Between legal and (non)responsible AI

Responsible Machine Learning Winter School, Umeå 21-13 March, 2025

Markus Naarttijärvi

Professor of law

UMEÅ UNIVERSITY

TAIGA

Transdisciplinary Artificial Intelligence
– for the Good of All

”It is essential that public administrations, hospitals, utility and transport services, financial supervisors, and other areas of public interest rapidly begin to deploy products and services that rely on AI in their activities.”

European Union White Paper On Artificial Intelligence
- A European approach to excellence and trust

”Aside from the many beneficial uses of artificial intelligence, that technology can also be misused and provide novel and powerful tools for manipulative, exploitative and social control practices. Such practices are particularly harmful and should be prohibited because they contradict Union values of respect for human dignity, freedom, equality, democracy and the rule of law…”

Proposed European Union Artificial Intelligence Act, Recital 40

We expected that at least some of the authors would argue that algorithmic threats to the rule of law were solvable, or that responsibly-implemented algorithms could even help the delivery of justice. None of the experts did.”

Meyer-Resende & Straub, Verfassungsblog, 2022

The New York Times

British Grading Debacle Shows Pitfalls of Automating Government

The uproar over an algorithm that lowered the grades of 40 percent of students is a sign of battles to come regarding the use of technology in public services.

Give this article
42
YOURE

AI skin cancer diagnoses risk being less accurate for dark skin – study

Research finds few image databases available to develop technology contain details on ethnicity or skin type

China is now using facial recognition cameras to monitor Uighur Muslims across the country, report claims

- Authorities in China are using AI cameras to track its Uighur Muslim minority
- CCTV cameras have been programmed to look for Uighurs based on appearance
- Beijing has long been criticised for its treatment of Uighurs in Xinjiang

: 22:32 GMT, 15 April 2019

Here's What Happens When Your Lawyer Uses ChatGPT

A lawyer representing a man who sued an airline relied on artificial intelligence to help prepare a court filing. It did not go well.

An automated policing program got this man shot twice

May 24, 2021 ... Robert McDaniel poses for a portrait at his home in Chicago's Austin neighborhood. Heat Listed. By Matt Stroud | May 24, 2021, 10:00am EDT.

UK police use of live facial recognition unlawful and unethical, report finds

deployment of technology in public by Met and South
e failed to meet standards

GOOGLE / TECH / ARTIFICIAL INTELLIGENCE

Google's AI 'Reimagine' tool helped us add wrecks, disasters, and corpses to our photos

FROM POLITICOPR

Dutch scandal serves as a warning for Europe over risks of using algorithms

The Dutch tax authority ruined thousands of lives after using an algorithm to spot suspected benefits fraud – and critics say there is little stopping it from happening again.

Google given access to healthcare data of up to 1.6 million patients

Artificial intelligence firm DeepMind provided with patient information as part of agreement with Royal Free NHS trust

The New York Times

'Deepfake Elon Musk' Replacing Meaningless Jobs Rapid Weather Forecasts A.I.'s Math Problem

LIBERAL CHATBOT
I don't know if it's possible for a conservative to be reasonable.

CONSERVATIVE CHATBOT
The left is trying to destroy our country. Conservatives are trying to save it.

See How Easily A.I. Chatbots Can Be Taught to Spew Disinformation

By Jeremy White May 19, 2024

Ahead of the U.S. presidential election this year, government officials and tech industry leaders have warned that chatbots and other artificial intelligence tools can be easily manipulated to sow

