# AI Governance and Regulation:

## Aspects of Ethics, Privacy, and Security

February 13, 2024
ISACA L.A. Chapter

Katharina Koerner, PhD
Responsible AI Advisor - Tech Diplomacy Network
Corporate Development Manager - Daiki

# Presentation Overview

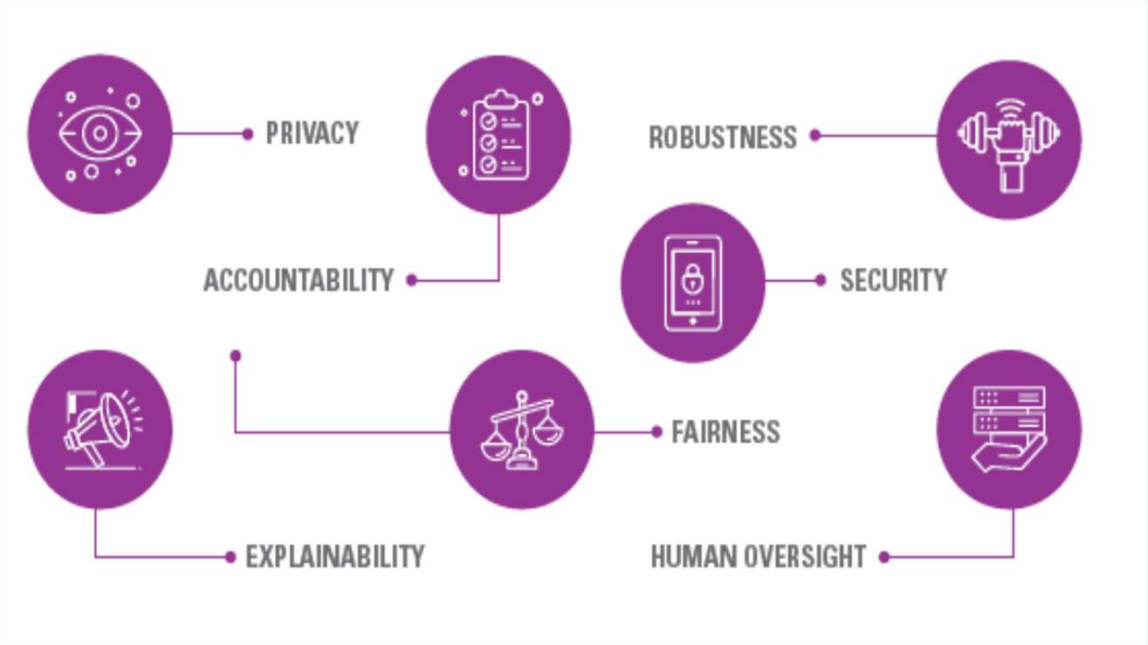
1. Responsible AI principles.
2. Existing and upcoming AI regulation.
3. AI management and governance frameworks.
4. AI security.
5. Privacy challenges and solutions.

# 1. Responsible AI Principels

The terms “Ethical AI”, “Trustworthy AI” and “Responsible AI” are often used interchangeably. For others, ethics goes beyond or is different from RAI.

Responsible AI frameworks are developed and implemented as self-regulatory initiatives, by international organizations, and standardization bodies.

Existing AI regulation can regularly be mapped to the principles of responsible AI.



# 2. AI Regulation

## US:

- AI regulated in sectoral approach, e.g., FTC, EEOC, CFPB, ..
- Executive Order on AI (not as regulation per se)
- State (privacy) laws

## EU + extraterritorial scope:

- GDPR for personal data
- EU AI Act (upcoming)
- EU Liability directive (upcoming)

## Other countries:

- Canada, China, Brazil,...

## Global AI Governance initiatives:

- UN, G7, OECD,...

# ENFORCEMENT EXAMPLES - US

## The FTC's biggest AI enforcement tool? Forcing companies to delete their algorithms

Algorithm disgorgement requires companies to remove products built on data they shouldn't have used in the first place.

BY TONYA RILEY • JULY 5, 2023

C. Within ninety (90) days after the issuance of this Order, delete or destroy any Affected Work Product, and provide a written statement to the Commission, sworn under penalty of perjury, confirming such deletion or destruction.

News

## EEOC settles first AI-bias case, but many more are likely

Employment attorneys caution that companies will be on the hook if a vendor's hiring software turns out to perpetuate bias. While such systems can be complicated and difficult to understand, ignorance is no excuse.

By Greg Andrews | August 16, 2023 at 11:13 AM

AP U.S. WORLD POLITICS VIDEO SPORTS BUSINESS SCIENCE FACT CHECK

Israel-Hamas war | Brussels shooting | Trump gag order | Colorado train de

## FTC investigating ChatGPT creator OpenAI over consumer protection issues

Generative AI refers to a class of artificial intelligence (AI) models that can create or generate new data, such as images, text, or music, that is similar to the data it was trained on.

patterns and relationships in the input data and then

# ENFORCEMENT EXAMPLES - EU

TC

Join TechCrunch+

Login

Search

TechCrunch+

Startups

Venture

Security

AI

## Uber still dragging its feet on algorithmic transparency, Dutch court finds

Natasha Lomas @riptari / 11:00 AM PDT • October 5, 2023

## OpenAI's ChatGPT breaches privacy rules, says Italian watchdog

Reuters

January 30, 2024 12:09 AM PST · Updated 15 days ago


REUTERS® World Business Markets Sust

Boards, Policy & Regulation | Data Privacy | Litigation

## Poland investigates OpenAI over privacy concerns

Reuters

September 21, 2023 2:10 PM PDT · Updated a month ago

August 13, 2021

## Italian data protection supervisory authority fines two food delivery companies for non-compliant algorithmic processing

LinkedIn Facebook Twitter Send Embed

# UPCOMING EU AI ACT: SCOPE

**Compliance with the EU AI Act is mandatory when targeting or impacting the EU market, regardless of the company's location.**

**Specifically applicable globally for:**

**Providers (Developers or Marketers of AI Systems) when**

- ➔ AI systems are introduced to the EU market.
- ➔ AI system's output is used in the EU under international law.

**Deployers (Users of AI Systems) when**

- ➔ International law applies or the system's output is used in the EU.

# UPCOMING EU AI ACT: requirement overview

## Prohibited Practices / Unacceptable Risk::

- Indiscriminate biometric data scraping.
- Public facial and emotion recognition systems.
- Social scoring by authorities.
- Behavior-manipulating AI.
- Profiling for delinquency.

## High Risk:

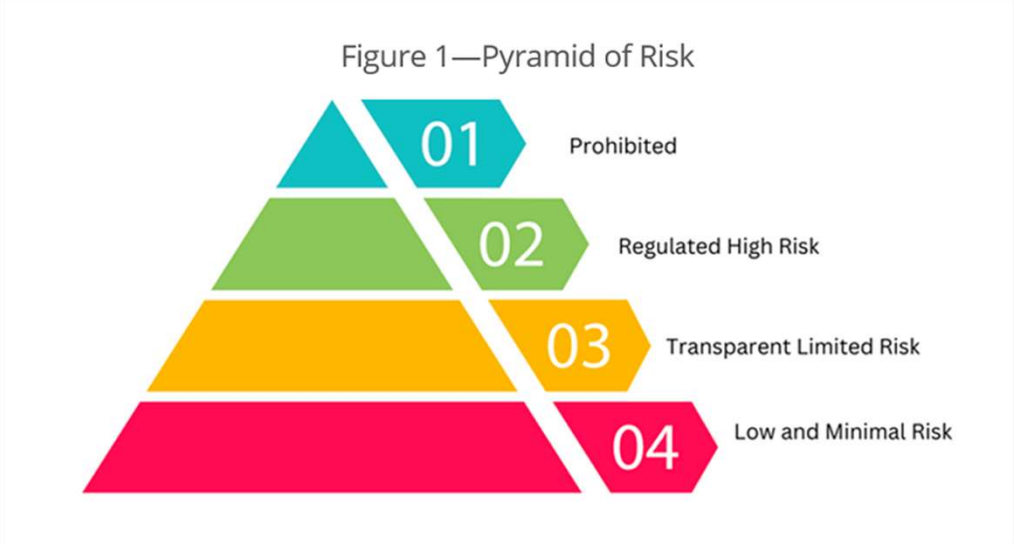
- Affects human safety and fundamental rights.
- Strict vetting and continuous monitoring.
- Mandatory EU database registration.

