Security and Privacy Concerns for AI

Eric Hibbard, CISSP, FIP, CISA
Samsung Semiconductor, Inc.

SNIA COMPUTE, MEMORY, AND STORAGE SUMMIT

Solutions, Architectures, and Community
VIRTUAL EVENT, MAY 21-22, 2024

Introduction

- Artificial intelligence (AI) systems are creating numerous opportunities and challenges for many facets of society.
- For security, AI is proving to be a power tool for both adversaries and defenders.
- Privacy is similar, but the societal concerns are elevated to a point where laws and regulations are already being enacted.

AI Ethical and Societal Concerns

- Ethical and societal concerns are a factor when developing and using AI systems and applications
- Taking context, scope and risks into consideration can mitigate undesirable ethical and societal outcomes and harms such as:
  - financial harm
  - psychological harm
  - harm to physical health or safety
  - intangible property (for example, IP theft, damage to a company's reputation)
  - social or political systems (for example, election interference, loss of trust in authorities)
  - civil liberties (for example, unjustified imprisonment or other punishment, censorship, privacy breaches)

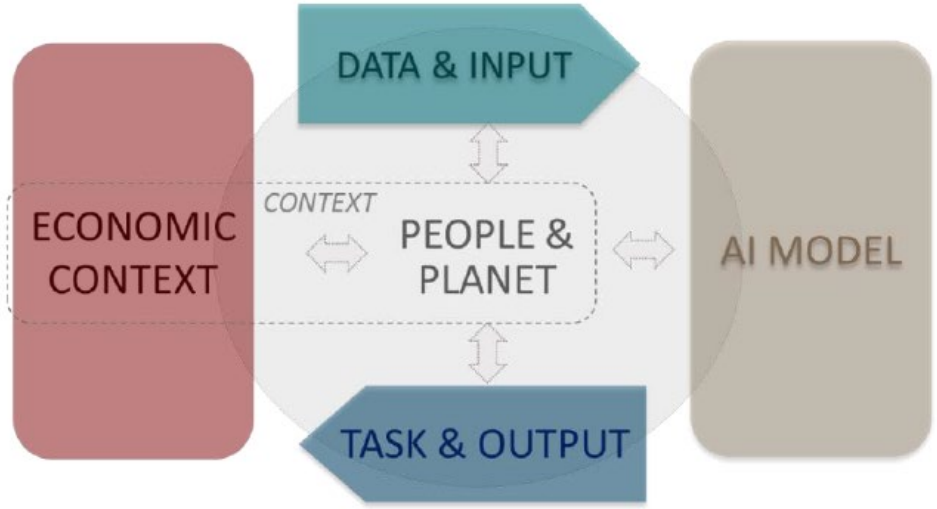
Source: ISO/IEC TR 24368:2022 Information technology – Artificial Intelligence – Overview of ethical and societal concerns

Examples of potential harms related to AI systems

Source: NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0)

OECD and Artificial Intelligence

OECD Framework for the Classification of AI Systems



The OECD has developed a framework for classifying AI lifecycle activities according to five key socio-technical dimensions, each with properties relevant for AI policy and governance, including risk management

The OECD AI Principles

Values-based principles for all AI actors	*Recommendations to policy makers for AI policies*
Principle 1.1. People and planet	*Principle 2.1.* Investment in R&D
Principle 1.2. Human rights, privacy, fairness	*Principle 2.2.* Data, compute, technologies
Principle 1.3. Transparency, explainability	*Principle 2.3.* Enabling policy and regulatory environment
Principle 1.4. Robustness, security, safety	*Principle 2.4.* Jobs, automation, skills
Principle 1.5. Accountability	*Principle 2.5.* International cooperation

