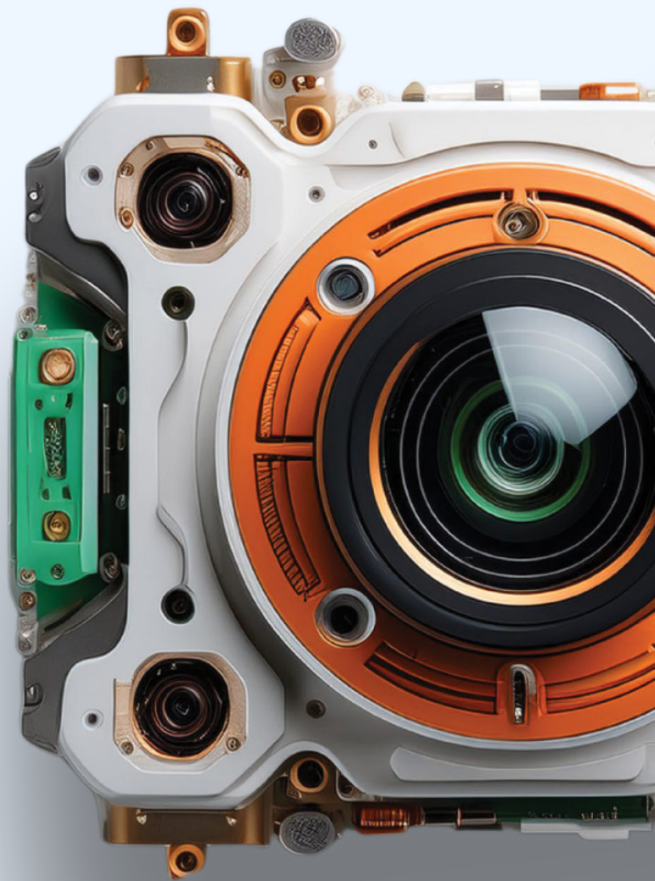
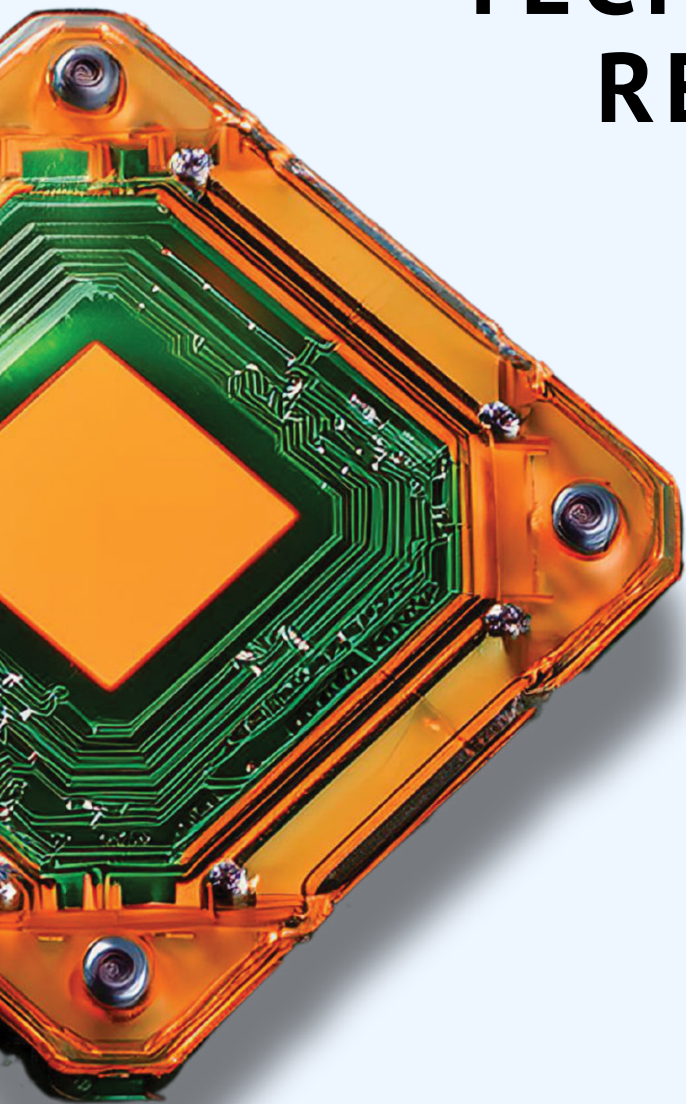




THE 2025 EDGE AI TECHNOLOGY REPORT



The guide to understanding the current state of the art in hardware & software for Edge AI.

Foreword	4
Introduction	5
About the Report	6
Chapter I: Industry Trends Driving Edge AI Adoption	7
The Safety Imperative: Real-Time Decision-Making in Autonomous Systems	8
Supply Chain Resilience: Harnessing IoT for Real-Time Optimization	9
Manufacturing and Industry 4.0: From Automated to Predictive	10
Overcoming the Challenges of Edge AI Adoption: Hardware, Algorithms, and Data	12
Smart Agriculture: Edge AI as the Catalyst for Precision and Sustainability	14
The Next Era of Healthcare: Personalized, Predictive, and Real-Time	15
Chapter II: The Role of Edge AI in Transforming Industry Trends	17
Enabling Instant Intelligence: The Role of Real-Time Edge AI in Industry	18
Why Real-Time Edge AI Matters in Autonomous Vehicles	19
How Edge AI is Enabling Advanced Manufacturing	19
Case Study: Stream Analyze's Edge AI Implementation in Manufacturing	21
The Power of Localized AI: Faster Decisions, Stronger Security, Smarter Operations	22
Healthcare and Diagnostics: From Reactive to Predictive and Personalized	23
Digital Health at the Edge: A Vision for Remote Patient Monitoring	24
Edge AI in Retail: Enhancing Operations, Personalization, and Security	26
Case Study: Amazon Go's Edge AI Implementation	26
Enhancing Security and Safety with Edge AI Efficiency	28
Scalability and Flexibility: Edge AI's Adaptive Framework	30
Scaling Intelligence Across Logistics Networks Through IoT	30
Edge AI in 2025: Scalability, Efficiency, and Real-World Impact	31
Smart Agriculture: Scaling Precision Farming for Global Food Demands	33
Chapter III: The Technological Enablers of Edge AI	34
Hybrid Edge-Cloud AI: Optimized Intelligence and Resource Management	34
The Next Generation of Specialized Edge Hardware	36
Scalable Edge NPU IP for SoC integration, from Embedded ML and Computer Vision up to Generative AI	37
Edge-Native Models and Algorithms	39
Moving LLMs and Generative AI to the Edge	40
The Role of Neuromorphic Chips	40
Explainability in Edge AI: Building Trust and Transparency	41
Privacy-Preserving Distributed Learning Paradigms for Edge AI	42
Chapter IV: Building an Edge AI Ecosystem	43
Edge AI Ecosystem & Architecture: A Multi-Layered Framework	44
Edge Devices: Real-Time Inferencing at the Source	44
Edge Servers: Local AI Execution & Aggregation	44
Cloud Platforms: Centralized AI Coordination & Model Training	45
Data Flow & Processing in Edge AI: From Collection to Insight Generation	45
The Edge AI Foundation: Unifying the Industry for Scalable Deployment	46
Accelerating The Edge AI Development Lifecycle	47
Strategic Industry Partnerships Driving Edge AI Adoption	49
Hardware and Cloud Collaborations	49
Google and Synaptics Collaborate on Edge AI for the IoT	50
Academic and Government Initiatives Supporting Edge AI	53
Challenges and Future Considerations in Edge AI Deployment	53
Energy Efficiency and Sustainability	53

Security and Data Privacy	53
Scalability and Infrastructure Management	54
The Path Forward	54
Chapter V: The Future of Edge AI	55
5 Emerging Trends in Edge AI	56
1. Federated Learning: Decentralized Intelligence at the Edge	56
2. Edge Quantum Computing and Quantum Neural Networks	57
3. Edge AI for Autonomous Humanoid Robots	59
4. AI-Driven AR/VR: The Next Evolution	61
5. Neuromorphic Computing: The Future of Energy-Efficient AI	63
New Approaches for GenAI Innovation at the Edge	65
Final Thoughts on Preparing for the Next Wave	66
References	67
Image Sources	75
About the Authors	77
About the Partner	78
Edge AI Foundation	78
About the Sponsors	79
embedUR systems	79
Ambiq	81
Edge Impulse	83
Axelera AI	84
Brainchip	85
Synaptics	85
Ceva	85
Ambient Scientific	86
About Wevolver	87

Foreword

What happens when intelligence isn't just something we access through screens or devices but something embedded in the world around us? When it's woven into our environments, shaping decisions, and unlocking new ways of working and living?

Edge AI is making intelligence feel present—alive in ways we're just beginning to grasp. It's shifting AI from something we access to something that moves with us, anticipates needs, and creates new opportunities across industries. Real-time patient monitoring in hospitals, smarter supply chains, and AI-powered creative tools are just a few examples. With this shift comes not only new possibilities but also new responsibilities.

In my work at IDEO, I've seen how emerging technologies reshape industries and redefine how we interact with the world. Edge AI is shifting the conversation from "How do we use AI?" to "How does intelligence exist around us?". It's moving beyond efficiency and automation, becoming something embedded into our environments in ways that feel seamless, responsive, and even alive.

Edge AI first gained traction in industries where real-time decision-making was essential. Autonomous vehicles, industrial automation, and healthcare couldn't afford to rely on cloud processing. What started as a solution for latency, bandwidth, and security challenges is growing into something much larger. Today, it is driving new business models, shaping more intuitive interactions, and transforming everything from adaptive healthcare systems to real-time retail.

Hospitals are already using edge AI-powered patient monitoring systems like Biobeat, which track vital signs without needing constant cloud connectivity. In manufacturing, companies like Stream Analyze are embedding AI-driven quality control directly into production lines, reducing defects and improving efficiency. In logistics, P&O Ferrymasters has increased load efficiency by 10% by using AI-driven, real-time tracking and automated decision-making. These aren't experiments. They are real, present-day innovations that make intelligence more immediate, responsive, and deeply integrated into everyday life.

This report comes at a moment when edge AI is shifting from a nice innovation to a foundational layer of technology. From next-generation AI hardware designed for low-power, high-performance edge computing to new breakthroughs enabling generative AI to run on-device, the landscape is shifting rapidly. As the technology evolves, leaders across industries will need to rethink how intelligence is designed, deployed, and experienced. This report offers insights into that transformation.