## Limited Risk:

- Transparency about AI use and data type.

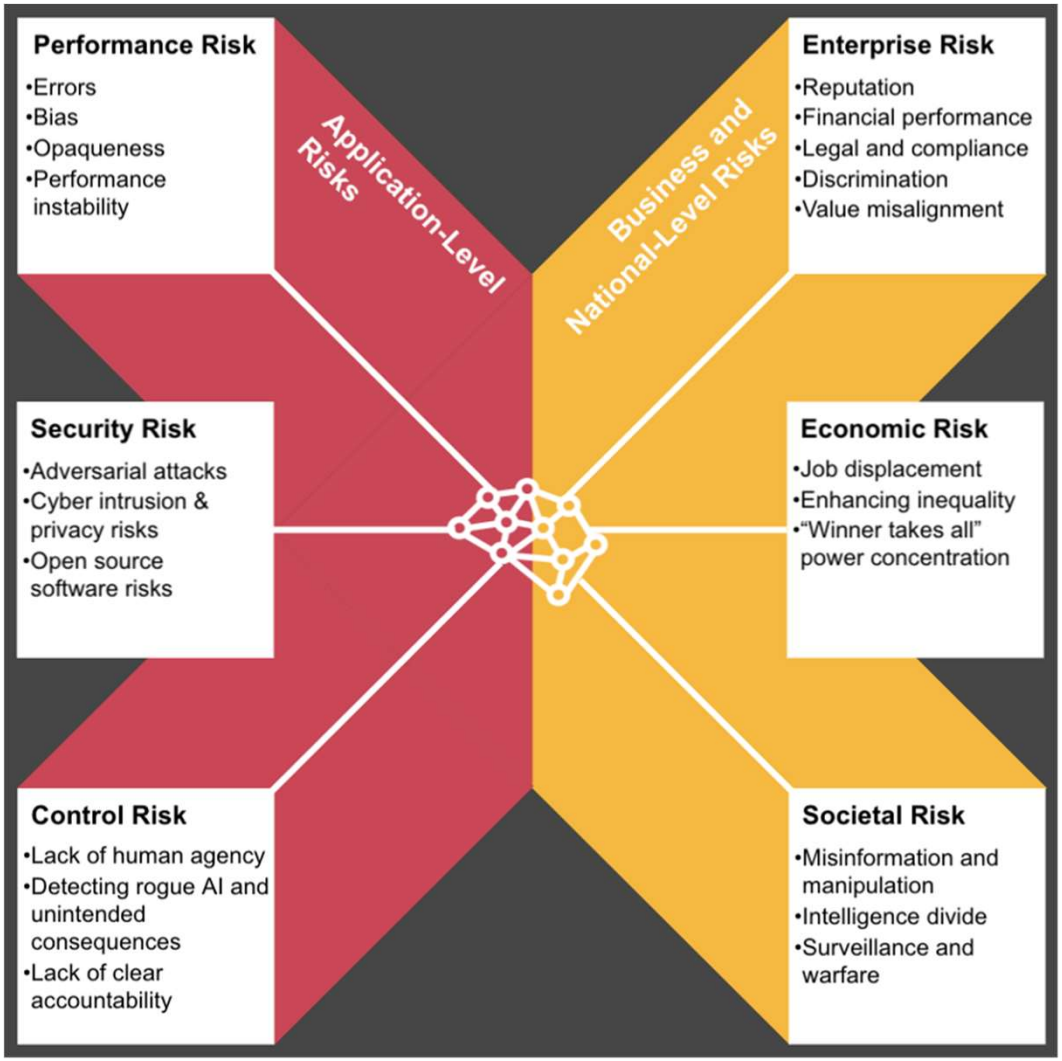
## Minimal Risk:

- Encourages ethical use codes.




Source: ISACA

# RISKS




SOURCE: PwC

# EXAMPLES - REGULATORY RISK

## Grading Foundation Model Providers' Compliance with the Draft EU AI Act

Source: Stanford Center for Research on Foundation Models (CRFM), Institute for Human-Centered Artificial Intelligence (HAI)

| Draft AI Act Requirements | GPT-4 (OpenAI) | Cohere Command (cohere) | Stable Diffusion v2 (stability.ai) | Claude 1 (ANTHROP\C) | PaLM 2 (Google) | BLOOM (BigScience) | LLaMA (Meta) | Jurassic-2 (AI21labs) | Luminous (Aleph Alpha) | GPT-NeoX (EleutherAI) | Totals |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Data sources | ●○○○ | ●●●○ | ●●●● | ○○○○ | ●●○○ | ●●●● | ●●●● | ○○○○ | ○○○○ | ●●●● | 22 |
| Data governance | ●●○○ | ●●●○ | ●●○○ | ○○○○ | ●●●○ | ●●●● | ●●○○ | ○○○○ | ○○○○ | ●●●○ | 19 |
| Copyrighted data | ○○○○ | ○○○○ | ○○○○ | ○○○○ | ○○○○ | ●●●○ | ○○○○ | ○○○○ | ○○○○ | ●●●● | 7 |
| Compute | ○○○○ | ○○○○ | ●●●● | ○○○○ | ○○○○ | ●●●● | ●●●● | ○○○○ | ●○○○ | ●●●● | 17 |
| Energy | ○○○○ | ●○○○ | ●●●○ | ○○○○ | ○○○○ | ●●●● | ●●●● | ○○○○ | ○○○○ | ●●●● | 16 |
| Capabilities & limitations | ●●●● | ●●●○ | ●●●● | ●○○○ | ●●●● | ●●●○ | ●●○○ | ●●○○ | ●○○○ | ●●●○ | 27 |
| Risks & mitigations | ●●●○ | ●●○○ | ●○○○ | ●○○○ | ●●●○ | ●●○○ | ●○○○ | ●●○○ | ○○○○ | ●○○○ | 16 |
| Evaluations | ●●●● | ●●○○ | ○○○○ | ○○○○ | ●●○○ | ●●●○ | ●●○○ | ○○○○ | ●○○○ | ●○○○ | 15 |
| Testing | ●●●○ | ●●○○ | ○○○○ | ○○○○ | ●●○○ | ●●○○ | ○○○○ | ●○○○ | ○○○○ | ○○○○ | 10 |
| Machine-generated content | ●●●○ | ●●●○ | ○○○○ | ●●●○ | ●●●○ | ●●●○ | ○○○○ | ●●●○ | ●○○○ | ●●●○ | 21 |
| Member states | ●●○○ | ○○○○ | ○○○○ | ●●○○ | ●●●● | ○○○○ | ○○○○ | ○○○○ | ●○○○ | ●●○○ | 9 |
| Downstream documentation | ●●●○ | ●●●● | ●●●● | ○○○○ | ●●●● | ●●●● | ●●○○ | ○○○○ | ○○○○ | ●●●○ | 24 |
| Totals | 25 / 48 | 23 / 48 | 22 / 48 | 7 / 48 | 27 / 48 | 36 / 48 | 21 / 48 | 8 / 48 | 5 / 48 | 29 / 48 | |

