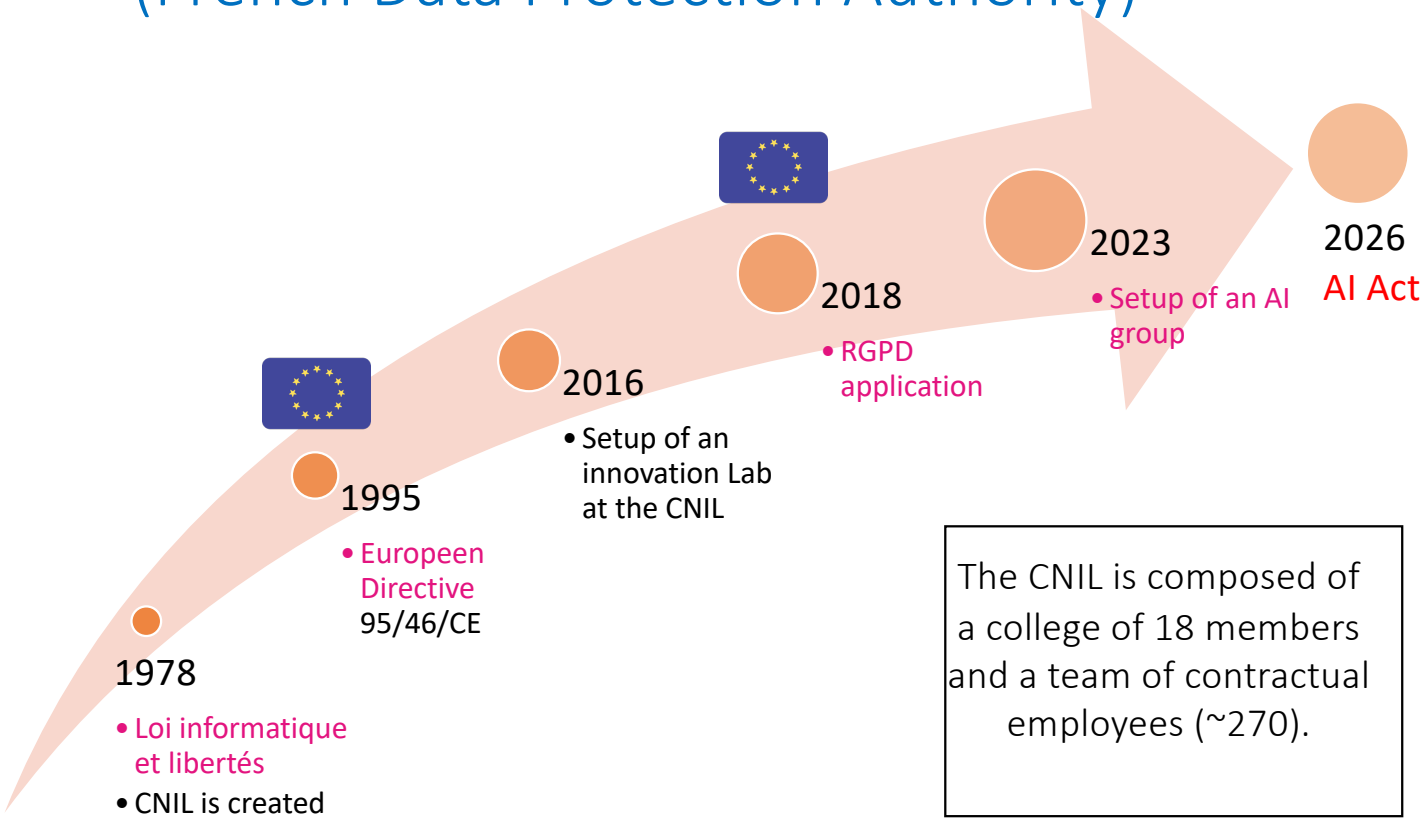
Applying GDPR to AI Models

Claude Castelluccia
INRIA & CNIL

Disclaimers

- This presentation reflects my opinions not the opinion of Inria or CNIL!
- This is work in progress and quite informal !
- This talk is more about questions/problems than answers/solutions ;)…

The CNIL
(French Data Protection Authority)