The edge has always been more than just a place where data is processed. It is where intelligence becomes immediate, responsive, and integrated into the world around us. Today, it is also where new ideas, interactions, and possibilities are taking shape.

Savannah Kunovsky, Managing Director of IDEO's Emerging Technology Lab

Introduction

While people have their eyes on the AI race of language models—from OpenAI’s ChatGPT o1 to DeepSeek’s R1, Anthropic’s Claude 3.5, and Google’s Gemini 2.0—some of the most transformative developments in AI are now occurring at the edge, where immediate, on-site processing is redefining business operations. Dubbed “the era of AI inference,” this next cycle of AI innovation is shifting inference increasingly onto edge devices, thus enhancing accessibility, customizability, and efficiency in AI applications^[1].

With 2025 underway, edge AI is rapidly changing how businesses operate by enabling real-time, localized data processing and decision-making. This shift is fueling significant trends across sectors such as autonomous vehicles, IoT, and computer vision. This report examines the evolution of edge AI from a niche technology to a mainstream driver of industry transformation, combining technical analysis with business insights.

The first chapter explores the evolving industry trends driving edge AI adoption. It analyzes how sectors like autonomous vehicles, healthcare, manufacturing, and agriculture are increasingly relying on immediate, localized intelligence to improve safety, operational efficiency, and overall performance. The analysis explains the demand for low-latency processing and reduced bandwidth

requirements, setting the stage for a shift in data processing and utilization.

The second chapter provides a detailed discussion of how edge AI is transforming operational models across industries. By processing data on-site, businesses achieve real-time analytics and decision-making capabilities that traditional centralized systems cannot offer. Specific applications, such as predictive maintenance in manufacturing and real-time patient monitoring in healthcare, illustrate the advantages of deploying AI directly at the source of data generation.

In the third chapter, the focus shifts to the technological enablers that support edge AI deployment. Advancements in specialized processors, ultra-low-power devices, and hybrid edge-cloud frameworks, along with software innovations such as edge-native algorithms and hybrid edge-cloud frameworks, are overcoming the challenges of limited processing power and scalability in resource-constrained environments. Moreover, the chapter addresses the critical topic of explainability in edge AI. By integrating lightweight, real-time explainability techniques, developers can ensure that AI decisions are transparent and verifiable, boosting trust in safety-critical applications and regulatory compliance.

The fourth chapter examines the collaborative efforts necessary to build a robust edge AI ecosystem. It explains how hardware vendors, software developers, cloud providers, and regulatory bodies are aligning their strategies to create standardized architectures and interoperable platforms. This section emphasizes the importance of partnerships and shared industry frameworks in ensuring that edge AI deployments are secure, scalable, and sustainable.

The final chapter presents a forward-looking perspective on the future of edge AI. It explores emerging technologies such as federated learning, quantum neural networks, neuromorphic computing, and the integration of generative AI models. These innovations will drive the development of autonomous systems capable of self-learning and real-time adaptation, reshaping competitive dynamics across industries.

This report promises actionable insights and thought leadership that empower decision-makers with a clear roadmap for harnessing edge AI innovation. Through rigorous analysis and industry-focused reporting, readers will gain a deep understanding of the challenges, opportunities, and practical strategies necessary to lead in the era of localized intelligence.

Samir Jaber, Report Editor

About the Report

This report is the latest installment in the Wevolver Edge AI Technology Reports series. It addresses the pressing need for actionable insights into the exploding field of edge AI and strongly focuses on the industry trends that are reshaping various sectors in 2025. It provides a clear view of the challenges and opportunities facing businesses today.

Editor-in-chief Samir Jaber led this initiative, drawing on the rigorous research and insights of co-authors John Soldatos and Deval Shah to form a cohesive narrative on the transformative impact of edge AI.

The Wevolver team has been instrumental in orchestrating discussions between contributors, synthesizing expert opinions, and steering the focus towards the most pressing questions. This collaborative effort ensures the report not only delivers actionable insights but also advances Wevolver's mission to equip engineers, developers, and decision-makers with authoritative analysis that catalyzes industry progress.

We extend our gratitude to our sponsors, whose generous support has made this insightful exploration possible. Their commitment to advancing edge AI technology underscores the shared vision of fostering a community poised to lead in the era of localized intelligence. This report stands as a testament to the power of partnership and shared knowledge in navigating the future of technology.

Chapter I: Industry Trends Driving Edge AI Adoption

The transformative power of edge AI lies in its ability to deliver localized intelligence where it is most critical, redefining how industries operate. From enabling real-time decisions in autonomous vehicles to driving predictive maintenance in manufacturing and advancing precision agriculture, edge AI has become such a cornerstone of innovation that researchers claim 2025 to be “the year of edge AI”^[2]. We are witnessing a value proposition extending beyond technological advancement; AI is serving as a strategic enabler for industries navigating the demands of speed, efficiency, and sustainability.

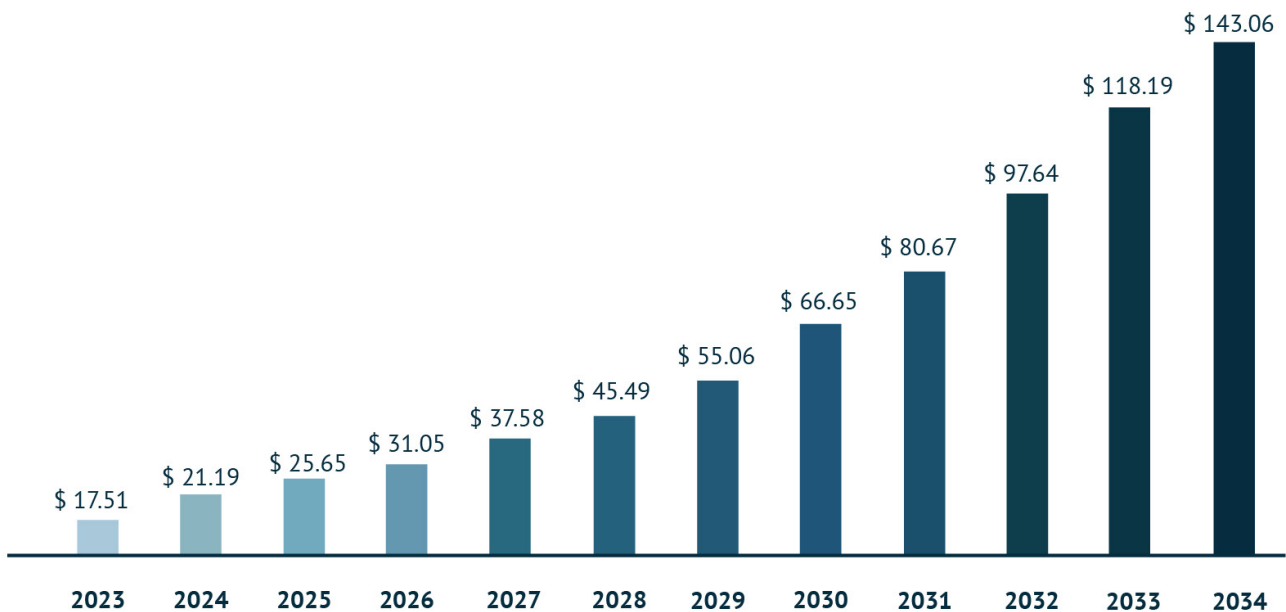
But what makes edge AI indispensable today? The answer lies in its capacity to solve two conflicting challenges: the need for instantaneous action and the imperative to reduce energy and

data waste. Traditional cloud-centric models, while powerful, still struggle with latency constraints, bandwidth bottlenecks, and environmental costs. Edge AI bridges this gap by embedding intelligence directly into devices, sensors, and machines, turning raw data into decisions at the source.

Edge AI's rapid ascent did not come merely as a response to technological curiosity; it is a direct consequence of major shifts across industries. These sectors are experiencing urgent, trend-driven demands, which are actively reshaping edge AI's evolution and adoption. Autonomous vehicles demand split-second safety decisions. Factories require predictive systems to avoid million-dollar downtime. Healthcare needs instant diagnostics to save lives. These are not isolated challenges but interconnected trends

propelling edge AI from a niche tool to an industrial imperative.

This chapter unpacks the why behind edge AI's rise: the industry-specific trends and cross-sector forces pushing intelligence closer to the source. We'll explore how trends like climate-driven resource scarcity, supply chain volatility, and regulatory mandates are rewriting the rules of innovation and why edge AI has emerged as the best viable solution.



Edge AI market size and forecast 2024 to 2034 (Image Credit: Precedence Research)^[1]

The Safety Imperative: Real-Time Decision-Making in Autonomous Systems

Two of the main drivers pushing the automotive industry toward autonomous systems are safety demands and technological momentum. According to research sponsored by the National Highway Traffic Safety Administration (NHTSA), vehicles equipped with advanced driver-assistance systems (ADAS), including blind-spot warnings (BSW), advanced cruise control, road warning systems and lane-keeping assistance, will increase substantially from 2020 to 2030, reaching near-full market penetration by 2050^[3].

Edge AI is critical to this transition, especially in aspects of safety like collision avoidance. Real-time decision-making is fundamental to autonomous vehicles navigating roads, obstacles, and other vehicles, which requires near-instantaneous data processing. Unfortunately, today's cloud-dependent processing introduces latency issues that are incompatible with collision avoidance due to factors such as the physical distance between network endpoints, the volume of network traffic, which can slow down data transmission, and the efficiency challenges across network infrastructure.

This necessitates local data processing onboard the devices collecting the data to ensure immediate decision-making, a requirement that edge computing can meet effectively.

As a result, edge AI technologies have enabled efficient and effective solutions that can accelerate the trend of autonomous systems across the automotive industry, including:

- Collision avoidance systems:** Edge AI enhances collision avoidance by processing data from multiple sensors (LiDAR, radar, cameras, and ultrasonic sensors) to analyze pedestrian movement, road conditions, and vehicle trajectories in real time. Unlike cloud-based solutions, edge AI minimizes latency by allowing the vehicle to instantly detect and respond to potential hazards, such as a pedestrian stepping into the road or an unexpected vehicle maneuver. This enables automated emergency braking, evasive steering, and predictive accident prevention.