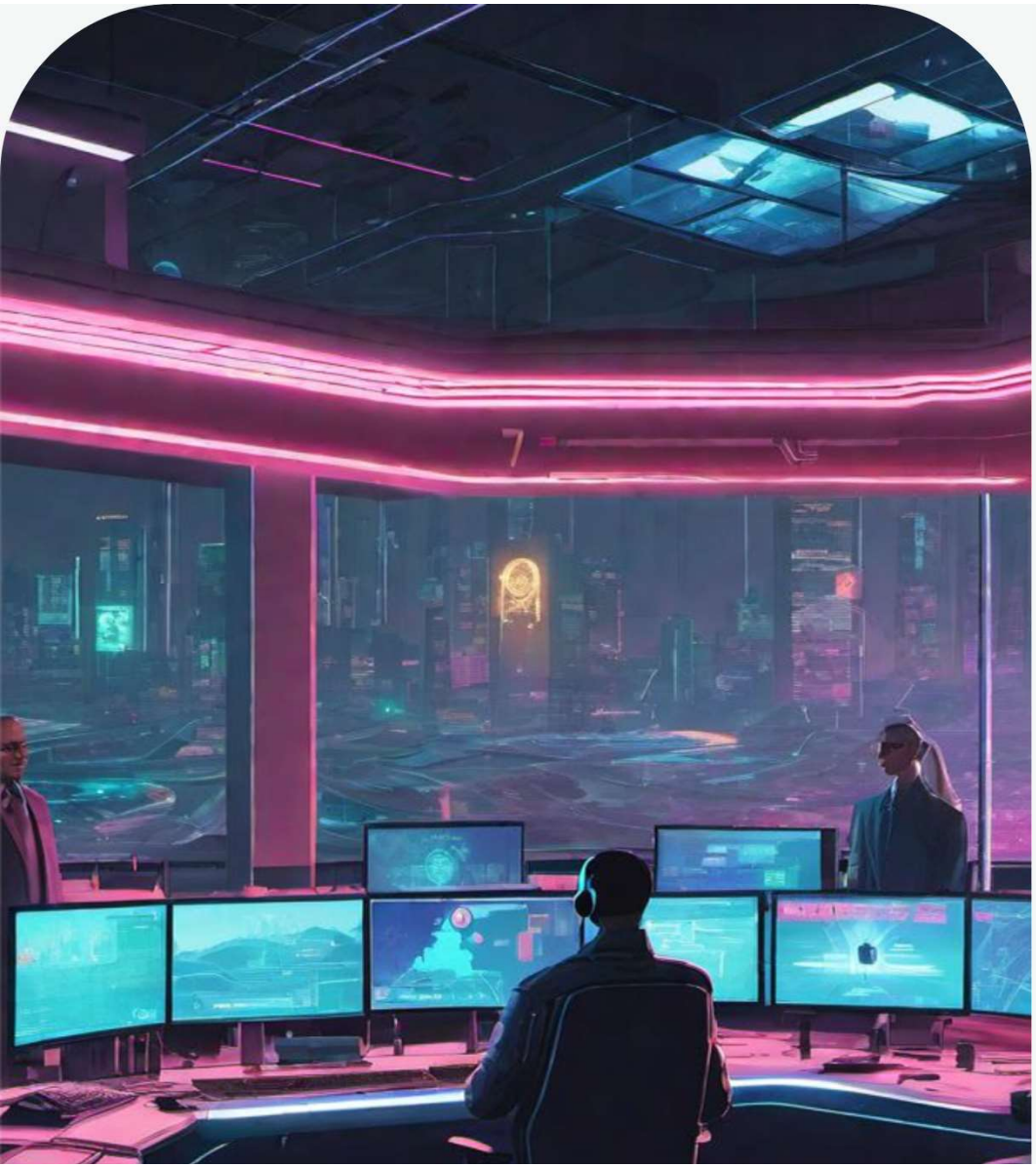
# EXAMPLES - REGULATORY RISK

**Foundation Model Transparency Index Scores by Major Dimensions of Transparency, 2023**

Source: 2023 Foundation Model Transparency Index

| Major Dimensions of Transparency | Meta Llama 2 | BigScience BLOOMZ | OpenAI GPT-4 | stability.ai Stable Diffusion 2 | Google PaLM 2 | ANTHROPIC Claude 2 | cohere Command | AI21labs Jurassic-2 | Inflection Inflection-1 | amazon Titan Text | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Data | 40% | 60% | 20% | 40% | 20% | 0% | 20% | 0% | 0% | 0% | 20% |
| Labor | 29% | 86% | 14% | 14% | 0% | 29% | 0% | 0% | 0% | 0% | 17% |
| Compute | 57% | 14% | 14% | 57% | 14% | 0% | 14% | 0% | 0% | 0% | 17% |
| Methods | 75% | 100% | 50% | 100% | 75% | 75% | 0% | 0% | 0% | 0% | 48% |
| Model Basics | 100% | 100% | 50% | 83% | 67% | 67% | 50% | 33% | 50% | 33% | 63% |
| Model Access | 100% | 100% | 67% | 100% | 33% | 33% | 67% | 33% | 0% | 33% | 57% |
| Capabilities | 60% | 80% | 100% | 40% | 80% | 80% | 60% | 60% | 40% | 20% | 62% |
| Risks | 57% | 0% | 57% | 14% | 29% | 29% | 29% | 29% | 0% | 0% | 24% |
| Mitigations | 60% | 0% | 60% | 0% | 40% | 40% | 20% | 0% | 20% | 20% | 26% |
| Distribution | 71% | 71% | 57% | 71% | 71% | 57% | 57% | 43% | 43% | 43% | 59% |
| Usage Policy | 40% | 20% | 80% | 40% | 60% | 60% | 40% | 20% | 60% | 20% | 44% |
| Feedback | 33% | 33% | 33% | 33% | 33% | 33% | 33% | 33% | 33% | 0% | 30% |
| Impact | 14% | 14% | 14% | 14% | 14% | 0% | 14% | 14% | 14% | 0% | 11% |
| **Average** | **57%** | **52%** | **47%** | **47%** | **41%** | **39%** | **31%** | **20%** | **20%** | **13%** | |

**Scores for 10 major foundation model developers across 13 major dimensions of transparency.**



# AI MANAGEMENT FRAMEWORKS

… help organizations navigate their involvement with AI systems, set up governance processes, and manage AI risks.

… streamline processes for using, developing, monitoring, or providing AI-related products and services.

… address trustworthiness concerns (security, safety, fairness, transparency, data quality).

… provide guidelines for deploying controls to support AI management processes.

… generate evidence of responsibility and accountability regarding the organization's role with AI systems.

# AI GOVERNANCE

… is a component of AI management.

… focuses on establishing policies, guidelines, and ethical considerations for the responsible development, deployment, and use of AI within an organization.

… emphasizes the need for ethical and responsible practices in the use of AI technologies (e.g., in defined AI Objectives).

