Security Risks of AI Hardware for Personal and Edge Computing Devices

Google

July 11, 2024

1 Table of Contents

2 Executive Summary

Synopsis

In April and May 2024, Google engaged NCC Group to conduct an analysis of potential security benefits and risks that hardware-for-Artificial Intelligence (AI) on personal and edge computing devices could have for AI companies, developers, users, Original Equipment Manufacturers (OEM), and other impacted stakeholders. This analysis includes an assessment of the global landscape from both a product development and regulatory perspective, such that Google and other companies can make informed decisions on a strategic approach to the cybersecurity of these technologies.

NCC Group adopted a threat modeling based approach to explore the differences in threats and overall risk profile to identified stakeholders, their data, and personal devices when AI is used in different modes of deployment. The different modes of AI integration deployment include vertically integrated and dis-aggregated models, as well as on-device and cloud-based compute.

The research was performed by a small team of specialists that leveraged expertise from distinct practices within NCC Group:

- Artificial Intelligence
- Hardware and Embedded Security
- Government Affairs
- Consulting and Implementation

Relevant examples of NCC Group's industry-leading cybersecurity research into AI, Machine Learning (ML), and Large Language Model (LLM) systems and hardware/supply chain security include:

- **Analyzing AI Application Threat Models** – An exploration of the threat actors, vectors, controls, and considerations for effective penetration testing with an emphasis on LLMs.[1]
- **Safety, Security, Privacy & Prompts: Cyber Resilience in the Age of AI** - A summary of how AI is changing the cyber security landscape, and an overview from our Government Affairs team of the regulatory landscape across different markets.[2]
- **Practical Attacks on Machine Learning Systems** – A taxonomy of the different types of cyber attacks against ML systems, with references to real world example attacks.[3]
- **Secure Device Manufacturing: Supply Chain Security Resilience** - A seminal and comprehensive whitepaper that explores the increased cybersecurity risk of manufacturing supply chains, and outlines best practices for device manufacturing and supply chain security.[4]

Scope

The research paper explores AI-related product design decisions, and compares the benefits and risks of open approaches (e.g., remappable buttons and accessible on-device compute) as well as closed approaches (e.g., hotwired buttons or locked and/or cloud

[1] https://research.nccgroup.com/2024/02/07/analyzing-ai-application-threat-models/
[2] https://www.nccgroup.com/media/reybxvtp/ncc-group-cyber-resilience-ai-whitepaper_updated.pdf
[3] https://research.nccgroup.com/wp-content/uploads/2022/07/practical-attacks-on-ml.pdf
[4] https://research.nccgroup.com/2015/11/09/secure-device-manufacturing-supply-chain-security-resilience/

compute power). NCC Group examined the following key questions and implications of dedicated hardware facilitating AI-powered assistance:

1. What are the security benefits and risks of such product integration to users, developers, OEMs and other stakeholders? Under what circumstances could such integrations harm the security of different stakeholders?
2. What concerns are there around security mono-culture? What lessons can be learned from historical case studies and precedents in the tech industry in or outside of an AI context?
3. Could the use of AI services affect users' exposure to malware attacks?
4. What are the security implications of using AI on edge devices in critical infrastructure environments?

In addition to modeling the security threats inherent to the different forms of AI integration, research covers common security, competition and other concerns of governments, legislators, and regulators in respect of vertically integrated models.

Key Findings

Some key concerns for threat modeling the use of specialized hardware in AI applications became apparent during the literature review:

- Based on recent vulnerability research explored in the AI Models Running Locally section, running AI models on dedicated hardware presents a complex attack surface in both edge computing and cloud computing environments. Vulnerabilities in components such as Graphics Processing Units (GPU) and Neural Processing Units (NPU) range from side-channel attacks that can leak data, to malware exploits of proprietary user space frameworks that interact with the GPUs where users can create custom GPU programs.
- Closed approaches to integrating AI capabilities in edge devices present different risks to taking a modular open approach. This includes risks around over-reliance on the vendors security practices, inflexibility when responding to security incidents and challenges around transparency and third party assurance. System owners should ensure they understand the security responsibilities they are implicitly delegating to vendors.
- Moving AI capabilities to the edge can result in performance and privacy improvements by performing at least some of the processing locally on user devices. However these privacy improvements become more nuanced when distributed deep learning is deployed as it becomes difficult to manage the privacy and integrity of data used in the distributed deep learning process.
- The context in which AI systems on edge devices are used affects their risk profile. Compliance requirements with regulations, such as the European Union (EU) AI act, will be driven by potential impacts which differ greatly in the context of an office or personal AI assistant versus capabilities deployed in safety or privacy critical industries.

Strategic Recommendations

Hardware Acceleration As AI services proliferate, edge devices delivering new AI-enabled user experiences will more than likely introduce hardware acceleration to support the large computational overhead needed to deliver those capabilities. As such, the hardware components (i.e. GPU, TPU) as well their respective drivers and user space Application Programming Interfaces (APIs) should be appropriately threat modeled from multiple perspectives that are covered in AI Models Running Locally. The proposed threat model is derived from a survey of recently published vulnerability research that outlines different ways that an attacker may exploit GPUs or TPUs and subsequently gain unauthorized access to sensitive data or achieve code execution.

Distributed Deep Learning offers significant advantages in terms of scalability and computational efficiency, making it a powerful tool for experimental purposes. However, we recommend deploying distributed deep learning only in environments where data privacy and integrity are not critical concerns, or when all participating nodes are owned by the same entity. Additionally, it is crucial that all supporting infrastructure shares uniform security policies and is managed consistently to ensure they remain within the same security boundary. This approach minimizes the risk of data breaches and integrity issues, ensuring that the distributed system operates securely and effectively.

3 Overview

We are currently experiencing a pivotal moment in AI, a trend that is expanding rapidly across various sectors and likely to have a monumental impact on society, businesses, and governments. This surge is primarily driven by the significant enhancements in performance nearly any profession can achieve by incorporating AI technologies. Consequently, entities that fail to adopt these capabilities may quickly find themselves at a competitive disadvantage. In response to this growing demand, various developers and companies are actively embedding AI into commonly used platforms such as desktop and mobile Operating Systems (OS). Some are even developing specialized hardware to improve the efficiency of this transformative technology, ensuring that AI tools are more accessible and effective for a broader audience.

In this report, we conducted a literature review focusing on the potential risks associated with using hardware technologies to deliver AI-powered services. Our review primarily examined various architectures, including cloud-based, edge-computing, and distributed deep learning (hybrid), with an emphasis on the latter two, given the lack of published research on the topics. Since AI security is a rapidly evolving field, some risks have not yet been thoroughly explored in the existing literature. Therefore, we sought to draw parallels with other analogous fields to identify and understand these emerging risks.

Finally, we analyze the security implications of utilizing fixed, on-hand AI capabilities as compared to more adaptable approaches that allow users and organizations to select their own AI providers. This comparison helps us assess how different deployment strategies impact security in various use cases.

4 AI Automations and Hot Buttons

AI Automations

Automation has seamlessly integrated into the workflows of end users, enhancing the efficiency of repetitive tasks. Automation can range from simple, software-only solutions to those requiring hardware enhancements. At their most basic level, software automations include shortcuts and macros that users can program within their existing software environments. These shortcuts allow users to execute complex sequences of actions with a single command, streamlining daily tasks without the need for additional hardware.

As the complexity of the task is increased, there are hardware-based automations that involve programmable buttons or switches integrated into devices. For example, some laptops come with special touch bars that allow users to execute customizable actions quickly. Additionally, external keypads are available, which are favored by professionals who require rapid access to a suite of commands, such as video streamers and content creators. These keypads can be programmed with a variety of commands and are adaptable to numerous applications, providing a tactile and intuitive interface for managing workflows.

AI is revolutionizing these traditional automation systems, adding layers of intelligence and adaptability. For example, the special touch bars and external keypads mentioned above may introduce dynamic understanding based on the user’s digital context, changing the button’s capabilities automatically to meet the needs of the user. When triggered, the programmable switches can activate AI assistants that understand natural language commands and execute complex tasks with fewer user interactions. This AI integration not only increases the potential for personalization but also enhances the capability of devices to perform a broader range of tasks autonomously, making technology more intuitive and aligned with individual user needs.

Indeed, the concept of AI-driven automation or assistance is not new. We have previously experienced the convenience of voice-activated assistants in various devices around our homes and offices, such as televisions, smartphones, and smart home devices. These systems respond to voice commands to perform tasks like changing channels, sending messages, or adjusting thermostats, demonstrating how AI can make everyday interactions more convenient and responsive.

However, the innovation of integrating AI capabilities directly into a hardware button on a general-purpose device like a personal computer marks a notable advancement. For the first time, this setup merges the tactile feedback and accessibility of a physical button with the intelligent, context-aware processing power of AI deployed on the device. This blend offers a novel interaction paradigm where users can instantaneously engage with AI without navigating through multiple menus or speaking out loud, thereby making the technology more accessible and discreet in professional and public environments. This development reflects a broader trend towards creating more deeply integrated and user-friendly AI solutions that cater to a wide array of tasks and workflows in everyday computing.

Proprietary vs Open Integrations

In the scenarios described above, two distinct approaches to integrating AI into devices emerge: proprietary and open integrations. In the proprietary model, the AI-powered functionalities are hard coded into the device, with the AI services and their capabilities tightly controlled by the device’s manufacturer. This means users cannot modify or choose different AI services, leading to a situation akin to vendor lock-in. This AI vendor lock-in restricts flexibility and can leave users dependent on a single provider for updates and security, which may expose users to various risks such as those described in the following subsections, depending on how mature the vendor is in terms of security.