- **Vehicle-to-Everything (V2X) Communication:** V2X communication allows vehicles to interact with their surroundings, including other vehicles, traffic lights, and road infrastructure, to optimize traffic flow and enhance safety. Edge AI enables ultra-fast processing of V2X data, helping vehicles anticipate collisions, adjust speed dynamically based on upcoming road conditions, and warn nearby cars about hazards like icy roads or sudden braking incidents. By decentralizing decision-making, edge AI improves reliability in congested or low-connectivity areas.
- **Adaptive Diagnostics and Predictive Maintenance:** Traditional vehicle diagnostics rely on periodic inspections or driver feedback, often delaying issue detection. Instead, edge AI continuously monitors vehicle components for aspects like brake wear, tire pressure, and engine health, using onboard sensors and machine learning models. By analyzing real-time data, AI detects early signs of failure, alerting drivers and fleet managers before minor issues escalate into costly breakdowns or safety hazards. This reduces downtime, extends vehicle lifespan, and enhances safety.

While AI-powered autonomous systems in passenger vehicles continue to mature, passengers still cannot take their eyes off the road yet—at least, not everywhere. The rise of Level 2+ semi-automated driving systems is effectively taking the industry a step further from partial automation (Level 2) toward conditional automation

(Level 3). In other words, we are one step away from “hands-off, eyes-off,” a state where the human driver no longer needs to be in driving or monitoring mode. This would shift the liability to the OEM during a level 3 operation^[4]. The time for widespread adoption of this shift remains to be seen as companies struggle with technological complexities, cost challenges, and server-level computational power that is not yet practical in everyday vehicles. Nonetheless, the technological trend is evidently heading in that direction.

Today, edge AI is more widely adopted in subsystems like blind-spot detection, driver behavior monitoring (drowsiness, distraction), autonomous emergency braking (AEB)^[4], and smart in-car climate control. In industrial settings, vehicles operating in controlled environments, such as forklifts in warehouses and autonomous haul trucks in mining operations, present a more immediate opportunity for edge AI deployment. In these environments, restricted and structured settings make AI modeling and decision-making more predictable, reducing the complexity of autonomous navigation. To further improve training, automakers like Hyundai Motor Group are also leveraging synthetic data platforms like NVIDIA Omniverse to simulate real-world conditions and generate high-quality training datasets without the need for physical test environments^[5].

Supply Chain Resilience: Harnessing IoT for Real-Time Optimization

Global supply chains have not been strangers to disruptions, especially since the early 2020s. These hyperconnected networks have been battling pandemics, geopolitical shifts, and climate volatility, all while trying to keep up with the demand for transparency, sustainability, and agility. Such challenges exposed fragilities and led to significant financial losses for businesses. Industry analysts say that the situation will not go back to the way that it was in 2019^[6]. This is why new supply chain trends and technology adoption have begun shaping up, ensuring higher robustness and agility. At the center of these trends is the Internet of Things (IoT).

IoT has already redefined logistics, from inventory tracking to last-mile delivery, but the next leap is turning raw data into instant, actionable insights. With the rise of digital supply networks (DSNs) and the emphasis on supply chain resilience and sustainability, AI-powered IoT has become a necessity in supply chain optimization. Edge AI enables IoT devices to process information right at the source, optimizing routes, minimizing losses, and countering disruptions as they occur.

In 2022, Gartner predicted that one in every four supply chain decisions will take place at the intelligent edge in 2025. Gartner analysts went on to describe supply chains as increasingly dynamic and covering larger networks where data *and* decisions take place at the edge^[7]. Today's market statistics further illustrate this trend, with a

projected market value growth of global IoT in the supply chain of about 13%, reaching over USD 41 billion by 2033^[8]. Here are three leading solutions that edge AI and IoT are offering in supply chain management.

- **Real-Time Visibility and Predictive Analytics:** Legacy tracking systems fail in low-connectivity ports, remote warehouses, or congested transit hubs. Edge AI eliminates these blind spots by processing data locally on IoT devices, enabling granular monitoring without cloud dependency. Such real-time data processing triggers instant alerts for anomalies such as temperature fluctuations, route deviations, or unexpected inventory shifts. Moreover, by leveraging predictive analytics at the edge, organizations can forecast bottlenecks and optimize routes, cutting downtime and operational costs.
- **Energy Efficiency Through Local Processing:** Local data processing

reduces reliance on constant cloud connectivity. By minimizing data transmission, edge AI cuts energy consumption and alleviates network strain. This localized approach supports sustainability goals while ensuring that large-scale supply chains operate with maximum efficiency. With edge AI-powered smart warehouses and low-power sensors, supply chains can become smarter and greener.

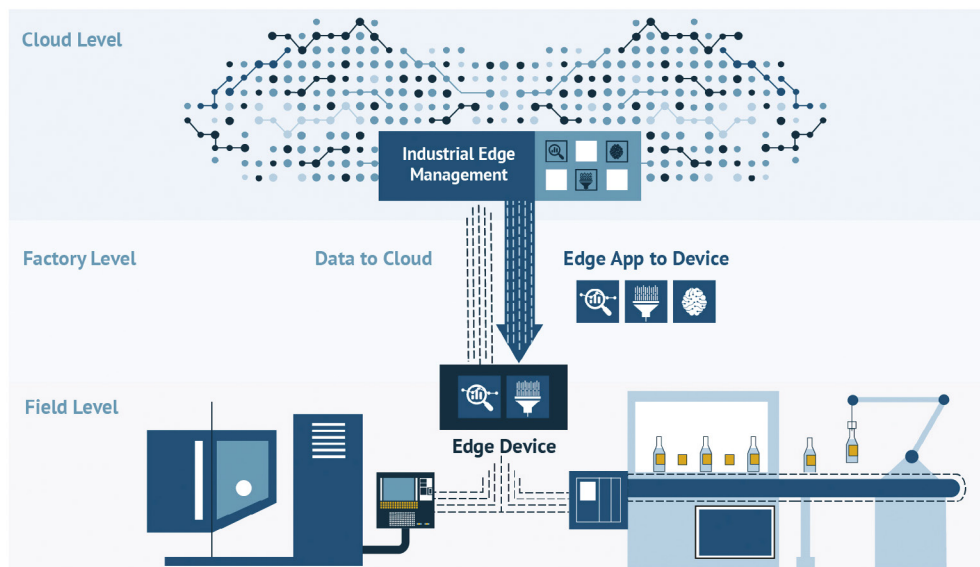
- **Automated Asset Tracking:** Modern supply chains require highly accurate asset tracking. Traditional tracking methods of assets like pallets, including manual counts or barcode scans, are relatively slow and prone to error. Advanced edge AI systems, integrated with digital twin technology and computer vision, can automate this process. For instance, a solution leveraging NVIDIA Omniverse, NVIDIA TAO, and the Edge Impulse platform included creating virtual replicas of warehouse environments to

generate synthetic data, train robust models, and deploy them in edge devices. This setup automated pallet detection and tracking, ensuring precise inventory management and streamlined operations^[9].

Integrating edge AI with IoT elevates supply chain management from reactive data collection to proactive, intelligent operations. Real-time monitoring, predictive analytics, energy efficiency, and automated asset tracking converge to create a supply chain ecosystem that is resilient, cost-effective, and prepared for future challenges.

Manufacturing and Industry 4.0: From Automated to Predictive

The fourth industrial revolution, or Industry 4.0, has been a major trend driving the manufacturing sector with smart automation, real-time analytics, and self-optimizing workflows. Yet,



Industry 4.0 using edge computing (Image Credit: Siemens)^[10]

as factories grow more complex, traditional centralized systems struggle to keep pace. Unplanned downtime costs manufacturers billions annually, while defects and inefficiencies erode margins in an era of razor-thin competitiveness.

Edge AI has emerged as a critical enabler of Industry 4.0's next phase, Predictive Manufacturing (PdM), where factories don't just react but predict, adapt, and optimize autonomously. PdM involves "gathering data from sensors embedded in manufacturing machinery, using advanced analytics to identify operational anomalies, and developing predictive models to forecast potential failures"^[10].

That is why edge AI is seeing a remarkable rise in adoption in the manufacturing sector. Surveys show that, in 2025, at least 93% of manufacturers will integrate AI into core operations, while 83% believe AI has already or will make a tangible impact^[11], driven by demands for resilience and sustainability. Edge AI meets these demands by embedding intelligence directly into machinery, sensors, and robotic systems, turning data into decisions at the source. Here are three trends enabled by edge AI and IoT in the manufacturing industry:

1. **Predictive Maintenance:** For modern manufacturing (and PdM) to take shape, reactive maintenance can no longer be the go-to methodology; predictive maintenance is essential to prevent unplanned downtime. AI-powered IoT systems enable predictive maintenance by analyzing real-time sensor data using dedicated ML algorithms to flag anomalies and anticipate

equipment failures. For example, vibration sensors installed on industrial machinery can detect early signs of wear or imbalance and trigger maintenance alerts before a potential failure occurs. This proactive approach minimizes production disruptions, prolongs equipment lifespan, and reduces downtime costs. According to Infosys^[10], "Predictive maintenance solutions enable cost savings of up to 40% over reactive maintenance and 8% to 12% over preventive maintenance. Additionally, predictive maintenance can decrease equipment downtime by up to 50% while increasing machine lifespan by 20%."

2. **Robotics and Cobots:** Industrial robots are no longer confined to repetitive tasks. Edge AI empowers collaborative robots (cobots) with real-time decision-making, enabling them to adapt to dynamic environments alongside human workers. This brings human-robot collaboration (HRC) closer to widespread application, with characteristics like:

- a. **Vision-guided assembly:** Cobots equipped with edge-powered computer vision adjust grip strength and trajectory mid-task, handling delicate electronics or irregularly shaped components with sub-millimeter precision.

- b. **Autonomous navigation:** AGVs (Automated Guided Vehicles) use LiDAR and edge AI to reroute around obstacles, reducing collision-related downtime significantly.

- c. **Safety enhancements:** Edge AI monitors human-robot interactions, halting machinery within milliseconds if a worker enters a hazardous zone.

Such advancements democratize automation, allowing even small manufacturers to deploy cobots for a fraction of traditional robotic system costs while achieving ultra-high accuracy in areas like pick-and-place operations.

