AI Privacy Frameworks

Madison Sveum, Dui Cao, Ruike Lin, Stefan Mihailescu.

Key Questions:

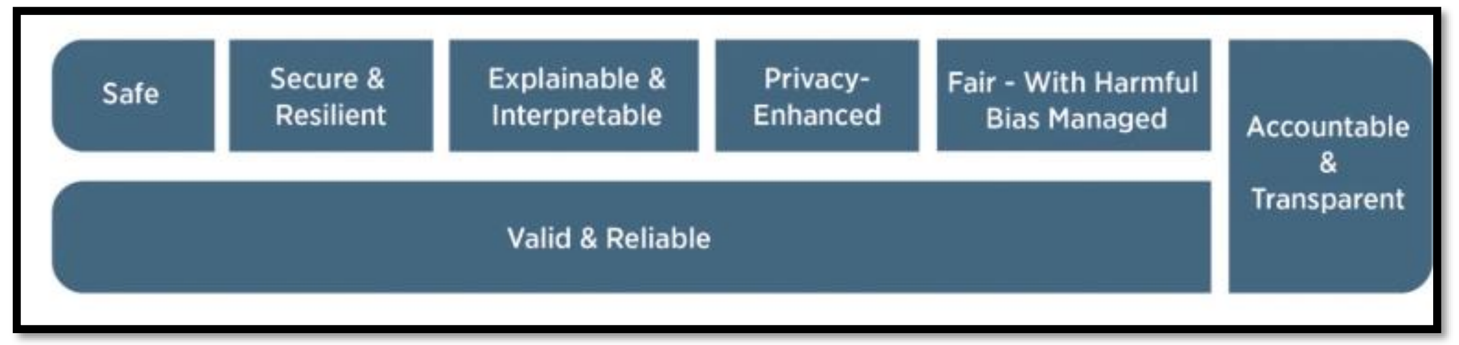
- Why are AI Privacy Frameworks important?
- How are AI Privacy Frameworks used?
- What are the key elements that are discussed in AI Privacy Frameworks?
- How do AI Privacy Frameworks protect users?
- What are the most popular AI Privacy Frameworks and what benefits do they bring to the table

Introduction and Objectives

Introduction: With the increase of AI systems, there has been an increase of concern regarding privacy, surveillance, and misuse of AI tools. This concerns arise from the large-scale data collection that AI companies use to train their LLMs and how unrestricted LLMs are leading to misuse by users. Due to this, AI privacy frameworks have been created to ensure that users practice safe use of AI tools so that their private remains secure

AI Privacy Framework Objectives:

1. Protecting Individual Privacy Rights
2. Ensure Transparency
3. Mitigating Risks and Harmful attributes
4. Promoting Accountability
5. Encouraging Ethical Use
6. Supporting Security and Data Integrity



Safe

Secure & Resilient

Explainable & Interpretable

Privacy-Enhanced

Fair - With Harmful Bias Managed

Accountable & Transparent

Valid & Reliable

Sample AI Privacy Framework

Disclaimer: This is a summarized version of the NIST AI privacy framework. The document to the whole framework can be found in the linked document in references.

Background: NIST stands for the National Institute of Standards and Technology. They created a framework to be used as a guideline and a helping tool for LLM tools.

Purpose: The NIST AI RMF is a voluntary guidance framework designed to help organizations develop, deploy, and use trustworthy AI systems by identifying and managing risks—including privacy, bias, safety, and security.

Structure of the Framework:

Description: The AI RMF has two main components that discuss the frameworks core and trust characteristics that should be analyzed and regulated within LLMs.

1. **Framework Core**

Govern - Establish organizational policies, culture, and processes for managing AI risks.

Map - Understand the context, use cases, and potential impacts of the AI system.

Measure - Analyze and assess risks quantitatively or qualitatively.

Manage - Take actions to mitigate and monitor AI risks in practice.

2. **Trustworthiness Characteristics**

The framework promotes the development of AI systems that are:

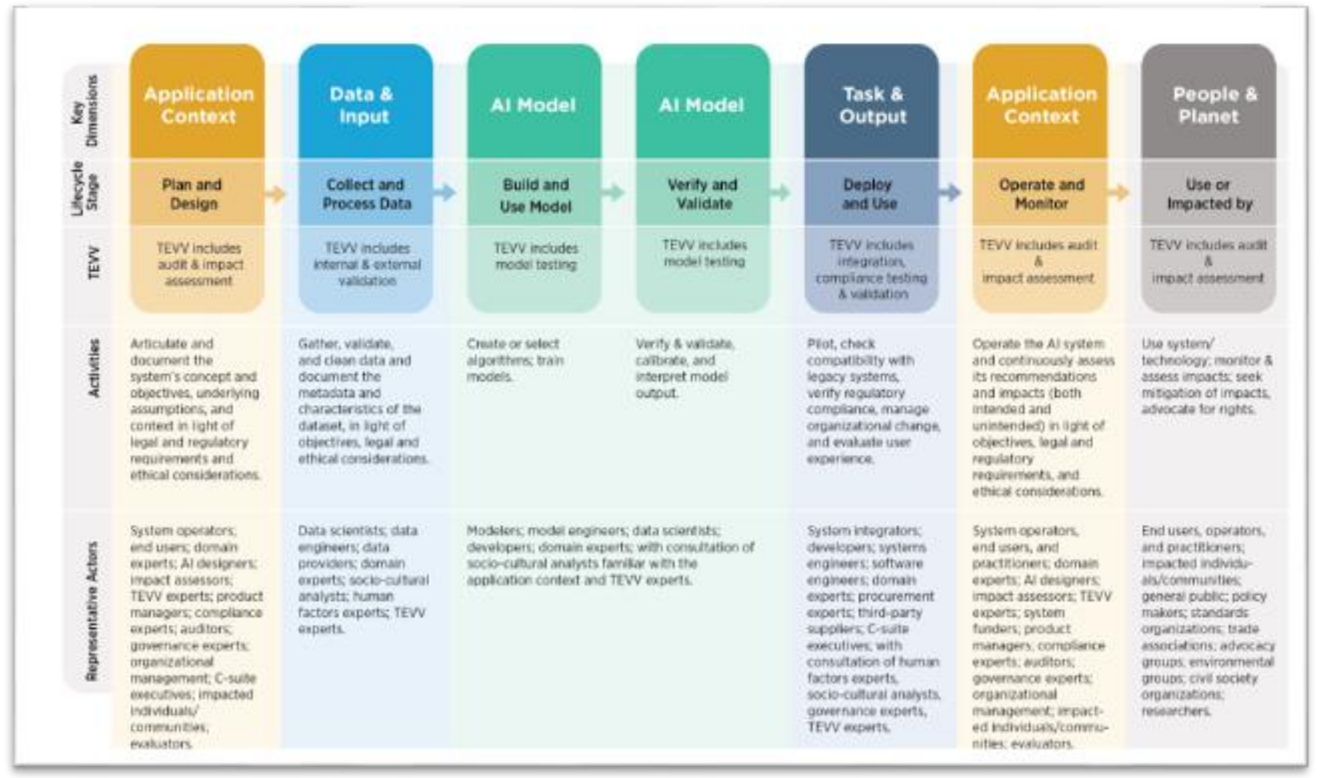
- Valid and Reliable
- Safe
- Secure and Resilient
- Accountable and Transparent
- Explainable and Interpretable
- Privacy-Enhanced
- Fair with Harm Mitigation

Key Findings

Key Findings: Analyzing NIST AI RMF we see that there are some important attributes that match our AI Privacy Framework

Objectives Met:

1. For your first objective we can see that framework ensures a privacy enhanced call to generative AI tools. It further pushes towards this objective with the secure resilient standard that is placed on AI generative tools.
2. The second objective is talked about in the framework as the framework puts a standard for accountability and transparency on companies that are developing and running LLM tools.
3. The third objective is met through standards that are listed under the harm and mitigation standard. These standards make sure that AI tools cannot be exploited and used in a harmful way.
4. The fourth objective is represented in the framework under the accountability and transparency section where user and LLM companies are held responsible for any wrong doings that the LLM does, or the user itself does.
5. The fifth objective is met through a multitude of standards ensuring that the LLM itself and users follow regulations to make LLM use and production ethical.
6. The last objective is met through the safe, secure, and resilient standard. It also includes the privacy enhancements that are measured on each LLM. These measures represents how safe using a certain LLM is and how to keep your data protected.



Key Dimensions	Application Context	Data & Input	AI Model	AI Model	Task & Output	Application Context	People & Planet
Lifecycle Stage	Plan and Design	Collect and Process Data	Build and Use Model	Verify and Validate	Deploy and Use	Operate and Monitor	Use or Impacted by
TEVV	TEVV includes audit & impact assessment	TEVV includes internal & external validation	TEVV includes model testing	TEVV includes model testing	TEVV includes integration, compliance testing & validation	TEVV includes audit & impact assessment	TEVV includes audit & impact assessment
Activities	Articulate and document the system's concept and objectives, underlying assumptions, and context in light of legal and regulatory requirements and ethical considerations.	Gather, validate, and clean data and document the metadata and characteristics of the dataset, in light of objectives, legal and ethical considerations.	Create or select algorithms; train models.	Verify & validate, calibrate, and interpret model output.	Pilot, check compatibility with legacy systems, verify regulatory compliance, manage organizational change, and evaluate user experience.	Operate the AI system and continuously assess its recommendations and impacts (both intended and unintended) in light of objectives, legal and regulatory requirements, and ethical considerations.	Use system/technology; monitor & assess impacts; seek mitigation of impacts, advocate for rights.
Representative Actors	System operators; end users; domain experts; AI designers; impact assessors; TEVV experts; product managers; compliance experts; auditors; governance experts; organizational management; C-suite executives; impacted individuals/communities; evaluators.	Data scientists; data engineers; data providers; domain experts; socio-cultural analysts; human factors experts; TEVV experts.	Modelers; model engineers; data scientists; developers; domain experts; with consultation of socio-cultural analysts familiar with the application context and TEVV experts.		System integrators; developers; systems engineers; software engineers; domain experts; procurement experts; third-party suppliers; C-suite executives; with consultation of human factors experts, socio-cultural analysts, governance experts, TEVV experts.	System operators, end users, and practitioners; domain experts; AI designers; impact assessors; TEVV experts; system funders; product managers; compliance experts; auditors; governance experts; organizational management; impacted individuals/communities; evaluators.	End users, operators, and practitioners; impacted individuals/communities; general public; policy makers; standards organizations; trade associations; advocacy groups; environmental groups; civil society organizations; researchers.