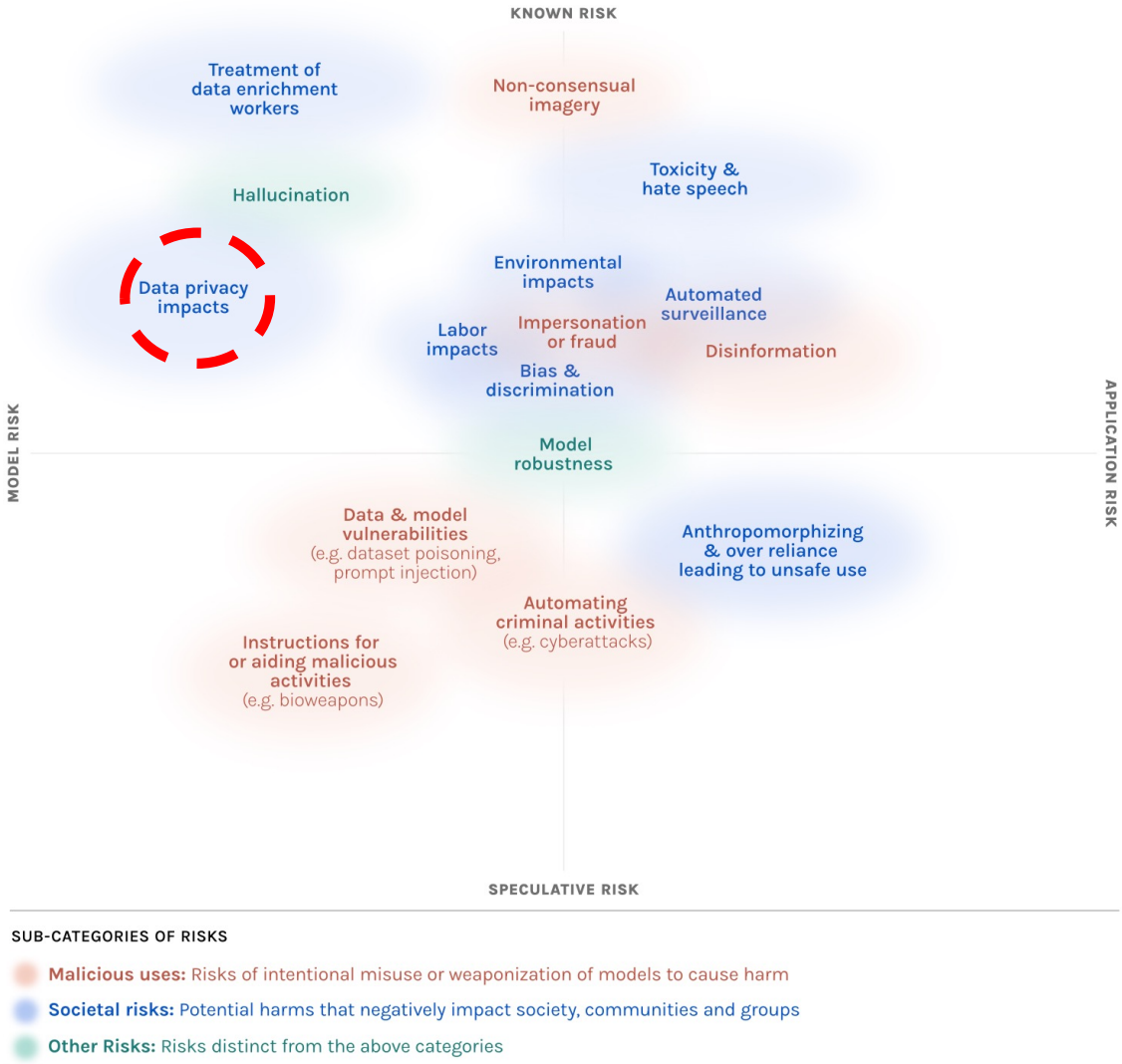
The CNIL is composed of a college of 18 members and a team of contractual employees (~270).

AI Promises and Risks

- The **promises of AI** are great : curing diseases, increasing productivity, helping to solve climate change.

- But AI comes with **some risks**… for example:
  - Discriminations/bias: Dutch Child Benefit Scandal
  - Security/Safety: automated fake news, automated cyber warfare, bio terrorism
  - Privacy: AI usually processes personal information
- Some are even pretending that AI could threaten the **existence of humanity**!

Risk Mapping



https://partnershiponai.org/modeldeployment/#learn_more

Controlling AI Systems

- It is important to **control** AI systems!
  - "We won't deploy electricity systems without electric breakers, It should be the same for AI" (Brad Smith, Microsoft)

- All **phases** should be considered: experimentation, conception, production.

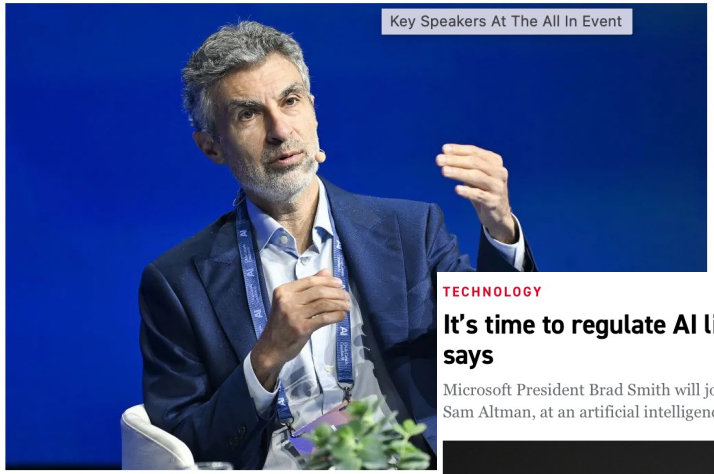
- In particular, it is essential to **evaluate/audit**:
  - Their performance and limitations
  - Their security/safety/privacy/robustness

  (beware of audit washing -> another talk ;)!)