3. **Automated Quality Control:** The demand for flawless production and the shift to zero-defect manufacturing is driving manufacturers to replace error-prone manual inspections with edge AI-powered quality control. Edge AI enables real-time, 100% inspection through embedded computer vision systems that analyze every component for defects as small as 0.1 millimeters. This shift is accelerated by synthetic data platforms, which simulate defects in virtual environments to train AI models without physical prototypes. The result is a paradigm where defects are detected and preemptively eliminated, reducing waste and aligning with global standards like ISO 9001 and zero-defect principles such as first-time-right (FTR) manufacturing.

By embedding intelligence at the edge, Industry 4.0 evolves from automated to predictive, enabling the factories of tomorrow^[12] to preempt challenges rather than merely respond to them, turning data into foresight and resilience into a strategic imperative.

Overcoming the Challenges of Edge AI Adoption: Hardware, Algorithms, and Data

The promise of edge AI is undeniable. From enabling real-time decision-making to reducing reliance on cloud infrastructure, edge AI has the potential to revolutionize industries like wearables, medical devices, and industrial automation. However, despite its transformative potential, widespread adoption of edge AI faces significant challenges across three critical areas: hardware limitations, algorithm optimization, and dataset availability. These challenges often create bottlenecks for developers and businesses looking to deploy AI on endpoint and edge devices.

The Hardware Challenge: Power and Performance Trade-offs

Edge AI applications demand hardware that can deliver high computational performance while operating within stringent power constraints. Traditional AI processors, while powerful, often consume too much energy to be practical for battery-powered devices. This creates a significant barrier for applications like wearables, medical sensors, and industrial IoT devices, where long battery life and compact form factors are non-negotiable.

Ambient Scientific's GPX10 processor addresses this challenge head-on. Leveraging the breakthrough DigAn® Analog In-Memory Compute technology, GPX10 delivers thousands of times more AI performance at the same power consumption—or thousands of times less power

for the same performance—compared to traditional AI hardware. Consuming as little as 100 microwatts of power for always-on AI applications, GPX10 enables AI on the smallest of devices, unlocking possibilities that were once considered nearly impossible.

The Algorithm Challenge: Optimization for Edge Devices

Even with the right hardware, developing AI algorithms optimized for edge devices remains a daunting task. Edge AI models must be lightweight, efficient, and capable of running on resource-constrained hardware without compromising accuracy. Many developers struggle to strike this balance, often spending months refining their models to meet the unique demands of edge deployment.

Ambient Scientific simplifies this process with a full-stack SDK that supports industry-standard AI frameworks like TensorFlow and Keras and comes with an AI model zoo with sample AI algorithms for various applications across voice recognition, image processing, and sensor fusion. Our custom AI compiler means developers are not limited to fixed neural network structures but rather empowered to create completely custom neural networks for edge devices, enabling product and software differentiation for product makers.

The Data Challenge: Collecting and Tagging Training Datasets


One of the most overlooked yet critical challenges in edge AI development is dataset availability. Training AI models requires large, high-quality datasets that are often difficult and time consuming to collect, especially for niche applications. Without the right data, even the most advanced algorithms and hardware cannot deliver meaningful results.

To address this challenge, Ambient Scientific has developed a unique training toolchain that simplifies data collection and tagging. Our Development Kit (DVK) allows users to easily gather and annotate data directly from onboard sensors on the development board or the edge device, ensuring that the training dataset is both relevant and representative of real-world conditions. This end-to-end solution accelerates the development cycle, enabling faster time-to-market for edge AI applications.

Enabling the Future of Edge AI

The convergence of hardware, algorithms, and data is essential for unlocking the full potential of edge AI. Ambient Scientific is uniquely positioned to address these challenges with a comprehensive solution that spans the entire development stack. From the ultra-low-power GPX10 processor to our full-stack SDK and training toolchain, we provide the tools and technologies needed to bring edge AI applications to life.

Whether it's enabling always-on voice detection in wearables, predictive maintenance in industrial systems, or real-time health monitoring in medical devices, Ambient Scientific is paving the way for a new era of battery-powered, cloud-free AI. By solving the critical challenges of edge AI adoption, we empower developers and businesses to create innovative applications that were once out of reach.



ambient scientific

AI Datasets

Smart Training Toolchain
Fast data collection & model training.

AI Algorithms

AI Models Ready to Deploy
Wake Words, Human Activity, FaceID & more.

AI Hardware

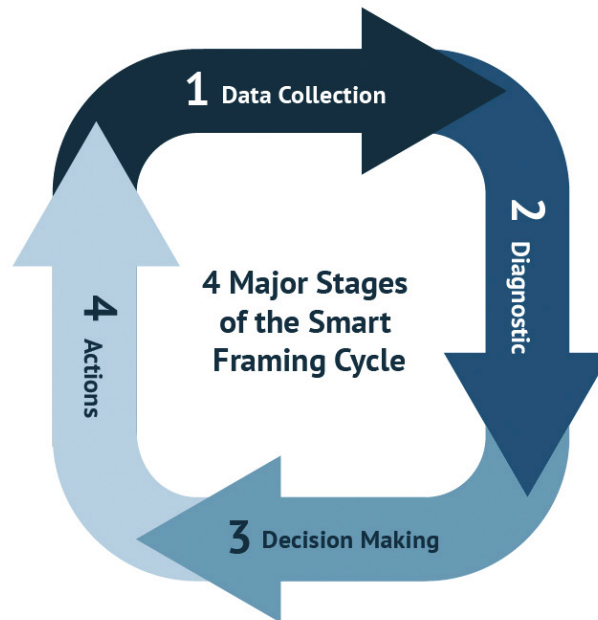
GPX10 – 512 GOPs | 100 μ W
Ultra-low-power AI for always-on edge computing.

● Data Collection

The key to smart farming is data. Therefore sensors are installed at all the strategic places in the farm, collecting real-time data about temperature, air quality, air humidity, soil moisture, weather condition and more.

● Actions

The activities planned in the previous stage are executed. Subsequently, sensors conduct new measurement on soil, air, moisture, etc., initiating a new cycle.



● Diagnostics

Edge AI solutions analyze the collected data to draw conclusions about the status of the object or process being monitored, which could help farmers identify potential problems or situations that require actions.

● Decision Making

Based on the analysis derived from the Diagnostic stage, Edge AI solutions and/or the people monitoring the system decides what actions to take.

The four major stages of the smart farming cycle (Image Credit: Aetina)^[iii]

Smart Agriculture: Edge AI as the Catalyst for Precision and Sustainability

Global agriculture is under immense pressure to increase productivity while reducing environmental impact, driven by climate volatility, labor shortages, and the urgent need to produce more food with fewer resources. By 2050, the world must feed a burgeoning 9.8 billion people^[13] while reducing agriculture's environmental footprint—a sector responsible for almost a fifth

of greenhouse gas emissions^[14]. For agriculture to become more efficient, more profitable, safer, and more environmentally friendly, technology integration is the best, if not the only, way to propel the sector forward.

As a result, concepts like smart agriculture, smart farming, and AgTech became prominent, encompassing technological trends driving the agricultural sector today. Precision agriculture, autonomous farming, and data-driven livestock welfare have become leading industry drivers by leveraging a transformative technology underlying them

all: edge AI. By processing data locally on drones, tractors, and soil sensors—and even in remote and resource-constrained areas—edge AI turns fields into intelligent, self-optimizing ecosystems. Today, edge AI-supported agriculture puts forward the following requirements: high-quality data, suitable algorithms, and computing hardware with high energy efficiency^[15]. That is why researchers, startups, and tech companies are putting on their thinking caps and coming up with edge AI-based solutions that are bringing the future of smart agriculture closer.

- **Precision Agriculture:** Traditional farming relies on uniform applications of water, fertilizers, and pesticides—a “one-size-fits-all” approach that wastes resources and harms ecosystems. Edge AI disrupts this model by enabling hyper-localized resource management, where every plant, soil patch, and livestock animal receives tailored care. A 2024 study in *Smart Agricultural Technology* notes that AI-driven precision systems reduce water use by 30% and chemical inputs by 20% while increasing yields by 15%^[16]. Edge AI achieves this by analyzing multispectral drone imagery, soil moisture sensors, and weather forecasts in real time, delivering millimeter-accurate irrigation or micronutrient dosing. For instance, AI models trained on edge devices can distinguish crop rows from weeds with up to 95% accuracy^[17], enabling targeted herbicide applications that preserve soil health. This precision is both efficient and regenerative, aligning with global standards like the EU’s Farm to Fork strategy to halve nutrient loss and cut pesticide usage by 20% by 2030^[18].
- **Autonomous Farming:** The global agricultural workforce is shrinking rapidly, with the average farmer’s age closing in on or exceeding 60 years old^[19,20]. Coupled with rising labor costs, these challenges are driving the demand for and adoption of autonomous farming systems. Ranging from self-driving tractors to robotic weeders, these edge AI-powered systems operate independently of human intervention, transforming how farms manage labor-intensive

tasks. Edge AI enables real-time decision-making at the source, allowing autonomous machinery to navigate fields, adapt to terrain changes, and avoid obstacles without relying on cloud connectivity. For example, robotic weeders equipped with edge-powered computer vision can identify and remove invasive plants with centimeter-level precision, eliminating the need for chemical herbicides^[21]. Similarly, fleets of drones coordinated by edge AI can plant seeds, monitor crop health, and apply micronutrients, reducing labor costs, especially in remote regions. These advancements are not limited to industrial-scale farms. Solar-powered edge devices and lightweight robots are democratizing automation, enabling smallholders to automate tasks like planting, pruning, and fruit picking at a fraction of traditional machinery costs.

- **Livestock Management and Sustainability:** Edge AI enables real-time livestock health monitoring and behavioral analysis without relying on manual inspections or RFID tags. Computer vision and biometric tracking identify and monitor animals individually, ensuring accurate record-keeping and seamless herd management across large farms. AI-powered behavioral analysis detects deviations in movement, feeding, and social interactions, flagging early signs of illness or distress^[22]. Farmers can intervene before issues escalate, reducing disease outbreaks and improving overall

herd health. By processing data locally, edge AI provides instant insights on weight loss, lameness, or abnormal breathing patterns, allowing for precise, timely veterinary care. Automated adjustments to feeding schedules, environmental conditions, and reproductive tracking further optimize livestock productivity while minimizing resource waste. Beyond health monitoring, edge AI contributes to traceability by integrating with IoT and blockchain systems, ensuring compliance with food safety regulations and strengthening supply chain transparency.