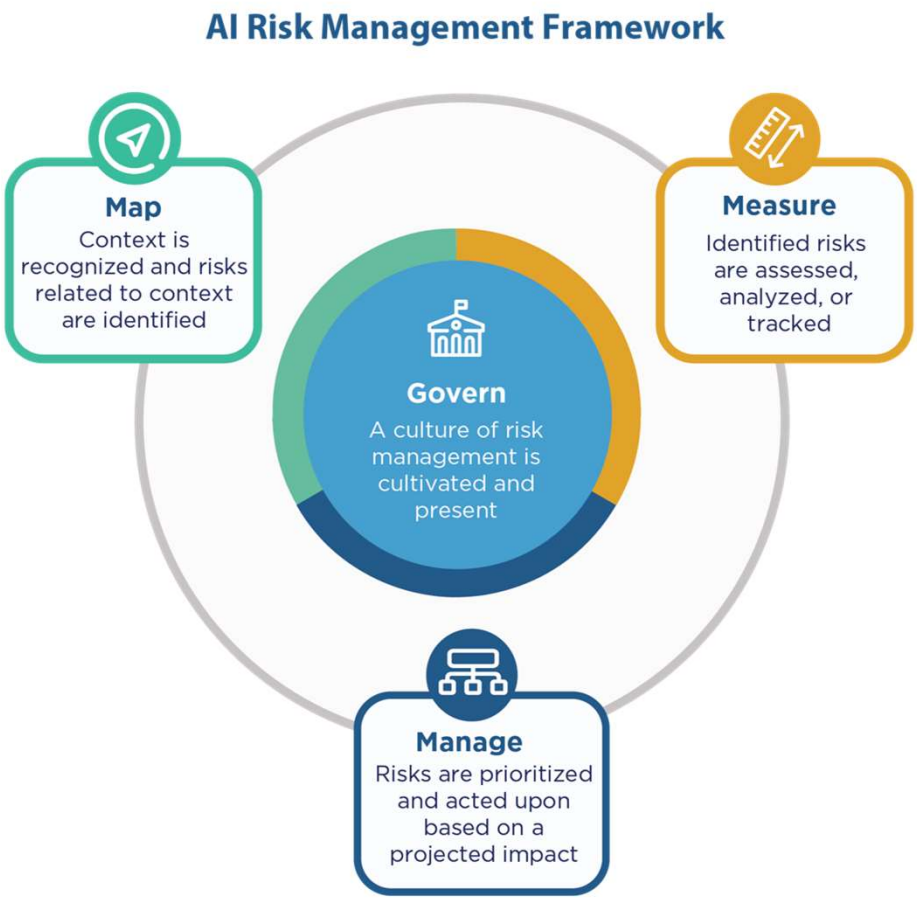
# ISO/IEC 42001:2023, Artificial intelligence Management system

ISO/IEC 42001 is a certifiable framework that helps to establish, implement, maintain, continually improve and document an AIMS.

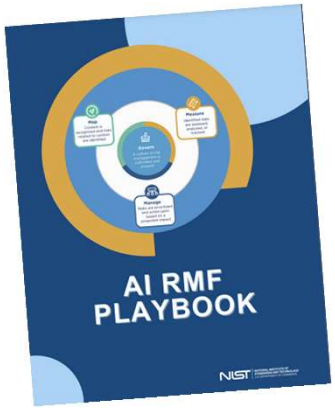
**Goal:**

Guidance for organizations of any size or type in the responsible development, provision, or utilization of AI systems to meet organizational objectives, applicable requirements, and obligations and expectations related to stakeholders.

## Contents




# NIST AI RISK MANAGEMENT FRAMEWORK




## MAP 1.5

Organizational risk tolerances are determined and documented.

### About

Risk tolerance reflects the level and type of risk the organization is willing to accept while conducting its mission and carrying out its strategy.

Organizations can follow existing regulations and guidelines for risk criteria, tolerance and response established by organizational, domain, discipline, sector, or professional requirements. Some sectors or industries may have established definitions of harm or may have established documentation, reporting, and disclosure requirements.

Within sectors, risk management may depend on existing guidelines for specific applications and use case settings. Where established guidelines do not exist, organizations will want to define reasonable risk tolerance in consideration of different sources of risk (e.g., financial, operational, safety and wellbeing, business, reputational, and model risks) and different levels of risk (e.g., from negligible to critical).

Risk tolerances inform and support decisions about whether to continue with development or deployment - termed "go/no-go". Go/no-go decisions related to AI system risks can take stakeholder feedback into account but remain independent from stakeholders' vested financial or reputational interests.

If mapping risk is prohibitively difficult, a "no-go" decision may be considered for the specific system.

### Suggested Actions

- Utilize existing regulations and guidelines for risk criteria, tolerance and response established by organizational, domain, discipline, sector, or professional requirements.
- Establish risk tolerance levels for AI systems and allocate the appropriate oversight resources to each level.



- Establish risk criteria in consideration of different sources of risk, (e.g., financial, operational, safety and wellbeing, business, reputational, and model risks) and different levels of risk (e.g., from negligible to critical).
- Identify maximum allowable risk tolerance above which the system will not be deployed, or will need to be prematurely decommissioned, within the contextual or application setting.
- Articulate and analyze tradeoffs across trustworthiness characteristics as relevant to proposed context of use. When tradeoffs arise, document them and plan for traceable actions (e.g.: impact mitigation, removal of system from development or use) to inform management decisions.
- Review uses of AI systems for "off-label" purposes, especially in settings that organizations have deemed as high-risk. Document decisions, risk-related trade-offs, and system limitations.

### Transparency & Documentation

**Organizations can document the following:**

- Which existing regulations and guidelines apply, and the entity has followed, in the development of system risk tolerances?
- What criteria and assumptions has the entity utilized when developing system risk tolerances?
- How has the entity identified maximum allowable risk tolerance?
- What conditions and purposes are considered "off-label" for system use?

**AI Transparency Resources**

- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities.
- WEF Model AI Governance Framework Assessment 2020.
- WEF Companion to the Model AI Governance Framework- 2020.



### References

Board of Governors of the Federal Reserve System. SR 11-7: Guidance on Model Risk Management. (April 4, 2011).

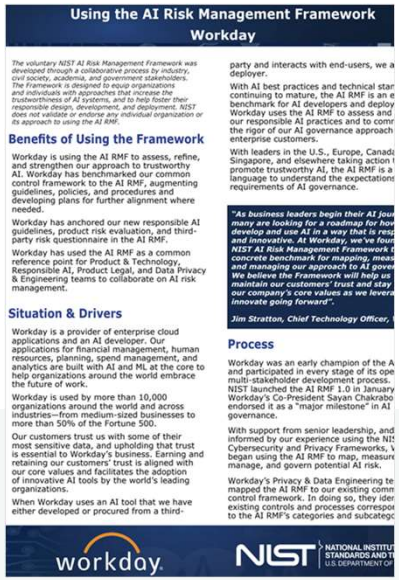
The Office of the Comptroller of the Currency. Enterprise Risk Appetite Statement. (Nov. 20, 2019).

Brenda Boultwood. How to Develop an Enterprise Risk-Rating Approach (Aug. 26, 2021). Global Association of Risk Professionals (garp.org). Accessed Jan. 4, 2023.

Virginia Eubanks, 1972-, Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor. New York, NY, St. Martin's Press, 2018.

GAO-17-63: Enterprise Risk Management: Selected Agencies' Experiences Illustrate Good Practices in Managing Risk. (See Table 3).

NIST Risk Management Framework.



## Using the AI Risk Management Framework Workday

*The voluntary NIST AI Risk Management Framework was developed through a collaborative process by industry, civil society, academia, and government stakeholders. The Framework is designed to equip organizations and individuals with approaches that increase the trustworthiness of AI systems, and to help foster their responsible design, development, and deployment. NIST does not validate or endorse any individual organization or its approach to using the AI RMF.*

### Benefits of Using the Framework

Workday is using the AI RMF to assess, refine, and strengthen our approach to trustworthy AI. Workday has benchmarked our common control framework to the AI RMF, augmenting guidelines, policies, and procedures and developing plans for further alignment where needed.

Workday has anchored our new responsible AI guidelines, product risk evaluation, and third-party risk questionnaire in the AI RMF.

Workday has used the AI RMF as a common reference point for Product & Technology, Responsible AI, Product Legal, and Data Privacy & Engineering teams to collaborate on AI risk management.