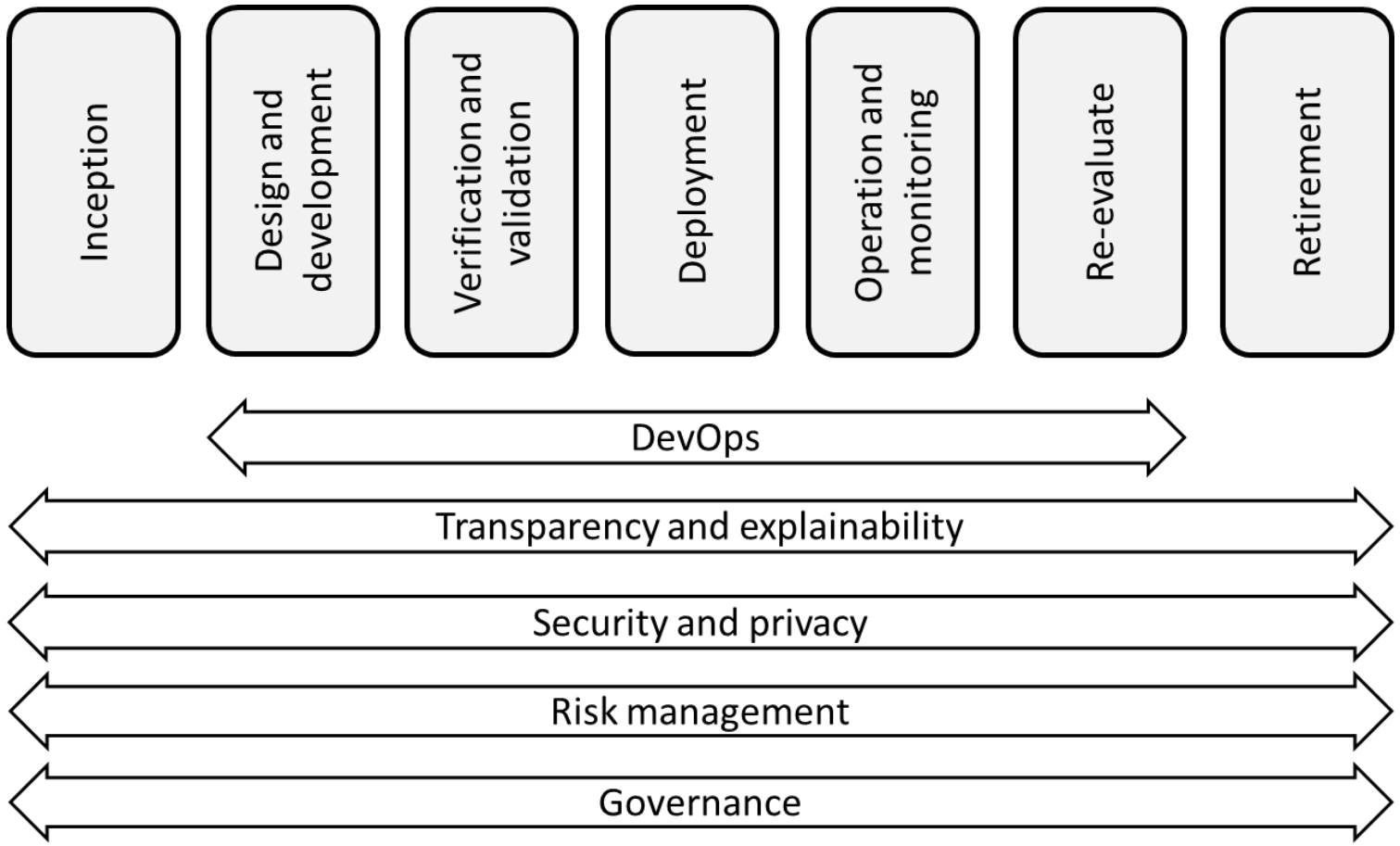
Source: OECD (2022) OECD Framework for the Classification of AI systems — OECD Digital Economy Papers

NIST AI RMF Lifecycle and Key Dimensions of an AI System

The NIST modification to OECD Framework highlights the importance of test, evaluation, verification, and validation processes throughout an AI lifecycle and generalizes the operational context of an AI system.

Source: NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0)

ISO/IEC AI Lifecycle and High-level Processes



Source: ISO/IEC 22989:2022 Information technology – Artificial Intelligence – Artificial intelligence concepts and terminology

AI Risks Differ from Traditional Software Risks

- The data used for building an AI system may not be a true or appropriate representation of the context or intended use of the AI system
- AI system dependency and reliance on data for training tasks
- Intentional or unintentional changes during training may fundamentally alter AI system performance
- Datasets used to train AI systems may become detached from their original and intended context or may become stale or outdated
- AI system scale and complexity (many systems contain billions or even trillions of decision points)

AI Risks Differ from Traditional Software Risks (cont.)