By processing data at the source, farms can operate independently of volatile labor markets and cloud infrastructure, ensuring resilience against climate shocks and supply chain disruptions. Looking forward, farms that adopt edge AI will lead the transition to Net-Zero Agriculture, where every input is optimized, every output is circular, and sustainability is the baseline. For agribusinesses, the question isn’t whether to adopt edge AI but how swiftly they can transform data into actionable foresight.

The Next Era of Healthcare: Personalized, Predictive, and Real-Time

Global healthcare systems are buckling under the dual pressures of aging populations and rising chronic disease burdens. By 2030, 1 in 6 people worldwide will be over 60, with 80% of older adults managing at least one chronic condition^[23]. At the same time, diagnostic errors contribute to

10% of patient deaths and 6–17% of hospital complications^[24]. Edge AI is emerging as the critical enabler of a paradigm shift from reactive treatment to preventive, personalized, and decentralized care. By processing data locally on wearables, imaging devices, and ambient sensors, edge AI delivers real-time insights without compromising patient privacy or relying on fragile cloud infrastructure. Such technology deployment is enabling trends like remote care, faster diagnostics, and real-time patient monitoring, which are shaping the future of healthcare across the world.

1. **Remote Patient Monitoring:**

Chronic diseases like diabetes, hypertension, and heart failure account for no less than 86% of US healthcare costs (CDC)^[25]. Traditional monitoring, including periodic clinic visits and manual vital checks, fails to capture critical fluctuations between appointments. Edge AI bridges this gap by enabling continuous, context-aware monitoring outside of hospitals. The rise of AI-powered wearable devices, smart homes, and telemedicine has expanded healthcare to homes, workplaces, and care facilities. Devices powered by ultra-low-power AI chips, like those developed by Ambiq, track vital signs such as heart rate, oxygen levels, and body temperature, enabling real-time health assessment^[26]. These systems reduce the burden on hospitals by allowing early intervention before conditions escalate. Similarly, edge AI-driven sensors embedded in smart home systems detect abnormal patterns in elderly patients, such as irregular

movement or prolonged inactivity. AI-enhanced fall detection devices provide immediate alerts, helping caregivers respond in time to prevent severe injury^[27].

2. **AI-Driven Symptom Identification and Early Diagnostics:**

Diagnostic errors affect at least 1 in 20 US adults annually^[28], often due to subjective symptom interpretation. Edge AI tackles this by embedding diagnostic intelligence into point-of-care devices, bringing diagnostics directly to the patient and allowing real-time symptom analysis and early disease detection. For instance, portable ultrasound devices with edge AI help diagnose cardiac anomalies in rural clinics, bypassing the need for specialist referrals. Edge AI in radiology and medical imaging accelerates disease detection by instantly analyzing X-rays, MRIs, and CT scans. These AI models improve workflow efficiency by prioritizing urgent cases and flagging anomalies for further examination. On the front lines, mobile diagnostic tools assist medical workers in identifying symptoms of infectious diseases. These tools process patient data in real time, reducing the need for lab-based testing in remote or resource-constrained environments.

3. **Predictive Healthcare and Preventative Medicine:**

Predictive analytics in healthcare is shifting treatment models from reactive to proactive. Edge AI enables this by continuously analyzing patient data, identifying risks, and facilitating early intervention. AI-powered devices tailor treatments

based on patient-specific data, optimizing medication dosage, therapy plans, and lifestyle recommendations. By leveraging local data processing, these systems adjust in real time, reducing side effects and improving outcomes. Machine learning models trained on real-time sensor data can detect early indicators of conditions like diabetes, hypertension, heart disease, and sepsis. For instance, sensors in hospital beds can detect sepsis early by monitoring body temperature, heart rate variability, and respiratory rate, flagging sepsis risks up to 6 hours earlier than traditional methods^[29]. Furthermore, by analyzing patterns and predicting deterioration, edge AI helps healthcare providers implement preventative measures before hospitalization becomes necessary. Edge AI is streamlining hospital operations by optimizing resource allocation, predicting patient admission trends, and automating administrative tasks. AI-driven scheduling tools reduce patient wait times, while smart hospital systems dynamically manage equipment usage, improving efficiency and reducing costs.

By embedding intelligence directly into medical devices, wearables, and hospital infrastructure, edge AI can redefine the speed, accuracy, and accessibility of healthcare. Faster diagnostics, real-time patient monitoring, and predictive healthcare solutions are converging to create a more responsive, efficient, and personalized medical ecosystem and ensure better patient outcomes and smarter, more sustainable medical practices.

Chapter II:

The Role of Edge AI

in Transforming

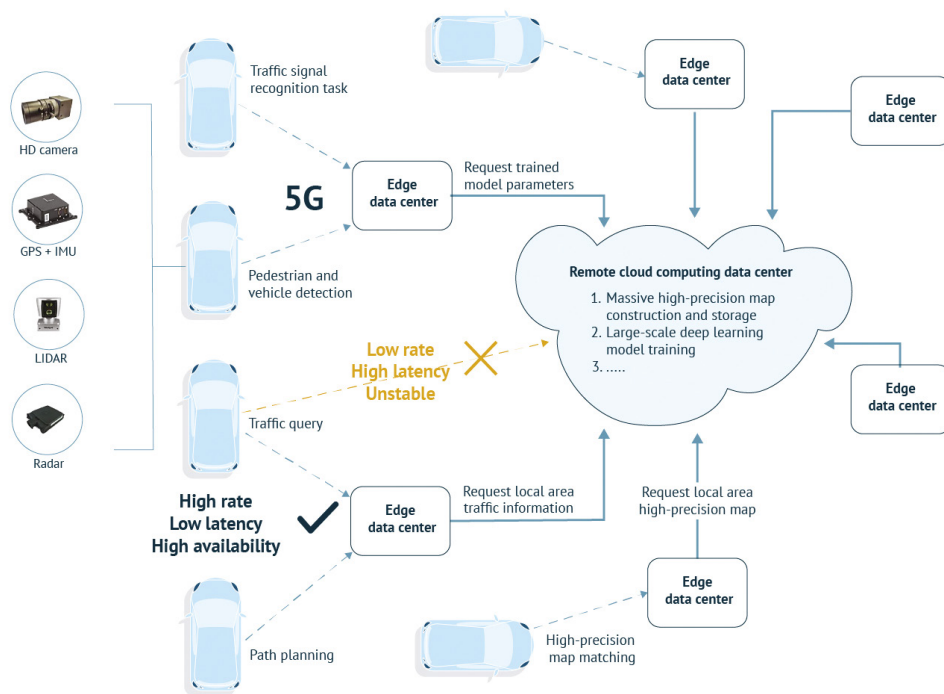
Industry Trends

In 2018, Gartner predicted that by 2025, 75% of enterprise-generated data would be created and processed outside a traditional centralized data center or cloud^[30]. Today, we are not that far from that. With AI taking 2024 by storm, the momentum going into 2025 has never been stronger for more data processing shifting to the network edge.

“*Agentic* will be the word of the year in 2025,” said John Roesse of Dell Technologies in his 2025 predictions with Forbes. But more importantly, he explains that the true potential of AI can be found when connected with other emerging technologies, such as the intelligent edge^[31]. And that is true. Today, we see a paradigm shift in how AI models are being deployed on edge devices and in edge architectures to make decisions in real time across various sectors.

This shift from so-called “isolated monolithic workloads” to end-to-end full-stack edge solutions is influenced by the need for effective, adaptable, and scalable edge environments capable of leveraging seamless interconnectivity and data movement across edge, core, and cloud infrastructures. That is why a rise in lightweight virtualization and containerization platforms is taking place to meet the demands of AI deployments at the edge. This is strongly supported by market sentiment: A 2024 survey showed that nine in ten professionals in IT, OT, and DevOps organizations believe more consistent edge application and infrastructure management would benefit them^[32]. The global edge AI market is also showing strong signs of growth, as it is projected to reach USD 84 billion by 2033, growing at a CAGR of 17.53% from 2025^[33].

Edge AI is transforming how industries collect, process, and act on data at the source. By reducing reliance on centralized systems, edge computing allows real-time decision-making, fueling everything from autonomous vehicles to predictive maintenance in manufacturing. This synergy of AI and edge technology empowers organizations to harness valuable insights faster, cut costs, and safeguard sensitive information. This chapter delves into how edge AI drives these industry shifts and spotlights key sectors that stand to gain the most. We will cover how real-time data analysis, localized model deployment, and emerging best practices converge to reshape automotive, manufacturing, retail, and beyond.



Comparison of cloud computing and edge computing (Image Credit: H. Zhao)^[iv]

Enabling Instant Intelligence: The Role of Real-Time Edge AI in Industry

Edge AI today has gone beyond on-device analytics to become a driving force that enables immediate, high-impact decisions across various sectors. The rapidly increasing global adoption of edge AI technologies underlines the growing momentum behind real-time data processing and localized intelligence. Today, the market is witnessing an increase in demand for low-latency, real-time processing, especially across the automotive, manufacturing, and smart cities sectors^[34].

In autonomous vehicles, where cameras now push into gigapixel

resolution, and LiDAR systems can fire millions of points per second, edge AI speeds up reaction times and bolsters safety^[35]. For example, Waymo has expanded simulation training and evaluations to handle rare edge cases effectively^[36]. At the same time, Li Auto expects its end-to-end model to learn from over 5 million driving data clips by this year's end^[37]. Similarly, with the AI manufacturing market forecast to grow, real-time edge AI capabilities have become a linchpin for boosting efficiency and minimizing downtime.

Picture a busy factory floor; intelligent sensors immediately flag heat spikes or mechanical stress, allowing teams to prevent disruptions before escalating. Drawing inspiration from the automotive sector, NIO's NWM (NIO World Model) demonstrates the power of ultra-fast AI predictions. Similarly, edge AI-based analytics can detect

micro-defects on production lines with remarkable precision.

By combining speed, reliability, and on-device intelligence, real-time data processing transforms standard practices for autonomous vehicles and industrial operations, paving the way for a more adaptive, efficient future across the board.