Many Calls/Open Letters for Regulation

TECH • ARTIFICIAL INTELLIGENCE

AI Experts Call For Policy Action to Avoid Extreme Risks



Yoshua Bengio, founder and scientific director of Mila at the Quebec AI Institute Canada, on Sept. 27, 2023. Graham Hughes/Bloomberg—Getty Images

← All Open Letters

Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures
33709

Add your signature

Published
March 22, 2023

TECHNOLOGY

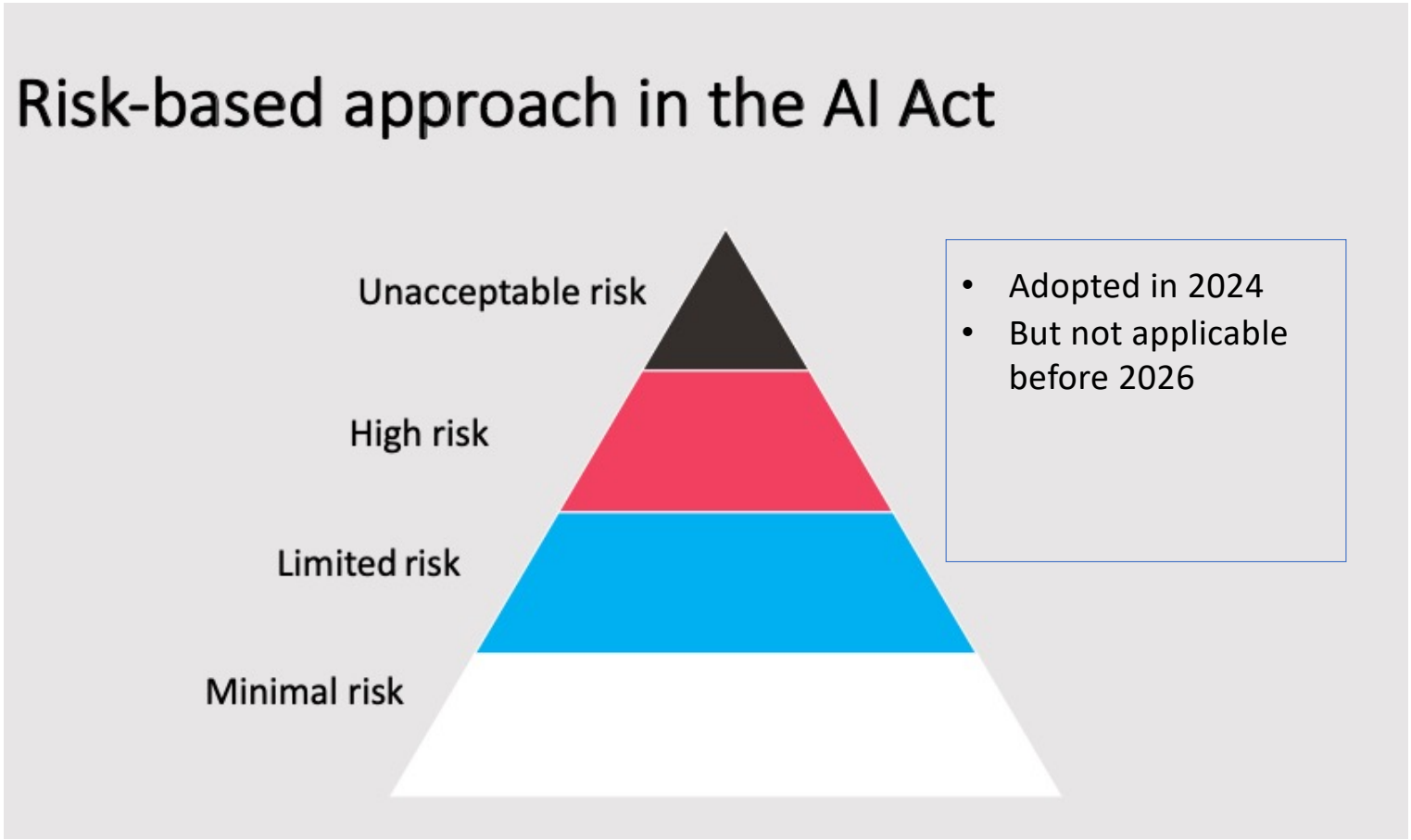
It's time to regulate AI like cars and drugs, top Microsoft exec says

Microsoft President Brad Smith will join other tech leaders, including Elon Musk, Mark Zuckerberg and Sam Altman, at an artificial intelligence forum today hosted by Senate Majority Leader Chuck Schumer.

Various “Regulation” Initiatives

- **G7** Hiroshima AI Process
- **US** AI Bill of Rights
- **UK** Safety AI Summit
- **UNESCO**’s Recommendations on Ethics of AI
- **OECD** AI Principles
- **EU AI Act**
- ….

