

**CATALYST**  
LEARN / THINK / CREATE

**AI Regulation, Risk Management and Responsible Practices: Overview & Opportunities**

KATHARINA KOERNER,  
TECH DIPLOMACY NETWORK

**BUILDING TRUST**  
**IN THE AGE OF AI:**  
ENSURING RESPONSIBLE INNOVATION

**OCTOBER 27-28, 2023**

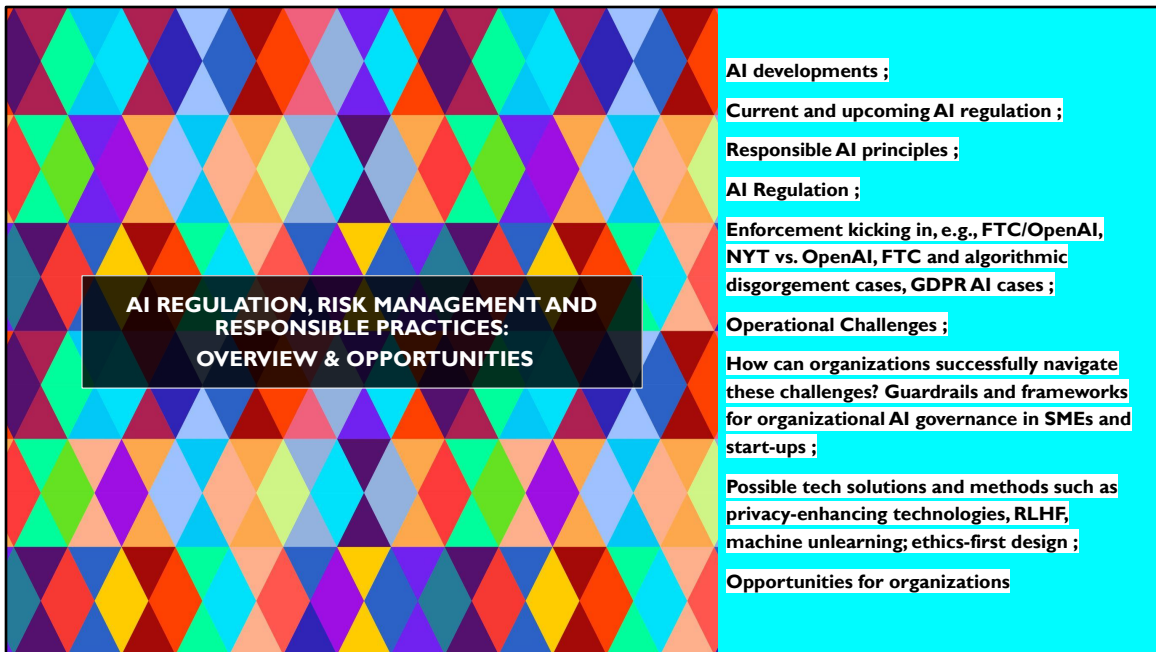
POWERED BY  
 Harvard Business School

My name is Katharina Koerner, I am currently with the Tech Diplomacy Network which is a think tank supporting diplomats on tech policy topics based in Silicon Valley. I also founded the AI Education Network to support AI Literacy for kids, and am working in corporate development for an AI enablement and governance platform in SV.

Before, I was principal research for technology at the international association of privacy professionals, a trade group with 80.000 members, where my focus areas were privacy engineering, privacy by design, privacy-enhancing technologies (PETs), and responsible AI governance.

My background is a PhD in EU Law with a number of certifications in info sec, privacy engineering, and ML.

It's an honor and pleasure to be here, and I will dive right in.



First, let me give you a short overview what I will cover:

**AI developments ;**

**Current and upcoming AI regulation ;**

**Responsible AI principles ;**

**Enforcement examples**

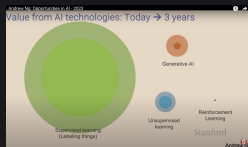
**How can organizations successfully navigate these challenges?**

**Operational Challenges**

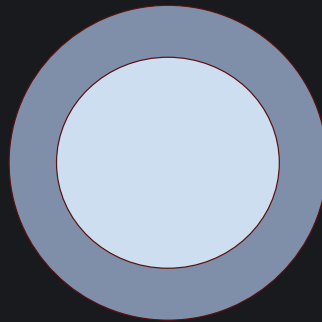
**Opportunities for organizations**

**Possible tech solutions**

## VALUE FROM AI TECHNOLOGIES: TODAY □ 3 YEARS



Andrew Ng, July 26, 2023, at Cemex Auditorium, Stanford University, hosted by the Stanford Graduate School of Business.



Supervised learning



Generative AI



Unsupervised learning

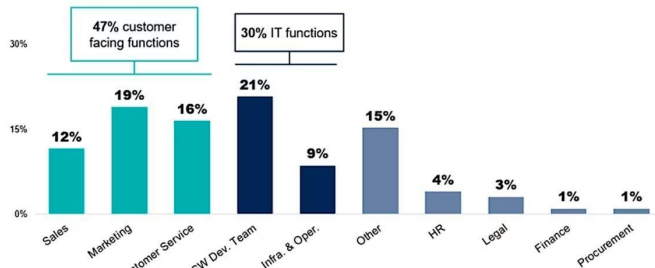


Reinforcement learning

With this slide adapted from Andrew Ngs presentation at Stanford in July, I want to demonstrate the state of the AI market and various ML techniques, With the inner circle being the current market or application adoption, and the outer one the projected growth, And we can see here that supervised learning is and will likely be by far the most important technique, with GenAI despite a lot of growth being comparatively small.

## GENERATIVE AI ADOPTION TRENDS: GARTNER INSIGHTS

Figure 1. Generative AI Investment by Business Function



Source: Gartner (October 2023)

Recent Gartner poll involving more than 1,400 executive leaders highlights accelerated Generative AI adoption.

- 45% of respondents are currently in the pilot phase for generative AI.
- An additional 10% have successfully transitioned generative AI solutions into production.

▣ Substantial increase compared to earlier findings from March/April 2023.

This month, Gartner predicted that by 2026, more than 80% of enterprises will have GenAI APIs or models, and/or deploy GenAI-enabled applications in production environments, up from less than 5% in 2023, in the areas of:

- Code Generation –
- Enterprise Content Management –
- Marketing Applications – chatbots and agents for contact centers
- Product Design & Engineering –

At the same time, Gartner warns that this is a very early stage and a hyped technology and advises to proceed, but don't over pivot.

And we see this in other report “Beyond the Buzz: A Look at Large Language Models in Production” by Predibase, which came to find that: enterprises are somewhat hesitant to embrace commercial LLMs, with 77% of the surveyed companies said they are not yet planning to use commercial LLMs in in real-world production applications,

primarily related to data privacy/protection of proprietary data concerns.


Additionally, a significant portion of organizations remains uncertain about the specific solutions they require, with 29% indicating that they don't yet know what they need.





**State of AI Report**  
October 12, 2023  
Nathan Benaich  
Air Street Capital

## KEY HIGHLIGHTS FROM THE STATE OF AI REPORT 2023

 <b>Research:</b> GPT-4 Dominance in Proprietary Models Challenges in Sustaining AI Scaling Trends Emerging Opportunities in Life Sciences Multimodality's Growing Significance	 <b>Industry:</b> NVIDIA's Market Cap Milestone Export Controls and Alternative Solutions GenAI Applications Fuel Investment	 <b>Politics:</b> Slow Progress in Global AI Governance Escalating Chip Wars AI's Impact on Sensitive Areas	 <b>Safety:</b> Existential Risk Debate in the Mainstream Challenges in Securing High-Performing Models Evolving Evaluation Difficulties
---	--	--	---

"State of AI Report 2023" by Air Street Capital is a great compilation of the most interesting things around the state of AI in the following key dimensions you see up there.