Why Real-Time Edge AI Matters in Autonomous Vehicles

Autonomous vehicles (AVs) are estimated to process approximately 11 to 152 terabytes of sensor data daily^[38]. AV operations are managed through sophisticated sensor fusion, integrating data from LiDAR, radar, cameras, and GPS systems directly onboard. This localized processing architecture addresses three fundamental challenges:

In 2024, the automotive industry prioritized edge AI hardware upgrades, such as Qualcomm's Snapdragon Ride Flex SoCs, integrating 5nm process nodes to process 150 TOPS (tera operations per second) locally^[39]. These systems reduced reliance on cloud relays to achieve sub-50ms response times for collision avoidance, which is

critical for handling sudden pedestrian crossings or highway debris.

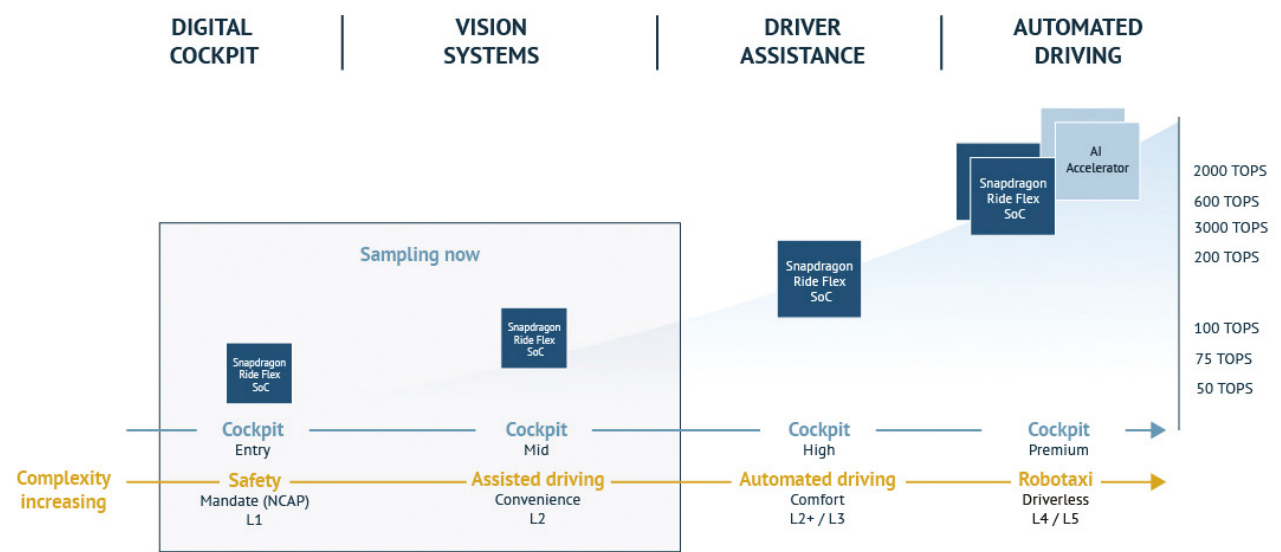
Interestingly, The 5G Automotive Association's (5GAA) updated cellular-vehicle-to-everything (C-V2X) technology roadmap emphasized hybrid V2X architectures that combined edge processing with 5G-V2X Direct Communication. This allows AVs or semi-AVs to maintain safety functions (e.g., lane-keeping, adaptive cruise control) even in cellular dead zones^[40]. Recent advancements in edge AI hardware and sensor fusion algorithms enabled autonomous vehicles to reduce decision-making latency by 30–40%, achieving response times as low as 20–50 milliseconds.

For instance, Innoviz's 2024 LiDAR upgrades incorporated edge-optimized neural networks to process point-cloud data at 20 frames per second,

minimizing delays in obstacle detection^[41]. Similarly, a Nature study highlighted multi-sensor fusion frameworks using DenseNet and YOLO V7 models, which improved real-time object tracking accuracy by 11% compared to existing techniques and in low-visibility conditions^[42]. Such examples emphasize that integrating data from edge devices like cameras, LiDAR, and radar enhances perception reliability, enables safe navigation, and helps self-driving vehicles take one step further toward actualization.

How Edge AI is Enabling Advanced Manufacturing

Advanced manufacturing lines can generate substantial amounts of data daily, depending on the complexity of operations and sensor deployment. A recent article from control engineering showed that smart factories generate



Snapdragon Ride Flex SoCs (Image Credit: Qualcomm)^[43]

more than 5 petabytes a week^[43]. Edge AI systems can process this information locally, delivering instantaneous insights and automated responses. Edge AI's impact is evident in three critical areas: predictive maintenance, quality control systems, and process optimization.

Predictive maintenance systems leveraging real-time sensor data analysis have been reported to reduce maintenance costs by up to 30% and decrease downtime by up to 45%^[44]. By continuously monitoring equipment performance, edge AI algorithms can detect subtle anomalies and potential failures before they occur, enabling proactive maintenance scheduling and minimizing unexpected downtime. A 2025 research study out of India showed how IoT-driven advanced predictive maintenance systems can reinforce maintenance strategies in industrial settings. By integrating

advanced ML techniques with IoT technologies, the researchers were able to notably enhance predictive accuracy and operational efficiency, outperforming conventional predictive maintenance methods by almost 40% in Mean Error Percentage reduction^[45]. In other words, such an integration would make nearly 40% fewer mistakes in predicting equipment failures compared to older methods. This improvement leads to more accurate forecasts, allowing for timely maintenance and reducing unexpected equipment breakdowns. Such improvements are why manufacturers are increasingly adopting predictive maintenance; a recent McKinsey report has projected a sharp increase in predictive maintenance adoption within this decade to reach up to 55% or even 70%^[46].

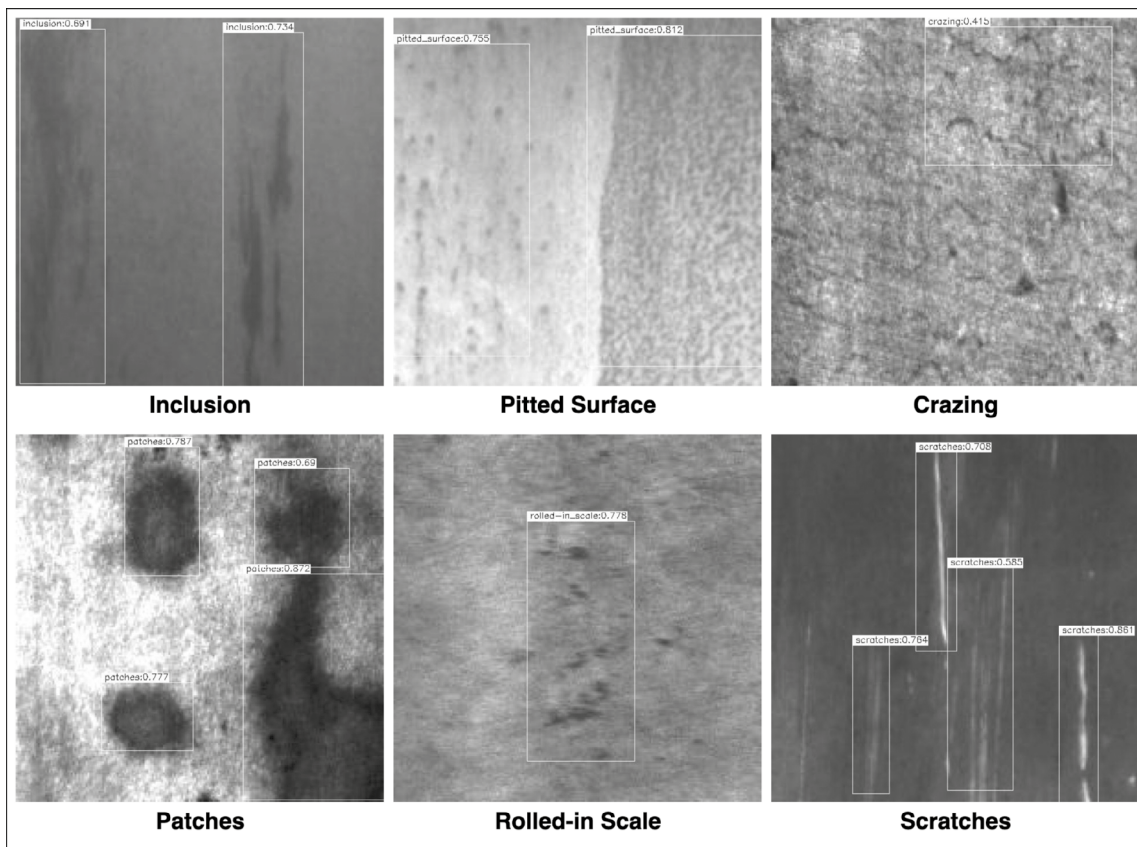
On the quality front, edge AI enhances quality control through real-time

inspection and defect detection. For instance, a major food and beverage manufacturer deployed Vision AI at the edge for quality inspection and closed-loop quality control. This system continuously monitors product variances and recommends equipment setting adjustments, improving inspection cycle times by 50-75% and enhancing accuracy^[47]. AI is increasingly embedded in industrial control systems to improve process efficiency and reduce the need for constant operator monitoring. As such, edge AI embedded in robots, sensors, and cameras enables real-time anomaly detection, root cause analysis, and immediate corrective action, reducing waste and rework.

In addition, process optimization through edge AI has demonstrated remarkable efficiency gains. Manufacturing operations utilizing edge computing have reported



Predictive maintenance architecture enabled by edge AI (Image Credit: Copernilabs)^[vi]



Quality assurance and defect detection using edge AI (Image Credit: LatentAI)^[vii]

significant reductions in data ingestion time for commercial analytics. Building on this concept, manufacturers now integrate localized AI for process optimization, reducing data ingestion time for analytics, monitoring resource utilization, and adapting production schedules in real time.

Case Study: Stream Analyze's Edge AI Implementation in Manufacturing

In this case study, Stream Analyze demonstrated the integration of edge AI in manufacturing by improving quality assurance processes through real-time, on-site data analysis^[48].