EU AI ACT (proposed)



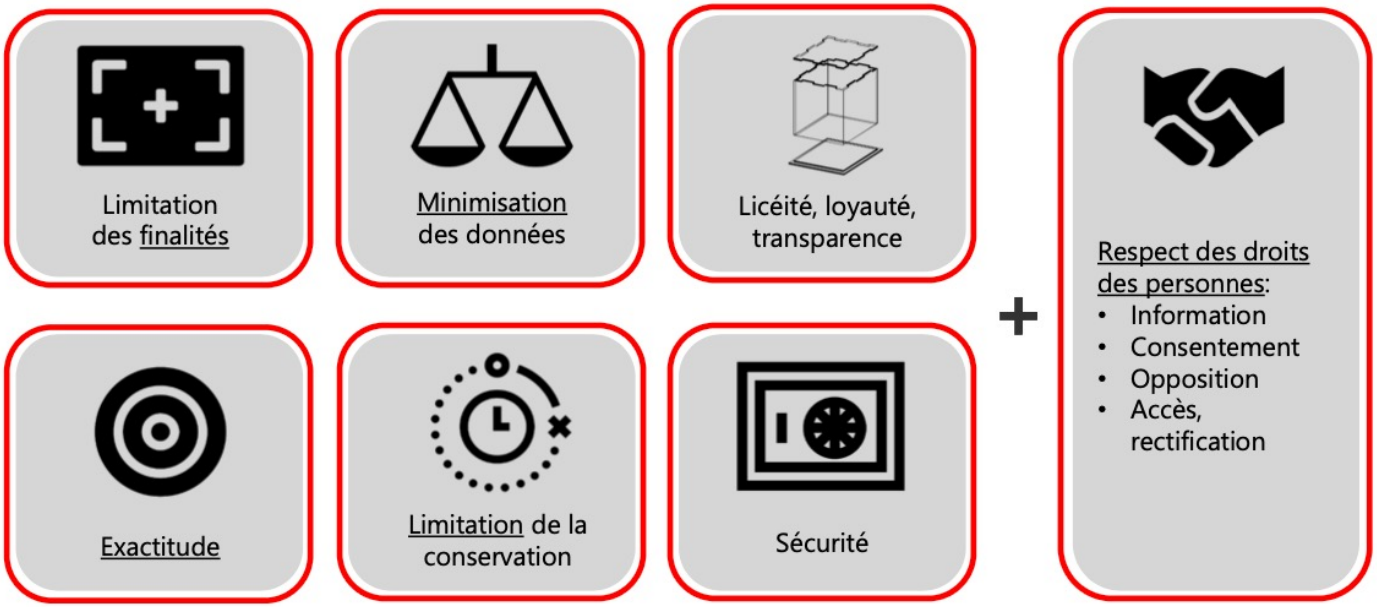
EU AI ACT (adopted)