- Higher degree of difficulty in predicting failure modes for emergent properties of large-scale pre-trained models
- Privacy risk due to enhanced data aggregation capability for AI systems
- Increased opacity and concerns about reproducibility
- Difficulty in performing regular AI-based software testing, or determining what to test, since AI systems are not subject to the same controls as traditional code development
- Computational costs for developing AI systems and their impact on the environment and planet.
- Inability to predict or detect the side effects of AI-based systems beyond statistical measures

Summary of Attacks Specific to AI systems

Attack Type	Overview
Poisoning attack	Malicious data is injected into the training or inference data of an AI system causing it to behave or learn incorrectly
Evasion attack	Inputs are entered into an AI system that may appear correct to humans, but are wrongly classified by the AI systems
Membership inference	An attacker is able to attribute training data membership, such as PII, and then recover this information through crafted input/output pairings
Model exfiltration	Copying of a model either through direct access, or through repeated inference
Model inversion	Crafted input is used to produce an output that mimics an input used in the original training set leading to unauthorized information disclosure
Scaling attacks	The scalability limitations of an AI system is exploited by overwhelming the AI system with requests

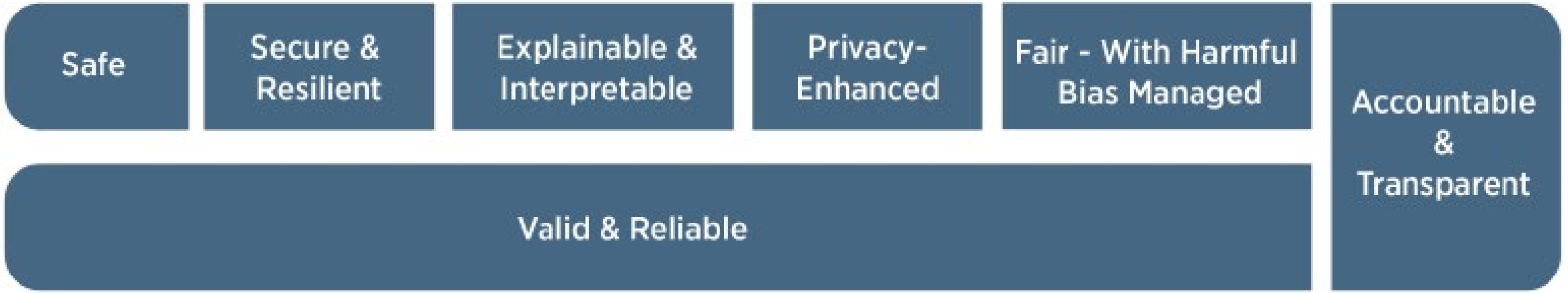
Source: ISO/IEC CD 27090 Cybersecurity – Artificial Intelligence – Guidance for addressing security threats and failures in artificial intelligence systems

Taxonomy of Privacy Threats

Privacy Threat	Privacy Threat Description
Linkability	Establishing the link between two or more actions, identities, and pieces of information
Identifiability	Establishing the link between an identity and an action or a piece of information
Non-repudiation	Inability to deny having performed an action that other parties can neither confirm nor contradict
Detectability	Detecting the PII principal's activities
Disclosure of information	Disclosing the data content or controlled release of data content
Unawareness	PII principals being unaware of what PII about them is being processed
Non-compliance	PII controller fails to inform the data subject about the system's privacy policy, or does not allow the PII principal to specify consents in compliance with legislation

Source: ISO/IEC WD 27091 Cybersecurity and Privacy – Artificial Intelligence – Privacy protection

AI Risks and Trustworthiness



- Valid & Reliable is a necessary condition of trustworthiness and is shown as the base for other trustworthiness characteristics.
- Accountable & Transparent is shown as a vertical box because it relates to all other characteristics.
- Accuracy and robustness contribute to the validity and trustworthiness of AI systems, and can be in tension with one another in AI systems.

Source: NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0)

Noteworthy AI Trends/Developments

- Governments have taken note of AI
  - EU AI Act
  - Executive Order (EO) on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence
  - Four US states — California (AB 302, 2023), Connecticut (SB 1103, 2023), Louisiana (SCR 49, 2023)and Vermont (HB 410, 2022)
- Privacy and safety are major areas of concern
- Intellectual property issues
  - In the US, human generated content can be protected
- Agentic AI systems emerging

Please take a moment to rate this session.

Your feedback is important to us.