Conversely, open integration allows users the freedom to select their AI provider and services based on their specific needs and preferences. This model promotes a more customizable and adaptable environment, enhancing competition and innovation among AI providers. In the next subsection, we will focus on traditional vendor lock-in scenarios such as those seen with cloud providers in enterprise or government contexts.

AI Vendor Lock-In Risks

As we consider the implications of vendor lock-in for AI services, particularly in the context of hardware-integrated AI functionalities, it becomes imperative to scrutinize the associated risks. These risks can have significant impacts on user autonomy, data security, and overall flexibility in utilizing AI technologies. Below is a list of potential challenges and concerns that arise when users are tethered to a single AI provider due to proprietary integration, emphasizing the importance of understanding and addressing these issues:

- **Single Point of Failure**: Over-reliance on a single component or system has historically led to significant outages and disruptions when those systems fail. For example, issues with central network switches or routers can cripple connectivity, causing extensive communication breakdowns and service interruptions across an organization. Such outages can significantly impact operational efficiency, resulting in substantial financial losses and damage to an organization's reputation due to lost productivity and customer dissatisfaction[5]. Also, a potential compromise of a massively used platform can result in sensitive user data being exposed[6] [7].
- **Risks in Homogeneous Environments**: Homogeneous IT environments create significant security risks by presenting attackers with widespread targets, a scenario highlighted by the XZ-Utils backdoor detected in 2024[8], where specific versions contained a sophisticated backdoor. The diversity in Linux distributions' software versions mitigated what could have been a broader security disaster, as not all distributions adopted the vulnerable versions. Vendor lock-in compounds these issues by restricting integration with other security solutions, which can force organizations to rely solely on a single vendor's security products. This lack of compatibility may necessitate duplicated security efforts or lead to gaps in cross-platform security policies, creating inconsistencies and potential vulnerabilities across an organization's security framework.
- **Dependence on Vendor's Security Practices**: If the vendor does not prioritize security, or if their security standards are not up to par, the customer's data and systems are at increased risk. The customer relies on the vendor's ability to respond to vulnerabilities, implement patches, and secure their products. The threat of state-sponsored attacks on critical infrastructure or supply chains, which have been extensively documented in recent years, underscores the potential risks involved in relying heavily on a single provider's security practices [9] [10].
- **Complicated Compliance**: Compliance with industry regulations can become more challenging when locked into a vendor. If the vendor's products do not fully comply with new or existing regulations, the customer may struggle to meet these requirements on their own. This has motivated the adoption of multi-cloud approaches in the cloud

[5] AT&T says service has been restored after massive, nationwide outage. Authorities are investigating: https://edition.cnn.com/2024/02/22/tech/att-cell-service-outage/index.html

[6] Cyber Safety Review Board - Review of the Summer 2023 Microsoft Exchange Online Intrusion: https://www.cisa.gov/sites/default/files/2024-04/CSRB_Review_of_the_Summer_2023_MEO_Intrusion_Final_508c.pdf

[7] The Hacker News - TeamViewer Detects Security Breach in Corporate IT Environment: https://thehackernews.com/2024/06/teamviewer-detects-security-breach-in.html

[8] Critical Vulnerability in XZ Utils: https://cert.europa.eu/publications/security-advisories/2024-032/

[9] NIST Updates Cybersecurity Guidance for Supply Chain Risk Management: https://www.nist.gov/news-events/news/2022/05/nist-updates-cybersecurity-guidance-supply-chain-risk-management

[10] Supply Chain Risk: https://www.nccgroup.com/media/co3o3psl/105503_ncc_insight_space_market_research_report.pdf

provider field, as organizations seek to mitigate these risks by diversifying their reliance on single providers. Utilizing multiple cloud services can enhance compliance flexibility and reduce the potential impact of any single vendor's compliance deficiencies[11].

- **Incompatibility with Other Security Solutions**: Vendor lock-in often restricts compatibility, as products may not be designed to integrate smoothly with other systems or security solutions. This limitation can lead to significant gaps in an organization's security architecture and complicate the management of cross-platform security policies. Additionally, being tied to a single vendor's security products can force a company to use only those specific tools, which might not align with the tools used across the rest of the organization. This discrepancy can necessitate the replication of security policies across different products that do not naturally integrate or replace each other, potentially creating inconsistencies and vulnerabilities in the company's overall security policy framework.
- **Reduced Flexibility in Responding to Security Threats**: Being restricted to a specific vendor's technologies can hinder an organization's agility in responding to new threats, as it limits their ability to integrate or adopt alternative security solutions. This dependency may slow down the implementation of necessary changes or updates critical for keeping pace with evolving cybersecurity challenges, thereby impacting the overall responsiveness and adaptability of the organization's security posture. For example, if a zero-day vulnerability is discovered in the corporate browser, an organization locked into a specific vendor's technology might struggle to respond swiftly. Ideally, they could switch temporarily to a different browser not affected by the vulnerability as a short-term containment action. However, vendor lock-in could complicate this switch, delaying the response and potentially increasing exposure to the threat. Such scenarios underscore the importance of flexibility in security strategies to adapt quickly to emerging risks.

These concerns typically associated with vendor lock-in also extend to AI services. While we have not yet observed these risks materializing specifically in the AI sector given the nascency of the industry, the potential for such issues is already evident. As reliance on AI technologies grows, organizations might face similar challenges as those seen with traditional software and cloud vendors, particularly in adapting quickly to new security threats or integrating diverse AI solutions. This emerging scenario underscores the importance of anticipating and preparing for potential risks as AI becomes increasingly embedded in operational and security processes.

History provides ample examples of emerging technologies encountering unforeseen security vulnerabilities during their initial design and adoption phases. For instance, WiFi technology, initially celebrated for its convenience and connectivity, soon revealed significant security flaws like Wired Equivalent Privacy (WEP) vulnerabilities[12], which were not anticipated by its creators. Similarly, the adoption of Internet of Things (IoT) devices brought about significant security challenges. These devices, while offering unprecedented connectivity and automation, have been plagued with issues such as weak default passwords[13] and lack of regular updates, leading to vulnerabilities that can be exploited by malicious actors. These precedents suggest that AI technologies are likely to face similar challenges. In fact, prompt injection[14] a vulnerability inherent to the design of LLMs,

[11] The Register - Motives of multi-cloud users? Compliance and dodging vendor lock-in top the list: https://www.theregister.com/2023/11/14/compliance_and_dodging_vendor_lockin/
[12] Arstechnica - A brief history of Wi-Fi security protocols: https://arstechnica.com/gadgets/2019/03/802-eleventy-who-goes-there-wpa3-wi-fi-security-and-what-came-before-it/
[13] SANS ISC - The Mirai Botnet: https://isc.sans.edu/diary/22786
[14] NCC Group - Exploring Prompt Injection: https://research.nccgroup.com/2022/12/05/exploring-prompt-injection-attacks/

illustrates how security issues can stem from fundamental design decisions. The security community continues to struggle with finding a perfect solution to this problem, highlighting the necessity for ongoing vigilance and proactive measures in the development and deployment of AI systems as the industry matures.

5 AI Models Running Locally

A locally running AI model is an AI runtime paradigm in which the computation is performed on the same device that receives the query. This configuration is in close alignment with edge computing models and the associated infrastructure can provide the framework required to enable robust locally running AI models.

It is worth noting that while all of the computation performed and subsequent results that are generated are tied to the device in which the model is running, it is not necessarily the case that any data is *restricted* to said device to be considered "locally running". For example, the model itself and its associated information may originate from sources external to the target device, but still fulfill all requests locally by running the model's computation on device. As such, locally running AI models are flexible and many of the benefits achievable with this paradigm require careful attention to the implementation details.

Benefits

Privacy

One of the primary advantages of using edge computing to run AI models locally is the potential for enhanced privacy. By processing data on devices within a local area network rather than sending it to remote cloud servers, edge computing ensures that sensitive information remains within the confines of its original environment. This is crucial for both individuals and industries like healthcare and finance, where privacy and compliance with regulatory requirements are paramount. For example, a locally administered AI model in a hospital can analyze patient data directly on-site, ensuring that confidential health records are not exposed during transmission to the cloud. This localized processing reduces the risk of data breaches and unauthorized access, providing a more secure framework for handling sensitive information.

Availability

Edge computing can improve availability guarantees, even in the absence of network connectivity or during periods of high demand in cloud services. By running AI models locally, the dependency on continuous internet access is eliminated, helping to ensure critical applications remain operational at all times. This is especially beneficial in scenarios where network reliability is a concern, such as in remote locations or situations where users are mobile.

Latency

By avoiding the latency and bandwidth issues associated with cloud computing during peak times, edge computing facilitates real-time data processing and decision-making. This ability to operate independently of the cloud not only enhances the responsiveness of AI-driven applications, but also ensures they are resilient to fluctuations in cloud computing resources.

Drawbacks

Scaling Performance

Edge computing favors computation at the "edge of the network"[15], preferably on devices closest to the end user. By its very nature, this means locally run models are restricted to a smaller subset of the overall network than a distributed approach would be. For rapidly improving models and increasing load, locally run models may quickly become a limitation in an organization's operations. In such circumstances, leveraging cloud resources or a hybrid approach may be an effective method to scale up.

Costs of Local Security

Edge computing also comes with significant added security costs. One major concern is the potential for side-channel attacks, such as Tempest[16] or power analysis, which can reveal

[15] https://ieeexplore.ieee.org/abstract/document/9083958

sensitive information being processed by the device. Access control and isolation are also critical challenges that operating systems must also address to keep these new capability risks under control. Implementing robust security mechanisms adds complexity and cost to the system, making it an additional hidden cost on using local AI models in edge computing.