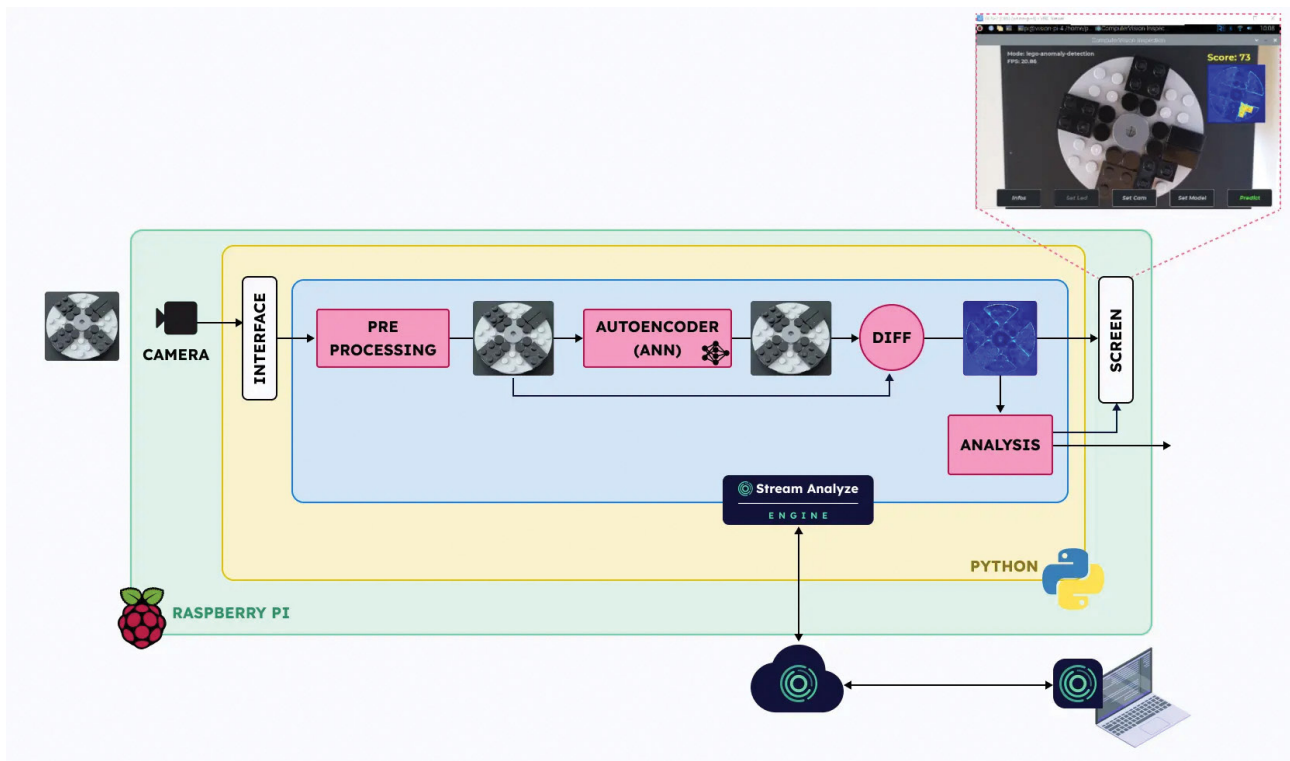
Faced with the need for faster, more reliable decision-making capabilities, an automotive manufacturer was seeking a robust solution to improve the accuracy and efficiency of its quality assurance processes. By implementing an edge AI solution, they managed to enhance product quality, boost operational efficiency, and maintain stringent security protocols.

Implementation Highlights

- **Real-time Data Processing:** By deploying AI models directly onto manufacturing lines using accessible hardware like the Raspberry Pi, Stream Analyze enables immediate production

data analysis, facilitating prompt decision-making and rapid response to quality issues.

- **Automated Quality Control:** The system automates the inspection process, identifying defects and inconsistencies without human intervention, thereby reducing the likelihood of errors and ensuring consistent product quality.
- **Data Security and Efficiency:** Processing data locally on the device ensures that sensitive information remains secure, eliminating the need for data transmission to external servers. This approach also reduces data



Stream Analyze Edge AI System (Image Credit: Stream Analyze)^[viii]

delay and storage costs, enhancing operational efficiency.

Key Improvements

- **Increased Inspection Speed:** The implementation resulted in a 100-fold improvement in inspection speed, significantly accelerating the production process and reducing bottlenecks.
- **Enhanced Data Security:** The system ensures complete data security by keeping data processing on-device, mitigating data breaches and unauthorized access risks.

- **Cost Savings:** Eliminating additional data storage requirements led to cost savings, as there was no need for investment in external data storage solutions.

Stream Analyze's solution emphasizes how integrated edge AI drives operational efficiency, accelerates production timelines, and protects sensitive data within factory walls.

The Power of Localized AI: Faster Decisions, Stronger Security, Smarter Operations

Unlike traditional AI deployments that require constant cloud connectivity, localized AI operates within defined environments, ensuring faster response times, enhanced security, and reduced operational costs. This shift is particularly critical in industries where latency, data privacy, and network reliability are major concerns. Custom localized AI models represent a strategic shift in edge computing

architecture, where AI systems are engineered explicitly for deployment within defined geographical or operational boundaries.

In financial services, for example, banks process millions of transactions per second, making fraud detection and risk assessment highly dependent on real-time AI inference. With the average cost of a financial data breach reaching an average of USD 5.17 million per incident^[49], institutions are integrating localized AI models to identify anomalies instantly, reducing exposure to cyber threats. Instead of transmitting sensitive financial data to external servers, AI models deployed within branch networks or ATMs detect suspicious activities in milliseconds, preventing fraud before transactions are finalized.

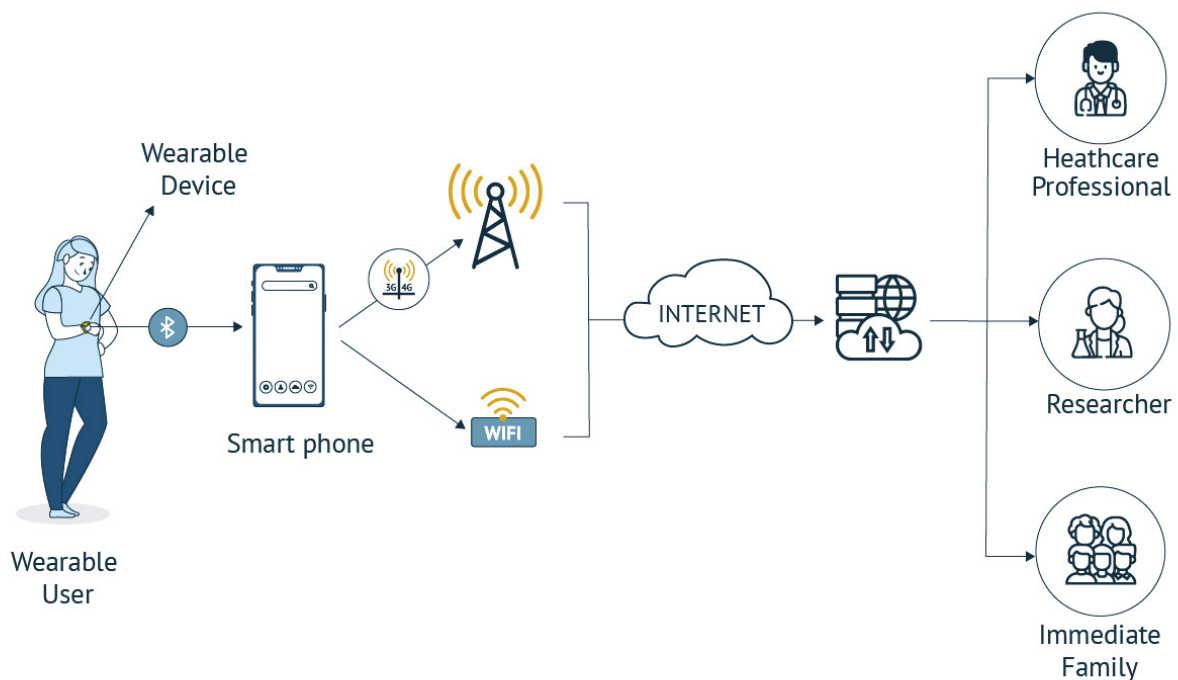
The impact of localized AI models can also be seen across many other industries, including automotive, manufacturing, and agriculture. Below, we explore how localized AI reshapes the healthcare and retail sectors with urgency and precision.

Healthcare and Diagnostics: From Reactive to Predictive and Personalized

In healthcare, localized AI is accelerating diagnostics and improving patient outcomes by processing medical data directly at the point of care. For example, edge AI-powered remote patient monitoring devices like portable ECG and blood pressure monitors can analyze heart rhythms and vital signs in real time. These devices, such as those developed by

Alive Cor^[50] and Biobeat^[51], enable clinicians to detect arrhythmias and other abnormalities without waiting for cloud-based analysis, cutting response times in critical situations.

Beyond diagnostics, localized AI enhances data security by encrypting patient information at the source, mitigating risks associated with cloud-based storage and transmission. Major hospitals are integrating AI-powered encryption within their infrastructure to maintain HIPAA compliance while ensuring uninterrupted access to critical health data^[52].



Edge AI-based wearable biosensors for patient monitoring
(Image Credit: Saifuzzaman, M. et al.)^[ix]

Digital Health at the Edge: A Vision for Remote Patient Monitoring

Healthcare is shifting to meet patients where they are, at home and in their daily lives. This change is stimulated by aging populations, a global shortage of healthcare professionals, and the growing demand for personalized care. Ambiq, a leader in ultra-low-power semiconductor solutions for edge AI, is leading this shift with innovative solutions in remote patient monitoring (RPM) and body-worn AI, setting a new benchmark for secure, energy-efficient, intelligent edge devices.

Tackling Sector-Wide Challenges with Body-Worn AI

RPM is changing how healthcare is delivered, helping patients stay safe at home and easing the load on overwhelmed healthcare systems. But its success hinges on solving two key challenges: making devices easy to use and addressing privacy concerns. Patients may forget to charge their devices or find bulky wearables inconvenient, while the thought of being monitored through intrusive video systems discourages adoption.

Ambiq's Apollo510 microcontroller addresses critical challenges in powering edge AI by enabling body-worn devices to operate seamlessly and securely. Leveraging advanced millimeter-wave radar technology, these devices can detect falls, monitor heart rates, and assess room activity without capturing sensitive visual information.

This non-invasive approach prioritizes user privacy while delivering meaningful and actionable health insights. Moreover, by processing data locally at the edge, the Apollo510 empowers devices to minimize dependence on cloud connectivity, thereby reducing latency, cutting costs, and enhancing overall system reliability.

Ambiq's focus on real-world usability sets its solutions apart. The Apollo510 powers smaller, lightweight devices that are comfortable to wear and require less frequent charging, making them practical for everyday use. By designing with patients in mind, Ambiq improves adoption rates and strengthens trust in the technology.