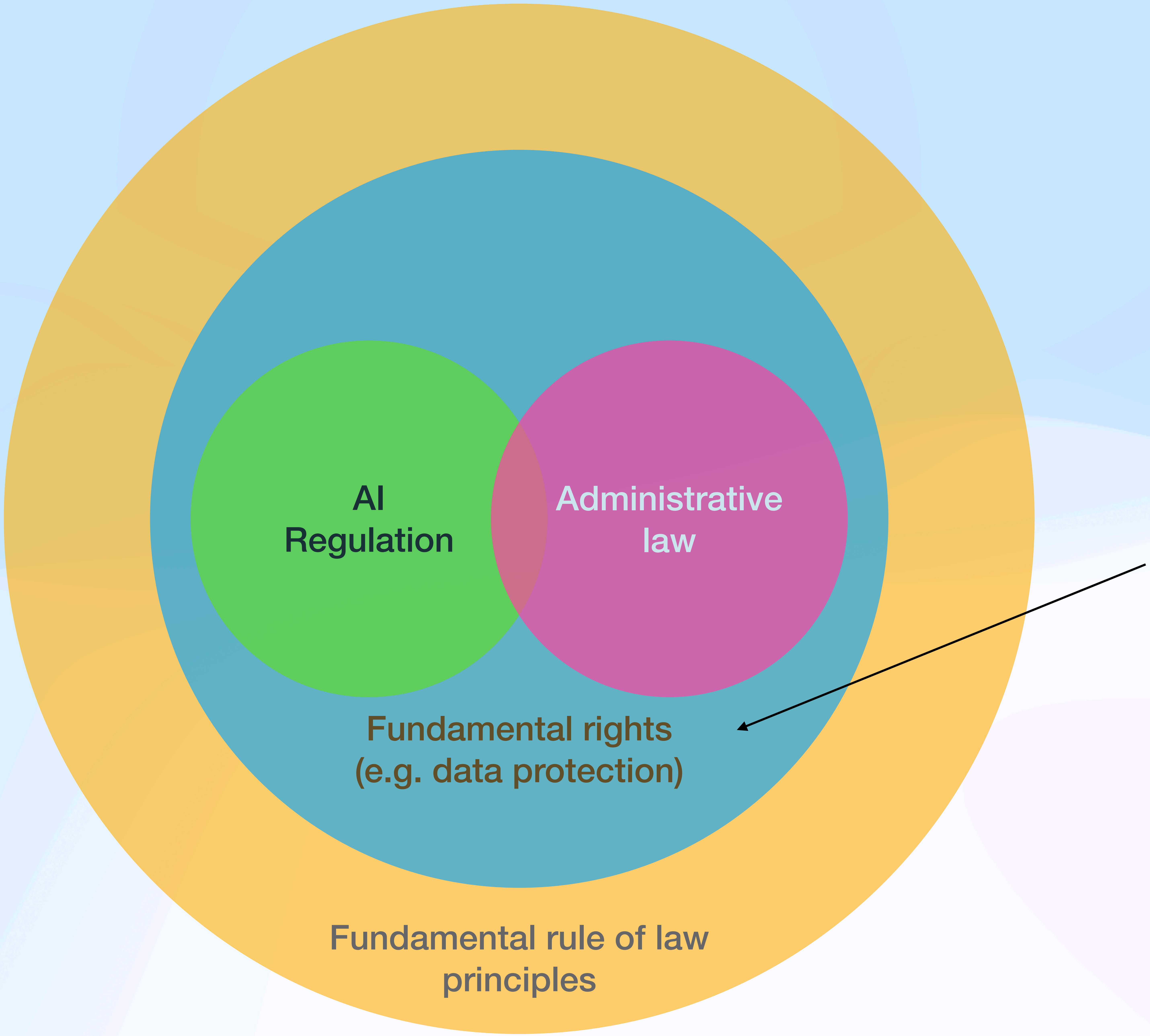
EU Law is the current gold standard of data protection and responsible AI regulation.

Ability to meet or exceed legal requirements in the EU can prove a competitive advantage in highly regulated sectors (medicine, critical infrastructure)

Data protection frameworks helps you make choices aligning with human centered AI

Potential sanctions under the GDPR, 20 000 000 EUR, or up to 4 % of the total worldwide annual turnover of the preceding financial year, whichever is higher:

Potential sanctions under the AI-Act administrative fines of up to 35 000 000 EUR or, up to 7 % of its total worldwide annual turnover for the preceding financial year, whichever is higher.



AI Regulation

Administrative law

Fundamental rights (e.g. data protection)

Fundamental rule of law principles

Today, we’ll be here

A (very) quick primer on data protection

- GDPR does not regulate machine learning as such.
- **But**. GDPR applies whenever personal data is *processed* (e.g. collected, transformed, consulted, erased) either within the union or relating to an EU resident.
- Fundamental question: **can a person be identified**, directly, or indirectly?
- **Data subject**: The person the data relates to.
- **Data controllers**: The entity that determines the purposes and means of the processing.
- Data can only be processed based on a lawful basis, for a specified and limited purpose, and for a limited period.