AI Hardware Acceleration

This section will be an overview on the potential harms of integrating AI into companies, developers and OEMs systems and workflows from a hardware perspective. In years past, AI has been enhanced with the advent of AI accelerators such as GPUs, Neural Processing Units and Google's Tensor Processing Units (TPUs). AI accelerators have increased computing efficiency and thus further enabled AI applications due to their ability to perform complex mathematical operations and multiple operations in parallel. This technological enhancement has enabled everything from AI applications to high-quality graphics within video games and televisions. However, like any technology, the use of AI and its associated hardware, as groundbreaking as they are, has a particular risk profile due to the attack surface of how these hardware acceleration devices interact with the host operating system and other hardware peripherals.

This risk profile, which will be covered in this section, is apparent in the context of how AI is integrated into personal computers, mobile handsets, and other networked IoT devices in edge computing ecosystems.

Enabling Edge Computing or Local AI Model Inference

Edge computing shall be constrained to specifically mean bringing computation of the data closer to the user in such a way that it lessens system latency between where data is being sourced and where it's being processed. In contrast, most modern system architectures, such as those involving cloud computing, rely on some external network connection to a database where the data is processed. Once processed, the appropriate response data is transmitted back to the user or source device which can be time-consuming.

Enabling edge computing, or stronger processing either on-device or as close to the device as possible, means that a stronger computational capability is necessary for the given device that performs most of the workload. Thus, AI accelerators offer a beneficial advantage such that these devices can effectively take on this computation overhead without affecting performance of the greater system. The applications for edge computing with GPUs, TPUs, and NPUs are numerous and span many different types of technologies and industries.

Terminology

The following are key concepts that will be helpful in understanding how the GPU generally works at a high level and interacts with other components, namely the OS. More detailed information on these concepts can be found here or with a GPU vendor such as NVIDIA.

- **CUDA:** The Compute Unified Device Architecture (CUDA) framework is a parallel computing platform and programming model developed by NVIDIA for general computing on its GPUs.
- **OpenCL:** Open Computing Language (OpenCL) allows developers to harness the computational power of various processing units in a system for parallel computing tasks.
- **Kernel:** Functions that are executed in parallel on the GPU. They are designed to perform small, independent tasks that can be executed simultaneously by multiple threads.
- **Kernel Grids:** Groups of thread blocks on the same kernel.
- **Threads:** Threads that are isolated to a given block.

[16] https://www.schneier.com/blog/archives/2016/02/practical_tempe.html

- **Registers:** Memory regions that are generally reserved for interacting with the GPU such as kernel manipulation.[17]
- **Shared Memory:** The GPU has a dedicated portion of Random Access Memory (RAM) with the host operating system. [18] This is meant to decrease latency in memory access between the host OS and the GPU.
- **Caches:** The GPU utilizes caches for optimized data retrieval. They store frequently accessed data and instructions. The L1 cache is the fastest but smallest in terms of memory capacity, while the L2 (or last level) cache has a larger memory capacity but is slightly slower with memory access time. The number of caches is extremely arbitrary and dependent on design decisions that primarily involve balancing performance and cost. However, generally speaking, varying memory capacities and access times will change between each cache.

Simply put, CUDA/OpenCL is a way for developers to create dynamic programs to interact with the GPU. The GPU is built with specialized computational components called kernels arranged in particular ways that enable highly efficient computation. Lastly, the shared memory and caches refer to typical hardware peripherals that pertain to faster memory access.

General AI Accelerator Threat Model

Even though the GPU has a unique threat model in and of itself, the importance of how these vulnerabilities and exploitation techniques propagate into AI applications (e.g., on-device AI) must be acknowledged to appropriately consider the underlying risks. If considering nation state level threat actors then it must be assumed that attackers are sufficiently capable to assume all threat actor roles identified below, perhaps even in combination for the most sophisticated attacks. When taking the above vulnerabilities into consideration the following can be considered when threat modeling an on-device or local AI application.

Protected Assets

- **Intellectual Property:** An attacker with the capability to leak data from an on-device AI application can extract information on the AI model that is presumably stored on the device. A notable example of such an AI application would be a proprietary LLM or closed-source AI model of which an attacker is able to extract or otherwise infer features or metadata of the stored model data.
- **User Data:** The security and privacy of user data on the device should always be taken into consideration, especially when interacting with LLMs. Under the right circumstances, users' data may be at risk if the correct precautions of confidentiality are not being taken with stored input when the user interacts with an LLM/AI agent. Given the amount of different uses that the modern handset is used for (e.g. banking, communication, personal information, etc.), the user's data may range from anything to credit card or banking information, to sensitive information/data such as passwords or photos.
- **Local Privilege Separation:** Depending on the infrastructure or architecture of on-device AI, there may be a level of isolation/restriction needed to determine who can access LLM capabilities if there is more than one user account on the device. In multi-user environments, there is a risk of using model stealing techniques to replicate the loaded model. For example, Android users can create more than one user on the device where each user has access to specific applications and other parts of the Android environment. Thus, if overly permissive privileges were granted by default to a new user

[17] https://developer.nvidia.com/blog/register-cache-warp-cuda/
[18] https://developer.nvidia.com/blog/using-shared-memory-cuda-cc/

pertaining to LLM access, an attacker could leverage this in an attempt to gain unauthorized access or tampering of the model.

Threat Actors

- **Malware or Spyware Developers:** Malware and spyware developers may aim to persistently or semi-persistently introduce malicious code into the GPU[19] with the goal of directly or indirectly influencing the Central Processing Unit (CPU) to execute unwanted instructions. Alternatively, the GPU could also be used to laterally exploit the host operating system by potentially leveraging vulnerable kernel driver or user space GPU framework code, though if the GPU is the primary initial target of exploitation it's assumed that the attacker would presumably need physical access. Though, the physical access may not apply to applications which are not fully on-device as data centers will more than likely have shared GPU resources for cloud environments (e.g. with SR-IOV, NVLink).
- **General Users:** Users themselves may seek to modify proprietary aspects of a trained AI model in such a way that gives the user complete control over how the AI model would normally respond. Such examples include on-device proprietary LLMs that have been trained to filter or respond to certain problematic prompts in a controlled way. The user would more than likely seek for methods of poisoning or disabling the filters in place.

Attack Surface

- **User Space:** Given that an AI application is often dependent on a GPU, the following user space or components of the operating system should be considered: the shared memory accessible by the host operating system, the instructions or APIs provided by the respective GPU vendor (such as CUDA), and the vulnerabilities present in the accompanying kernel module.
- **Device Storage:** Data that is either indirectly produced or stored by LLM applications should be appropriately considered as a sensitive component of the greater system. The following items are just a few data sources that should be considered: AI/LLM Models, Vector Databases, any other generated metadata which may contain user Personally Identifiable Information (PII).
- **Physical Access:** AI applications with associated hardware such as GPUs must take into consideration that there are several different types of security mechanisms that should be enforced between hardware components or peripherals. Namely, the use of an Input-Output Memory Management Unit (IOMMU) or some intermediary component that enforces restrictions or memory isolation between what memory space the GPU has access to.

Hardware Acceleration Communication

While AI accelerators offer specialized or targeted AI computational power, the remainder of this section will analyze hardware acceleration by focusing on the GPU and its general architecture and communication protocols along with the security implications of their current design paradigms due to many similarities with how AI accelerators communicate with the host operating system. For instance, in a whitepaper detailing several use cases of Edge TPUs, the researchers mention that their general-purpose programming interface for TPUs, OpenCtpu, "shares similarities with popular GPU programming models like CUDA and

[19] This was reviewed in the "Demonstrating the Security of Discrete GPUs" Whitepaper within the "GPU Vulnerability Demonstrating IOMMU Bypass" bullet point above. In it, the researchers note that they were able to achieve introducing a "long-lived" payload within GPU. Based on the context of the proposed attack, NCC Group assesses that this most likely refers to a persistent malware capability.

OpenCL".[20] To that end, we will assume any security observations or threat modeling will closely resemble other AI accelerators such as the TPU and NPU.

The modern GPU consists of numerous key concepts that enable optimized computation. These computations can be applied to various applications outside of graphical processing, which is in large part why GPUs are also used in AI applications.[21] Below are several concepts or components that comprise the GPU which will be imperative in understanding how we apply threat modeling analysis. In addition to these concepts, there is a basic understanding that the GPU is "fundamentally [a] single instruction multiple data (SIMD) stream architecture" where this division of labor enables GPUs to perform multiple mathematical operations in parallel.

Although this section will predominantly focus on an on-device GPU setting, we noted that there are other uses for the GPU (or other hardware accelerators) that go outside of the realm of the basic consumer device. Data centers that offer the computational power of the GPU from a cloud service (i.e. multi-tenancy) are similarly comprised of technology that is found on consumer devices. However, the configuration of how the GPUs may be arranged, as well as the interconnected nature of multiple servers in a data center may add more complexity to the system design, which may in turn broaden the attack surface.

Bus Communication

There are two predominant types of GPUs that are most often used which is either an Integrated GPU (iGPU) or Discrete GPU (dGPU). However, as complex or different as they may be, they have similar functionality and interactivity with hardware peripherals such as the CPU, IOMMU, and bus interaction with memory retrieval (i.e. PCIe). The following are general components or hardware peripherals that are generally involved with the GPU.