https://ai-regulation.com/visualisation-pyramid/

Some of AI CNIL Initiatives

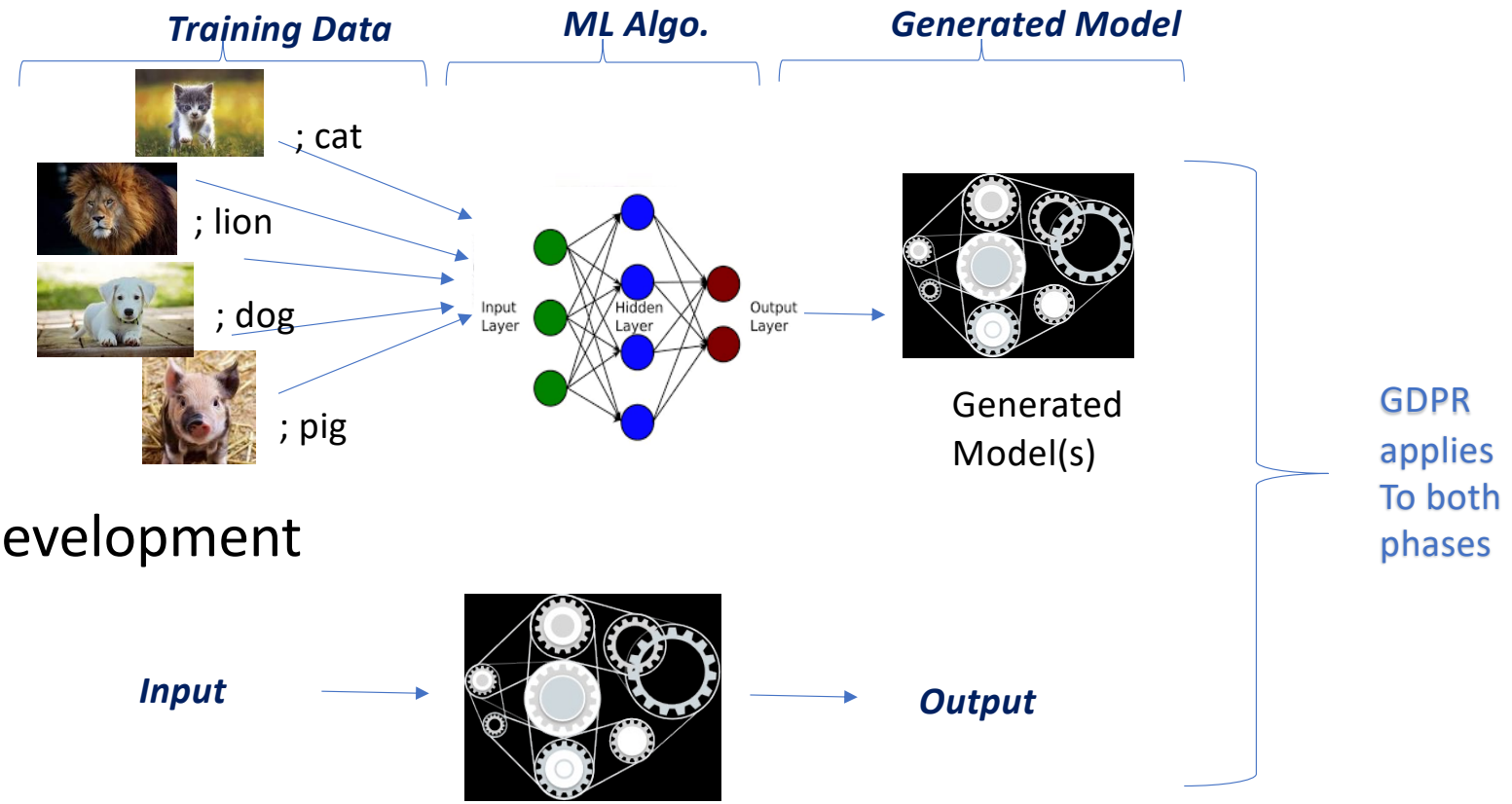
- **Support** Companies, Organizations and State
  - Support of several companies using AI
  - **Enhanced Support Program** for closer supports with selected companies (e.g. Hugging Face)
  - **Sandboxes** for closer supports with administrative organizations that want to use/develop AI (list still confidential)
  - Support/control Olympic Game smart videos
- Initial **guidelines** on how to apply GDPR when designing Machine Learning systems
  - **Consultations!**

AI and GDPR


Limitation des finalités
Minimisation des données
Licéité, loyauté, transparence
Exactitude
Limitation de la conservation
Sécurité
+
Respect des droits des personnes:
• Information
• Consentement
• Opposition
• Accès, rectification

AI (or ML) Systems

- **Phase1**: Conception



- **Phase2**: Development

AI & GDPR Guidelines

- For each of these phases, we are trying to answers various questions such as:
  - Which legal basis? Is legitimate interest an option?
  - How to apply the minimization principle?
  - How to implement the user rights?
  - Data governance
  - …

Some Open Questions

- Q1: Should an AI model be considered Personal Data?
- Q2: If yes, when could an AI model be considered as “anonymized”?
- Q3: Should the answer of the previous depend on the model?
  - should we treat models differently, according to the nature of the training dataset (medical vs location) ?
  - I.e. should we adopt a risk-based approach?

Why Do These Questions Matter?

- If (or when) a model is considered as personal data (i.e. it is possible to extract training data)
  - Then a model can not be “freely” distributed
  - The GDPR applies!

- This makes, for example, Open Models more challenging
  - Important Economical/competition/safety consequences
    - But this could be the topic of another talk!

- The data controller has also the GDPR obligations…

Why Do These Questions Matter? (2)

	Model is NOT Anonymous	Model is Anonymous
Verify Lawfulness of the processing (uploading, manipulation, deployment, distribution,…)	**YES** (processing the model require, for example, identification of Data Controller, legal base, and purpose)	**NO**
Inform People about (whose data has been used to train the model)	**YES**, for all processing related to the model.	**NO.** (People needs to be informed if their data is used to train the model)
User Right Obligation (access, deletion, modification,…)	**YES,** except for some cases (i.e. the data controller cannot retrieve/re-identify the data of the requester)*	**NO.** Except for the training data

* Q: How do you erase the data of a user in a model? Re-train the whole model?

Why Do These Questions Matter? (3)

	Model is NOT Anonymous	Model is Anonymous
Security Obligation	**YES**	**NO**. Guarantying the security of the model is not required by GDPR, however security of the training data is…
PIA requirement	**YES**, when the risk level is high enough	**NO**. The risk associated to the processing of the training data or inputs of the model can however require a PIA.
Data Transfer outside of Europe	**YES**	**NO**, except for training data
IA Act Conformity (such as documentation, transparency, certification,…)	**YES**	**YES**