### Situation & Drivers

Workday is a provider of enterprise cloud applications and an AI developer. Our applications for financial management, human resources, planning, spend management, and analytics are built with AI and ML at the core to help organizations around the world embrace the future of work.

Workday is used by more than 10,000 organizations around the world and across industries—from medium-sized businesses to more than 50% of the Fortune 500.

Our customers trust us with some of their most sensitive data, and upholding that trust is essential to Workday's business. Earning and retaining our customers' trust is aligned with our core values and facilitates the adoption of innovative AI tools by the world's leading organizations.

When Workday uses an AI tool that we have either developed or procured from a third-party and interacts with end-users, we a deployer.

With AI best practices and technical star continuing to mature, the AI RMF is an e benchmark for AI developers and deploy Workday uses the AI RMF to assess and our responsible AI practices and to comr the rigor of our AI governance approach enterprise customers.

With leaders in the U.S., Europe, Canad Singapore, and elsewhere taking action promote trustworthy AI, the AI RMF is a language to understand the expectations requirements of AI governance.

***"As business leaders begin their AI jou many are looking for a roadmap for how develop and use AI in a way that is resp and innovative. At Workday, we've foun NIST AI Risk Management Framework concrete benchmark for mapping, meas and managing our approach to AI gove We believe the Framework will help us maintain our customers' trust and stay our company's core values as we lever innovate going forward".***

***Jim Stratton, Chief Technology Officer,***

### Process

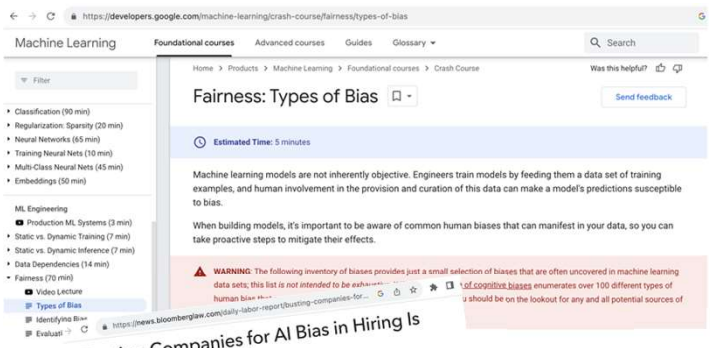
Workday was an early champion of the A and participated in every stage of its ope multi-stakeholder development process. NIST launched the AI RMF 1.0 in January Workday's Co-President Sayan Chakrabo endorsed it as a "major milestone" in AI governance.

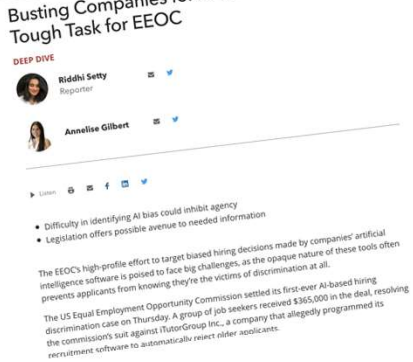
With support from senior leadership, and informed by our experience using the NI Cybersecurity and Privacy Frameworks, began using the AI RMF to map, measure manage, and govern potential AI risk.

Workday's Privacy & Data Engineering te mapped the AI RMF to our existing comm control framework. In doing so, they ider existing controls and processes correspo to the AI RMF's categories and subcatego

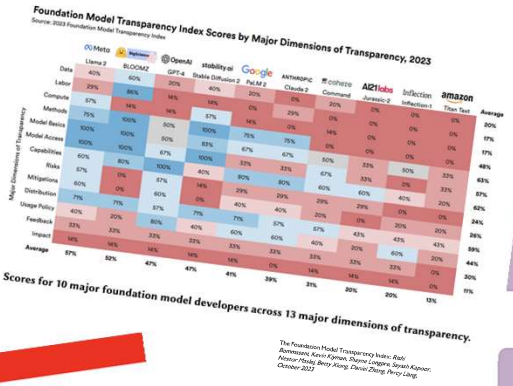
workday. NIST NATIONAL INSTITU STANDARDS AND T U.S. DEPARTMENT OF

# PRACTICAL IMPLEMENTATION





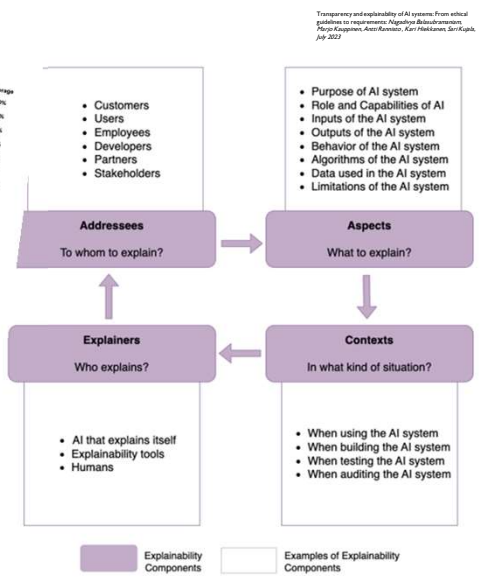




Fig. 2. A model of explainability components.

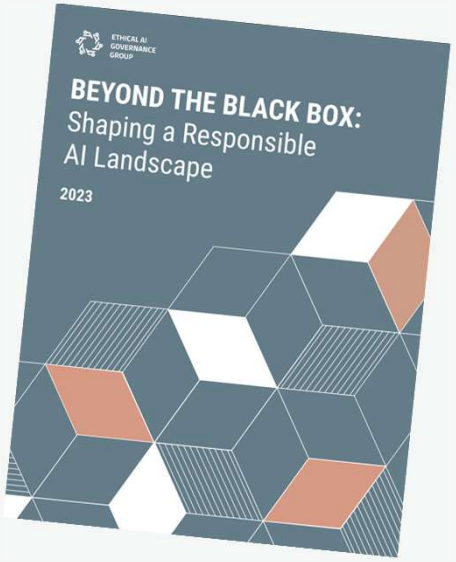
# OPERATIONAL CHALLENGES


# AN EMERGING ECOSYSTEM AROUND RESPONSIBLE AI

The landscape of AI startups providing **services for responsible AI** operationalization is growing fast.

Includes:

- Identifying AI Deployment Across Organizations
- Ensuring Compliance with Regulatory Requirements during development
- Privacy-first data and processing



# 4. AI Security

Field is in development.