- **Memory Mapped IO (MMIO)**: a technique in computer architecture where hardware devices, such as input/output (I/O) devices, communicate with the CPU and memory by directly accessing memory locations. Instead of using separate I/O instructions, MMIO assigns memory addresses to control and data registers of the device
- **IO Memory Management Unit (IOMMU)**: component in a computer system that manages input-output memory mapping. It's primarily used in virtualized environments to provide memory address translation services for peripherals such as network cards, graphics adapters, and storage controllers
- **Peripheral Component Interconnect Express (PCIe)**: high-speed interface standard used to connect various components like graphics cards (GPUs), network cards, and Solid State Drives (SSD) to a computer's motherboard. It works by providing a high-bandwidth, low-latency pathway for data transfer between the CPU and these peripherals, utilizing serial communication lanes to transmit data packets bidirectionally. PCIe slots on the motherboard allow for the insertion of compatible expansion cards, enabling the computer to expand its capabilities and performance.
- **Direct Memory Access (DMA)**: DMA enables devices like GPUs to transfer data directly between themselves and the system memory, bypassing the CPU. This process enhances system performance by reducing CPU involvement in data transfers, facilitating faster communication between devices and memory.
- **Caches**: high-speed memory units used by computer processors to store frequently accessed data and instructions. L1 cache, the fastest but smallest, is directly integrated into the processor core, while L2 cache, larger but slightly slower, sits between L1 cache and main memory. These caches help reduce the time needed to access data, improving

[20] https://dl.acm.org/doi/pdf/10.1145/3458817.3476177
[21] https://www.cherryservers.com/blog/everything-you-need-to-know-about-gpu-architecture

overall system performance. In GPUs, similar cache hierarchies exist to enhance graphics processing by efficiently storing and retrieving frequently used data, such as textures and vertex information, contributing to smoother and faster rendering of graphics.

There are other types of bus communication methods that are used with GPUs but may not be as pertinent from an on-device or local AI model perspective. Such examples include NVLink,[22] which is typically used in a data center where the GPUs may need to be interconnected to communicate with one another and handle higher bandwidth data transfer than PCIe allows.

User Space Communication

GPUs from different vendors may have their own vendor-specific features and libraries related to computation. However, there are many similarities from a high-level architecture perspective that will apply to most GPUs:

- **User Space Framework**: There are various frameworks that directly interact with the respective vendors' GPU. One such example of how the user space interacts with the GPU is the use of the CUDA C/C++, the OpenCL C frameworks or AMD's ROCm / HIP API. This is typically done using functions provided by the GPU-accelerated computing framework, such as CUDA's `cudaMemcpy()` function. The developer writes one or more kernels, which are functions that execute in parallel on the GPU. Kernels are written in a GPU-accelerated computing language such as CUDA C/C++ or OpenCL C. These kernels are then launched from the CPU and executed by multiple threads on the GPU.
- **Driver**: The GPU's accompanying driver is an interface between the host operating system's user space and the physical GPU. Such examples of the graphics driver include the NVIDIA Open Source Kernel Module, which contains a variety of transport methods including DMA and PCIe API to transfer data between the OS/CPU to the GPU.[23]
- **Shared Memory**: The host operating system will have a dedicated region of RAM that is specifically reserved for the graphics processor. This reserved memory is shared between the respective kernel/user space and physical GPU. The shared memory paradigm allows for faster memory access, which consequently allows for data to be transported very quickly.

GPU Vulnerability Landscape

The following articles demonstrate the complex attack surface of GPUs. While most of the following threats and exploitation techniques are not unique from a GPU exploitation technique context, they can be directly applied to AI workloads on a GPU.

- LeftoverLocals:

  The LeftoverLocals vulnerability exposes the inadequacy of memory isolation between GPU kernels, allowing a device-resident attacker to extract sensitive information. This vulnerability underscores the inherent risks associated with shared GPU resources, particularly in distributed edge computing scenarios. Attackers exploiting LeftoverLocals can intercept data (i.e. users' input prompts) intended for other GPU kernels, posing a significant threat to data security and privacy. However, as novel as the Large Language extraction was, the lack of memory isolation between processes is reminiscent of the "Mali GPU" example (below) which was able to map kernel memory into the GPU's address space using a kernel API that controls the GPU's cache memory.

[22] https://www.nvidia.com/en-us/design-visualization/nvlink-bridges/

[23] It's important to note that in the realm of GPUs, the term "kernel" is overloaded, and in this context refers to the operating system kernel rather than a function that executes on the GPU (mentioned above).

- **GPU.zip Vulnerability**:

  The GPU.zip vulnerability exploits the iGPU-based compression channel to execute cross-origin pixel stealing attacks, even bypassing constant time implementations such as SVG filters. This vulnerability highlights the susceptibility of GPUs to malicious manipulation, enabling attackers to extract sensitive data from web browsers. In an edge computing context, where GPUs are utilized for local computation, the GPU.zip vulnerability poses a direct threat to user privacy and data confidentiality. Specifically, this vulnerability could potentially expose web browser image data containing sensitive information.

- **Multi GPU Vulnerability Research on P100**:

  Research on multi-GPU vulnerabilities, particularly on the NVIDIA P100, reveals the potential for covert channel attacks across interconnected GPUs. Attackers can exploit the cache hierarchy and cause contention on the L2 cache of remote GPUs, enabling them to discern or “fingerprint” applications running on other GPUs. This multi-GPU vulnerability underscores the complexity of securing distributed edge computing environments, where interconnected GPUs create additional attack surfaces for malicious actors. Though the act of “fingerprinting,” or discerning what application is running, is not necessarily as invasive as eavesdropping, using this vulnerabilty in combination with others may be detrimental to a user’s security on their device.

- **GPU Vulnerability on Mobile Handset**:

  GPU vulnerabilities extend beyond traditional computing environments, such as mobile devices, which includes Qualcomm Adreno GPUs. The discovery of a side-channel timing attack on GPU performance counters highlights the pervasive nature of GPU vulnerabilities across diverse hardware platforms. In the context of edge computing on mobile handsets, GPU vulnerabilities pose a significant risk to user data security, as attackers can exploit hardware-level weaknesses to extract sensitive information. In this example, they were able to accurately discern keys that were being pressed, which may lead to high-impact security concerns such as leaked password and entered information.

- **Mali GPU Vulnerability**:

  This blog post gives an overview over how a researcher was able to gain kernel code execution on an ARM Mali GPU by exploiting its cache memory management, which is referred to as JIT memory. The researcher eventually gained kernel code execution by chaining together several flaws in the kernel module that ultimately enabled the GPU to maliciously map kernel code into the GPU’s address space from an unprivileged context. Moreover, what made this exploit particularly concerning was that “of the seven Android 0-days that were detected as exploited in the wild in 2021– five targeted GPU drivers”. At the very least, this type of privilege escalation may put the users’ data at risk. At its worst, the kernel code execution could completely compromise the security posture of the device.

- **GPU Vulnerability Demonstrating IOMMU Bypass**:

  In this whitepaper, the researchers detail several attacks and general weaknesses on a closed-source NVIDIA GPU driver. Most notably, the researchers were able to prove out several attacks centered around the ability to bypass IOMMU memory access checks to “create stealthy, long-lived GPU-based malware”. Additionally, they enabled arbitrary CPU memory mapping into an unprivileged GPU program. Considering the permanence and pervasiveness of this exploit, there is a potential for a persistent compromise of the device’s security. If this is achieved, it may completely undermine the security within a device and put the user’s sensitive data at risk.

The identified vulnerabilities underscore the critical importance of addressing hardware-level weaknesses in GPU-accelerated edge computing AI environments. GPUs, while integral to AI processing and edge computing, present prime attack surfaces for malicious actors to exploit known vulnerabilities.

Security Impacts and Considerations

There are several further areas of consideration that were not explored in depth previously, and these warrant a brief discussion regarding potential areas of additional security concerns and mitigation strategies.

One area of consideration is potential threats to device storage. Considering all the aforementioned vulnerabilities and threat model which gain access to other parts of the device, the device storage is also at risk. For example, while AI chat assistant applications operate on real-time user input, the majority of how this technology is enabled is also dependent on the context of the user's prior conversations as well as the trained model of the AI chat assistant. These components of the AI chat assistant system are crucial as they represent a user's potentially sensitive data and proprietary technology from the respective AI vendor. To mitigate this, both users and vendors will inevitably need to rely on edge-computing AI devices to have robustly secure environments that preclude an attacker from accessing their device's storage by securely storing or encrypting this content.

Another consideration includes the methods by which attackers can gain initial access to a device. Each device will have an associated risk profile depending on where/how the device is deployed, or its respective environment (e.g. multi-user computers, third-party software, devices confiscated by law enforcement, devices taken by thieves, etc.). Thus, adequate threat modeling by device manufacturers and AI vendors would be required in order to better enumerate and anticipate how attackers may try to exploit AI applications and hardware accelerators. If the device's sensitive assets within an AI application are not adequately protected, then the more access that these presumed AI assistants have (e.g. access to photos, screenshots, phone contacts), the more the user is at risk of leaking or compromising their sensitive information.

In terms of potential resolutions, there are several ways of securing hardware components within an on-device AI system that range from hardening memory access permissions, to designing a robust framework that is tamper/leak resistant. NCC Group agrees with the recommendations listed by NVIDIA from a physical security perspective, most notably the importance of rooting critical software to hardware security features and encrypting data.

Overall, on-device AI enablement could greatly improve users' experience. For example, users would have the ability to do things ranging from native LLM Chat Assistants without a dependency of an internet connection, to natively using applications whose user interface is enhanced with the use of on-device AI frameworks. However, the risks associated with AI, specifically with the hardware components and their attack surface, should be appropriately threat modeled by respective vendors' security teams such that they aim to consider keeping users' data secure.