### Research

GPT-4's capabilities are so advanced and superior that there is a significant gap or difference between what it can do and what other proprietary or open-source models can achieve

While Efforts grow to beat proprietary model performance with smaller models, better datasets, longer context

- Unclear how long human-generated data can sustain AI scaling trends and what the effects of adding synthetic data are.
- Opportunities by LLMs and diffusion models for life science (drug discovery).
- Multimodality becomes the new frontier; excitement around agents grows substantially.

### Industry –

it is predicted that GenAI apps will have a breakthrough year, driving \$18 B of VC and corporate investments.

In Politics – it is predicted that Progress on global AI governance remains slow.

But what I want to highlight and am seeing clearly is that AI security and Safety is moving center stage. One example being The UK **AI Safety Summit** taking place on the 1 and 2 November.

Against the background that that Many high-performing models are easy to 'jailbreak', meaning using techniques or vulnerabilities that allow users to modify or misuse the model in ways that were not intended or authorized. including using the model for malicious purposes, bypassing security measures, or extracting sensitive information.

## RESPONSIBLE AI PRINCIPLES

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.

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



Against this backdrop of rapid developments, as we know, the concern that these developments will happen in a responsible manner have grown over the past few years.

In fact over the last few years literally hundreds of good governance guidelines on **Responsible or trustworthy or ethical AI** were published.

While these terms can sometimes seem fluffy or undefined, in fact responsible AI today is a set of common principles that include privacy and data governance, accountability, robustness, security, transparency and explainability, non-discrimination, and human oversight.

Some **prominent examples of Guidelines by public institutions or regulators** include UNESCO’s Recommendation, the OECD AI Principles, or the work by the Council of Europe and the High-Level Expert Group on AI set up by the European Commission, as well as the **White House Office of Science and Technology** issuing the Blueprint for an AI Bill of rights in October 2022.

Beyond that, one can find **countless self-regulatory initiatives** by companies.

e.g., the Microsoft Responsible AI Standard, with sarah being here which is amazing, and microsoft definitely being a leader in this space with a lot of public guidelines, Google’s Responsible AI practices,



Salesforce's Trusted AI principles  
or Facebook's five pillars of Responsible AI

Plus, we have fantastic collaborations of **industry, academia and nonprofits**, for example, the Partnership on AI initiative or the Global Partnership for AI.

And additionally, standardization bodies such as ISO/IEC, IEEE and NIST offer guidance.//

While those ethical or trustworthy AI principles are for the most part still not legally binding,

**they have in fact a big overlap with existing regulation, including privacy regulations,**

**meaning that privacy regulations or anti-discrimination law is covering many of those principles already when it comes to the processing of personal data.**

**United Nations** Office of the Secretary-General's Envoy on Technology

Home - High-Level Advisory Body on Artificial Intelligence

## High-Level Advisory Body on Artificial Intelligence

**The Global AI Imperative**  
Globally coordinated AI governance is the only way to harness AI for humanity while addressing its risks and uncertainties as AI-related services, algorithms, computing capacity and expertise become more widespread internationally.

**The UN's Response**  
To foster a globally inclusive approach, the UN Secretary-General is convening a multi-stakeholder High-level Advisory Body on AI to undertake analysis and advance recommendations for the international governance of AI.

**Calling for Interdisciplinary Expertise**  
Bringing together up to 32 experts from relevant disciplines from around the world, the Body will offer diverse perspectives and options on how AI can be governed for the common good, ensuring internationally interoperable governance with human rights and the Sustainable Development Goals.

**A Multistakeholder, Networked Approach**  
The Body, which will comprise experts from government, private sector and civil society, will engage and consult widely with existing and emerging initiatives and international organizations to bridge perspectives across stakeholder groups and networks.

**Supporting the Body**  
The UN is calling for support to the Body's operations and the secretariat based in the Office of the Secretary-General's Envoy on Technology (OSET). Through their support, contributors will strengthen stakeholder cooperation on governing AI in the face of pressing technical breakthroughs, and thereby contribute to better-governed AI globally.

For more info, contact [techenvoy@un.org](mailto:techenvoy@un.org)

**AI REGULATION**

**Roadmap - Toward Inclusive Governance**

- AUG 2023**  
CALL FOR EXPERTS  
1000+ nominees from across 128 countries
- NOV 2023**  
ANALYSIS & ENGAGEMENT  
Initial consultations
- Q1 2024**  
FURTHER CONSULTATIONS  
Across stakeholder groups and ongoing initiatives
- SEP 2024**  
SUBMIT OF THE FUTURE  
Member States consider Global Digital Compact

**US: AI regulated in sectoral approach, e.g., FTC, EEOC, CFPB, State privacy laws**

**EU: GDPR for personal data, upcoming: EU AI Act & EU Liability directive, extraterritorial scope**

**Canada, China, Brazil, ...**

**Global AI Governance initiatives: UN, G7, OECD, ...**

I am very much opposed to saying that while witnessing such immense technological progress, legally, we are in a wild west right now. Because in fact, there is plenty of law already in force regarding AI with more coming down the line.

So, in general in the US, we currently have a sectoral approach when it comes to regulation of AI:

With the primary enforcement regulators being:

- first, the Federal Trade Commission with its Section 5 of the FTC Act, Which has the authority to protect against “unfair and deceptive” practices related to AI systems which are affecting consumers.

The FTC has published several blog posts over the past 2-3 years about its expectation how companies have to build and deal with AI for consumer protection

to avoid discriminatory impacts, and the FTC announced very clearly to take action against companies that make claims about AI that are not substantiated, or to deploy AI before taking steps to assess and mitigate risks.

Besides the FTC, the Equal Employment Opportunity Commission (EEOC) is another example of a very active sectoral regulator for AI.

The EEOC can impose transparency requirements for AI, demand a non-AI alternative for individuals with disabilities, and enforce non-discrimination in AI hiring.

Furthermore, the Consumer Financial Protection Bureau (CFPB) mandates explanations for credit denials from AI systems and has the potential to enforce non-discrimination requirements.

And then there is The Department of Justice's Civil Rights Division (which I have not mentioned) which enforces constitutional provisions and federal statutes prohibiting discrimination across many facets of life, including in education, the criminal justice system, employment, housing, lending, and voting, including related to AI and automated systems,

In April, these four federal agencies also [released](#) a joint statement pledging to increase "enforcement efforts to protect the public from bias in automated systems and artificial intelligence" ("AI").

Apart from these sectoral approaches on a federal level, US states' interest in regulating AI services and products is on the rise.

Several U.S. states have enacted AI-related regulations.

These regulations address various aspects, including prohibiting AI in ballot processing (Arizona), establishing an Office of Artificial Intelligence and protecting personal data (Connecticut), allowing consumers to opt-out of AI-driven profiling (Delaware, Indiana, Montana, Oregon, Tennessee, Texas), regulating automated eye assessments (Georgia), creating an advisory council to study AI effects (Texas).

and we have **consent and notice requirements for using AI in the employment context.**

With a very prominent example being NYC law 144, requiring that employers that use an automated employment decision tool to confirm that such tools have undergone a "bias audit."

which has to be made publicly available,

- **to** notify employees and candidates that these tools will be used, and
- making available an **alternative selection process** for those who do not want to be reviewed by such tools.

or – last example of many – with California having introduced AB 331, a law specifically targeting automated decision tools, including AI, mandating developers and users to submit annual impact assessments.

And, despite the EU AI Act being so prominently on the horizon, we should not forget that also in the EU a variety of existing laws already apply to AI applications, e.g., the General Data Protection Regulation (GDPR) to personal data protection, the NIS Directive that covers AI systems in critical infrastructure sectors, or the Medical Devices Regulation (MDR) which applies to AI-based medical devices,

And all of this will be complemented by the proposed AI Act which will be the EU's first comprehensive horizontal, cross-sectorial regulation focusing on AI, currently under negotiations between the 3 EU regulatory bodies – EP, Council, and Commission.

As well as the **AI Liability Directive which will** address civil claims for damages in case any harm occurs due to AI systems.

The AI Liability Directive will establish legal and financial accountability for harms resulting from AI systems.

Additionally, a revision of sectoral safety legislation, including for machinery and general product safety, is underway.