AI Models vs Algorithms

- We are focusing on Neural Networks (NN)
- A NN Model is not simply an algorithm
  - It is composed of a ”database” of data (weights) and an inference algorithm
    - The weights (w1, w2) are the result of the training algorithm on the training dataset that often contains PI (weights are then aggregation of PI)
    - The inference algorithm uses the model’s input together with the weights to compute the output (inference/prediction)
- So a Model is built from PI…
  - But can these PI leak from the model?
  - i.e. can an ADV retrieve the PI from the model?

```
/**
 * Simple HelloButton() method.
 * @version 1.0
 * @author john doe <doe.j@example.com>
 */
HelloButton()
{
  JButton hello = new JButton( "Hello, wor
  hello.addActionListener( new HelloBtnList

  // use the JFrame type until support for
  // new component is finished
  JFrame frame = new JFrame( "Hello Button"
  Container pane = frame.getContentPane();
  pane.add( hello );
  frame.pack();
  frame.show();              // display the fr
}
```

An Algorithm

(coded by an human or chatGPT;))

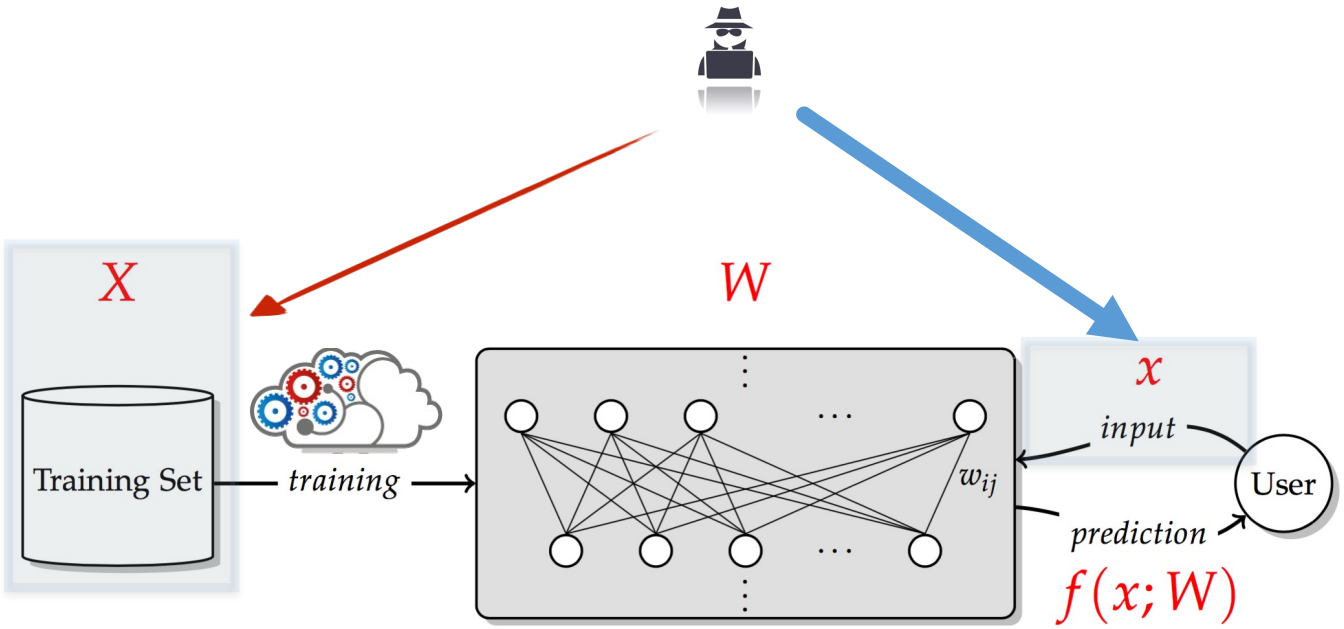
```
def NN(m1, m2, w1, w2, b):
    z = m1 * w1 + m2 * w2 + b
    return sigmoid(z)

def sigmoid(x):
    return 1/(1 + numpy.exp(-x))
```

A Model (generated)

How Can An AI System Leak Information from Training Dataset?

- **Direct Information Leakage**
  - From the training phase or from inputs of the model
- **Indirect Information Leakage**
  - From the model
    - **Model inversion** attacks
    - **Membership** attacks
    - …

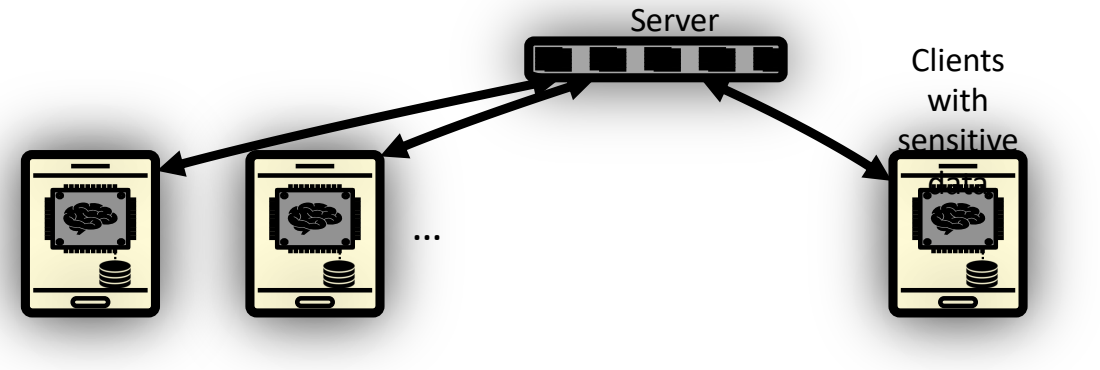
Direct Information Leakage



Some Proposed Solution to Direct Leakage

- **Encrypt** the collected data
  - Training on encrypted data is difficult!