6 Distributed Deep Learning

Benefits

Distributed deep learning represents a transformative approach in harnessing the collective power of various computing devices, especially when individual devices might not possess sufficient computational resources to host and process complex AI models independently. One of the key benefits of this method is the ability to leverage edge computing devices, like smartphones, IoT devices, and embedded systems, in a synergistic network. These devices, while individually limited, can collaboratively handle parts of the computation for AI inference or even participate in the training phase of machine learning models. This approach not only reduces the load on a central server but also diminishes the dependency on the health of any specific network resource, thereby enhancing the system’s overall reliability and operational efficiency, as more devices join the network.

This approach is particularly beneficial for devices that lack AI-specific hardware, such as GPUs or certain CPU-integrated capabilities. These devices can tap into the AI capabilities they would otherwise be unable to utilize by leveraging a network of interconnected devices that share computational tasks. This collaborative framework allows each device to contribute to and benefit from advanced AI functionalities without the need for high-end, individual hardware upgrades. As a result, a broader range of devices can employ sophisticated AI-driven applications, from real-time analytics to enhanced decision-making processes, thereby democratizing access to cutting-edge technology across various platforms and industries. Although this is a novel approach in AI, distributed computing initiatives such as SETI@Home and Folding@Home have consistently contributed to the scientific community, and they are an inspiration for other distributed and collaborative scenarios, such as those described in this section.

Challenges in Distributed Deep Learning

Distributed deep learning involves training complex neural networks across multiple computational units, often spanning different geographical locations. This approach leverages the power of parallel processing to handle vast amounts of data and intricate computations more efficiently than a single machine could. However, this method introduces several challenges, including managing data exchange and synchronization across components, balancing computational loads, and maintaining data privacy and security. Additionally, scaling up the system, adding more elements to increase processing power, does not always result in proportional gains in performance, due to complexities in network communication and data management. These issues are crucial to understand as they significantly impact the efficiency and effectiveness of distributed deep learning systems.

Terminology

The following are key concepts that will be helpful in understanding how distributed deep learning functions:

- **Node:** Refers to an individual computational unit within a larger network. There are three types of nodes: an index server where other nodes register, servers which share their computational power with the network, and clients which use that computational power to solve their requests.
- **Swarm:** Refers to a group of nodes working together in a collaborative way. Public swarms are accessible to many users without strict access controls, whereas private swarms are protected by some form of authentication, allowing only authorized users to participate and access the network.
- **Index Server:** Acts as a central directory that helps in locating and connecting to other peers within the network. This server maintains a list of active peers (nodes) and their corresponding network addresses, facilitating the discovery process by answering queries from peers seeking to establish connections. Although they are not directly

involved in the data transfer between peers, they are in charge of task coordination and, to some degree, of data management.

Most of these definitions align with their usage in the Hivemind architecture[24], but they can also be applied to other distributed deep learning approaches.

As a real-world example of a widely used distributed network, consider the vast BitTorrent network as an intricately connected distributed computing system for file sharing. Each participant (**node**) in this global community downloads and uploads pieces of a file to multiple other nodes, forming a decentralized **swarm** (group of interconnected nodes) that collaboratively distributes content. When someone wants to obtain a specific file, they first search the BitTorrent **index servers**, which maintain lists of active torrents (distributed datasets) and their associated metadata. The **index servers** provide the locations of the nodes possessing the desired pieces, enabling efficient and distributed data retrieval from the swarm – replicating the essential functions of a distributed computing system.

General Distributed Deep Learning Threat Model

Although a distributed deep learning environment inherits many potential risks from both cloud-based and edge computing approaches, its hybrid nature introduces additional inherent risks, such as those described in further subsections. If considering nation state level threat actors then it must be assumed that attackers are sufficiently capable to assume all threat actor roles identified below, perhaps even in combination for the most sophisticated attacks. When considering these risks, the following aspects should be taken into account when threat modeling a distributed deep learning AI application:

Protected Assets

- **Intellectual Property**: Given its distributed nature, swarm members can download portions of the model design and weights, or even the entire model, depending on how the specific protocol is designed. This poses a significant intellectual property risk when using proprietary models in collaborative environments where not all server nodes are trusted. In such scenarios, the potential for unauthorized access and dissemination of the model's sensitive information is heightened, leading to concerns over the protection of proprietary algorithms and data.
- **User Data**: Distributing user data across a swarm of nodes for inference purposes can represent a significant privacy risk. In such a distributed environment, nodes processing the initial or latter layers of the models are particularly vulnerable, as the inputs and outputs in these layers can be easily guessed or reconstructed. This makes sensitive user information susceptible to unauthorized access and misuse.
- **Index Server**: Controlling an index server, even when it may not process data or contain sensitive information, can result in the modification of natural data flows to redirect traffic to malicious or rogue nodes. This redirection can facilitate unauthorized data access, manipulation, or leakage, thereby compromising the integrity and security of the entire network. The ability to alter routing paths and network behavior underscores the critical need to protect these central nodes from cyber threats, ensuring the reliability and trustworthiness of the distributed AI infrastructure.
- **Node**: While not as prime a target as their index counterparts, compute nodes are still a worthy asset to acquire. At a minimum, they are fundamental components that perform the required functionality; disabling a significant number of compute nodes can significantly hamper performance, potentially to the denial of service. In more advanced scenarios, surreptitiously compromising compute nodes can potentially yield favorable

[24] Hivemind - decentralized deep learning in PyTorch: https://github.com/learning-at-home/hivemind

positions for attackers and present the opportunity to collect data that may belong to any client using the swarm for inference purposes, given its distributed nature.

- **Model Storage Location**: Nodes in a distributed AI network need to share portions of the model among themselves, or alternatively, these portions should be stored on a specific server. Nodes then download the required segments for the task they are performing. This approach ensures that each node has access to the necessary components of the model while maintaining overall efficiency and scalability within the distributed environment.
- **Service Integrity**: As is the case in all types of distributed environments, the output of an operation relies on the proper functioning of multiple components. Consequently, a failure in the integrity of any node processing a request can lead to unexpected, or even harmful, outputs. Each node in the distributed network plays a critical role in the overall operation, and any compromise or malfunction can propagate errors, distort results, or introduce vulnerabilities. This interdependence highlights the necessity for robust security measures and integrity checks across all nodes to ensure reliable and safe outputs in distributed computing systems.

Threat Actors

- **Unauthorized External Users:** particularly in swarms widely distributed over the Internet, can create denial of service conditions by overwhelming nodes in the network, especially index servers. By flooding these critical components with excessive requests, attackers can disrupt the network’s functionality, rendering it unable to process legitimate tasks efficiently.
- **Authorized External Users:** who control a swarm member can exploit internal functionalities available in the network to create denial of service conditions. By leveraging their access privileges, they can generate excessive internal requests or manipulate data flows, overwhelming nodes and disrupting the network’s normal operations.
- **Malware and Spyware Developers:** Malware or spyware developers may seek to persistently or semi-persistently introduce malicious code into nodes within the swarm to directly or indirectly influence its behavior. This could involve backdooring responses or collecting sensitive data, thereby compromising the integrity and security of the network.

Attack Surface

- **External Denial of Service (DoS)**: External attackers may execute DoS attacks using traditional techniques to overwhelm the most sensitive nodes in the distributed network, particularly in networks that are widely distributed and accessible from the internet.
- **Rogue Index Servers** pose a significant threat to distributed networks because they can manipulate the flow of operations within the system. These compromised servers could, for instance, assign the input or output layers of the model to other rogue compute nodes. This manipulation could lead to unauthorized access to sensitive user information or the generation of manipulated outputs.
- **Rogue Compute Nodes** also present a significant threat to distributed networks. If these nodes can exploit the protocol to position themselves as input or output nodes, they can gain unauthorized access to sensitive user information or manipulate the output. Even if they cannot directly choose their roles, the risk remains considerable. The impact of a rogue compute node will depend on the specific task it is assigned.
- **“Supply Side” Brownout Scenario**: One of the major benefits of distributed AI is the access to massive amounts of compute power over a broad spatial context; however, malicious agents with control over a significant portion of the swarm can easily negate

this by restricting or rate limiting access to the network resources. The protocol should account for this scenario and other potential threats to identify and remove unresponsive or malicious nodes promptly.

Distributed Deep Learning Node Communication

Communications play a critical role in distributed deep learning systems, particularly as these systems extend beyond local computational resources to include wider network interactions across Local Area Networks (LANs) or the Internet. While traditional hardware communication channels like buses are still utilized, the expansion to distributed environments introduces more complex communication layers. This not only increases the overall network traffic but also brings about new challenges and risks not present in scenarios restricted to local resources.

The shift to distributed networks in deep learning is somewhat akin to a hybrid approach, merging aspects of cloud-based AI capabilities with the utilization of local computing power. In this model, the communication risks associated with both local and cloud environments become relevant. For example, data in transit needs robust encryption to prevent interception, and systems require secure authentication mechanisms to verify the identity of network nodes. These are common requirements in many internet-based services, such as web browsing and email, but they gain additional complexity in a peer-to-peer network or other architectures used in distributed deep learning. Here, every node potentially communicates with multiple other nodes directly, rather than through a central server, which necessitates comprehensive security measures across all points of the network to safeguard against vulnerabilities and ensure data integrity and privacy.