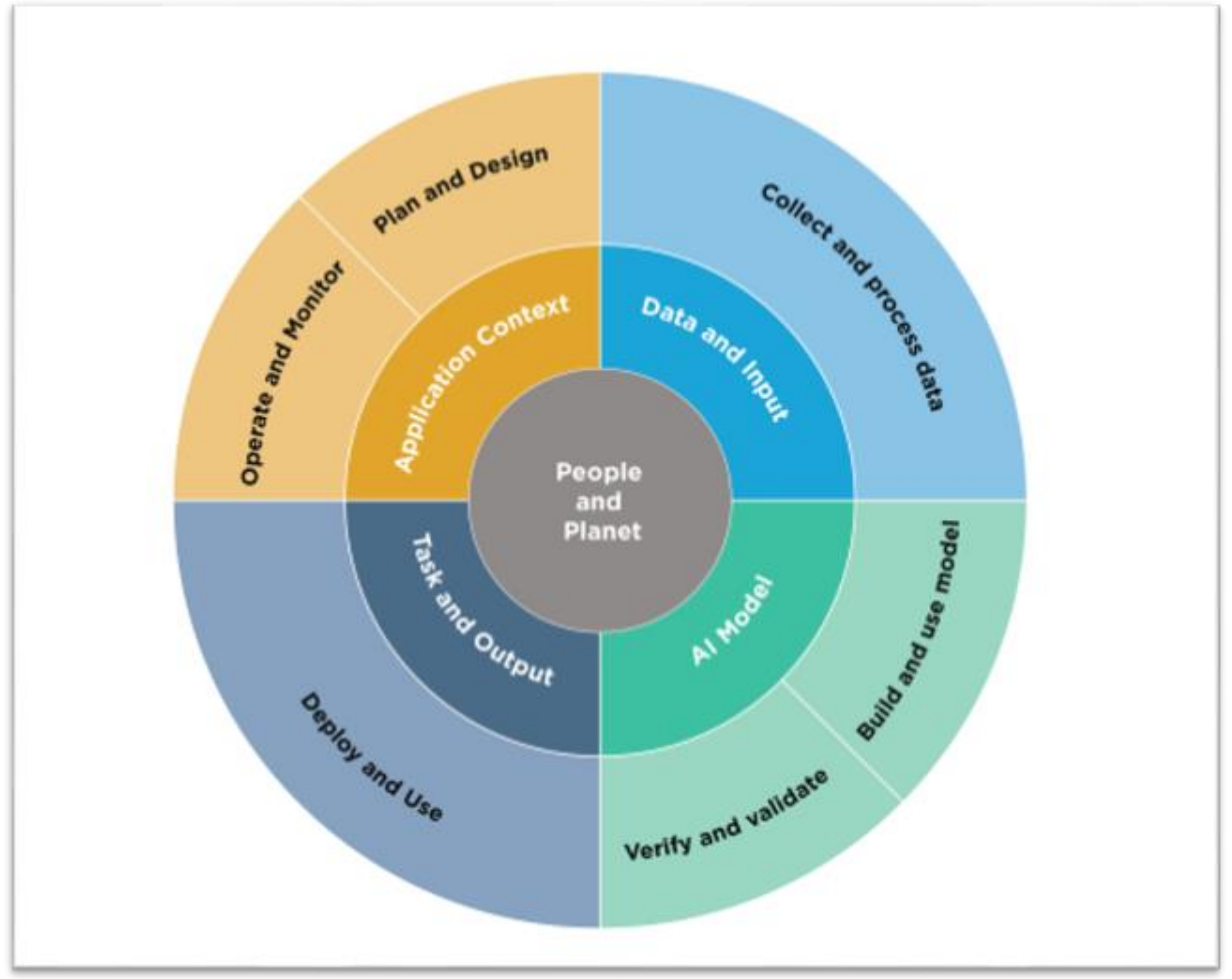
Conclusion

Conclusion:

NIST AI Privacy Framework is one of many AI privacy frameworks out there. The importance of these frameworks continue to increase day by day as LLM technology rapidly advances. The fundamentals and moral guidelines will maintain users' privacy while providing a list of responsibilities that the user themselves has to. As LLMs become more prominent there will have to be more descriptive frameworks to keep users safe. The NIST AI Privacy Framework is a great place to begin for many users to fully comprehend what future LLM training and usage should look like.

References

- https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf
- https://www.nist.gov/video/introduction-nist-ai-risk-management-framework-ai-rmf-10-explainer-video


Plan and Design
Collect and process data
Build and use model
Verify and validate
Deploy and Use
Operate and Monitor
Application Context
Data and Input
AI Model
Task and Output
People and Planet