects

- Do we avoid discrimination when using AI?
- Do humans (really) keep control over the use of AI?
- Does our AI generate output we can justify/explain?
- Do we tell people how we use AI where it may be unexpected and allow them to opt-in or opt-out?
- Do we have adequate testing, monitoring and risk management of AI?

[vischerlnk.com/ai-compliance-short](https://vischerlnk.com/ai-compliance-short)

Copyright is often not an issue when using common sense

AI Act is focused on product safety for higher risk use cases

Other stuff that you will be asked to provide answers to

The usual stuff when dealing with personal data – make sure you keep control of it, in particular when using third-party providers

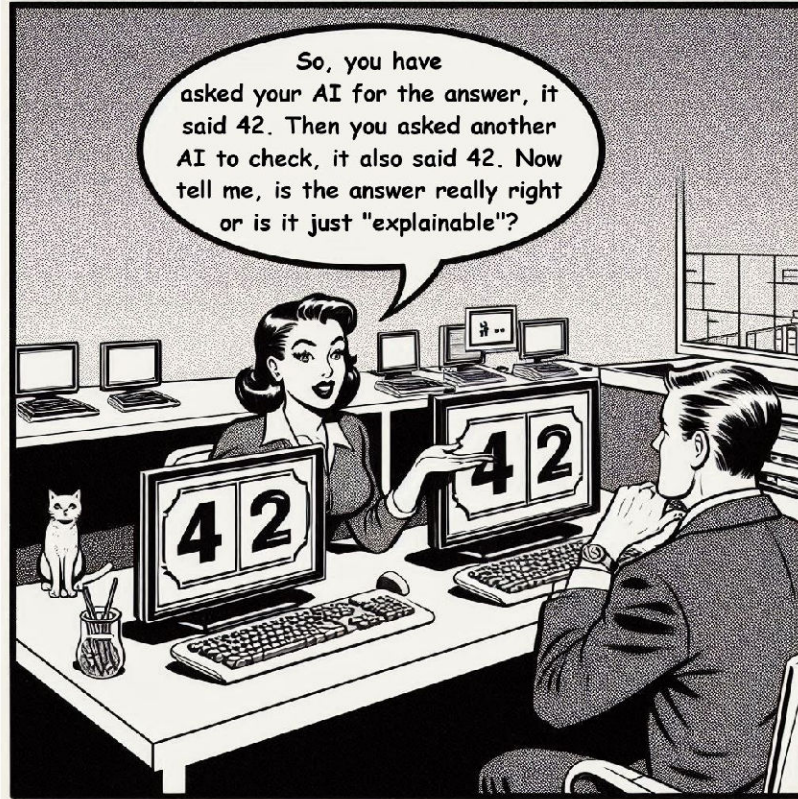
This is critical – to whom do you disclose highly confidential customer data?

Author: David Rosenthal ([david.rosenthal@vischer.com](mailto:david.rosenthal@vischer.com)) All rights reserved. For information purposes only (focused on European law), 16.5.24 Updates: [vischerlnk.com/ai-compliance-short](https://vischerlnk.com/ai-compliance-short)



VISCHER  
AN DER UNIVERSITÄT ZÜRICH





Accept that for some questions,  
there are no clear, established  
answers ...

Example: Explainability

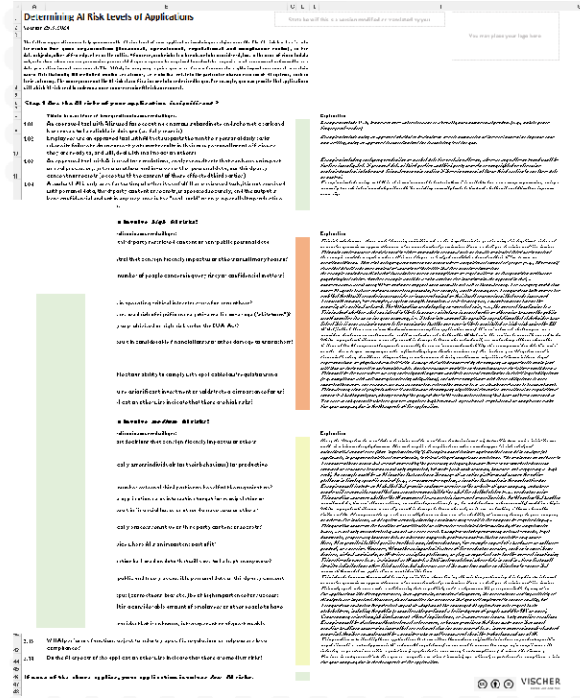
# Help building a risk-based approach

- It's primarily about **risk management**
- Do not expect or go for full compliance with all laws