In these architectures, initial network setup often necessitates visibility across all participating nodes, which can become problematic, especially when these nodes communicate through gateways that implement Network Address Translation (NAT). NAT can obscure the Internet Protocol (IP) addresses of individual nodes, complicating direct node-to-node communication. This requirement not only poses functional challenges but also introduces additional security risks. For nodes to effectively participate in the network, certain ports may need to be exposed to the Internet, which increases vulnerability to unauthorized access and attacks. An alternative solution to manage these challenges involves using index servers, which help in organizing and facilitating node discovery and communication without requiring direct exposure of each node’s network details. This approach, while potentially mitigating some risks associated with port exposure, adds another layer of complexity and requires robust security measures to protect the index server itself from cyber threats.

An example of a protocol designed to address these communication risks is libp2p[25]. In frameworks such as Hivemind or Petals[26], libp2p plays a crucial role in enhancing the security of communication layers within Peer-to-Peer (P2P) networks[27]. It supports a variety of transport protocols, including QUIC[28], which incorporates built-in encryption at the transport layer, while others require a security handshake after establishing the transport connection. This protocol also covers the authentication by utilizing peer identity verification mechanisms. In libp2p, each peer is uniquely identified by a Peer ID derived from a private cryptographic key. This approach enables the secure authentication of remote peers,

[25] libp2p - Documentation Portal: https://docs.libp2p.io
[26] Petals - Run LLMs at home, BitTorrent-style: https://petals.dev
[27] libp2p - Security Considerations: https://docs.libp2p.io/concepts/security/security-considerations/
[28] QUIC - A UDP-Based Multiplexed and Secure Transport: https://datatracker.ietf.org/doc/html/rfc9000

ensuring that communications and data exchanges are with the intended and verified entities and not imposters.

Model Secure Storage and Distribution

The practice of storing and loading machine learning models using serialization mechanisms like Python's `pickle` module poses significant security risks[29]. Serialization involves converting an object into a format that can be easily stored or transmitted, and deserialization is the reverse process. Unfortunately, many machine learning frameworks rely on serialization formats that are vulnerable to attacks, particularly through untrusted deserialization vulnerabilities[30]. These vulnerabilities occur when unvalidated or malicious data is deserialized, potentially leading to Remote Code Execution (RCE) or other security breaches.

Moreover, while Static Application Security Testing (SAST) tools are employed to detect deserialization issues in source code[31], they often miss vulnerabilities in machine learning frameworks. This oversight can occur because these frameworks might implement their own interfaces for model storage, using underlying libraries known for their weaknesses, but these are obscured by the framework's interfaces. Thus, the actual usage of vulnerable libraries may not be apparent, remaining undetected by security analyzers, leaving user data exposed for extended periods since vulnerabilities might not be identified as swiftly as under different circumstances.

Addressing these concerns, initiatives like Hugging Face's `safetensors`[32] aim to establish a universal and secure format for storing and distributing models. This method is designed to mitigate the risk of RCE vulnerabilities associated with traditional model serialization methods. However, despite these advancements, many models and data scientists continue to use insecure storage formats. This issue is particularly risky in distributed and collaborative environments, where controlling the integrity of models is crucial. If a model stored in an insecure format is shared across a network, it could potentially compromise all connected peers, leading to widespread security breaches.

The persistence of these risks underscores the need for continued vigilance and the adoption of secure practices in machine learning model management, especially as the field grows and more entities rely on shared, distributed systems for their operations.

Privacy and Integrity Considerations in Distributed Deep Learning

In distributed deep learning environments, particularly those involving P2P architectures, the exchange of data between nodes introduces significant privacy and integrity risks. Each participant in a swarm contributes to the computational process, which involves sending and receiving portions of the data, such as model inputs and outputs. This sharing mechanism can potentially allow a malicious participant to access and possibly infer sensitive information from the data being processed. Moreover, there is a risk that an adversarial node could manipulate the data it is supposed to process, intentionally skewing results or injecting false data, which could lead to incorrect model outputs.

Ensuring the integrity and confidentiality of data in such settings is challenging. The architecture must incorporate robust security measures to prevent unauthorized data access and to verify the authenticity and integrity of the data being exchanged. Mechanisms like end-to-end encryption of data in transit and rigorous validation checks before data is processed can help mitigate some of these risks, potentially at the cost of the stated

[29] Exploiting Python pickles: https://davidhamann.de/2020/04/05/exploiting-python-pickle/
[30] Practical Attacks on Machine Learning Systems - Models are Code: https://research.nccgroup.com/2022/07/06/whitepaper-practical-attacks-on-machine-learning-systems/
[31] Semgrep Serialization Rules: https://semgrep.dev/r?q=serialization
[32] Safetensors - ML Safer For All: https://github.com/huggingface/safetensors

benefits of localized computing like privacy and latency. However, the decentralized nature of P2P networks makes implementing comprehensive security solutions more complex compared to centralized systems. Therefore, designing these systems requires a careful balance between accessibility and security to protect against both data leakage and manipulation.

Homomorphic encryption, which has been proposed as a potential solution to address privacy concerns in distributed machine learning[33], allows computations to be performed on encrypted data, enabling the processing of sensitive information without exposing it. However, its practical application remains a significant challenge. The complexity and computational overhead associated with homomorphic encryption techniques have made it difficult to implement effectively in real-world distributed settings. As such, it continues to be an active area of research within the academic community, with ongoing efforts to refine its feasibility and efficiency[34].

Distributed Deep Learning Vulnerability Landscape

Distributed deep learning, predominantly explored and utilized within academic and research environments, has not been extensively subjected to security research and bug hunting efforts. This relative lack of exposure has resulted in an absence of a comprehensive list of well-known vulnerabilities specific to this field. Consequently, references to potential security weaknesses are primarily based on the known issues already discussed in previous subsections, rather than a robust catalog of documented vulnerabilities. This underscores the nascent state of security considerations in the realm of distributed deep learning.

[33] Securing Machine Learning Workflows through Homomorphic Encryption: https://defence.ai/ai-security/homomorphic-encryption-ml/

[34] Google Scholar - privacy machine learning: https://scholar.google.es/scholar?q=privacy+machine+learning

7 Regulatory Landscape

To understand the regulatory context of this analysis, we reviewed approaches to specific regulation potentially affecting dedicated buttons, keys, or compute power to access AI assistants on devices. In addition, we considered what approaches regulators might take based on their historic and current views of relevant technologies and their characteristics including: hardware security; open/proprietary technology systems; and AI, including vertically-integrated AI in which single companies have outsized control over the development, implementation, integration, and distribution of their AI products.

We have taken a global view. While our assessment includes many references to EU legislation, the EU has historically set stringent compliance benchmarks, which are often considered the most challenging scenario that global businesses might need to meet. For example, the EU's General Data Protection Regulation (GDPR) is frequently regarded as the 'gold standard' for data privacy requirements around the world.

Product-specific regulation

While authorities have reviewed the vertical integration of hardware and software (see 'Views on proprietary systems' below), we found no evidence of governments or regulators (specifically) scrutinizing dedicated hardware facilitating AI-powered assistance on devices.

That said, there is a global movement to require developers and OEMs to embed higher levels of cybersecurity in their hardware and software products. The most significant piece of legislation in this area is the EU's Cyber Resilience Act (CRA). The new law is due to be enacted later this year and will require all hardware and software sold into the EU to meet cybersecurity requirements[35] – as outlined in Annex I of the draft Act[36].

[35] EU Cyber Resilience Act: https://digital-strategy.ec.europa.eu/en/policies/cyber-resilience-act

[36] Regulation (EU) 2019/1020 - Annexes: https://eur-lex.europa.eu/resource.html?uri=cellar:864f472b-34e9-11ed-9c68-01aa75ed71a1.0001.02/DOC_2&format=PDF

In spotlight: EU's Cyber Resilience Act (CRA)

The CRA seeks to establish the parameters for the creation of secure products with digital components by ensuring that hardware and software products are released onto the market with fewer vulnerabilities and that manufacturers and developers take security seriously throughout a product's lifecycle. It includes both real and intangible digital items, such as linked devices and software products integrated into such devices.

The Act's purpose is to regulate **all** hardware and software, with very few exceptions. Thus, the integration of AI systems (software) within physical devices (hardware) is likely to be in scope and subject to the requirements. The Act splits covered products into three categories: Class I; Class II; Unclassified or Default. Under the current draft of the legislation, Class I and II products will be required to demonstrate conformance through third-party assessments or the application of a standard.

As the Act is yet to be finalized, it remains unclear whether AI-powered assistance systems will fall into scope of the Class I or Class II categories. Nevertheless, security and privacy risks associated with such systems are likely to become focal points in the compliance journey no matter whether third-party assessments or the application of a standard is required, as the CRA prioritizes enhancing the resilience of digital infrastructure against cyber threats for all products it will regulate. It states that developers must navigate these regulatory expectations by implementing robust security measures and safeguarding user privacy to mitigate risks effectively.

The CRA is being developed in tandem with other recent EU regulatory frameworks, including the AI Act. While the Act does not contain specific rules for AI systems, it makes heavy references to the EU AI Act, specifically Article 6 ("High Risk AI Systems") (see further information on the AI Act below). This is discussed more thoroughly later in this report.

Any developer considering integrating dedicated buttons, keys, or compute power to access AI assistants on devices must consider whether doing so will affect their ability to comply with evolving cybersecurity rules for hardware and software products.

Looking beyond the security requirements placed on all hardware and software, we may also be able to take learnings from policymakers' broader position on, and historic regulation of, proprietary models and AI to determine how AI-driven keys or compute power might be perceived and regulated over time.

Views on proprietary systems

Broadly speaking, regulators' and policymakers' stated aims for governing technology are to promote competition, innovation and consumer choice, while also ensuring the security, privacy, and safety of users. We have seen these considerations play out in the global debate on proprietary systems.

