Artificial Intelligence: Cohabitating with Security/Privacy

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

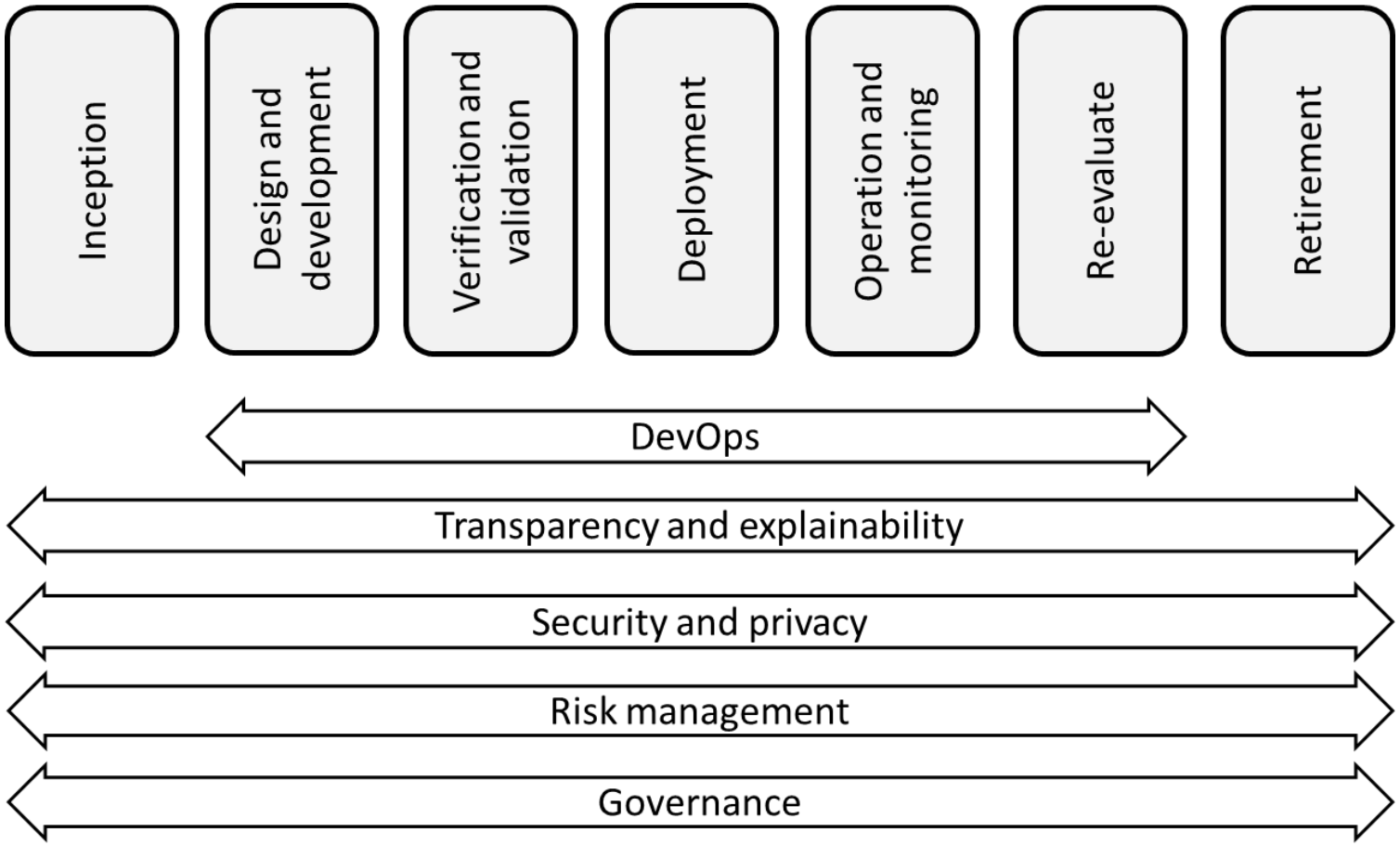
REGIONAL SDC 24
BY Developers FOR Developers
APRIL 24, AUSTIN, TX