On a global level, we have more flagship AI regulations, with the draft AI law in Brazil which prioritizes users' rights by requiring AI providers to provide information about their AI products, including explanations for AI decisions. Users have the right to contest AI decisions and request human intervention, and AI developers having to conduct risk assessments before launching AI products.

China has published a draft regulation for generative AI where generative AI must reflect “Socialist Core Values.” but also prohibiting the generation of fake news including deepfakes, and requiring synthetically generated content to be labeled.

And Canada has introduced a proposed legislation regulating AI systems with its *Artificial Intelligence and Data Act (AIDA)*, as well as has published a code of practice for generative artificial intelligence development and use.

On a non-binding global level, we have initiatives by the UN which just announced the creation of a 39-member advisory body to address issues in the international governance of artificial intelligence.

G7 nations are collaborating through the Hiroshima AI process to establish global

guidelines for advanced AI systems, including foundational models and generative AI. These guidelines will serve as the basis for a Code of Conduct to provide direction to AI tool developers, emphasizing safety and trustworthiness.  
In May 2019, the OECD released the OECD AI Principles, and now is trying to assist countries in implementing these principles through the OECD AI Policy Observatory

15 min

## EU AI ACT: COMING!



The Act will include a definition of AI systems and a classification system based on a risk-based approach.

The act aims to prohibit AI systems with unacceptable risks, authorize high-risk systems with specific requirements, and subject low-risk systems to minimal transparency obligations.

Having initially adopted the soft-law approach of ethical and responsible AI principles as mentioned before, the European Commission (EC) pivoted towards a comprehensive legislative strategy with the introduction of the draft AI Act in April 2021.

The Act will address fundamental rights and safety risks stemming from the development, deployment, and utilization of AI systems within the EU, with extraterritorial effect, similar to the GDPR.

AI systems will be categorized into 4 categories:

AI systems with unacceptable risks that will be banned, high-risk systems with specific requirements, limited-risk systems with specific transparency obligations, and low-risk systems with minimal transparency obligations.

In case of persistent non-compliance with the act, EU Member States will need to take appropriate actions to restrict or withdraw the high-risk AI systems from the EU market.

Fines are planned to go up to 30 million euros or 6% of worldwide annual turnover.

The Act is currently in the negotiation phase, with draft versions being

reviewed by the EC, the Council, and the European Parliament through trilogues, expected to be agreed upon end of this year or beginning of the next, and getting into force after a 18-24 months period, so end of 2025 or 2026.

**Substantial negotiations are currently still taking place regarding the definition of AI systems, the expansion of the list of prohibited AI systems, and the obligations for general-purpose AI and generative AI models like ChatGPT.**

## SCOPE: EU AI ACT APPLIES TO..

**Providers** (natural or legal person, public authority, agency or other body that develops an AI system or that has an AI system developed with a view to placing it on the market or putting it into service under its own name or trademark, whether for payment or free of charge):

- **regardless of their location**, when they introduce AI systems to the EU market, international law applies, or the AI system's output is used in the EU.  
when located in the EU, also when they introduce high-risk systems outside the EU, either directly or through a distributor.

**Deployers** (natural or legal person, public authority, agency or other body using an AI system under its authority, except when used during a personal non-professional activity):

- **regardless of their location**, when international law applies or when the system's output is used in the EU.
- located within the EU.

**Importers** (natural or legal person established in the Union that places on the market or puts into service an AI system that bears the name or trademark of a natural or legal person established outside the Union):

- located in the EU.

**Distributors** (natural or legal person in the supply chain, other than the provider or the importer, that makes an AI system available on the Union market without affecting its properties):

- located in the EU.

Under all circumstances, the new rules of the upcoming EU AI Act will have extraterritorial effect such as the GDPR. In simple terms:

**Providers** (natural or legal person, public authority, agency or other body that develops an AI system or that has an AI system developed with a view to placing it on the market or putting it into service under its own name or trademark, whether for payment or free of charge):

regardless of their location, when they introduce AI systems to the EU market, or the AI system's output is used in the EU.

when located in the EU, also when they introduce high-risk systems outside the EU, either directly or through a distributor.

**Deployers** (natural or legal person, public authority, agency or other body using an AI system under its authority, except when used during a personal non-professional activity):

regardless of their location, when international law applies or when the system's output is used in the EU.

located within the EU.

**Importers** (natural or legal person established in the Union that places on the market or puts into service an AI system that bears the name or trademark of a natural or

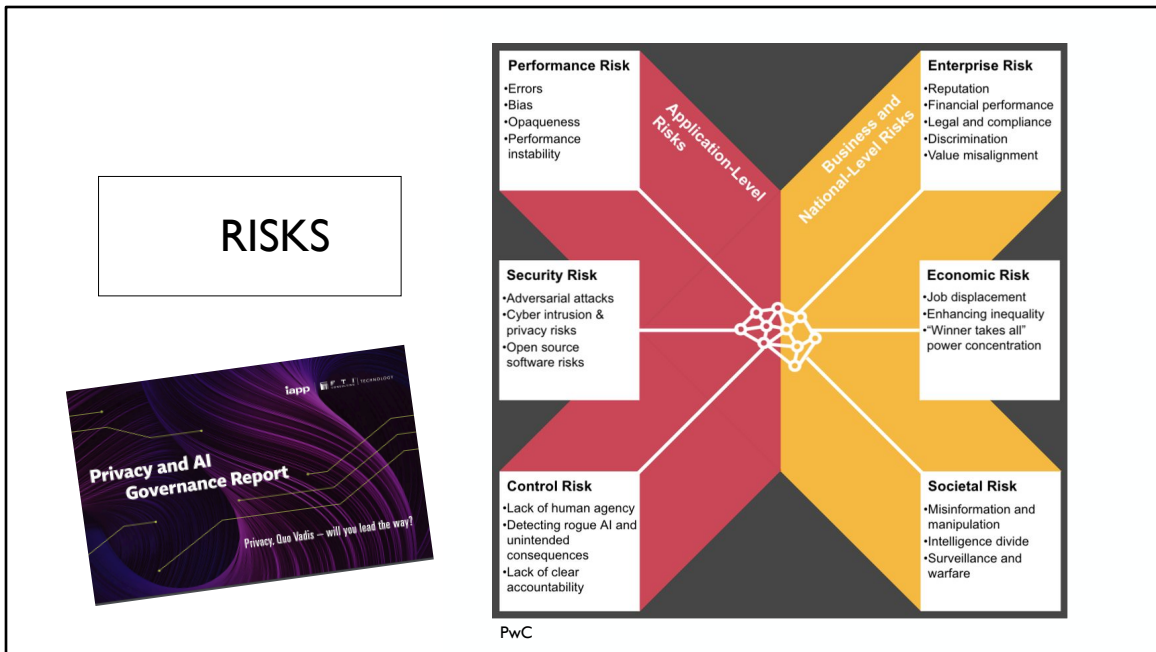


legal person established outside the Union;):

located in the EU.

**Distributors** (natural or legal person in the supply chain, other than the provider or the importer, that makes an AI system available on the Union market without affecting its properties):

located in the EU.



Regarding risks related to AI, we can see here in this graphic by PwC that there are definitely plenty.

These risks come in various flavors, and they're spreading across different levels,

This infographic by PwC I find very useful for a general overview and your reference.

But I actually want to refer explicitly to a study I conducted last year with FTC consulting and the IAPP, and here, the top three new risks for Organizations sampled within our study cited were: → First, Harmful bias. → Risk 2: Bad governance. → Risk 3: Lack of legal clarity brought by the changing regulatory environment.

In addition to the primary AI risks, organizations also named as their primary a lack of in-house AI skills, concerns about unintended consequences in the absence of clear requirements, increased liability and difficulties with third-party vendors, potential privacy issues in data use, and elevated security risks associated with networked AI systems.