- **Decentralize** the training phase…
  - See Federated Machine Learning



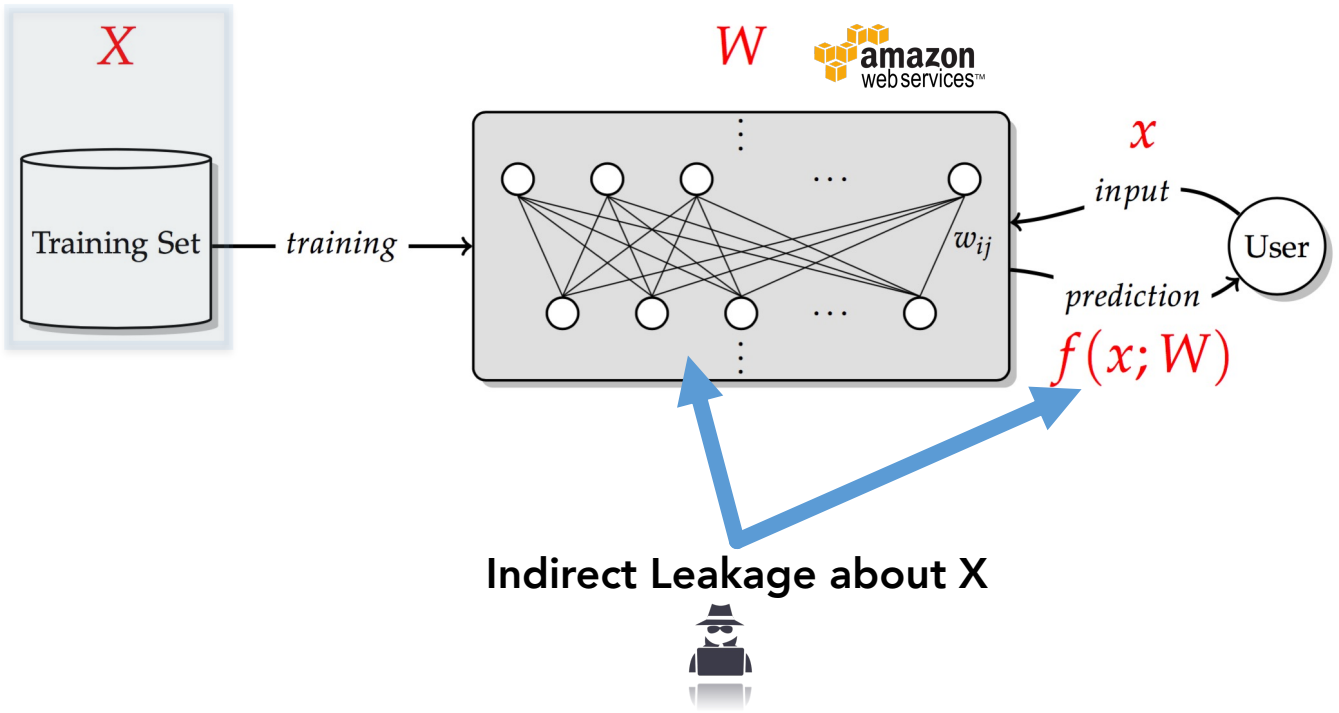
Ref: https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Indirect Information Leakage

- The Adversary goal is to recover some data of the training dataset…
- **Direct Information Leakage**
    - From the training phase or from inputs

- **Indirect Information Leakage**
    - From the (black-box or white-box) model
        - **Model inversion** attacks
        - **Membership** attacks
        - …

Indirect Information Leakage from the Model



Data Leakage from the Models

- Some models can have enough capacity to memorize training examples
- Multiple types of attacks exist (under the black box adversary model):
  - **Membership** attacks: Given a data record with its true label, and black-box access to the model, determine if the record was in the model's training dataset
  - **Model inversion attacks**: Model inversion recovers the average of training samples **within a given class (i**f this average relates to a single individual, then it can be considered as privacy breach)

Some Proposed Solution to Indirect Leakage

- **Anonymize** the collected data (or use synthetic data)
  - Anonymization is difficult and costly (in term of utility)!
  - Use Anonymized Synthetic data (still tricky!)
- **Privacy-Preserving Machine Learning**
  - Embed “anonymization” in the ML algorithm
  - **Example**:
    - Use Federated Machine Learning
    - Intermediate Updates are “noised” (Differential Privacy)
  - Other examples might involved **secure hardware** and/or **homomorphic** encryption

*Ref: **Constrained Differentially Private Federated Learning for Low-bandwidth Devices**, Raouf Kerkouche, Gergely Ács, Claude Castelluccia and Pierre Genevès , The Conference on Uncertainty in Artificial Intelligence (UAI'21), 2021.*

That’s the Theory…the Practice is more Difficult !