<u>New resources:</u>

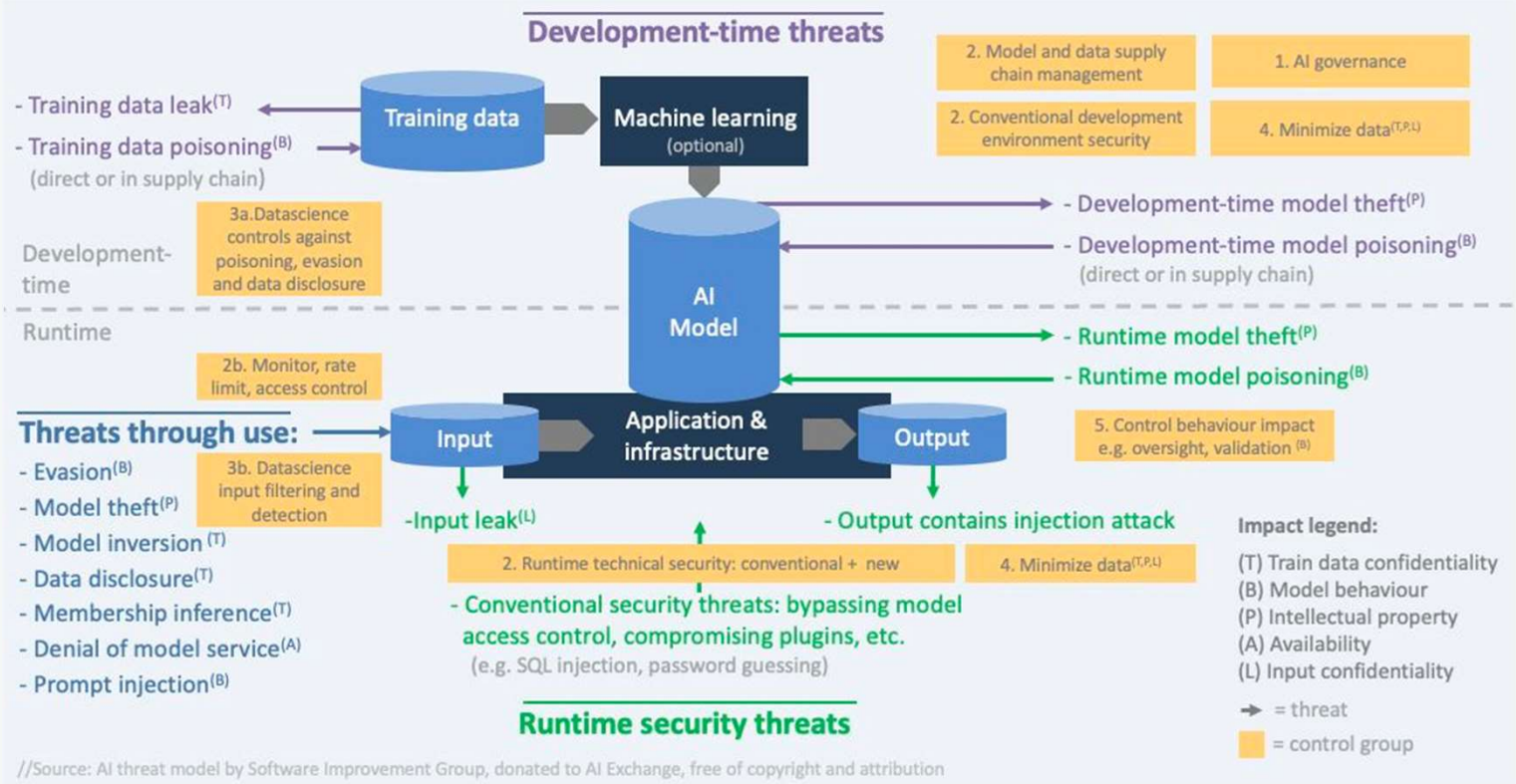
NIST Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations

ENISA Securing Machine Learning Algorithms

Microsoft Failure Modes in Machine Learning

Google Secure AI Framework (SAIF)

The OWASP AI Exchange



# Essentials of AI Security

1. Well-established **information security management systems** form the foundation of AI security.
1. Create explicit and **supplementary AI security measures** and incorporate them into current risk management protocols.
1. Effective **communication** between cybersecurity experts and data scientists is essential for success in this arena.

# 5. Privacy Challenges in AI

General Privacy and Data protection principles apply to AI systems and processing of personal data.

**AI privacy risks include:**

- *Bias and Discrimination*
- *Harms from Inferences*
- *Problematic Data Actions: (unexpected data collection, storage, and use)*
- *Lack of Processing Basis: Processing data without a clear legal basis*
- *Secondary Data Use: Data repurposed for new, undisclosed uses*
- *Jurisdictional Issues and Data Scraping: Data scraping across jurisdictions*
- *Re-identification Risks in Anonymized Data*

Effective management of these risks requires adherence to privacy principles, clear regulatory frameworks, mechanisms for accountability, and engineering privacy into AI system development.

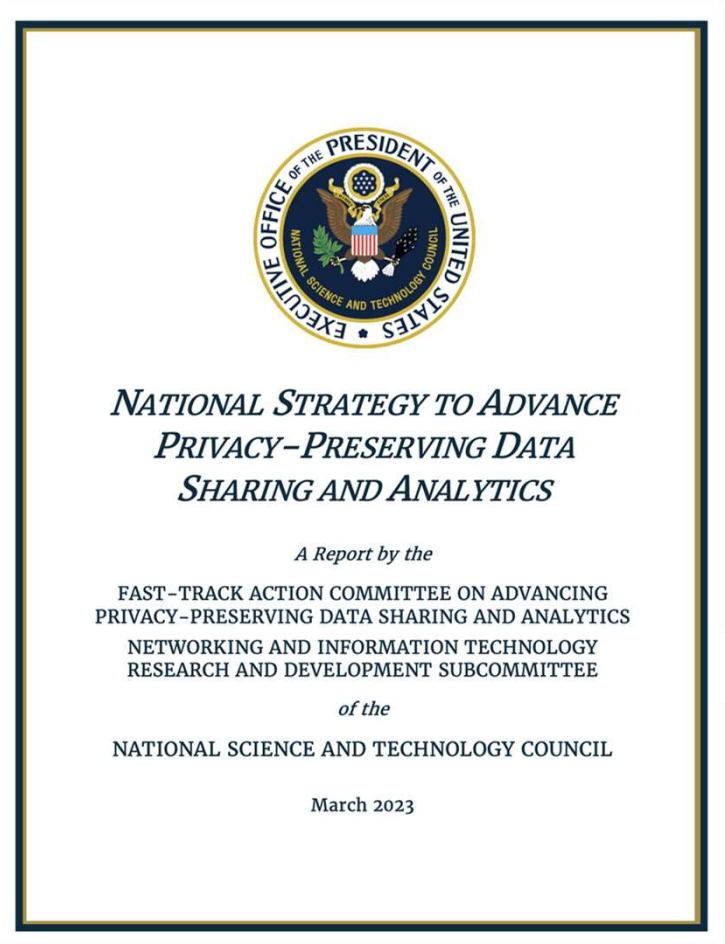
# PRIVACY ENHANCING TECHNOLOGIES IN AI/ML

PETs that support privacy and security in the context of AI/ML and responsible AI are also referred to as **Privacy-Preserving ML** (PPML) or Privacy-Preserving Data Sharing and Analytics (PPDSA).

*"Privacy-preserving data sharing and analytics (PPDSA) methods **utilize cryptographic techniques,** which **inherently satisfy the confidentiality objective.***

*The distinctive aspect of PPDSA approaches is their ability to achieve dissociability, **preventing authorized entities from establishing linkages between data and individuals' identities**, thereby enhancing privacy even with authorized data usage.*

*Such technologies currently include, but are not limited to, secure multiparty computation, homomorphic encryption, zero-knowledge proofs, federated learning, secure enclaves, differential privacy, and synthetic data generation tools."*



EXECUTIVE OFFICE OF THE PRESIDENT OF THE UNITED STATES

NATIONAL SCIENCE AND TECHNOLOGY COUNCIL

NATIONAL STRATEGY TO ADVANCE PRIVACY-PRESERVING DATA SHARING AND ANALYTICS

*A Report by the*

FAST-TRACK ACTION COMMITTEE ON ADVANCING PRIVACY-PRESERVING DATA SHARING AND ANALYTICS

NETWORKING AND INFORMATION TECHNOLOGY RESEARCH AND DEVELOPMENT SUBCOMMITTEE

*of the*

NATIONAL SCIENCE AND TECHNOLOGY COUNCIL

March 2023

## Privacy Preserving Machine Learning (PPML) @ SciPy 2023

Privacy guarantees are **the** most crucial requirement when it comes to analyse sensitive data. These requirements could be sometimes very stringent, so that it becomes a real barrier for the entire pipeline. Reasons for this are manifold, and involve the fact that data could not

Event | June 2022

## Apple Privacy-Preserving Machine Learning Workshop 2022

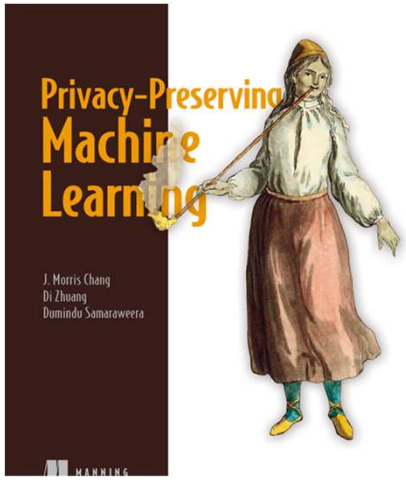
Earlier this year, Apple hosted the Workshop on Privacy-Preserving Machine Learning (PPML). This virtual event brought Apple and members of the academic research communities together to discuss the state of the art in the field of privacy-preserving machine learning through a series of talks and discussions over two days.