## 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)

	OpenAI	cohere	stability.ai	ANTHROPIC	Google	Microsoft	Meta	AI21labs	ALPHA	EleutherAI	Totals
Draft AI Act Requirements	GPT-4	Cohere Command	Stable Diffusion v2	Claude 1	PaLM 2	BLOOM	LLaMA	Jurassic-2	Luminous	GPT-NeoX	
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	

As on good example of regulatory risks, how challenging this will become, is when we look at a first evaluation of foundation model providers for their compliance with the proposed EU AI Act, conducted by Stanford University's Center for Research on Foundation Models in June this year.

The results indicate a significant variation in compliance across providers, with some scoring less than 25% and only one provider scoring at least 75% at present.

Challenges were especially identified in 4 areas:

- (i) unclear liability due to copyright,
- (ii) unclear compute/energy,
- (iii) unclear risk mitigation, and
- (iv) lack of evaluation/standards/testing.

## EXAMPLES - REGULATORY RISK

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

Source: 2023 Foundation Model Transparency Index

	Meta	BigScience	OpenAI	stability.ai	Google	ANTHROPIC	cohere	AI21labs	Inflection	amazon	Average
	Llama 2	BLOOMZ	GPT-4	Stable Diffusion 2	PaLM 2	Claude 2	Command	Jurassic-2	Infection-1	Titan Text	
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.

Another even newer example is the collaborative effort by researchers from [Stanford University](#), [Massachusetts Institute of Technology](#), and [Princeton University](#), who published the Foundation Model Transparency Index about 2 weeks ago.

The index comprises 100 social and technical indicators to evaluate the transparency of developers' practices throughout the development and deployment of foundation models.

fanned out into three major practice domains of foundation model development:

- upstream transparency (which includes the resources used to build a foundation model),
- model-level transparency (encompassing the model's capabilities, risks, and evaluations),
- downstream transparency (encompassing distribution, usage policies, and affected regions)..

Some example findings from their research:

- Lack of Downstream Impact Disclosure: None of the assessed developers

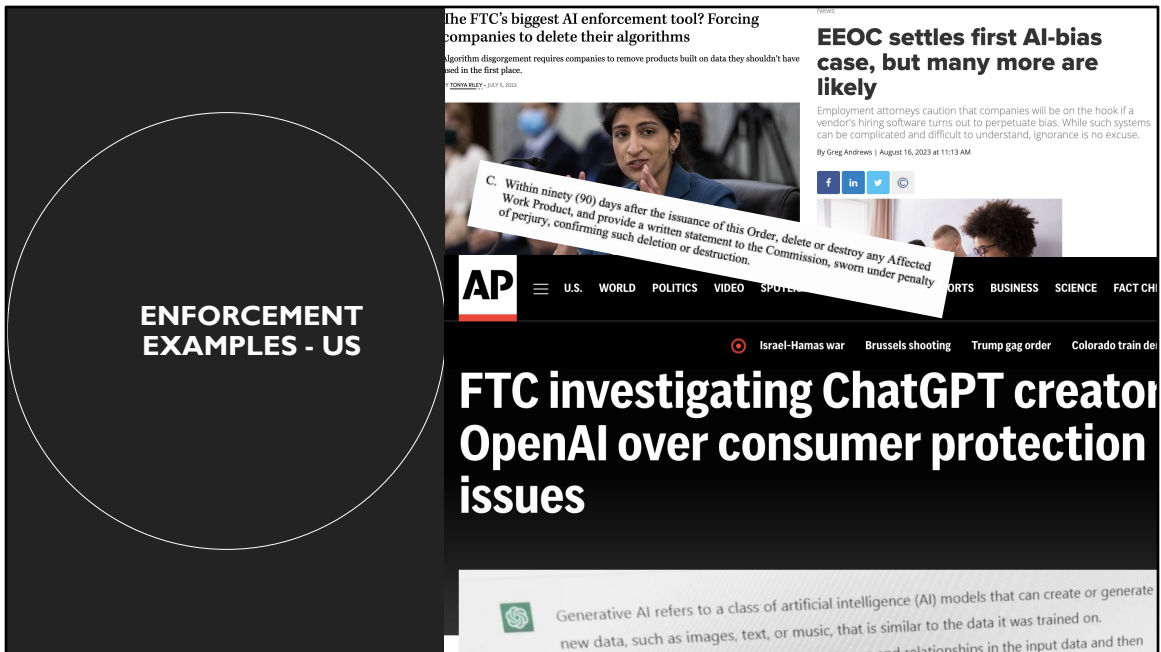
currently provide significant information about the downstream impact of their flagship models, such as the number of users affected, market sectors impacted, or mechanisms for addressing harm caused by the models.

- Developer transparency regarding a model's capabilities doesn't extend to transparency concerning its limitations, evaluating potential harms their models could enable, and none developer had externally reproducible or third-party assessments of mitigation effectiveness.

- Open developers, who share model weights and possibly data, demonstrate a distinct advantage in transparency over their closed counterparts, such as API providers.

- Open developers excel in transparency regarding upstream resources and maintain comparable transparency regarding downstream use compared to closed developers.

The paper also includes a long and precise analysis on calls for transparency by regulators worldwide, industry itself, academia and non-profits.



Where does that leave us right now? As mentioned, in the U.S., the FTC holds AI developers and companies accountable under Section 5 of the FTC Act, the US Fair Credit Reporting Act as well as the Equal Credit Opportunity Act.

And we do have enforcement examples here already with a very severe enforcement mechanism:

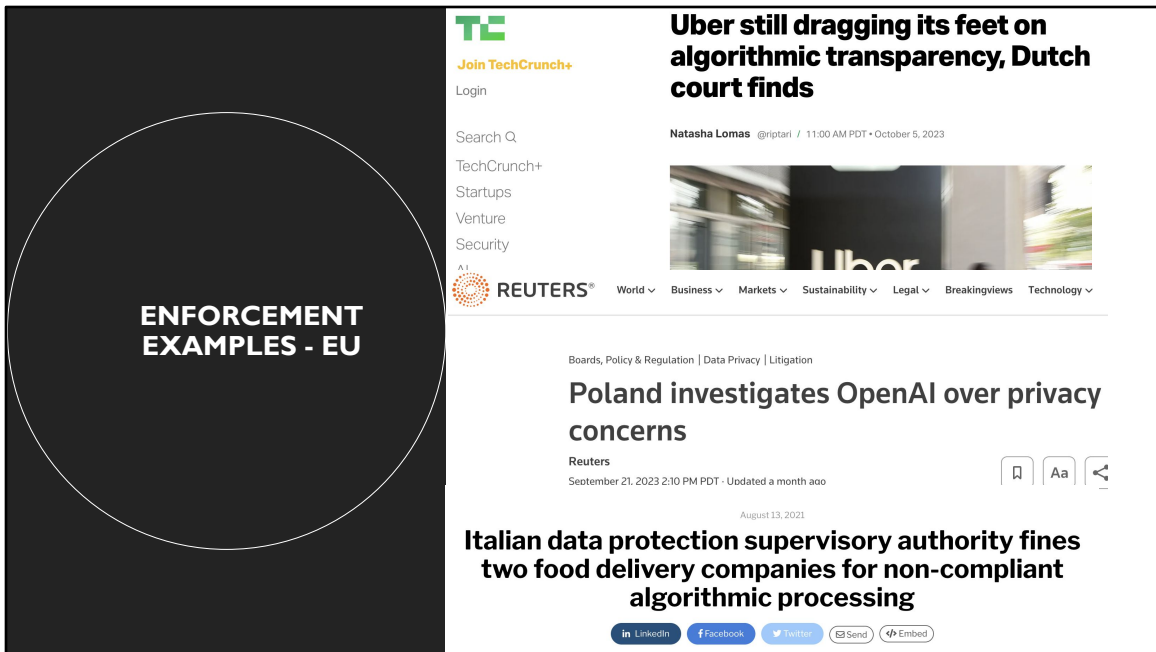
In the matter of [Everalbum](#), the **first facial recognition misuse settlement**, the FTC not only focused on the obligation to disclose the collection of biometric information to the user and obtain consent, it also demanded that the illegally attained data, as well as models and algorithms that had been developed using it, be deleted or destroyed. – something also referred to as “algorithmic disgorgement”

The same thing the agency also demanded in March 2022 from WW International — formerly known as Weight Watchers — ...it had to destroy the algorithms or AI models it built using personal information

collected through its healthy eating app from kids as young as 8 without parental permission, in addition to a fine of \$1.5 million and the order to delete the illegally harvested data.