- In theory, you could have an anonymized data… by making some (not always realistic) assumptions (adversarial model, parameters i.e. epsilon in Diff Priv…)
- In practice, anonymization is not a binary process: it is not “anonymized or it is not anonymized”…and the border is fuzzy!
- GDPR states that the anonymized data should not be re-identified with “reasonable effort”
- How to interpret “reasonable effort”?
  - Should the value of the data be considered when considering the ”reasonable effort”?
  - I.e. the required effort be higher for valuable data (such as sensitive data, i.e. medical data) and lower for less sensitive one (i.e. data that expire in a short term or that is available elsewhere)?

Should We Consider All Models Equally?

- Should we handle models built from medical data the same than models built from Location data (less sensitive)
- How to interpret “reasonable effort”?
  - Should the value of the data be considered when considering the ”reasonable effort”?
  - I.e. reasonable effort is higher for valuable data (such as sensitive data, i.e. medical data) and lower for less sensitive one (data that expire in a short term or that is available elsewhere?).
    - Current interpretation is that a data anonymity is independently of the data’s nature.

What Would be a Good Model's Anonymisation Metric and Test?

1. The ADV is not able to recover/extract some of the training data?
   - Memorization Attacks in LLM?
2. The ADV is not able to tell whether a given input I was part of the training data ?
   - Membership Inference Attacks (info. Theory approach)?
3. The "standard" Data anonymization tests (G29 guidelines)?
   - Singling-out, linkability, inference?

- This still needs more research…

Some Open Questions

- Q1: Should an AI model be considered Personal Data?

  => YES

- Q2: If yes, could an AI model be considered as “anonymized”?

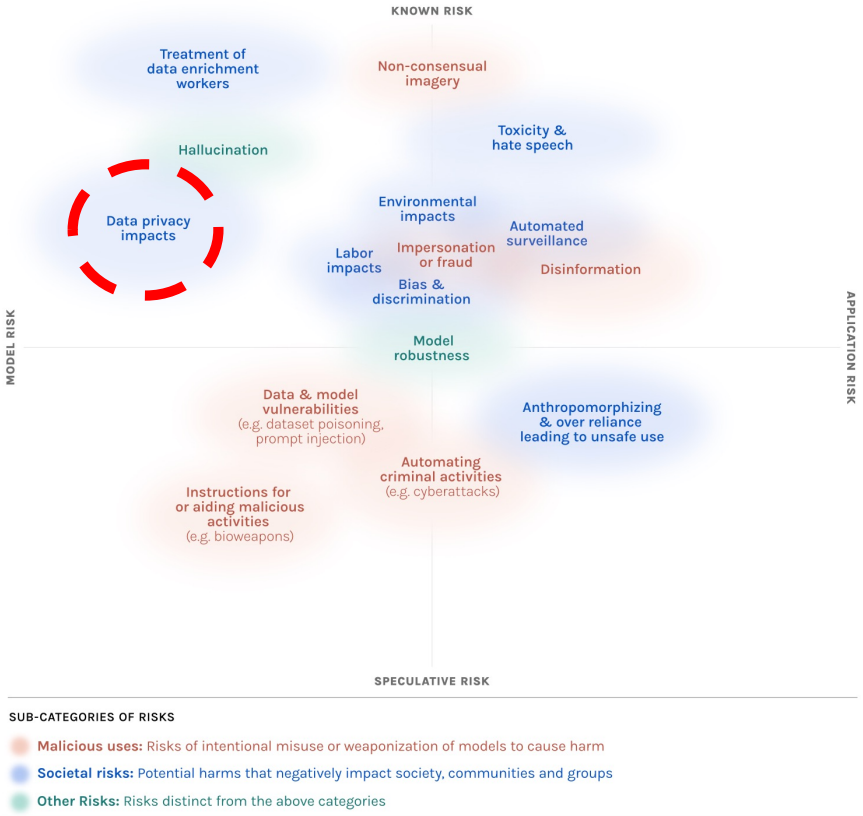
  => YES, if indirect information leakage is hard enough….but more research needed.

- Q3: Should the answer of the previous depend on the model?
  - should we treat models differently, according to the nature of the training dataset (medical vs location) ?
  - I.e. should we adopt a risk-based approach?

  => This is what I think, but this is debatable!

Conclusion

- Regulating AI is complex…. And I only considered a single issue of Privacy!



Conclusion (2)

- *“ while billions of dollars are spent each year to make AI more powerful, funding for research to make AI understandable, free from bias, and safe is tiny in comparison.”* – J. Bengio

AI Regulation Schizophrenia