Responsible AI in the Hardware Ecosystem

PTC Annual Conference

January 19, 2025

Jenifer Sunrise Winter
Professor, School of Communication and Information
University of Hawai‘i at Mānoa

Introduction

- Much of the discourse about Responsible Artificial Intelligence (RAI) focuses on data and algorithms, with less attention focused on foundational AI hardware (Winter et al., 2024).
- Computing power is an essential point of intervention for responsible governance of AI (Sastry et al., 2024).
- Compute increasingly recognized as a node for AI governance (Heim, 2023).

Overview

- Introduction to Responsible AI
- Role of Hardware in AI Systems
- Key Principles of Responsible AI
- Challenges in the Hardware Ecosystem
- How AI Hardware Can Enable New Governance Mechanisms
- Conclusions and Future Opportunities

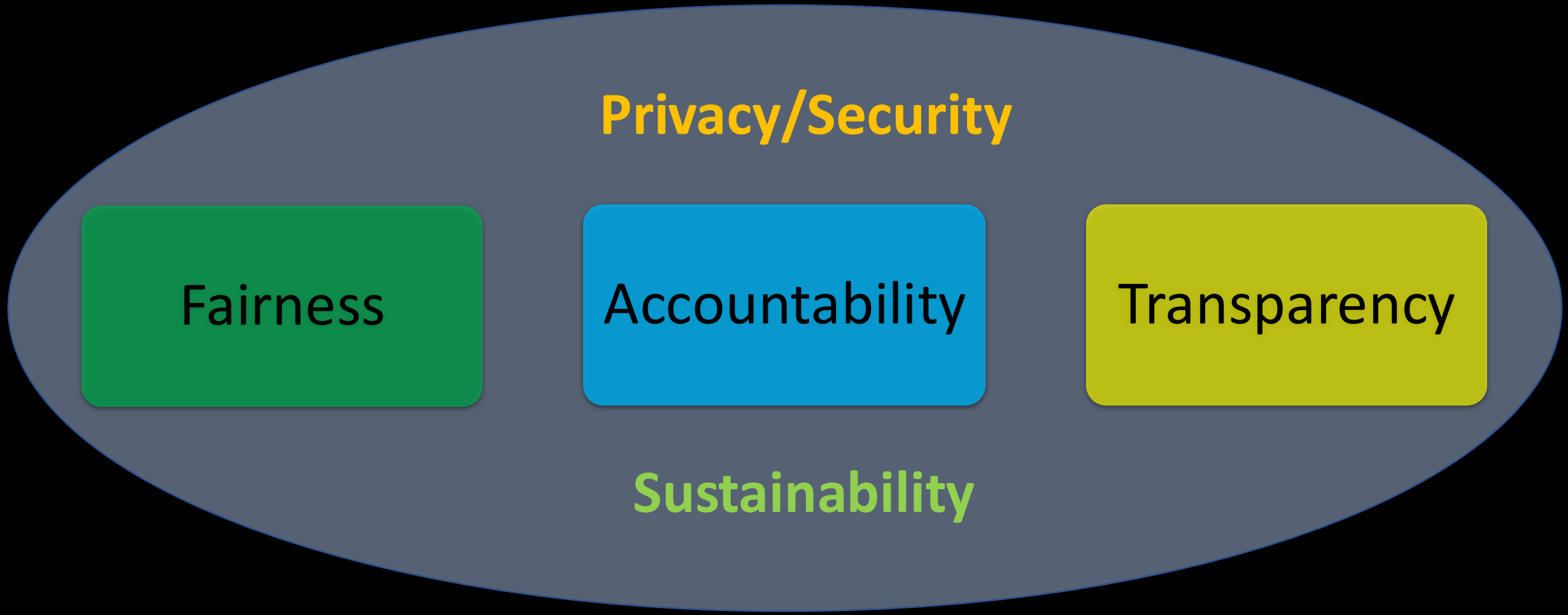
Responsible AI (RAI)

- *“Responsible AI is an approach to developing and deploying artificial intelligence from both an ethical and legal standpoint. The goal is to employ AI in a safe, trustworthy and ethical way. Using AI responsibly should increase transparency while helping to reduce issues such as AI bias.”* –ISO, 2024

Role of Hardware in AI Systems

- Hardware provides the compute needed to handle large amounts of processing.
- GPUs, TPUs, and other AI-related hardware components enable AI systems, and new technical innovations are emerging (e.g., quantum computing and neuromorphic chips).
- The compute used to train large AI models has doubled approximately every six months since 2010, and the “largest AI models now use 350 million times more compute than thirteen years ago” (Lewsey, 2024).
- Optimized for the parallel processing and intensive computational demands of machine learning tasks, these innovations can drastically reduce the time required for training and running AI models, facilitating the deployment of sophisticated AI applications on a broad scale and in real-time environments.