Training data as personal data

Special category data

’Responsible’ training data

Are models anonymous?

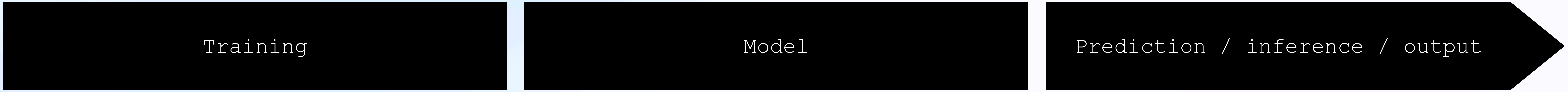
Regurgitation and inversion attacks

Legal risk assessment & mitigation

Contestability

Opacity

Meaningful information



Training

and use of training data

Training data as personal data

Points of departure from EU data protection law

- Definition of personal data under EU data protection rules is extremely wide – *any information relating to an identified or identifiable natural person.*

- Persons can be identified through aggregated data.

- Everything is processing: collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction…

- Fully anonymised data is not personal data, but:

- Avoidance of personal data in large training sets unless carefully controlled is… unlikely.

Sensitive category data in training sets

Difficulties in avoiding prohibited practices

- Processing of certain categories of data are **prohibited by default** – Personal data *revealing:*

  Racial or ethnic origin; political opinions; religious or philosophical beliefs; trade-union membership; data concerning health or sex life; genetic data; biometric data for the purpose of uniquely identifying a natural person; data concerning a natural person's sexual orientation; + “data relating to offences, criminal convictions or security measures”

  - Where a set of data containing both sensitive data and non-sensitive data is collected en bloc without it being possible to separate the data items from each other at the time of collection, the processing of that set of data must be regarded as being prohibited under the GDPR if it contains at least one sensitive data item – unless an exception applies.

- Possible exception for ’data manifestly made public by the data subject’ themselves, but:

  - ’it is important to ascertain whether the data subject had intended, explicitly and by a clear affirmative action, to make the personal data in question accessible to the general public’

  - the mere fact that personal data is publicly accessible does not imply that the data subject has manifestly made such data public

— CJEU judgment of 4 July 2023, Case C-252/21, Meta v. Bundeskartellamt

Legal basis for processing training data

A baseline requirement…

- Legal basis necessary for processing training data – usually these two become relevant:
  - **Consent from the data subject** with explicit and informed consent for the use in training.
    - Clear requirements on a separate approval of this purpose of processing, with far-reaching information to be provided to data-subjects + right to revoke consent.
  - **Legitimate interest of the data controller**
    - This entails a three step test; 1) is it legitimate, 2) is it necessary, and 3) do the rights and interests of the data subject override the interest of the controller?

Legitimate interest to process training data?

Principles from data protection

1. Is it a legitimate interest?

- Is it lawful, clearly and precisely articulated, and real and present?

2. Is it necessary?

- Are there no less intrusive way of pursuing this interest? Is the amount of personal data processed proportionate to the interest at stake, in light of the data minimisation principle?

3. Do the rights and interests of the data subject override the interest of the controller?

- Are there specific risks to fundamental rights that may emerge either in the development or the deployment phases of AI models?

- What is the nature of the data processed by the models, the context of the processing and the possible further consequences of the processing?

- What are the reasonable expectations of individuals whose data is being processed?

Mitigating measures when using web-scraping

Recommendations from the EDPB

1. **Excluding data content from publications which might entail risks for particular persons** or groups of persons (e.g. individuals who might be subject to abuse, prejudice or even physical harm).

2. **Ensuring that certain data categories are not collected** or that certain sources are excluded; e.g., websites that are particularly intrusive due to the sensitivity of their subject matter.

3. **Excluding collection from websites** (or sections of websites) **which clearly object** to web scraping by respecting robots.txt or ai.txt files or or similar expressions of objections.

4. Imposing other relevant limits on collection, possibly including criteria based on time periods.

5. **Creating an opt-out list**, managed by the controller and which allows data subjects to object to the collection of their data on certain websites or online platforms, including before the data collection occurs.