In this post we will introduce a new dataset for community benchmarking in PPML, and share highlights from workshop discussions and recordings of select workshop talks.

amazon | science

Research areas Blog News and features Publications

Privacy-preserving machine learning




Northeastern University
College of Engineering

About Academics & Experiential Learning Research

/ NEWS & EVENTS / NEWS /
MAKING AI MORE SECURE WITH PRIVACY-PRESERVING MACHINE LEARNING

## Making AI More Secure with Privacy-Preserving Machine Learning

July 6, 2023

## The Fourth AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI-23)

February 13, 2023

PPAI is an in-person event at: Walter E. Washington Convention Center - Washington, DC, USA. Room 147A

PPAI will also be live-streamed at underline.io

Microsoft | Research Our research Programs & events Blogs & podcasts About Sign up: Research Newsletter

Return to Blog Home

Microsoft Research Blog

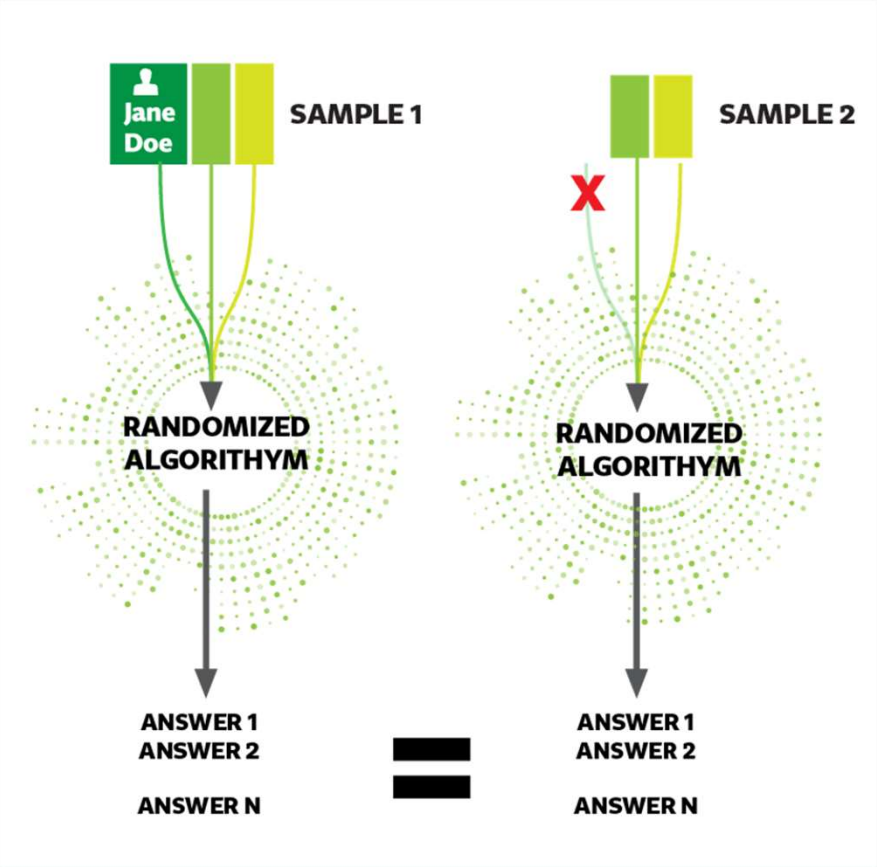
## Privacy Preserving Machine Learning: Maintaining confidentiality and preserving trust

Published November 9, 2021

By Victor Ruehle, Principal Research Manager; Robert Sim, Principal Research Manager; Sergey Yekhanin, Sr Principal Research Manager; Nishanth Chandran, Principal Researcher; Melissa Chase, Principal Researcher; Daniel Jones, Senior Applied Researcher; Kim Laine, Principal Research Manager; Boris Köpf, Principal Researcher; Jaime Teevan, Chief Scientist & Technical Fellow; Jim Kleewein, Technical Fellow; Saravan Rajmohan, Partner Director AI and Applied Research

