# Bias, Fairness and Unlearning

*in Artificial Intelligence*

Lecture 12

AI Risk | Bias | Fairness | Machine Unlearning | Governance

# 01 AI Risk

## Malicious Use

- Bioweapons acceleration (e.g. AlphaFold)
- Bioterrorism enabled by AIs that can help humans create deadly pathogens
- AI-enhanced cyberattacks
- The use of AI capabilities for propaganda, censorship, and surveillance.

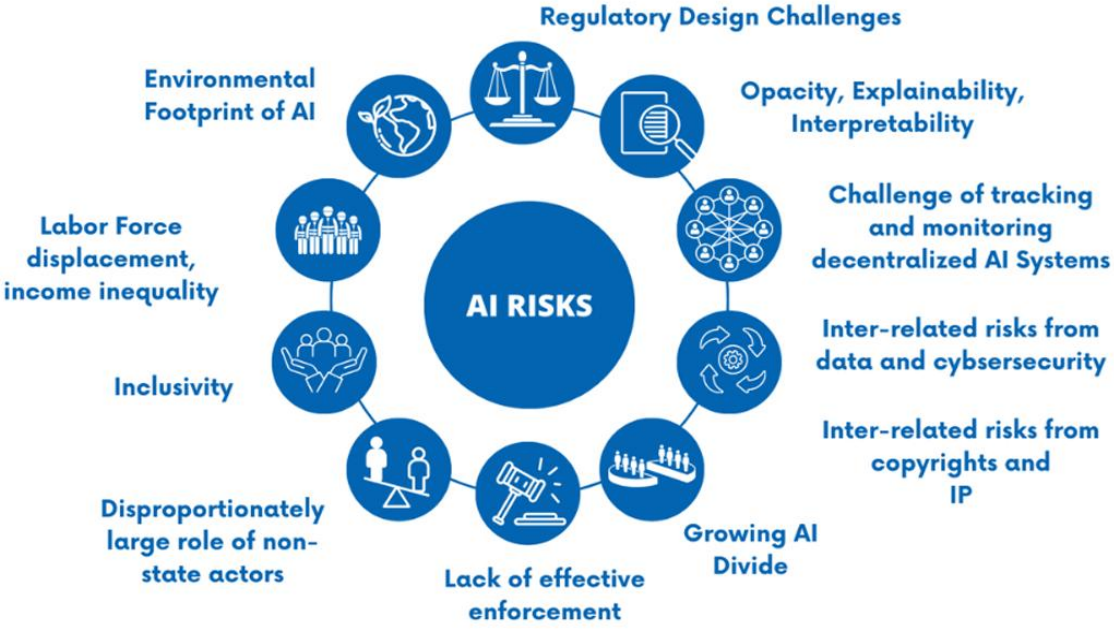
## AI Race Dynamics

- Competition could pressure nations and corporations to rush the development of AIs and cede control to AI systems
- Labor displacement risks
- Data & model leakage

## Environmental Cost

- 1 ChatGPT query = 1L of water
- Equivalent of 14 LED bulbs/hour
- Billions of queries daily
- Growing infrastructure demand

# 01 AI Risk




Welcome to the Artificial Intelligence Incident Database

# 02 Types of Bias

*Bias: systematic errors in decision-making producing uneven outcomes*

**Sampling Bias**

Issues in how population data is sampled

**Algorithmic Bias**

Design choices that prioritize certain features

**Representation Bias**

Data source not representative of target distribution

**Confirmation Bias**

System confirms pre-existing biases of developers/users

**Measurement Bias**

Data collection systematically over/under-represents groups

**Interaction Bias**

AI interacts with users in a biased manner

**Generative Bias**

Generative models over-represent certain attributes from training data

# 03 Bias Mitigation Strategies

## 1 DATA PREPROCESSING

- Oversampling minority groups
- Undersampling majority groups
- Synthetic data generation
- Equal group representation

## 2 MODEL SELECTION

- Regularization techniques
- Ensemble methods
- Combine models to counteract

individual model biases

## 3 POST-PROCESSING

- Equalize positive/negative

decisions across groups

- Similar outcome proportions

for all protected groups

# 04 Types of Fairness

***Fairness: a deliberate attempt to eliminate negative bias in automated decision processes***

### Group Fairness

How different groups as a whole are treated. Includes demographic parity, disparate mistreatment, and equal opportunity.