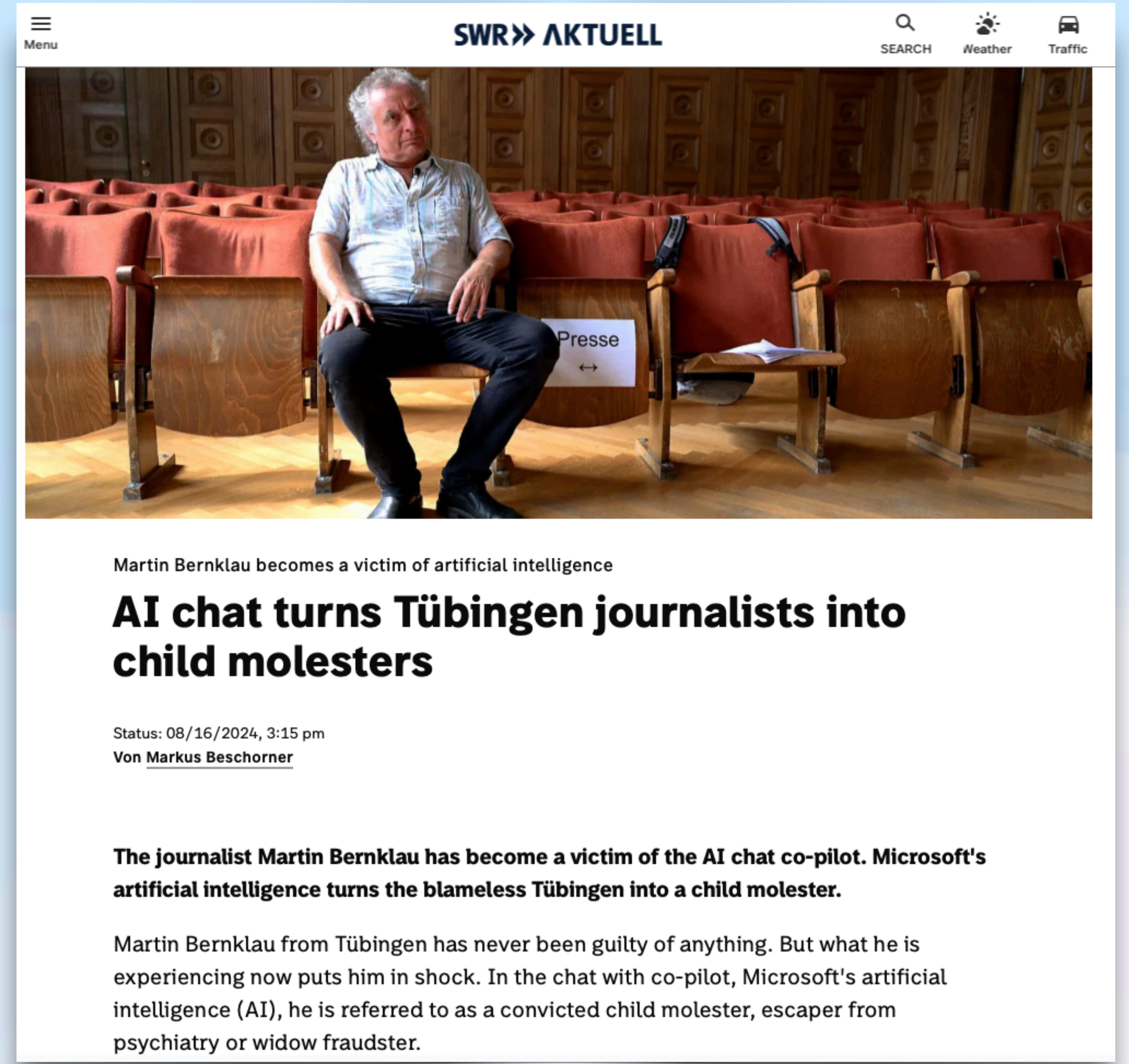
Model

and deployment

Points of departure

Do models contain personal data?