By way of example, closed mobile app ecosystems have been subject to scrutiny over the past few years and may offer a case study on how regulators' views on closed AI-driven keys or compute power on devices could evolve:

- The **EU's** Digital Markets Act (DMA)[37] was the first notable intervention in the mobile app ecosystem (while also impacting many other aspects of digital markets). It aims to promote end user choice by ensuring that gatekeepers[38] allow third parties to

[37] Digital Markets Act: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R1925

[38] Under the DMA, the European Commission can designate digital platforms as "gatekeepers" if they provide an important gateway between businesses and consumers in relation to core platform services.

interoperate with their services in specific situations, while also ensuring they do not treat services and products offered by the gatekeeper itself more favorably than those offered by third parties.

Responding to concerns that “opening up” digital markets in this way could create security risks, the DMA stipulates that gatekeepers can take measures to ensure third parties “do not endanger the integrity, security and privacy of its services, provided that such measures are strictly necessary and proportionate and are duly justified by the gatekeeper.” In practice, the Act has resulted in gatekeepers changing the way their app stores operate[39] to achieve compliance.

- Reports from January 2022 stated[40] that the **US** Federal Government expressed concerns that “certain [DMA] obligations could create vulnerabilities online”, particularly “regarding the practice of ‘sideloading’ a process that allows for the distribution of apps outside closed systems.” According to the reports, the Government urged the EU to ensure the DMA does not create “inadvertent cybersecurity risks or harms to technological innovation”.

  However, a Department of Commerce paper on the mobile application ecosystem[41] - published the following year - recommended that Congress should enact laws that “open up distribution of lawful apps, by prohibiting anti-competitive restrictions or barriers to the direct downloading of applications” while also “retaining appropriate latitude for legitimate privacy, security, and safety measures”. The report summarized the debate about closed and “semi-closed” mobile app ecosystems as follows:

  “Consumers and the economy in general are best served when there is a level playing field for companies to compete. This fosters innovation and allows the competitive process to select winners. That said, the need for security is obvious and pressing, and it is clear that maintaining acceptable levels of security on mobile devices operating systems is not a small or simple task. The question is, as it has always been, how best to allow for openness at the same times [sic] as addressing these concerns.”

- In their market study into mobile ecosystems[42] published in 2022, the **UK** Competition and Markets Authority’s (CMA) came to the following conclusion on app stores:

  “There could be significant potential benefits from interventions that open up choice for users, such as new curated app stores [...] But these are not straightforward interventions. Significant concerns have been raised about their implications, particularly on security and privacy. We agree that these are important and that sufficient safeguards need to be in place. Our view is that these security concerns are likely to be surmountable, although will need to be given further consideration.”

- A 2021 report from the Australian Competition & Consumer Commission (ACCC) on app marketplaces stated that there are security and privacy concerns “associated with releasing functionality to third parties in certain circumstances”, but noted that “vertical integration and the accompanying risk of self-preferencing can give rise to potential competition concerns.”[43]

[39] Apple announces changes to iOS, Safari and App Store in the EU: https://www.apple.com/uk/newsroom/2024/01/apple-announces-changes-to-ios-safari-and-the-app-store-in-the-european-union/

[40] US pushes to change EU’s digital gatekeeper rules: https://www.politico.eu/article/us-government-in-bid-to-change-eu-digital-markets-act/

[41] Competition in the Mobile App Ecosystem: https://www.ntia.gov/sites/default/files/publications/mobileappecosystemreport.pdf

[42] Mobile ecosystems market study: https://www.gov.uk/cma-cases/mobile-ecosystems-market-study

In summary, while policymakers recognize that opening up closed mobile app ecosystems can present security and privacy challenges, there is a general view that these are not insurmountable – in other words, the choice between competition and ensuring technology is safe and secure is not a binary one.

For many reasons – including the fact that, as this report finds, closed AI models present notable security risks such as a single point of failure and risks associated with homogeneous environments, while open integration can present benefits such as promoting an adaptable and customizable environment - the mobile app case study is not a direct parallel to the ecosystem under consideration in this report. However, it does indicate a general view that security and competition considerations are both critical parts of the regulation of proprietary systems and integrations.

The regulation of AI-powered systems

While legal and regulatory frameworks governing AI are continuing to be shaped and developed, there is global agreement - within both the democratic world and beyond it - about the high-level principles that should underpin the development of AI systems and technologies.

As captured in the table below, these principles reflect the aforementioned approach to governing technology: promoting competition, innovation, and consumer choice, while also ensuring the security, privacy, and safety of users. In addition, we see a focus on transparency and accountability.

[43] Digital platform services inquiry: https://files.lbr.cloud/public/2021-04/Digital platform services inquiry - March 2021 interim report.pdf?WFqpIsj8vT7N5umf3Ksgg7YOjFkweamR=

	OECD AI principles[44]	U.S. Biden-Harris Executive Order[45]	EU's AI Act[46]	UK – a pro-innovation approach to AI regulation[47]	Australia's interim response – safe and responsible AI[48]	South Korea – Digital Bill of Rights[49]
Safety and security	AI systems should be robust, secure and safe throughout their entire lifecycle	AI must be safe and secure	Achieve appropriate levels of accuracy, robustness, and cyber security [*requirement for high-risk systems]*	AI systems should function in a robust, secure and safe way throughout the AI life cycle	The government will balance the need for innovation and competition with the need to protect community interests including privacy, security and public and online safety	Build a safe and trustworthy digital society. Digital threats [should] be managed through systematic structures
Privacy	AI actors should respect [...] privacy and data protection	Americans' privacy and civil liberties must be protected as AI continues advancing	The right to privacy and to protection of personal data must be guaranteed throughout the entire lifecycle of the AI system	Processing of personal data [...] should be compliant with [existing] requirements		Guarantee of access and control over personal information

[44] OECD AI principles: https://oecd.ai/en/ai-principles

[45] Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[46] References taken from the European Parliament's position published in April 2024: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf

[47] A pro-innovation approach to AI regulation: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

[48] Safe and responsible AI in Australia consultation: Australian Government's interim response: https://storage.googleapis.com/converlens-au-industry/industry/p/prj2452c8e24d7a400c72429/public_assets/safe-and-responsible-ai-in-australia-governments-interim-response.pdf

[49] South Korea presents a new digital order to the world: https://www.msit.go.kr/eng/bbs/view.do?sCode=eng&mId=4&mPid=2&pageIndex=&bbsSeqNo=42&nttSeqNo=878&searchOpt=ALL&searchTxt=

	OECD AI principles[44]	U.S. Biden-Harris Executive Order[45]	EU's AI Act[46]	UK – a pro-innovation approach to AI regulation[47]	Australia's interim response – safe and responsible AI[48]	South Korea – Digital Bill of Rights[49]
Competition, innovation and consumer choice	AI should be developed consistent with [...] consumer rights and commercial fairness.	The Federal Government will promote a fair, open, and competitive ecosystem and marketplace	Unfair commercial practices leading to economic or financial harms to consumers are prohibited under all circumstances	AI systems should not [...] discriminate unfairly against individuals or create unfair market outcome		Guarantee of fair access to and equitable opportunities in the digital; promote fair access to and equitable opportunities in the digital
Transparency	AI Actors should commit to transparency and responsible disclosure regarding AI systems		*High-risk AI systems* shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent	AI systems should be appropriately transparent and explainable	The government will consult further on options for introducing new regulatory guardrails with a focus on testing, transparency and accountability	
Accountability	AI actors should be accountable for the proper functioning of AI systems	It is necessary to hold those developing and deploying AI accountable	*[The draft Act establishes responsibility for multiple actors in the supply chain]*	Clear lines of accountability [should be] established across the AI life cycle		The Bill defines the universal rights of citizens and the responsibilities of different entities

We note that across these jurisdictions the only new legislation that has been passed setting additional rules for AI systems is the Colorado Artificial Intelligence Act (SB 24-205) (CAIA) which focuses primarily on addressing issues of bias within AI systems - while the EU's comprehensive AI Act is very close to entering into force. Indeed, while there is agreement on the high-level principles, governments' policy approach does differ - with some, like the EU, moving more quickly than others to set regulatory guardrails.

In spotlight: EU's AI Act

In March 2024, the European Parliament adopted[50] the world's first regulatory framework on AI – the EU AI Act. Member States have since given their final agreement. Once it enters into force, the Act will provide requirements and obligations regarding specific uses of AI. The Act regulates numerous organizations involved in the development, distribution, and deployment of AI systems in the EU market, including providers[51], importers[52], and distributors[53].

The EU AI Act is part of a wider plan set out by the EU Commission, which also includes the updated Coordinated Plan on AI. Together, both the regulatory framework and the Coordinated Plan will ensure the safety and rights of people and businesses. The wider plan also aims to increase investment and innovation in AI across EU member states.

The Act defines 4 levels of risk in AI:

- The highest level of risk within the regulation is **unacceptable risk**. This encompasses any and all AI systems that are considered a clear threat to the safety, livelihoods and rights of people. Any AI system designated as unacceptable will be banned in the EU.
- The second-highest level of risk within the regulation is **high-risk**. This encompasses the use of AI in Critical National Infrastructure and administration of justice and democratic processes etc. Any AI system designated as "High-Risk" will be subject to further stringent requirements before they can go out to market.
- The third-highest level of risk within the regulation is **limited risk**. AI systems classed as limited risk are those with specific transparency obligations, such as chatbots and virtual assistants.
- The lowest level of risk within the regulation is **minimal or no risk**. The majority of AI use in the EU will fall under this category, and developers/users will have minimal obligations within the AI Act.