**Step 1: Are the AI risks of your application insignificant?**

This is the case if one of these questions is answered with yes:

- 1.01 An approved tool with AI is used for executing a narrow, subordinate and schematic task and has proven to be reliable in doing so (we fully trust it)
- 1.02 Employees use an approved tool with AI that supports them in their personal daily tasks where its failure to do so correctly at worst results in their own personal loss of efficiency they are ready to, and will, deal with (no impact on others)
- 1.03 An approved tool with AI is used for simulations, analyses and tests that can have no impact on real processes, systems or others and involve neither personal data, nor third party content nor secrets (except with the consent of these affected third parties)
- 1.04 A tool with AI is only used for testing whether it can fulfill an envisaged task, it is not provided with personal data, third-party content or secrets, is operated securely, and the output is kept confidential and not in any way used in the "real world" or only in parallel to productive systems and not relied upon.



# Help building a risk-based approach

- Know the **"legal"** triggers
  - Biometrics, emotions, employees
  - (Sensitive) personal data
  - Professional and company secrets
  - Copyrighted content
  - (Semi-)automated key decisions
  - Third-party service providers
  - Serving third parties with AI
- Try to establish **"sandboxes"** for AI testing and experimenting
- Come up with **risk mitigation measures** on your own

The screenshot displays a comprehensive risk assessment tool. At the top right, a 'Risk Radar' chart visualizes risk levels across various dimensions. Below it, a table titled 'Operational risks to be assessed' lists risks such as 'Data security' and 'Data integrity' with associated scores and mitigation measures. The bottom section features a 'Compliance housekeeping questions' table with columns for 'Assess', 'Reason for concern', 'Mitigation', 'Assessment', 'Justification Comment', 'By whom?', and 'Risk Rating'. The interface includes navigation icons on the right side.

## Push for proper governance

- **Five steps** to take
  - Set forth the responsibilities and procedures in relation to AI
  - Set forth the substantive rules to apply when using of AI
  - Train for safe, legal and responsible use of AI, and provide for AI literacy – up to the board
  - Map and track your use of AI – and assess it
  - Include AI in your risk management, include all stakeholders
- Good governance is **half the battle**
  - Document what was considered by whom and reason decisions
- The **business decides** – lawyers ("2<sup>nd</sup> line") only advise

# Get them involved early on ...

Privatkunden Unternehmenskunden Institutionelle Anleger Über uns DE

**baloise**

Versichern Firma gründen Konten, Karten & Finanzierung Anlegen Nachhaltigkeit Kontakt & Service

**Der Chief LOL Officer**  
Laut lachen – gesund arbeiten

- Erkennen**  
Mit KI erkennt der Chief LOL Officer die Stimmung der Mitarbeiter.
- Messen**  
Anhand der Intensität des Lachens über einen bestimmten Zeitraum wertet das Gerät die Stimmung im Büro aus.
- Lachen**  
Wenn nicht gelacht wird, teilt der Chief LOL Officer lustige Inhalte aus dem Internet mit den Mitarbeitenden, um allen eine Lachpause zu ermöglichen.

The following AI practices shall be **prohibited**:  
... the placing on the market, the putting into service for this specific purpose, or the **use of AI systems to infer emotions** of a natural person **in the areas of workplace** and education institutions, except where the use of the AI system is intended to be put in place or into the market for medical or safety reasons;

Maybe they have some good ideas on how to deal with such hurdles ...

# VISCHER

Thank you for your attention!

Questions: [david.rosenthal@vischer.com](mailto:david.rosenthal@vischer.com)

## **Zürich**

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

[www.vischer.com](http://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

For in-depth  
materials on the  
topic visit us at  
[vischer.com/ai](http://vischer.com/ai)