### Individual Fairness

Similar individuals should be treated similarly, regardless of group membership.

### Counterfactual Fairness

Same individual would be treated similarly even if their attributes changed.

### Procedural Fairness

The decision-making process itself must be fair and transparent to all parties.

### Causal Fairness

Verifies the system does not maintain historical biases and structural inequalities.

# 05 The Trustworthy ML Model

*Beyond accuracy: four pillars for responsible AI evaluation*

## Accuracy

Precision, recall, F1 (traditional performance metrics)

## Fairness

Equal treatment across groups and individuals

## Privacy

Protecting confidential data and limiting exposure

## Robustness

Stable performance across varied inputs and attacks

### Trade-offs

**Privacy ↔ Fairness:** more privacy often reduces model explainability **Robustness ↔ Fairness:** individual fairness requires algorithmic robustness

# 06 Attacks on ML Fairness

## Fairness Techniques

- GAN are used to sanitize synthetic data [1]
- Domain adaptation techniques

[1]Wang and H. Huang, “Approaching machine learning fairness through adversarial network,” 2019, arXiv:1909.03013

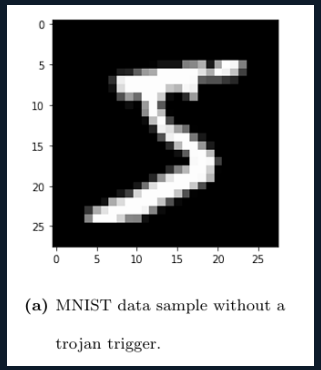
## ⚠ Data Poisoning Attacks

- Fairness techniques don't consider robustness of synthetic data generation
- Adversarial examples can be injected into original dataset
- Ex: Label fliping [2]

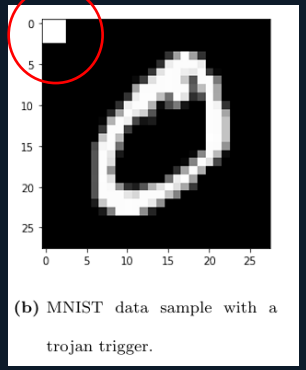
## 🎭 Backdoor Attacks

- Un-Fair Trojan attack: targets model fairness, hard to detect [2]
- Benign usage: backdoor triggers can identify and reduce model bias [3]

## Trojan Trigger Attacks



(a) MNIST data sample without a trojan trigger.



(b) MNIST data sample with a trojan trigger.

[2]https://digitalcommons.njit.edu/cgi/viewcontent.cgi?article=3010&context=theses

[3] Shangxi Wu, Qiuyang He, Yi Zhang, Dongyuan Lu and Jitao Sang. Debiasing backdoor attack: A benign application of backdoor attack in eliminating data bias. INS 119171, 2023

# 07 AI Governance & Regulation

## EU AI ACT RISK LEVELS

**Unacceptable Risk**

*BANNED (e.g. social scoring, real-time surveillance)*

**High Risk**

*Health, justice, hiring, loans (obey strict obligations)*

**General Purpose AI**

*Transparency and copyright compliance required*

**Limited Risk**

*Must inform users they're interacting with AI*

**Minimal Risk**

*Video games, spam filters (mostly unrestricted)*

## KEY FRAMEWORKS

- OECD AI Principles (2019)
- UNESCO Ethics of AI (2021)
- IEEE Ethically Aligned Design (2023)
- EU Ethics Guidelines (2019)
- Council of Europe AI Treaty (2024)
- EU AI Act (21 May 2024)

## TRUSTWORTHY AI PILLARS

| | |
|---|---|
| **Lawful** | Respecting applicable laws |
| **Ethical** | Respecting ethical principles and values |
| **Robust** | Technically and socially reliable |