The Act's emphasis on promoting trustworthy AI underscores the significance of addressing security, privacy, and other relevant risks associated with these systems. For example, Article 14 Section 4 states that "High-Risk AI systems shall be resilient as regards attempts by unauthorized third parties to alter their use or performance by exploiting the system vulnerabilities." Competition and consumer choice also emerge as vital considerations, reflecting the Act's overarching goal of fostering innovation whilst safeguarding fundamental rights. Compliance efforts must navigate the complex interplay between regulatory requirements and evolving industry standards, ensuring that AI-powered assistance systems meet rigorous standards of transparency, accountability, and user empowerment prescribed by the EU AI Act.

[50] MEPs adopt landmark law: https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law

[51] A 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: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

[52] Any 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: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

[53] Any 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: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

Dedicated buttons, keys, or compute power that are integrated with AI systems may fall into the scope of the AI Act, once passed (though this is subject to how the law will come to be interpreted in the coming months/years).

Despite the Act (in its current form) not making reference to specialized hardware, it may be inferred that such systems would currently fall within the “Limited Risk” category for AI assistants. As such, these systems would be subject to transparency obligations.

In addition, it’s worth noting that the Act has been developed in such a way that it supports future development and implementation plans, and will therefore take into account any new developments within the AI space. This therefore will allow the Commission to adjust risk categories as the threats to AI becomes greater. This means that its likely current inclusion in scope at “Limited Risk” level could potentially change and evolve over time.

The Commission has already stated that generative AI models will not be classified as “High-Risk”[54], however, high-impact general-purpose models may be subject to enhanced scrutiny, and would be subject to evaluations by the Commission. While not designated as “High-Risk”, generative AI models will need to comply with the aforementioned transparency requirements, as well as EU copyright law, ensuring that consumers are aware the content has been generated by AI.

Despite AI-specific regulatory frameworks being still in development, competition authorities have indicated their intention to scrutinize vertically-integrated AI models under existing law:

- In an ‘AI strategic update’ published in April 2024, the **UK** CMA[55] stated:

  “On the competition risks around [foundation models (FMs)], our strongest concerns arise from the fact that a small number of the largest incumbent technology firms, with existing power in the most important digital markets, could profoundly shape the development of AI-related markets to the detriment of fair, open and effective competition. [...] Some of these incumbent firms have strong upstream positions in one or more critical inputs for FM development as well as control over key access points or routes to market for FM services (including downstream AI-powered applications).”

  Speaking about the CMA’s approach in a November 2023 speech[56], Chair Marcus Bokkerink stated that there should be “genuine ‘flexibility’ to switch, or use multiple models or environments according to need”, adding that this “might require interoperability and ease of porting data.”
- The **U.S.** Biden-Harris Administration’s AI Executive Order stated that agencies should address risks “arising from concentrated control of key inputs, taking steps to stop unlawful collusion and prevent dominant firms from disadvantaging competitors.”[57] In particular, the Federal Trade Commission (FTC) is encouraged to use its existing powers to “ensure fair competition in the AI marketplace”, while reports have stated that they reached an agreement with the Department of Justice on how it “will divide antitrust oversight of critical AI players”[58].

[54] EU AI Act: first regulation on artificial intelligence: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence
[55] CMA AI strategic update: https://www.gov.uk/government/publications/cma-ai-strategic-update/cma-ai-strategic-update
[56] Consumers, Competition and Artificial Intelligence: https://www.gov.uk/government/speeches/consumers-competition-and-artificial-intelligence
[57] Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

According to a write-up by law firm White & Case, at the FTC's inaugural public summit on AI policy, the FTC "signaled its concern that incumbent technology companies will quickly consolidate control of the AI sector [...] including through vertical integration between AI developers and the infrastructure stack that supports the technology." The report stated:

"FTC Chair Lina Khan argued that these companies already have power over the internet and digital media, and that the government should not allow that to happen in AI development. She suggested that this risk may require a policy solution to ensure open markets. The focus on risks from vertical integration also suggest that the FTC is looking beyond AI development to consider stakeholders deeper within the technology supply chain. The FTC's 2023 review of the cloud computing market fits this pattern, and the role of cloud computing companies appears to be an area of focus for the FTC's examination of the AI sector."

While not covered within this report of the summit, these considerations could, in future, be extended to the vertical integration of programmable buttons within devices.

- The **European Commission** launched a call for contributions[59] in January 2024 on competition in virtual worlds and generative AI. It stated fast moving digital markets "can result in entrenched market positions and potential harmful competition behavior that is difficult to address afterwards." It is therefore engaging in "a forward-looking analysis of technology and market trends to identify potential competition issues that may arise in these fields."
- A working paper on LLMs[60] by the **Australian** Digital Platform Regulators Forum - a cross-regulator forum -considered whether there are likely to be any competition issues arising from the development of generative AI models. It noted that LLM "models are likely to have features common to digital platform services that make them tend towards concentration." As such, "new entrants could find it difficult to compete with digital platform services that use LLMs as part of new and existing services."

In summary

When considering integrating hardware that facilitates AI-powered assistance onto devices, developers will need to be conscious of the principles that are - or are likely to - frame regulation and associated activity in this space. Across both hardware and AI, governments and regulators have stated their aims to foster competition and consumer choice, enhance security and safety, promote privacy, and ensure transparency. Across several jurisdictions, these principles are framed in the context of fundamental rights in the digital age. The products under consideration in this report are either already, or likely to be in future, subject to multiple overlapping global regulations and laws that promote or require compliance with these principles.

Even where dedicated hardware or compute power enabling the use of AI assistants may not be subject to regulatory compliance obligations today, that does not mean they won't be in future. For example, as noted above, the EU's AI Act has been designed to allow the European Commission to adjust risk categories as the technological and threat landscapes evolve - meaning additional systems could be brought in scope of "high-risk system"

[58] Federal Trade Commission Convenes Technology Summit Exploring AI Policy: https://www.whitecase.com/insight-our-thinking/laws-books-continue-apply-federal-trade-commission-convenes-technology-summit
[59] Commission launches calls for contributions on competition in virtual worlds and generative AI: https://ec.europa.eu/commission/presscorner/detail/en/ip_24_85
[60] Examination of technology – Large Language Models: https://dp-reg.gov.au/publications/working-paper-2-examination-technology-large-language-models#_edn74

requirements over time. Indeed, we have seen security and safety regulation develop in this way in other sectors of the economy, such as critical infrastructure where governments around the world are strengthening rules and expanding them to new industries. At the same time, many of these new laws and regulations are yet to be formally adopted, and therefore may be subject to change in the coming months. Developers must continually monitor the evolution of these rules to understand whether they are, or will be, in scope.

8 Conclusions

Incorporating closed AI capabilities into general-purpose equipment often resembles a vendor lock-in situation, presenting numerous challenges, especially in terms of security. Unlike customizable AI solutions, which allow users to modify or enhance their systems according to changing needs and threats, closed systems restrict this flexibility. This limitation can hinder the ability to respond to emerging security vulnerabilities and reduces the opportunity for third-party validations or audits, potentially increasing the overall risk profile of the technology.

Utilizing edge computing for AI introduces distinct advantages over traditional cloud-based AI systems, particularly in terms of privacy and availability. By processing data locally, edge computing can significantly reduce latency and enhance data privacy, if such benefits are prioritized by developers, as sensitive information does not need to traverse the internet to reach a central server. However, this approach is constrained by the computational power available on edge devices, which is typically less than that available in cloud environments.

Moreover, shifting AI capabilities to the edge transfers certain infrastructure risks, such as:

- The potential deployment of malicious AI models directly onto edge devices, necessitating robust security measures at the device level.
- Theft of intellectual property such as model weights via insufficient privilege management and memory protections, and training data via membership inference attacks.
- Theft of potentially personal data such as prompt and usage history.
- Malicious model loading and prompt injection.

One of the specific risks associated with edge computing arises from the management of specialized hardware like GPUs or processors with GPU-like capabilities, such as NPUs or TPUs. These components introduce new shared resources that must be meticulously managed on each device. They also expose devices to novel threats, such as those targeting shared hardware resources, which can be particularly challenging to mitigate because of the more limited set of security resources available compared to more centralized approaches.

Securely harnessing distributed learning capabilities to augment computational resources in edge computing remains an unresolved challenge. While distributed learning can theoretically pool resources from multiple edge devices to mimic more powerful computational frameworks, significant hurdles persist, especially in privacy. Ensuring that data remains protected while being processed across multiple nodes in a potentially insecure network highlights the complexity of deploying distributed learning outside controlled academic or research settings.

Lastly, it is crucial to consider that the use of AI technologies, whether in edge, cloud, or distributed scenarios, must comply with applicable laws and regulations. This includes ensuring data protection as mandated by privacy laws, adhering to cybersecurity regulations, and maintaining transparency in AI operations to meet regulatory and ethical standards. Legal compliance not only safeguards against regulatory repercussions but also builds trust with users and stakeholders in the deployment and evolution of AI systems.

9 Contact Info

The team from NCC Group has the following primary members:

- Gage Polonsky – Project Manager
  gage.polonsky@nccgroup.com
- Robert Herrera – Consultant
  robert.herrera@nccgroup.com
- Verona Johnstone-Hulse – Consultant
  verona.hulse@nccgroup.com
- Jose Selvi – Consultant
  jose.selvi@nccgroup.com
- Rafi Mueen – Consultant
  rafi.mueen@nccgroup.com
- Josh Waller – Consultant
  josh.waller@nccgroup.com