With these orders, the FTC [followed](#) its approach in its Cambridge Analytica order from 2019, where it had also required the deletion or destruction not only of the data in question but all work products, including any algorithms or equations that originated in whole or in part from the data.

Another example is The Equal Employment Opportunity Commission's (EEOC) which just recorded its first-ever settlement in a case involving AI discrimination in the workplace in August with a tutoring company which's AI-powered hiring selection tool automatically rejected women applicants over 55 and men over 60.



Examples of AI enforcement cases in the EU are currently based on the GDPR

e.g., In December 2021, the Dutch Data Protection Authority fined the Dutch Tax and Customs Administration 2.75 million euros for a GDPR violation involving a discriminatory ML algorithm that flagged double citizenship as high-risk, which eventually also led to the government having to step back.

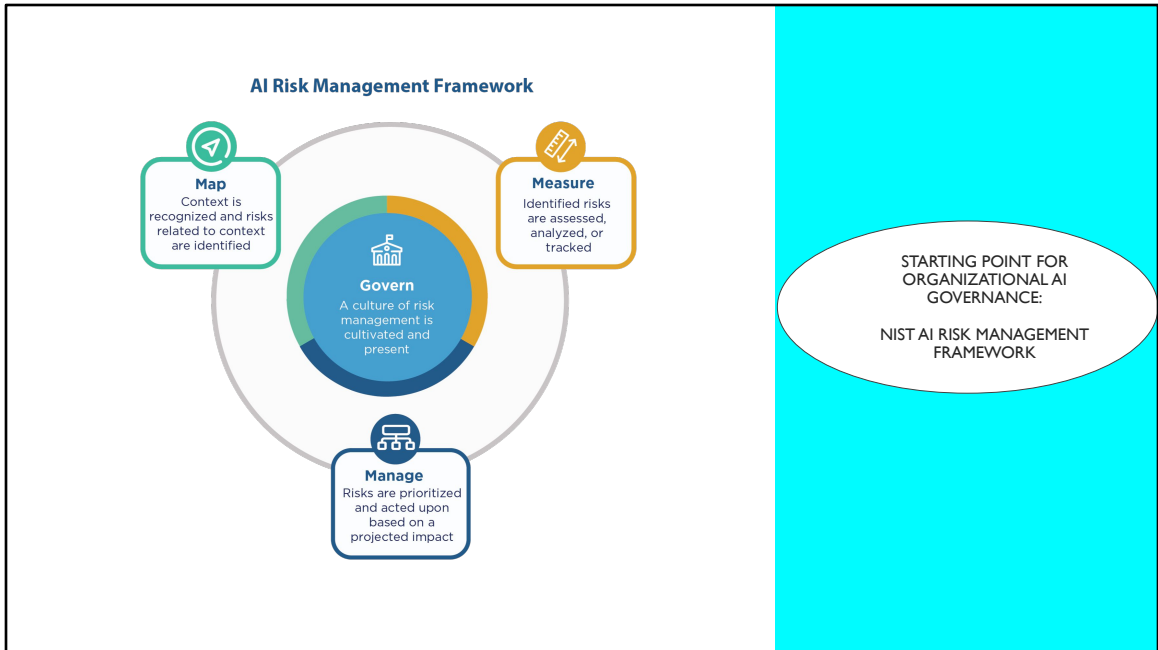
Or in August 2021, Italy's Data protection authority fined food delivery companies Foodinho and Deliveroo \$3 million each for GDPR breaches related to lack of transparency, fairness, and accurate information about their algorithmic management of riders.

We also all know about Italy's ban of Chatgpt and then being satisfied with increased transparency on data processing, more opt-out rights, including being able to toggle off the option for conversations to be used **for training ChatGPT's algorithms.**

but there is also a newer example with Poland's data protection authority



starting investigating OpenAI following a GDPR complaint from an applicant in September.

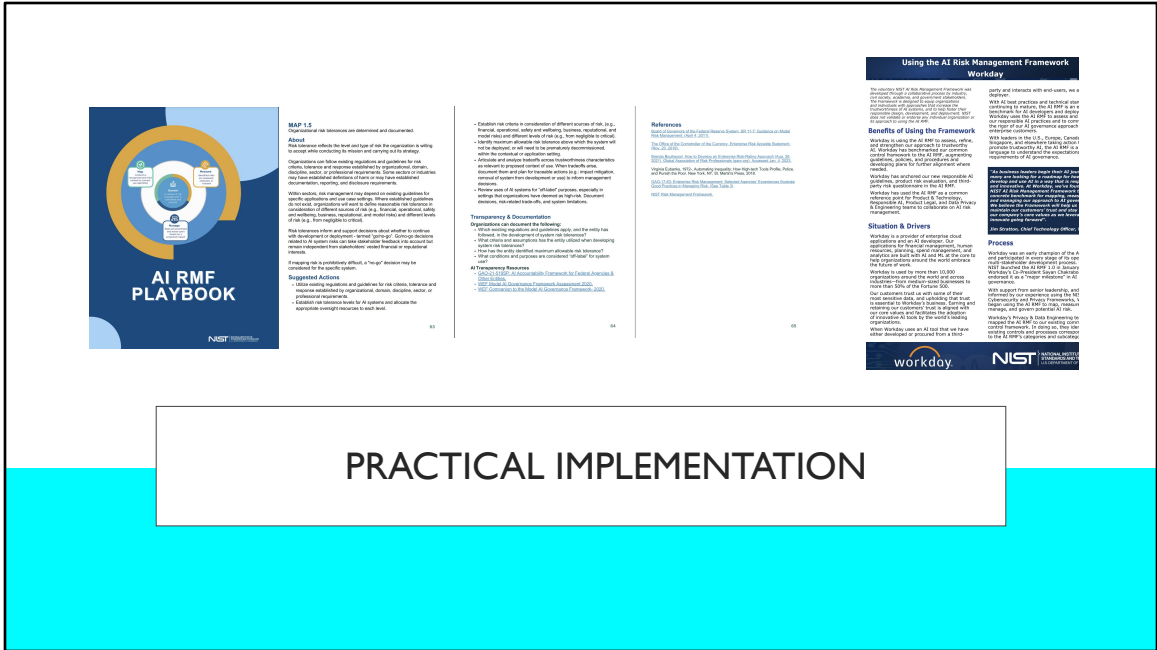


- So, what can we do to avoid this?
- I am a big supporter of the NIST AI Risk Management framework, which was released in its version 1 in January this year.
- The NIST AI RMF is a voluntary guidance document aimed at fostering trust in AI, by promoting innovation, and risk mitigation at the same time.
- Unlike the upcoming EU AI Act, it lacks binding legal requirements, or enforcement mechanisms,
- Nevertheless, it's gaining significant traction in the U.S., with endorsements from tech companies like Microsoft and support from the National Artificial Intelligence Advisory Committee, which advises the U.S. President and the White House National AI Initiative Office.

The framework consists of two main sections:

1. **Identifying Risks and Trustworthy AI Qualities:** Organizations need to assess AI-related risks, considering the extent of harm and likelihood of events. Challenges include measuring risks related to third-party software, hardware, and data, as well as tracking emergent risks. The framework doesn't prescribe

1. risk tolerance.
2. **Core Functions of Governance, Mapping, Measurement, and Management:**  
These functions are flexible and adaptable for different AI lifecycle stages.
  1. **Govern:** Establishes accountability structures, encourages diversity and safety-first AI practices.
  2. **Map:** Helps categorize AI systems based on capabilities, goals, and potential impacts.
  3. **Measure:** Supports risk analysis, assessment, and monitoring using appropriate methods and metrics.
  4. **Manage:** Involves prioritizing risks, allocating resources, and implementing monitoring and improvement mechanisms, especially for third-party sources.