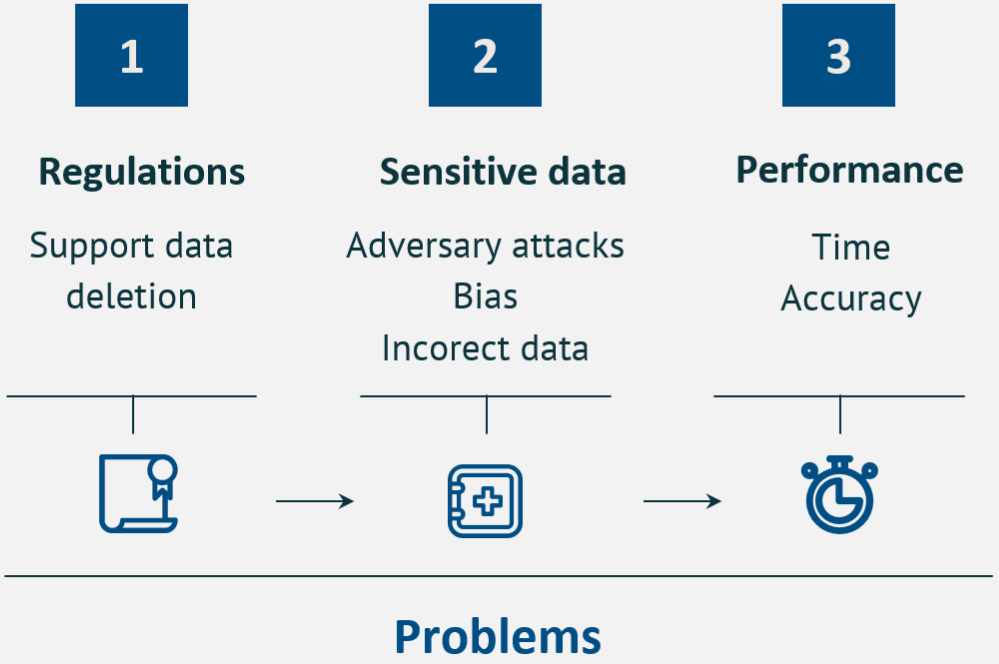
# 08 MACHINE UNLEARNING

## How data privacy challenges machine learning ?

### Current situation

ML models has to obey legislative regulations on the use of consumer personal data (e.g., GDPR, HIPAA).

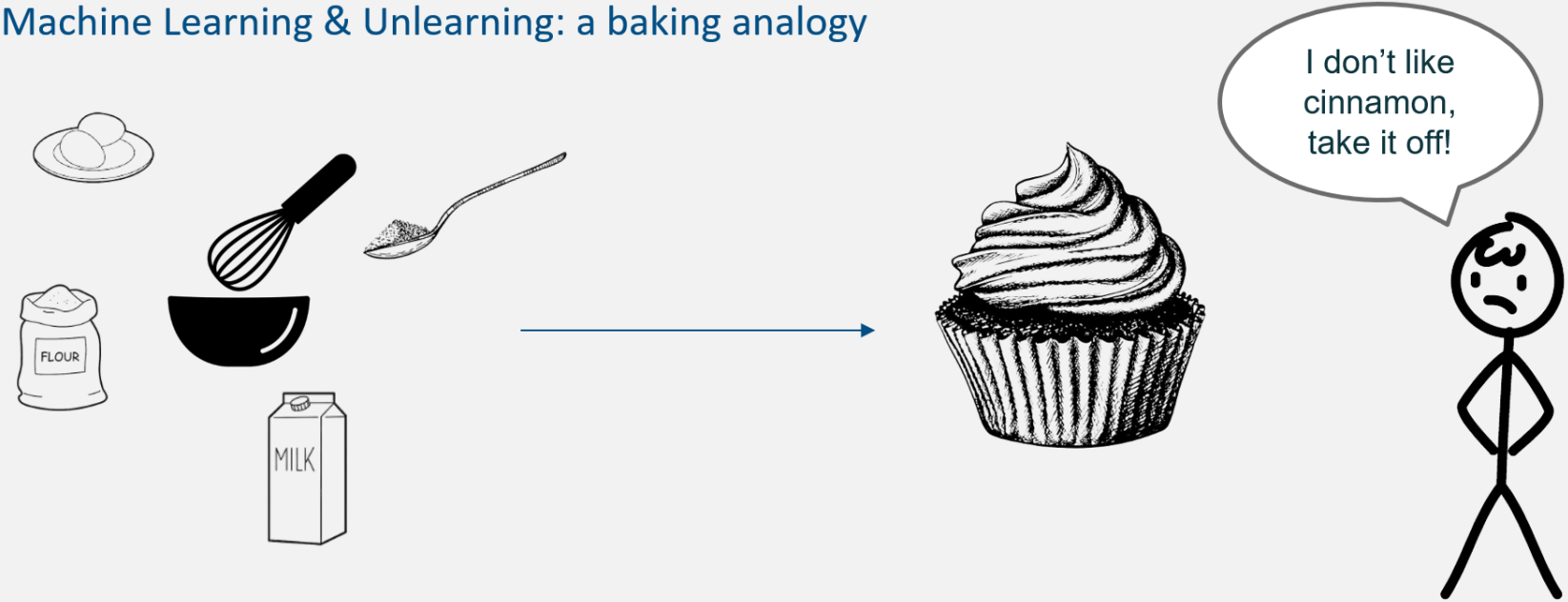
According to Art.17- **„Right to be forgotten”** from GDPR, users have the right to withdraw their consent for their sensitive or personal data to be used by the service they consented to