- EDPB: ”even when an AI model has not been intentionally designed to produce information relating to an identified or identifiable natural person from the training data, **information from the training dataset, including personal data, may still remain ‘absorbed’ in the parameters of the model**, namely represented through mathematical objects.”
- They may differ from the original training data points, but may **still retain the original information** of those data, which may ultimately be extractable or otherwise obtained, directly or indirectly, from the model.
- Whenever information relating to identified or identifiable individuals whose personal data was used to train the model may be obtained from an AI model with means reasonably likely to be used, it may be concluded that **such a model is not anonymous**.



Martin Bernklau becomes a victim of artificial intelligence

AI chat turns Tübingen journalists into child molesters

Status: 08/16/2024, 3:15 pm
Von Markus Beschorner

The journalist Martin Bernklau has become a victim of the AI chat co-pilot. Microsoft's artificial intelligence turns the blameless Tübingen into a child molester.

Martin Bernklau from Tübingen has never been guilty of anything. But what he is experiencing now puts him in shock. In the chat with co-pilot, Microsoft's artificial intelligence (AI), he is referred to as a convicted child molester, escaper from psychiatry or widow fraudster.

Regurgitation and inversion attacks

Assessing the data protection risks

- EDPB highlights risks of extracting personal data from the model through **membership inference** attacks or **model inversion attacks**.
- Regurgitation of personal data from training can lead to the model being seen as processing of personal data independently of the processing of training data.
- ”Consultation” of personal data is seen as processing, so end users being able to make a model regurgitate personal data through prompts or attacks is enough.
- Knowing a person exist in a training data set can be sensitive, if belonging to the category the model is trained on is sensitive (e.g. crime data, mental illness)
- ”AI models trained on personal data cannot, in all cases, be considered anonymous. Instead, **the determination of whether an AI model is anonymous should be assessed, based on specific criteria, on a case-by-case basis”**. (EDPB)

Legal risk assessment for models

EDPB recommendations

- **Ensure a proper risk evaluation has been made** (and documented) assessing the identification risk.
- Take into account all means reasonably likely to be used by (anyone) to identify individuals or extract personal data. (include: characteristic of training data; context of deployment; additional information needed; cost and time needed; available technology and developments)
- AI models are likely to require a thorough evaluation of the likelihood of identification to reach a conclusion on their possible anonymous nature, and should also consider unintended (re)use or disclosure of the model.
- For example: A publicly available model will likely imply a much higher risk than a private model only accessible to a limited number of employees.

Risk mitigation in AI model design

EDPB recommendations

- Take steps to avoid or **limit the collection of personal data** and document these steps, such as selection criteria; the relevance and adequacy of the chosen sources; and how inappropriate sources have been excluded.
- Has anonymous or pseudonymised data been considered? If not, document the reasons for this decision, taking into account the intended purpose.
- **Document data minimisation strategies** and techniques employed to restrict the volume of personal data included in the training process; inducing data filtering processes implemented prior to model training intended to remove irrelevant personal data.
- Choose robust methods for AI model development, using methods that reduce or eliminate the identifiability, including regularisation methods to improve model generalisation and **reduce overfitting and use effective privacy-preserving techniques** (e.g. differential privacy).
- Implement and document measures added to the AI model itself which might lower the likelihood of obtaining personal data related to training data from queries.

The crux of non-anonymous models

Difficult rights to live up to

- **Information** – data subjects have a right to know what information about them you process. Might be impossible to compile.
- **Correction** – data subjects have a right to have erroneous information corrected. Could be impossible to fix without retraining model for each correction.
- **Erasure** – data subjects have a right to get their data erased (with certain limitations). Might be impossible without retraining model and purging training data.
- Fixes at the output or prompt end (correcting specific inputs/outputs) not a full-proof solution.