Introduction to Responsible AI (RAI)



Fairness

- Ensure that individuals or groups are not unjustly discriminated against – are there disparate impacts? (see Barocas & Selbst, 2016).
  - *"For supervised systems, consider the relationship between the data labels you have, and the items you are trying to predict. If you are using a data label X as a proxy to predict a label Y, in which cases is the gap between X and Y problematic?"* (Google, 2024)
- Employ human-centered design (iterative feedback from wide array of users and use cases).
  - Consider inclusion of multiple stakeholders (in the broadest sense).

Fairness

- Justify algorithmic goals – "*Set goals for your system to work fairly across anticipated use cases: for example, in X different languages, or to Y different age groups. Monitor these goals over time and expand as appropriate*" (Google, 2024).
- Avoid over- or underrepresentation of some sources or material.
- Assess your data (errors, omissions, etc.) – is sampling appropriate?
- Make classification *more flexible* to account for different contexts.
- Seek to understand the norms underlying the data and how these might vary.
- Provide support for impacted populations who may not have a voice (Washington, 2023).

Accountability

- Understand that standards for accountability vary – in some cases there may be legal compliance requirements.
- Establish provenance of data and algorithms.
- Acknowledge that, even where there are no legal requirements, we are responsible for outcomes.
- Document and evaluate how datasets and models are created/obtained, trained, and evaluated.
- Choose multiple metrics for evaluation and ensure that these align with context and goals.
- Recognize that oversight mechanisms are not always technical – e.g., data use committees, oversight board, governance boards.

Transparency

- Make it explainable– both developers and those impacted should be able to understand how decisions were made (e.g., XAI)
  - “How the machine ‘thinks’: Understanding opacity in machine learning algorithms” (Burrell, 2016).
- Document or use other tools to make AI/ML auditable.
- Understand that the goal of accountability is to build trust with stakeholders/public.

Privacy/Security

- AI systems should be designed to prevent data leaks and disclosures (Microsoft, 2023).
- Privacy by Design and Security by Design principles.
- Data minimization enables access to only the data required to perform specific tasks, thereby reducing exposure to risk.
- Use end-to-end encryption.
- Ensure compliance with laws and regulations such as GDPR and HIPAA.
- Use access controls and network monitoring.
- Implement encryption at the hardware level (e.g., data in transit, data at rest) and secure enclaves.
- Include hardware-level support for secure boot and firmware integrity checks.
- Use unique identifiers for chips to prevent cloning or spoofing.
- Add tamper-resistant features.

Sustainability

- Data centers may account for as much as 21% of global energy use by 2030 (Stackpole, 2025).
- Hardware and software efficiencies are needed.
- Reduce compute by rethinking model training.
- Example: Biden’s January 14, 2025 executive order emphasizes clean energy plants, facilitation of grid connections and electricity transmission, and possible user of nuclear and geothermal power sources.

Bridging RAI and Hardware Development

- RAI frameworks are about mitigating the risks of AI but largely focus on algorithms and data.
- At the forefront of AI's rapid evolution, there is a notable absence of a cohesive and comprehensive agenda for responsible AI design and governance that keeps pace with hardware innovations. This disconnect between the ethical imperatives for RAI and the relentless drive of AI advancements presents a pressing challenge.
- Three regulatory challenges have emerged from AI (Heim, 2024):
  - Safety of deployment
  - Unexpected capabilities
  - Proliferation

RAI Challenges in the Hardware Ecosystem

- High power requirements of AI systems – how can we minimize energy use through both chip design and algorithms?
- How can we source rare earth elements responsibly?
- How can we ensure hardware design and limitations do not affect AI model fairness?
- Accessibility – who should have access to AI chips, and what types of risk mitigation might be designed into the hardware?

Regulatory Approaches

- Increased focus on ensuring AI hardware is developed/deployed in trusted and secure environments
  - EU – AI Act
  - US – Executive Order on AI, and in 2025 has just limited export of GPUs by establishing country caps on all but 18 countries (Freifeld, 2025).
  - China – Generative AI Regulation
  - UK AI Safety Institute

Opportunities for RAI in Hardware Systems

- Growing focus on compute for AI governance – “Computing hardware is visible, quantifiable, and its physical nature means restrictions can be imposed in a way that might soon be nearly impossible with more virtual elements of AI” (Lewsey, 2024).
- Hardware offers an important opportunity for AI governance because it is physically trackable: “By observing, regulating, and influencing an entity's access to compute, [we] can roughly predict and modulate an actor's access to AI ecosystems” (Heim, 2023).

Because compute has these properties...

Detectability

Large-scale AI training and deployment is highly resource intensive, often requiring thousands of specialized chips in a high-performance cluster hosted in a large data center consuming large amounts of power.

Excludability

The physical nature of hardware makes it possible to exclude users from accessing AI chips. In contrast, restricting access to data, algorithms, or trained models is much more difficult.

Quantifiability

Computational power can be easily measured, reported, and verified.

Supply chain concentration

AI chips are produced via a highly inelastic and complex supply chain, several key steps of which (e.g. design, EUV lithography, and fabrication) are dominated by a small number of actors.

It can enable these critical governance capacities...

Visibility

The ability to track and assess the development and use of advanced AI.

Allocation

The ability to influence which AI systems are built, when, and by whom.