At first glance, in general, I find NIST in general not super accessible. But in fact, they become very concrete when really taking advantage if them and studying all of the materials that come along with the basic framework. So, here as well, The functions that mentioned are meant to be customized by organizations to align with their specific needs, legal requirements, and objectives.

For instance, the "Map" function includes five categories, such as understanding the context of AI systems, interdisciplinary collaboration, organizational goals, and risk tolerances. and then, there are associated subcategories to those categories which provide detailed recommendations.

To assist organizations in customization, NIST also provides a comprehensive Playbook that explains the subcategories and offers specific actions.

For example, the subcategory "MAP 1.5" is about **determining organizational risk tolerances, and there are 3 pages in the Playbook with concrete action points, which I have put up on the slide just to get an expression. Here, the Playbook recommends**

formalizing risk acceptance levels by raising and documenting the questions/answers like:

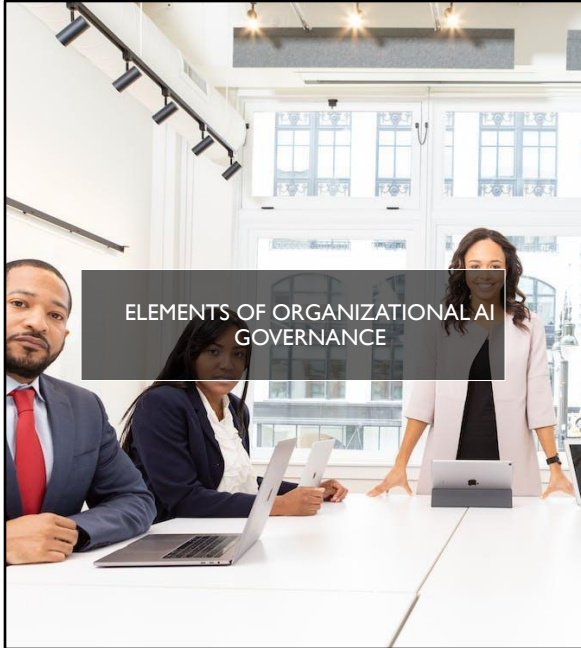
Which existing regulations and guidelines apply, and the entity has followed, in the development of system risk tolerances?

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?

So, to promote responsible AI practices, it is tremendously helpful when stakeholders within organizations familiarize themselves with the framework's core functions, categories, and subcategories. to identify gaps in essential elements for AI risk management and prioritize actions. The Playbook's actionable steps can also just be used to guide discussions and planning within specific AI/ML projects. NIST is also planning to update the Playbook approximately twice per year.

NIST also shares **best practice examples** for successful AI RMF implementations which you see here - which explains which approach workday has taken.



ELEMENTS OF ORGANIZATIONAL AI GOVERNANCE

- A corporate function is appointed or established to define standards, requirements, internal guidelines, processes and execution of responsible AI (e.g., "Office of Responsible AI").
- Conduct a thorough survey of your business operations to identify where AI is being used in decision-making processes and assess associated risks.
- The Office for Responsible AI can serve as sparring partner for business units (e.g., HR, operations, marketing, sales, finance, R&D) who plan an AI deployment.
- The Office can help with an initial assessment of the projects in terms of compliance and responsible development.
- After this first assessment, inputs and recommendations for an AI project can be given by an ethics board, privacy steering committee or enterprise data council.
- The committee assesses pros and cons of specific AI projects along the internal guidelines about technologies, processes and best practices, and points out risks that needs to be mitigated.
- Additional committees coordinated through the Office for Responsible AI can focus on rules, processes, and tooling for the practical implementation of responsible AI in engineering.

**With the FTI/IAPP study we asked organizations who have already operationalized their RAI guidelines how they went about it, and what their recommendations are, and in general it is:**

- They establish a corporate function to define standards, requirements, internal guidelines, processes and execution of responsible AI (e.g., "Office of Responsible AI"). this can, e.g., sit in security, or legal/privacy as well as an AI ethic board or group.
- Conduct a thorough survey of your business operations to identify where AI is being used in decision-making processes and assess associated risks. so basically, a comprehensive survey of your business operations to identify AI usage and assess associated risks.
- plus, for new projects, the Office for Responsible AI can serve as sparring partner for business units who plan an AI deployment, and help with an initial assessment of the projects in terms of compliance and responsible development.
- After this first assessment, inputs and recommendations for an AI project can be given by an ethics board, privacy steering committee or enterprise data council, which can point out risks that needs to be mitigated.
  1. An AI Ethics board can also Advise the board of directors on various aspects, such as research priorities, commercialization, partnerships, and fundraising.
- 2. Oversee model releases and publications, including staged releases.

- 3. Support risk assessments by using a risk taxonomy, providing comments on risk heatmaps, and commissioning third-party audits or evaluations.
- 4. Review the company's risk management practices to ensure compliance with regulations, standards, and internal policies.
- 5. Interpret AI ethics principles in both abstract and concrete cases to influence risk-related decisions.
- 6. Serve as a trusted contact point for whistleblowers.

- When selecting members for an ethics board, companies should consider factors such as expertise, diversity, seniority, and public perception. Board members should possess technical, ethical, and legal expertise, and diversity should be ensured in terms of gender, race, and geographical representation to incorporate diverse perspectives in risk assessment.

- Additional committees coordinated through the Office for Responsible AI can focus on rules, processes, and tooling for the practical implementation of responsible AI in engineering, e.g., in an "Ethics by Design" playbook for workflows. Then, it is also important to make these documents and processes accessible and train on them, e.g., by having RAI champions in business units and having a "hub and spoke" approach with centralized governance and stakeholders from various business units.

**Fig. 1. The challenge of managing AI bias**

**The AI Register**  
 Funnily enough, AI models must follow privacy law – including right to be forgotten  
 Well jr · · · take a look at the training data, off... wait  
 The 13th of 2023, 10:27 PM

**Right to be Forgotten in the Era of Large Language Models: Implications, Challenges, and Solutions**  
 Dawen Zhang<sup>1,2</sup>, Pamela Finckenberg-Bromer<sup>1</sup>, Thong Hoang<sup>1</sup>, Shikong Pan<sup>1,2</sup>, Zhenchang Xing<sup>2</sup>, Mark Staples<sup>1</sup>, Xuei Xu<sup>1</sup>

**Fig. 2. A model of explainability components.**

**Fig. 3. Complexity of bias.** The first node entry in each category in the diagram are hyperlinked in the Glossary. Clicking them will bring up the definition in the Glossary. To return, click on the current page number (1) printed right after the glossary definition.

nevertheless, with all these efforts, Operational Challenges still remain a real concern, as speaker before me have pointed out.

In the FTI/IAPP study that I conducted and mentioned before, Responses from organizations indicate their focus on how to avoid bias in AI systems, and being definitely challenged by that.

They expressed uncertainty how to address bias risk effectively, and the need for clear definitions of harm, fairness guidelines, established risk thresholds, common bias detection tools, and benchmarking for specific use cases.

How complex bias is also demonstrated by NISTs March 2022, report titled "Towards a Standard for Identifying and Managing Bias in AI." which lists numerous sources for bias, which one can basically be bucketed into:

1. Statistical Context: where Bias is primarily emerging as a statistical phenomenon in technical systems,
2. Legal Context: where bias refers to discriminatory practices, including disparate treatment in areas like consumer finance, housing, and



employment, as well as discrimination based on disabilities, race, gender, or age.