Output

Inferences, predictions, decisions

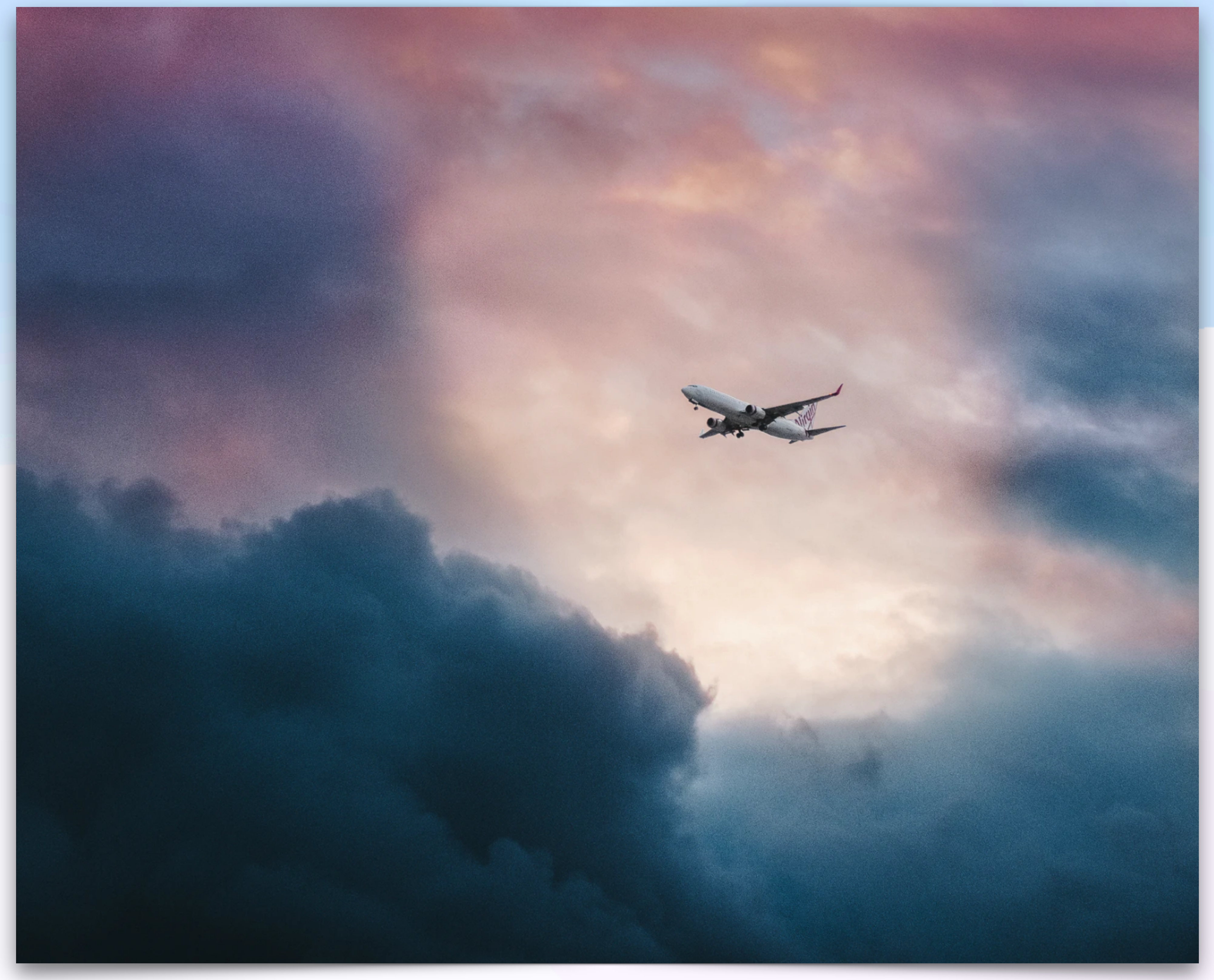
”Applying a norm to a human individual is not like deciding what to do about a rabid animal or a dilapidated house. It involves paying attention to a point of view. As such it embodies a crucial dignitarian idea — respecting the dignity of those to whom the norms are applied as beings capable of explaining themselves.”

Jeremy Waldron, 2020

EU Court of Justice approach to machine learning

C-817/19, Ligue des droits humains mot Conseil des ministres (21 June 2022)

- EU Court of Justice case on the directive on passenger name records (PNR data)
- PNR data is used to create risk profiles on flight passengers.
- AI system to predict risk.
- According to the directive, the risk assessment should be made based on pre-determined and non-discriminatory criteria.
- Any output that determines a risk is manually reviewed and no decisions with negative legal consequences are allowed to be taken based solely on the automatic processing of PNR data or sensitive personal data.



On the opacity of machine learning

Position of the Court of Justice of the European Union

- A law requiring pre-determined criteria precludes the use of machine learning capable of modifying the assessment process without human review. In particular, the assessment criteria on which the result of the application of that process is based as well as the weighting of those criteria.