A SNIA Event

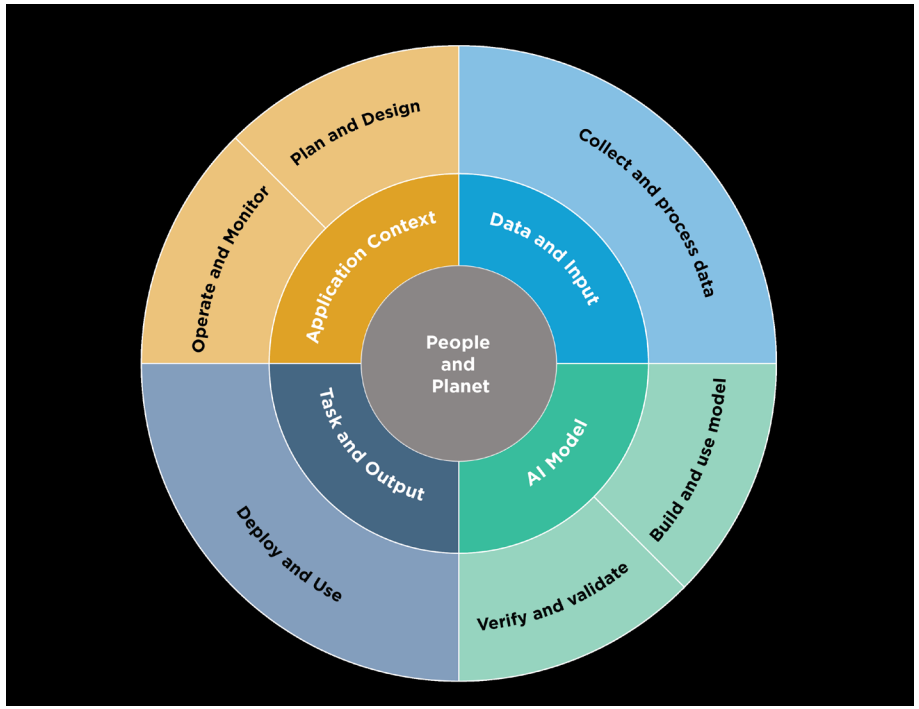
Introduction

- 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 Lifecycles



Source: ISO/IEC 22989



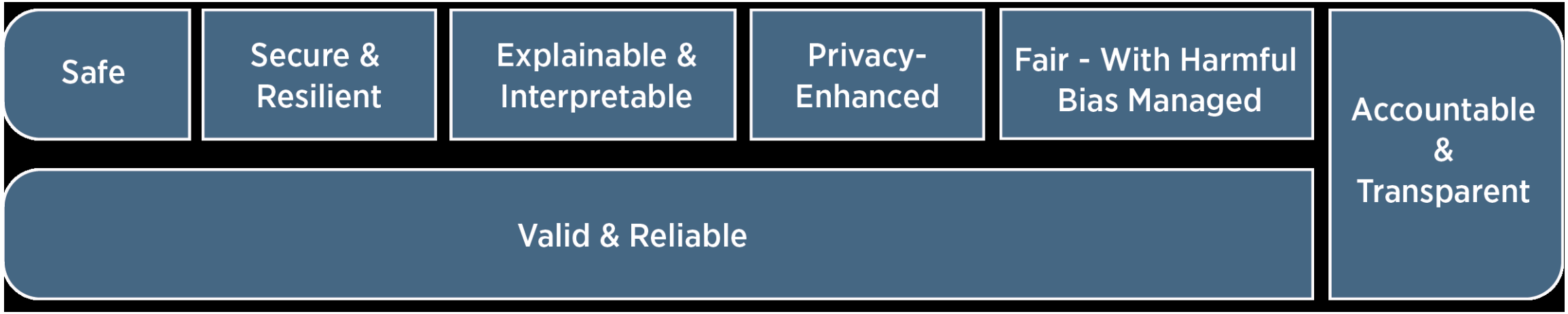
Source: NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0)

REGIONAL SDC 24

Examples of potential harms related to AI systems

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

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)

Examples of Undesirable AI Solutions

- Highly secure but unfair systems
- Accurate but opaque and uninterpretable systems
- Inaccurate but secure, privacy-enhanced, and transparent systems

- Note: For any given AI system, an AI designer or developer may have a different perception of the characteristics than the deployer.

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

Taxonomy of Privacy Threats

Privacy Threat	Mitigation Objective	Privacy Threat Description
Linkability	Unlinkability	Establishing the link between two or more actions, identities, and pieces of information
Identifiability	Anonymity	Establishing the link between an identity and an action or a piece of information
Non-repudiation	Plausible deniability	Inability to deny having performed an action that other parties can neither confirm nor contradict
Detectability	Undetectability and unobservablity	Detecting the PII principal's activities
Disclosure of information	Confidentiality	Disclosing the data content or controlled release of data content
Unawareness	Content awareness	PII principals being unaware of what PII about them is being processed
Non-compliance	Policy and consent 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

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

Useful resources

- NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0)
- Organization for Economic Co-operation and Development's (OECD's) 2019 Recommendation on Artificial Intelligence
- ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology
- ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management
- ISO/IEC TR 24028:2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence
- ISO/IEC TR 24368:2022 Information technology — Artificial intelligence — Overview of ethical and societal concerns
- ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system

REGIONAL SDC 24

THANK YOU

Please take a moment to rate this session.

REGIONAL SDC 24

BY Developers FOR Developers