# 08 MACHINE UNLEARNING

## Why erasing data from a ML model is hard?

Machine Learning & Unlearning: a baking analogy



# 08 MACHINE UNLEARNING

## Types of unlearning

### Exact unlearning

Completely removing the data and retrain (i.e. not adding cinnamon from the beggining)

### Aproximative unlearning

The data impact on model weights is diminished by various methods (i.e adding something else to cancel the cinnamon taste)

Xu, Jie, et al. "Machine unlearning: Solutions and challenges." *IEEE Transactions on Emerging Topics in Computational Intelligence* 8.3 (2024): 2150-2168.

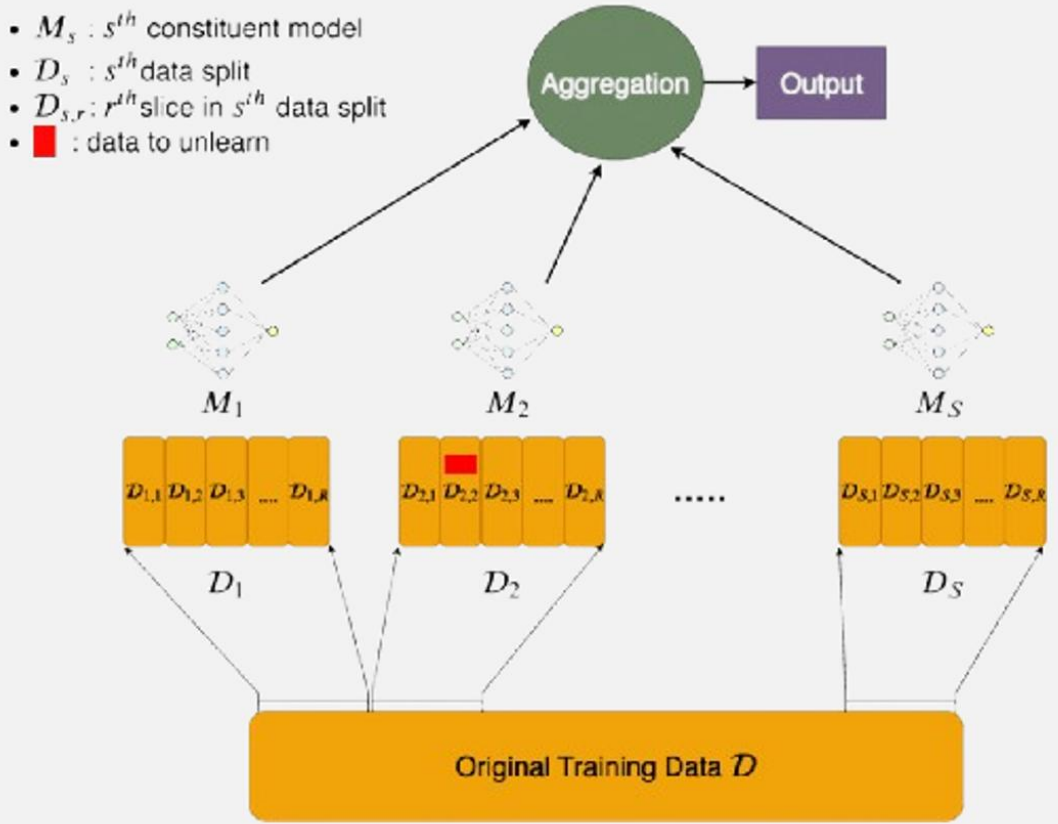
CASE STUDY

# SISA++

## *Extending Machine Unlearning in Medical Imaging*

*New aggregation strategies for privacy-preserving unlearning on unbalanced datasets*

# **SISA:** Sharding, Isolation, Separation and Aggregation




- Faster unlearning
- Maintains privacy
- Reduced retraining cost
- Scalable to large datasets
- Supports batch processing

[1] L. Bourtoule et al., 'Machine Unlearning', arXiv [cs.CR]. 2020.

# 09 SISA — Limitations in Medical Imaging

### Low Performance for Minority Classes

SISA's performance drops significantly for underrepresented classes regardless of how data is sliced. Standard imbalance techniques don't transfer well.

*Critical in medical imaging, rare diseases are the minority.*

### Sensitive to Data Imbalance

Real medical datasets are highly imbalanced by nature. SISA's random sharding amplifies this imbalance across constituent models.

*HAM10000 skin lesion dataset: <5% samples for rare classes.*

### Weak Learners After Unlearning

After multiple unlearning requests, constituent models are trained on progressively less data, each becomes a weaker predictor over time.

*Performance degrades cumulatively with repeated deletions.*

# 10 SISA++ — New Aggregation Strategies

*6 new voting strategies replacing simple majority/soft vote to improve minority class performance*

### Accuracy Weighted Vote

Shards with more data → higher accuracy → more weight