- The use of such technology would be liable to render redundant the human review of decisions and the control of lawfulness of decisions.

- Given the opacity which characterises the way in which AI technology works, it might be impossible to understand the reason why a given program arrived at a positive match.

- The use of such technology may deprive the data subjects also of their right to an effective judicial remedy, in particular in order to challenge the non-discriminatory nature of the results obtained.

Meaningful information on logic

Automated decision-making and profiling

- GDPR Art. 15 establishes a right to ’meaningful information’ about the logic underpinning automated decisions.
- ’Decisions’ is interpreted broadly, may include probability values that form the basis for decisions on e.g. credit scoring, or e-recruitment practices.
- EU Court of Justice has established that ’meaningful’ information implies both good intelligibility and the value/usefulness of the information.
- Covers all relevant information concerning the procedure and principles relating to the automated use of personal data with a view to obtaining a specific result.
- In addition, information concerning the importance and the intended consequences of the processing for the data subject.
- Data and information should be provided in a concise, transparent, intelligible and easily accessible form, using plain and clear language – should be accompanied by ‘real, tangible examples’.

See Case C-203/22, Dun & Bradstreet; Case C-817/19. Ligue des Droits Humains; Case C-634/21, Chufa holdings

- Cannot be satisfied either by showing a complex mathematical formula, such as an algorithm, or by the detailed description of all the steps in automated decision-making – would not constitute a sufficiently concise and intelligible explanation.

- Should enable data subjects to ensure the correctness and lawfulness of processing – connected to the right of rectification and erasure, as well as the importance of effective judicial remedies.

- Should describe the procedure and principles actually applied in such a way that the data subject can understand which of his or her personal data have been used in the automated decision-making at issue, and how a variation in the personal data taken into account would have led to a different result.

- Complexity of the operations to be carried out in the context of automated decision-making cannot relieve the controller of the duty to provide an explanation.

See Case C-203/22, Dun & Bradstreet; Case C-817/19. Ligue des Droits Humains; Case C-634/21, Chufa holdings

Mitigating measures at the deployment phase

Recommendations from the EDPB

- Output filters to prevent the storage, regurgitation or generation of personal data.
- Digital watermarking of AI-generated outputs.
- Measures to allow for the exercise of the right to erasure of personal data from model output data or deduplication.
- Post-training techniques that attempt to remove or suppress personal data.

Summary

Responsible (and legal) machine learning is hard.

- While compliance with legal frameworks can feel daunting, it’s a *fundamental requirement* for responsible machine learning.

- It’s also a baseline, the aim should probably be a lot higher.

- Legal rules and principles are often expressions of granular and fine-tuned ethical and moral judgments, as such they can guide you towards asking relevant questions and finding proper solutions.

- Better to incorporate at the design and concept stage, as most issues are difficult to fix unless data protection was considered by design and by default.

Further reading

- EDPB Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en
- Veale M., Binns R., Edwards L., 2018, Algorithms that remember: model inversion attacks and data protection law. Phil. Trans. R. Soc. A 376: 20180083, available at http://dx.doi.org/10.1098/rsta.2018.0083
- CJEU Case law:
    - Case C-203/22, *Dun & Bradstreet*
    - Case C-817/19. *Ligue des Droits Humains*
    - Case C-634/21, *Chufa holdings*
- Some of my own work on automated decision-making and the rule of law:
    - Naarttijärvi, M. (2023). Situating the Rule of Law in the Context of Automated Decision-Making. In M. Suksi (Ed.), The Rule of Law and Automated Decision-Making: Exploring Fundamentals of Algorithmic Governance (pp. 15–31). Springer International Publishing. https://doi.org/10.1007/978-3-031-30142-1_2
    - Enqvist, L., & Naarttijärvi, M. (2022). Discretion, Automation, and Proportionality. In M. Suksi (Ed.), Rule of Law and Automated Decision-Making. Springer.
    - Naarttijärvi, M. (2023). "Chapter 65: Exploring critical dichotomies of AI and the Rule of Law". In Handbook of Critical Studies of Artificial Intelligence. Cheltenham, UK: Edward Elgar Publishing. Retrieved Mar 14, 2025, from https://doi.org/10.4337/9781803928562.00076