## The Second IACR School on Privacy-Preserving Machine Learning 2023

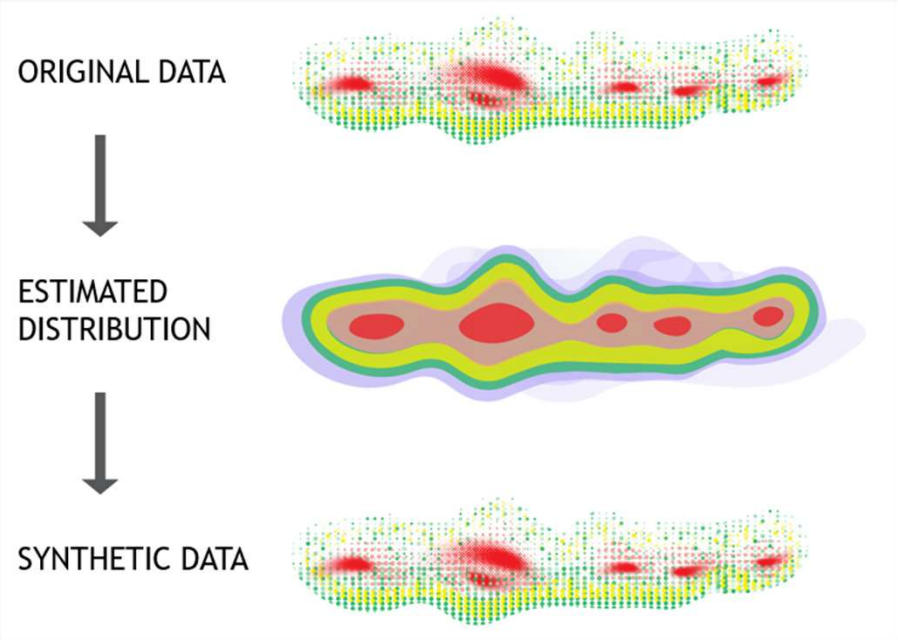
# PRIVACY PRESERVING MACHINE LEARNING



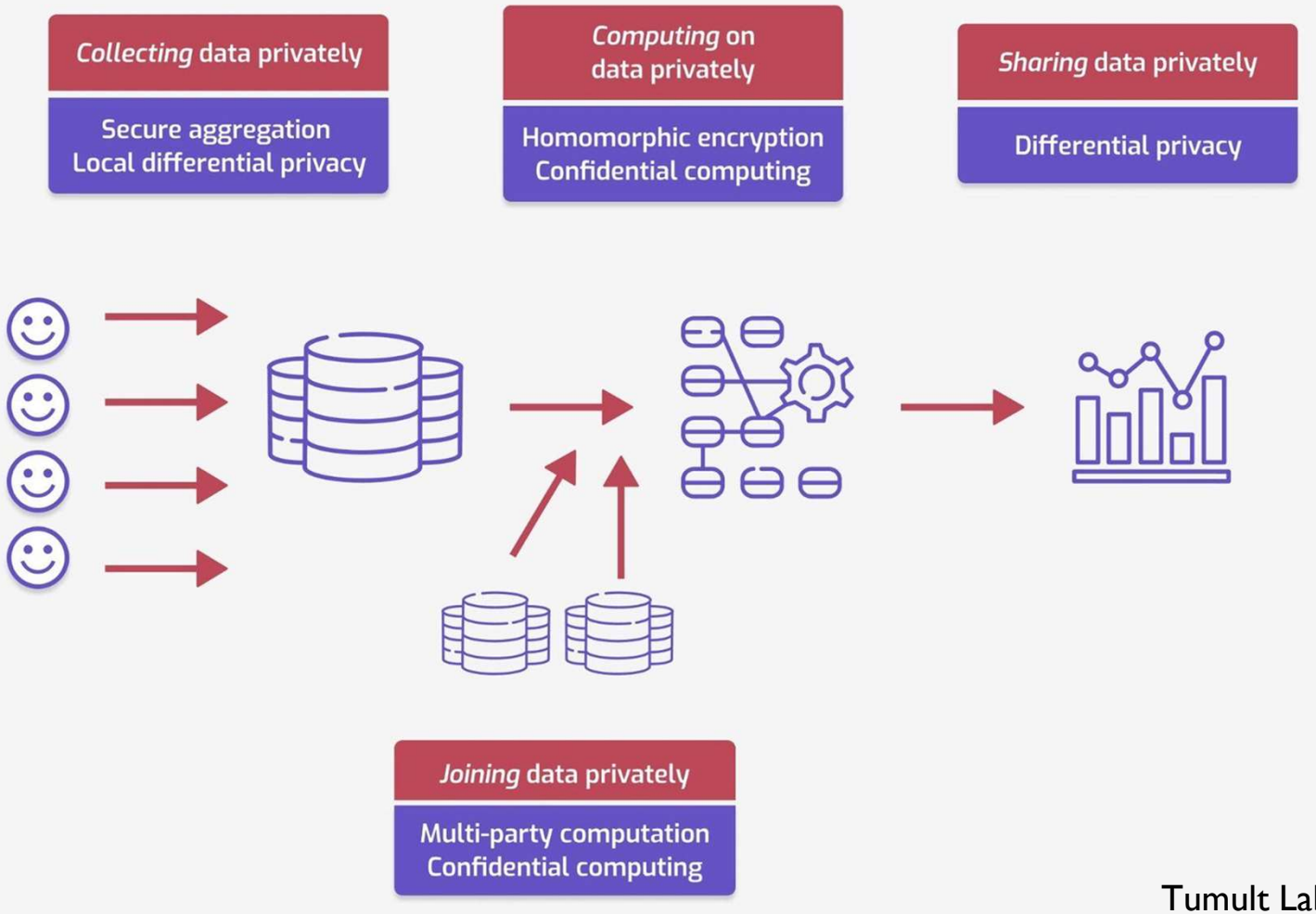

# DIFFERENTIAL PRIVACY

- Mathematical definition for privacy-preserving data analysis.
- **Goal:** Learn as much as possible about a group while learning as little as possible about any individual who is part of it.
- Outcome of analysis is essentially equally likely, independent of whether any individual joins, or refrains from joining, the dataset (Dwork, 2017).
- Achieved by adding “noise” (randomized responses) to data set.

# SYNTHETIC DATA




- Artificially generated data by an algorithm trained on a real data set.
- Goal: Preserving privacy in testing systems or training data for machine learning algorithms.
- Replaces original data while reproducing the statistical properties and patterns of the original set.


*Collecting* data privately
Secure aggregation
Local differential privacy
*Computing* on data privately
Homomorphic encryption
Confidential computing
*Sharing* data privately
Differential privacy
*Joining* data privately
Multi-party computation
Confidential computing

Tumult Labs

# PERMANENT PROGRESS

*Some examples:*

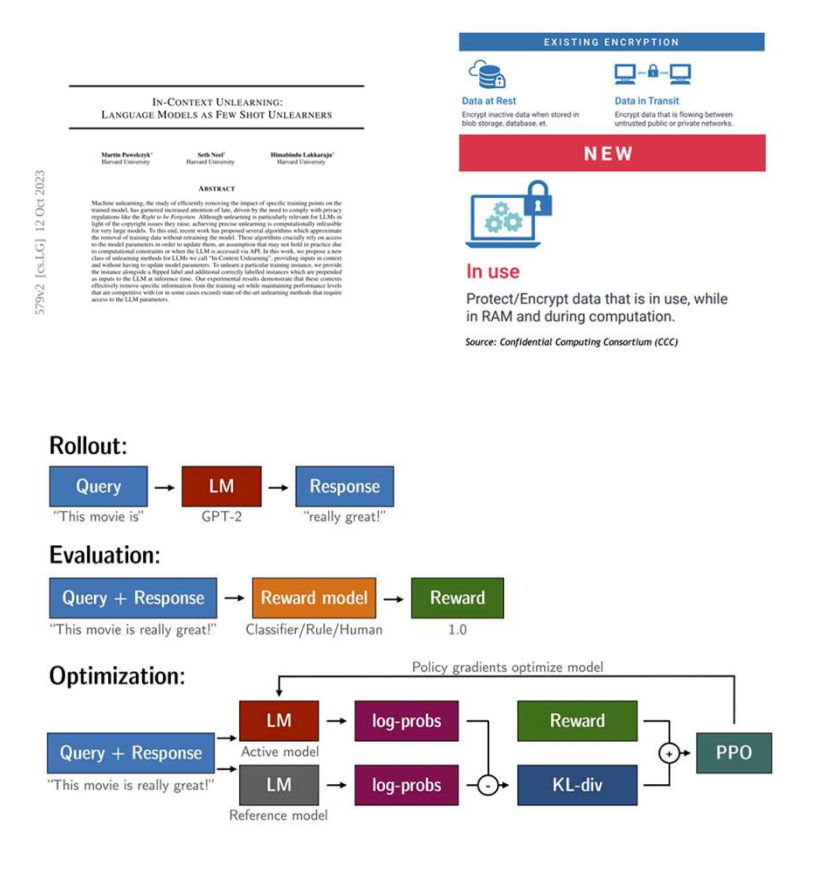
**Privacy rights (right to be deleted) & IP: Machine unlearning**.

**Hallucination: Retrieval Augmented Generation (RAG) -** supplementing prompts with external data from an external data source (internet, APIs, databases, or documents).

**Explainability/Interpretability:** Decomposing groups of neurons into interpretable features… (*Anthropic*)

**Model/Inference/Prompt Confidentiality: Trusted Execution Environments (TEEs) / Confidential computing**

**Best practices: Du-duplication, auditing, red teaming..**



# OUTLOOK: RESPONSIBLE BY DESIGN

- Current AI systems will likely not meet all regulatory requirements.
- Attention will shift from merely understanding IF models comply with regulation to understanding how to BUILD models that comply with regulation.
- Harnessing Advanced ML with Transparency and Explainability


## LET'S DISCUSS:

1. What do you perceive as the biggest challenge in implementing AI governance in your organization?

2. How prepared do you feel your organization is to integrate ethical considerations into its AI strategies?

3. How familiar are you with privacy-preserving machine learning techniques, and do you see them as a necessity in your current AI projects?

4. In your opinion, how well does your organization's current data governance framework support the effective management of AI systems?

# Thank you for your attention!

Follow me on Linkedin and get in touch for learning more about Daiki - AI governance & enablement