Enforcement

The ability to ensure compliance with AI regulations and standards.

Figure from: Sastry et al., 2024

Governance Mechanisms (Sastry et al. 2024)

Note all are possible but not necessarily desirable

Visibility examples:

- Use public information about compute quantities to estimate actors' AI capabilities
- Require reporting of training compute usage from cloud providers and AI developers
- International AI chip registry

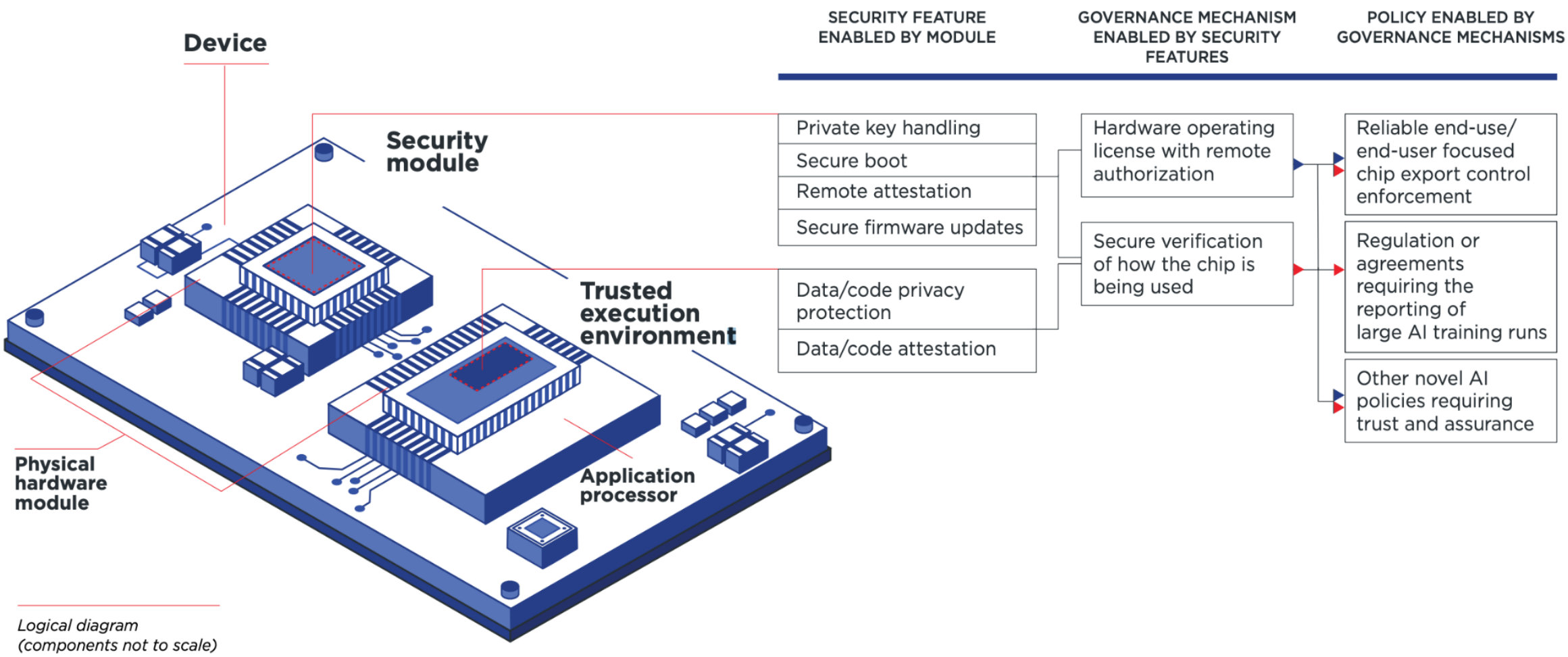
Allocation examples:

- "Redistributing AI development and deployment across and within countries"
- Collaborations on joint AI megaprojects

Enforcement examples:

- Adding "compute caps" via physical limits on chip-to-chip networking
- Hardware-based remote enforcement
- Multiparty controls

On-chip Governance Model (Aarne et al., 2024)



“Sovereign AI”

- “Every country needs to own the production of their own intelligence” – democratizing AI (NVIDIA’s Jensen Huang in Caulfield, 2024)
- “Sovereign AI refers to a nation’s capabilities to produce artificial intelligence using its own infrastructure, data, workforce and business networks” (Lee, 2024) – so, both the physical and data infrastructures.

Conclusions and Future Opportunities

- There is a need to map Responsible AI principles to hardware design – We need to bridge the gap between abstract responsible AI principles (such as fairness, transparency, accountability) and practical hardware design decisions.
- Further exploration of the potential for bias to be embedded within hardware itself, and strategies for its identification and eradication, are needed.
- We can focus on enhancing transparency and explainability through hardware.
- Security and Privacy by Design should be emphasized in AI hardware design.
- Promoting environmental sustainability in AI hardware development – In response to the growing environmental footprint of AI systems, there is need for energy-efficient, sustainable hardware designs.

Mahalo

Jenifer Sunrise Winter
jwinter@hawaii.edu