### Shard Size Weighted Vote

Larger shards get proportionally higher voting weight

### Macro F1 Weighted Vote

Favors shards with better overall F1 score

### Median Vote

Less sensitive to individual shard prediction extremes

### Class F1 Weighted Vote

Weights by per-class F1 to boost minority classes

### Greedy Vote

Highest-confidence shard gives the final prediction

# SISA++ Training Structure

## Knowledge-aware distribution

### Strategy

**SISA++ Training**

- Slice union
- Slice no-union

**Data distribution**

**Random**

Random unlearning probability and uniform distribution

**Class based**

Placing minority classes in earlier/later slices

**Experimental setups**

Searching for methods to overcome the data unbalanced issues

# 11 Experimental Results

**Architecture:** ResNet18 (pretrained ImageNet) **Setup:** 5 shards × 3 slices **Unlearning:** 5%, 10%, 15% of training data **Datasets:** HAM10000, OrganMNIST, PathMNIST

## HAM10000 — *Skin Lesion Classification*

| Baseline | SISA++ |
|---|---|
| 0.85 | 0.80 |

*Weighted F1*

- Confidence-max & weighted aggregations improved minority class precision
- Unlearning 10–15% (shards 4 & 5) improved performance vs. uniform distribution
- Strategy differences became more distinct post-unlearning

| Dataset | Model(s) | Training time |
|---|---|---|
| HAM10000 | Baseline | 818.32 s (00:13:38) |
| | SISA union | 501.56 s (00:08:21) |
| | SISA no-union | 266.52 s (00:04:26) |
| organMNIST | Baseline | 1704.11 s (00:28:24) |
| | SISA union | 1231.66 s (00:20:31) |
| | SISA no-union | 498.07 s (00:08:18) |
| pathMNIST | Baseline | 4884.62 s (01:21:24) |
| | SISA union | 2638.75 s (00:43:58) |

## OrganMNIST — *Organ Classification*

| Baseline | SISA++ |
|---|---|
| 0.99 | 0.99 |

*Weighted F1*

- Near-perfect parity between baseline and SISA++
- All strategies performed similarly (±0.01–0.03) with uniform data
- Best-balanced dataset, unlearning had minimal impact

Top 3 strategies: weighted-shard-size · weighted-f1-class · soft vote with a gap of ±0.1–0.2 among top 3, up to ±0.8 vs. others