3. Cognitive and Societal Context: meaning Cognitive biases within AI teams, both individually and collectively, which influence decisions related to data usage, AI model development, system placement, and the necessity of AI. Institutional-level systemic biases also affect organizational structures and decision-making.

other examples is security, For example,

[OWASP®](#) recently listed prompt injection as the top security vulnerability for AI applications built on LLMs which fall into two categories:

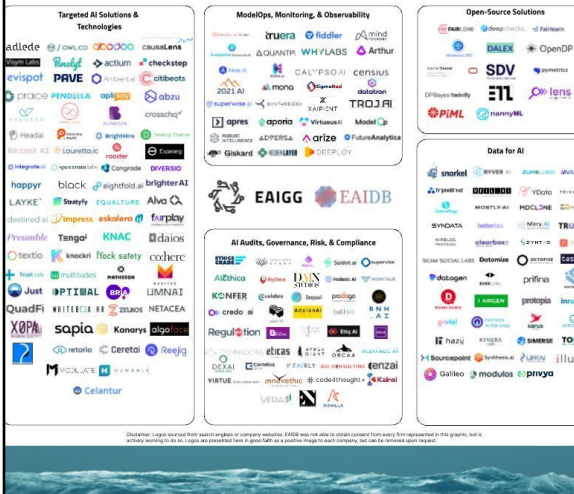
- Harmful prompts are injected as inputs to the application.
- or to recover previously input prompts at the service provider end.

and pointed out that currently there are no established systematic techniques to prevent prompt injection in LLM-integrated applications.



Now, I want to give some pointers what all of this means in regards to opportunities and what reserach we are seeing addressing some of the issued that we mentioned over the course of the past 2 days.

## Ethical AI Startup Landscape



### AN EMERGING ECOSYSTEM AROUND RESPONSIBLE AI

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

Identifying AI Deployment Across Organizations  
Ensuring Compliance with Regulatory Requirements

First Line of Defense in AI Governance



Startups in the field of responsible AI have a good chance to succeed because of new rules and trends.

Most of these tools focus on questions like 'Where is AI deployed across my organization?' and 'Do my AI systems meet technical requirements set by regulators around bias and fairness?'.

As AI continues to proliferate across enterprises, these tools will serve as an essential first line of defense in AI governance, helping enterprises map potential risks across the organization and understanding where more attention needs to be directed.

But identifying potential AI risk is just the start.

As regulation matures and enterprises will see that many of their AI programs do not meet regulatory requirements, attention will likely shift from merely understanding if models comply with regulation to understanding how to BUILD models that comply with regulation.

E.g., in credit model management, lenders face challenges in harnessing

advanced ML techniques, because models are overly complex and lack transparency and explainability. New services [|](#) offer explainable credit models.

If you want to get an overview of this space, Since May 2022, the ethical AI DB aims for systematic categorization and regular updates for a comprehensive picture of start-ups in the responsible AI ecosystem.

Their website provides a curated compilation of currently 260 responsible AI enablement startups. It is updated semiannually.

Some observations by EAIDB on key trends in the in the responsible AI market :

- During the first half of 2023, the focus has predominantly been on generative AI.
- The progress of the draft EU AI Act has further spurred movement within this ecosystem. The EU AI Act is expected to do for the responsible AI market what GDPR did for privacy.
- The AI Security sector has witnessed significant growth,

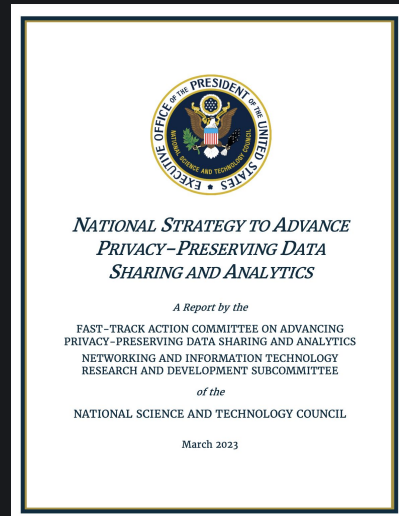
## PRIVACY ENHANCING TECHNOLOGIES IN AI/ML

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 disassociability, **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.”*



So in this context I would like to highlight PETs. PETs can be seen as advanced risk mitigation methods to demonstrate state-of-the-art “privacy by design” for data processing, with 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).

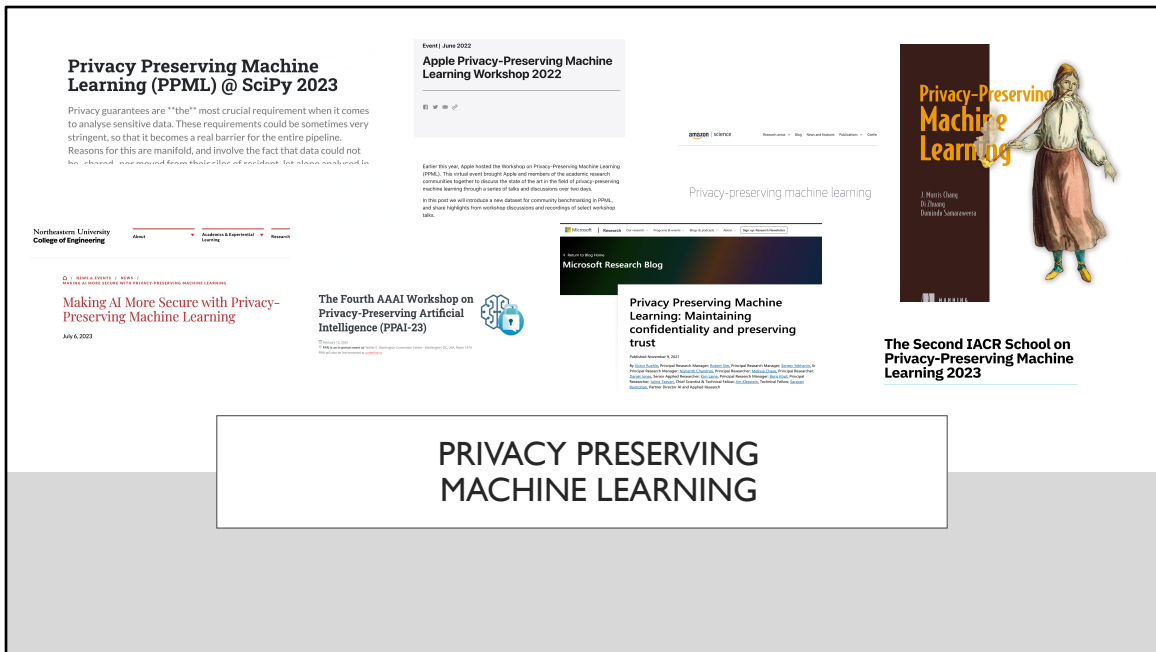
PETs are getting wide policy support by governments around the world, for example,

in March 2023, the US National Science and Technology Council (NSTC), reporting to the White House, published its US National Strategy to Advance Privacy-Preserving Data Sharing and Analytics, where they define

“Privacy-preserving data sharing and analytics (PPDSA) as methods which utilize cryptographic techniques, which inherently satisfy the confidentiality objective. with The distinctive aspect of PPDSA approaches being their ability to achieve disassociability, preventing authorized entities from establishing linkages between data and individuals' identities, thereby enhancing privacy even with authorized data usage.

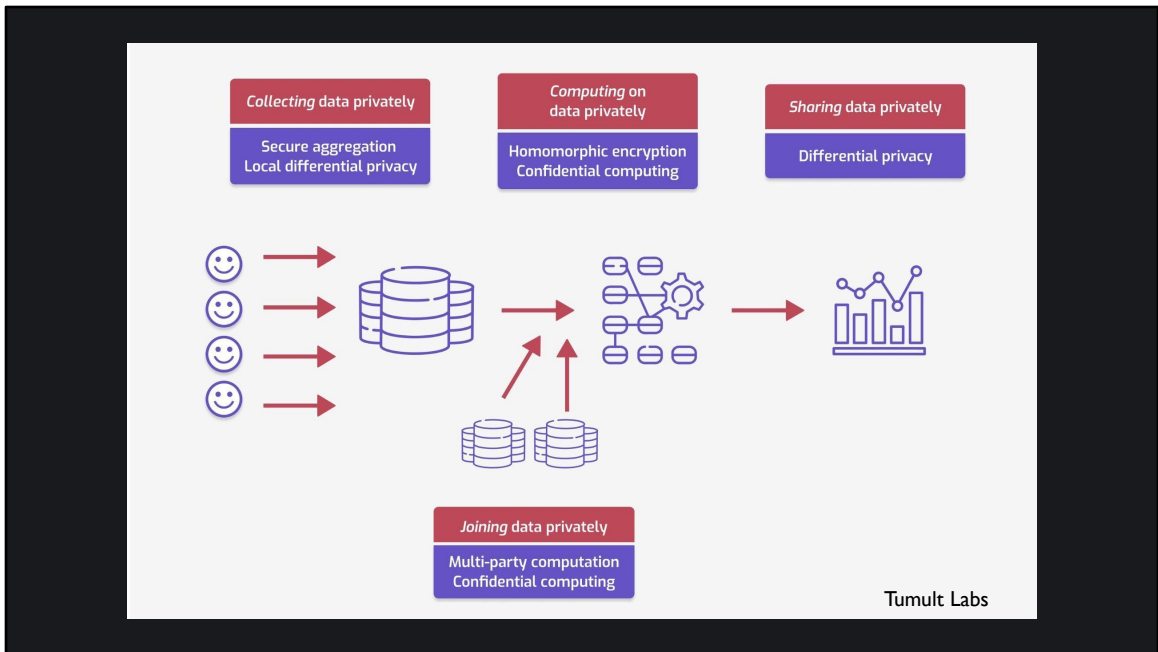
They continue and emphasize:

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.



**PRIVACY PRESERVING MACHINE LEARNING**

This slide I just put together to demonstrate what a huge field of research and education PPML has become. There are books just released, we have Apple, Microsoft, conferences, universities being involved in research, And when seeing this, I think there is a tremendous opportunity to establish PETs in the space of responsible AI using the term ppml more widely.



Here you see one of the best infographics on some key concepts and techniques within privacy-preserving machine learning and when to apply them:

1. **Differential Privacy:** Differential privacy introduces randomness or noise to query responses to protect privacy so that **the output of a differentially private analysis will be roughly the same, without being able to tell if a single data point contributed your data.** You can apply it to input or output data.

3. **Homomorphic Encryption:** This cryptographic technique allows computations to be performed on encrypted data without decrypting it. So that means it enables privacy-preserving data analysis and machine learning on encrypted data.

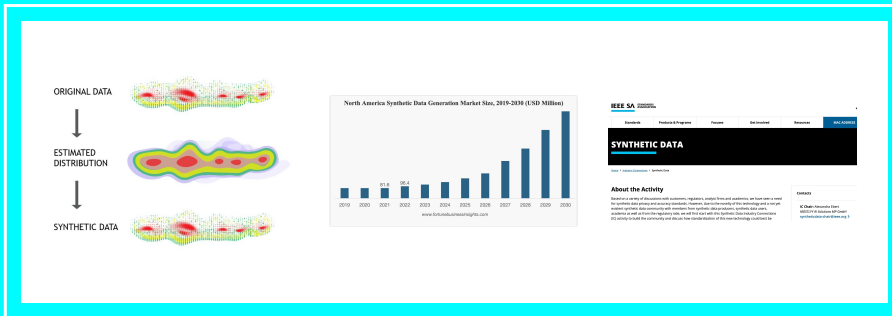
4. Confidential computing / = TEEs / secure enclaves refers to a technology that can isolate data within a CPU while it is being processed. It allows AI models and data to be shared without exposing IP and **sensitive** data, with a common example being **Intel® SGX** enclaves.

4. **Secure Multi-Party Computation (SMPC):** allows multiple parties to jointly compute a function over their inputs while keeping those inputs private. It ensures that no party learns more than what is required from the computation.

Regarding homomorphic encryption, multiparty computation, and trusted execution environments mentioned here, I want to emphasize that we have a totally new concept coming up: Because They protect “data in use” - We all know about the



importance of securing data when it is stored (protection of data at rest), when it is sent (protection of data in transit). And we now have techniques that protect data in use, meaning while the data is processed and computed upon. And that's huge and why a lot of people are very excited about those new techs.



## SYNTHETIC DATA

In this context, let's also not forget about synthetic data.

A synthetic data set is generated by taking a relational database, creating a generative machine learning model for it, and generating a second set of data which has the same mathematical properties as the real-world data set it's standing in for,

The result is a data set that mimics the general patterns and properties of the original along with enough "noise" to mask the data itself. so usually combined with differential privacy

Synthetic data can be used to test machine learning models or build and test software applications without compromising real, personal data.

## PERMANENT PROGRESS

Some examples:

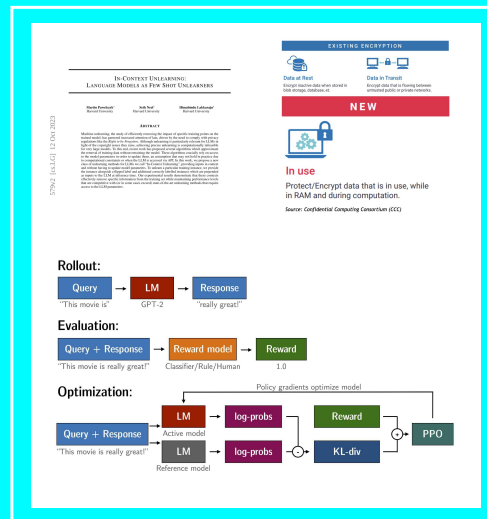
**Privacy rights (right to be deleted) & IP: Machine unlearning.**

**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..**



Apart from these techniques, and regarding other challenges that we have mentioned over the course of the last 2 days, I find it super exciting and encouraging what we see in terms of ongoing efforts to make AI and generative AI more secure.

### Privacy Rights & IP Protection: Machine Unlearning

- One significant aspect of AI security is safeguarding privacy rights like the Right to be forgotten and IP.
- To address this, researchers are exploring the concept of 'machine unlearning.' This involves techniques to erase or modify information stored in AI models.
- The two main approaches currently discussed are Data Reorganization and Model Manipulation.
- Data Reorganization includes methods like Data Obfuscation, Data Pruning, and Data Replacement, while Model Manipulation involves techniques like Model Shifting, Model Replacement, and Model Pruning.
- **Hallucination: Retrieval Augmented Generation (RAG)**
- As we know, Hallucinations in AI responses are a common challenge.
- Here, Retrieval Augmented Generation (RAG) - mentioned yesterday already -

- is a technique that mitigates this issue. It supplements AI prompts with external data sources like the internet, databases, or documents. So, instead of immediately generating a response, RAG instructs the model to retrieve accurate information, reducing knowledge gaps and hallucinations.

### **Explainability & Interpretability**

- Another critical aspect is the explainability and interpretability of AI models. Researchers are working on methods to break down complex AI models into interpretable features.
- That's very challenging of course as The understanding of neural networks' mathematical operations, doesn't explain their observed behaviors fully.
- Here, we are also seeing constant progress, e.g. led by Anthropic, introducing decomposing language models into interpretable features instead of individual neurons, suggesting that these features offer a more understandable unit of analysis. but scaling it to larger models remains a future challenge.

### **Model, Inference, and Prompt Confidentiality**

Trusted Execution Environments (TEEs) and confidential computing technologies li have already mentioned. There are plenty of start ups which apply these to LLMs, using it as access controls, or with the vision to have the service provider running TEEs so that

- The user encrypts the data using a key not known to the service provider
- They send the encrypted data to the service provider
- The service provider then uses an enclave on the data in a way that prevents the service provider from seeing the data in clear
- Finally, the transformed data is re-encrypted, and the encrypted result is sent back to the client

- **Best Practices**

- Best practices in AI security encompass various measures, including deduplication of training data, robust auditing, and ongoing research into emerging threats and ramping up of red teaming.

These strategies and innovations are part of an ongoing effort to enhance AI security.

Like we also see with initiatives like the UK AI summit next week.



## 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

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.

So I think we will see many new services and probably retraining of LLMs with clear data set transparency, and all in all, I'm very optimistic and excited about these developments.

Thank you



LET'S DISCUSS

- Which areas are currently more relevant to you: AI Enablement or AI Governance/AI Risk Management?
- What are opportunities & challenges you encounter in these areas?
- Where do you see most potential for responsible innovation?