The Apollo510 MCU: A Game-Changer in Digital Health

At the heart of Ambiq's innovation, the Apollo510 is a microcontroller purpose-built to meet the demanding needs of digital health applications. Known for its groundbreaking "10x efficiency" advantage, it delivers up to 90% energy savings compared to similar technologies. This allows health monitoring devices to perform more complex tasks without sacrificing battery life, making it ideal for RPM devices and other wearable applications.

The Apollo510 combines power efficiency with robust performance. Its Arm® Cortex®-M55 CPU with Arm

Helium™ technology enables processing speeds up to 250 MHz, delivering up to 10 times better latency than its predecessor. Enhanced memory capabilities (4 MB of non-volatile memory and 3.75 MB of SRAM) facilitate real-time data processing and storage, ensuring devices can analyze complex datasets directly on the device. This combination of high performance and ultra-low power consumption enables manufacturers to develop smarter, longer-lasting healthcare solutions.

Ambiq's proprietary Subthreshold Power Optimized Technology (SPOT®) further ensures efficiency and reliability. By optimizing power consumption at the transistor level in real time, SPOT® enables the Apollo510 to deliver stable performance across various operating conditions, from wearables to embedded systems in cars or homes.

Privacy and Security by Design

In digital health, trust starts with keeping patient data private and secure. The Apollo510 integrates the secureSPOT® platform and Arm TrustZone® technology to provide a trusted execution environment, ensuring that sensitive health data remains protected. By processing information locally and transmitting only encrypted, minimal datasets, Ambiq significantly reduces attack surfaces while complying with stringent healthcare regulations.

This approach is evident in applications such as speech pattern analysis for early dementia detection. By analyzing data locally, the Apollo510 ensures patient confidentiality while delivering critical insights. This capability builds trust among users and healthcare providers, addressing privacy concerns that often hinder RPM adoption.

Enabling a Smarter Healthcare Ecosystem

Ambiq's vision extends beyond individual devices to a broader ecosystem of interconnected healthcare solutions. Its technologies pave the way for integration with smart homes and buildings, enabling seamless, secure health monitoring in everyday environments. While millimeter-wave radar can monitor room activity and detect irregular behaviors, Ambiq's emphasis on privacy and usability

ensures these solutions are designed for real-world adoption.

Applications like hearing aids further demonstrate Ambiq's transformative impact. The Apollo510 enables hearing aids to adapt to environmental contexts, filter background noise, and implement cutting-edge features like semantic hearing, intelligently isolating and enhancing specific voices in noisy settings. These innovations improve user experiences and set the stage for future advancements in healthcare technology.

Edge AI in Retail: Enhancing Operations, Personalization, and Security

Using localized processing, edge AI is transforming the retail sector by optimizing in-store operations and enhancing customer experiences through real-time behavioral analytics. AI-driven smart shelves and checkout systems process customer interactions locally, analyzing purchasing patterns and adjusting inventory forecasts without relying on cloud synchronization. Retailers are deploying AI-powered video analytics to detect anomalies in foot traffic, monitor stock levels, and reduce checkout times, leading to increased efficiency and reduced operational costs.

On the operational front, AI-based optimization is already showing promise going into 2025 with solutions like autonomous checkout systems. AI-powered computer vision now enables fully contactless transactions, reducing average checkout times by up to 30%^[53]. Retailers like Amazon Fresh use shelf-mounted cameras or trolley-mounted cameras to automatically bill customers as they exit or provide real-time spending previews.

In terms of customer experience enhancements, edge AI provides behavior-driven personalization by analyzing in-store movement patterns (via 3D LiDAR sensors) and purchase history to deliver hyper-targeted promotions. For example, a customer lingering in the skincare aisle receives instant mobile coupons for their preferred brands. Moreover, retailers are implementing AI-enabled dynamic

pricing displays, where digital shelf labels adjust prices in real time based on different factors like demand signals. About a third of e-commerce companies today are using dynamic pricing strategies, which has also been met with acknowledgment from consumers; 7 in 10 consumers are happy to see dynamic pricing as long as they perceive the pricing to be transparent and fair^[54].

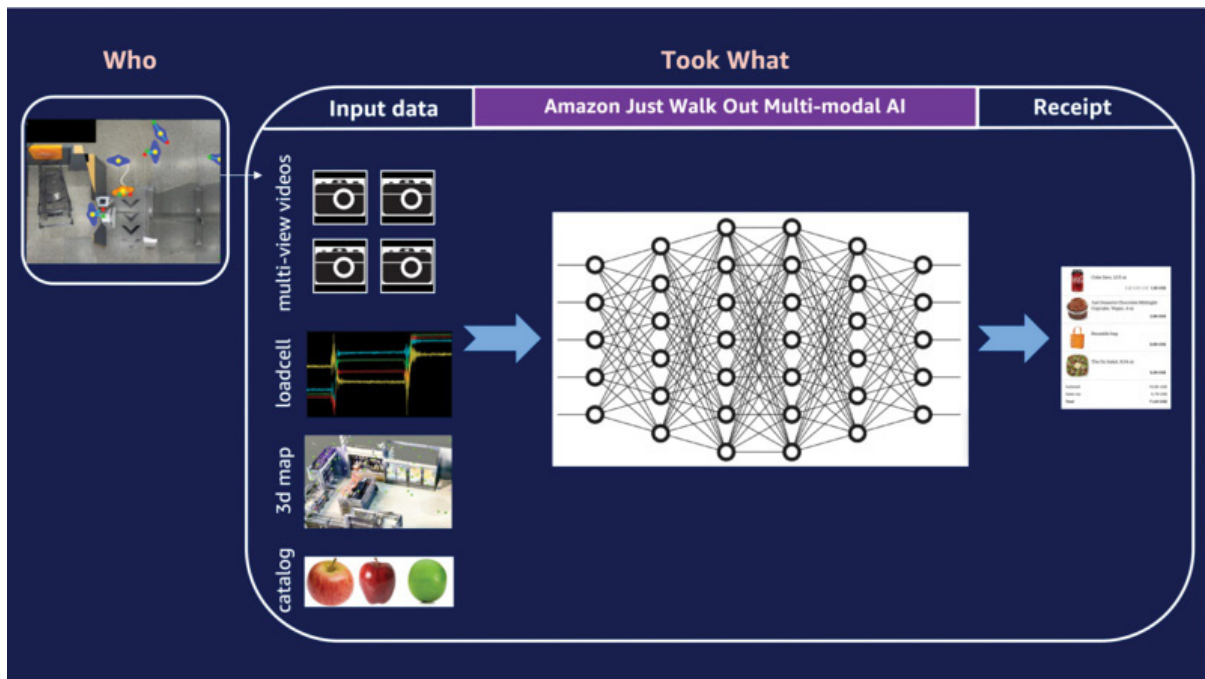
Retail theft and inventory shrinkage remain pressing challenges, costing retailers in the US, for example, over \$31 billion annually^[55]. Traditional surveillance systems rely on manual monitoring, making it difficult for security personnel to detect theft in real time across large retail spaces. Edge AI-powered video analytics transform loss prevention by enabling real-time threat detection without cloud dependence. AI algorithms analyze camera feeds locally, identifying suspicious behaviors such as fidgeting, unusual movement patterns, or repeat offender presence through facial recognition watchlists. Immediate alerts allow security teams to respond proactively, preventing losses while maintaining an open shopping environment. Additionally, AI-enabled drones and automated inventory tracking reduce human error in stock management, ensuring better control over shrinkage. By integrating edge AI into retail security strategies, businesses can minimize theft-related losses while streamlining security operations with greater efficiency and accuracy.

Case Study: Amazon Go's Edge AI Implementation

Amazon's Just Walk Out (JWO) system is a prime example of retail-focused edge AI, integrating sensor arrays, on-device analytics, and an advanced ML model^[56]. It demonstrates the integration of Edge AI to create a seamless, cashier-less shopping experience at scale. All computations are processed locally on custom edge hardware, enabling real-time decision-making by processing data locally and enhancing customer convenience and operational efficiency.

Implementation Highlights

- **Just Walk Out Technology:** Amazon Go utilizes a combination of computer vision, sensor fusion, and deep learning algorithms to monitor customer interactions with products. This system tracks when items are taken from or returned to shelves, allowing customers to simply pick up products and leave the store without traditional checkout processes.
- **Edge Computing Infrastructure:** The technology relies on edge computing to process real-time data from in-store sensors and cameras. By performing computations locally, the system reduces latency and ensures immediate updates to virtual shopping carts, providing a frictionless shopping experience.
- **Advanced AI Models:** Amazon has enhanced the system's accuracy by implementing transformer-based machine learning models. These



Amazon's JWO Receipt Generation using Edge AI
(Image credit: Amazon)^[x]

models analyze data from various sensors simultaneously, improving the system's ability to handle complex shopping scenarios and making it more adaptable to different store layouts.

with fewer staff dedicated to cashier duties, allowing employees to focus on other areas such as customer service and store maintenance.

Key Improvements

- **Enhanced Customer Experience:** The cashierless model eliminates checkout lines, reducing wait times and streamlining the shopping process. This convenience has been well-received by customers, contributing to increased satisfaction and repeat visits.
- **Operational Efficiency:** By automating the checkout process, Amazon Go stores can operate

- **Scalability:** Using edge computing and advanced AI models enables Amazon to deploy this technology across various store formats and locations, demonstrating its scalability and adaptability in retail.

Amazon Go's implementation of Edge AI showcases how localized data processing can revolutionize retail operations, offering valuable insights for industry leaders and technology developers aiming to enhance customer experiences and operational efficiencies through advanced technologies.

Enhancing Security and Safety with Edge AI Efficiency

Edge AI applications demand high accuracy and real-time performance while operating within strict computational constraints. Axelera AI's Metis AI Processing Unit addresses these challenges by enabling multi-model inference workflows that optimize resource utilization without sacrificing precision. A prime example of this capability lies in security applications where object detection is critical; high-security environments, such as airports, require image classification and object detection, but traditional methods often struggle to balance resolution and processing efficiency. This is where Metis AI Processing Unit (AIPU) can be a game-changing solution.

The Challenge: Detecting Small or Concealed Threats

In security screening, 4K cameras capture high-resolution footage to identify potential threats. However, most AI systems downscale input frames to 640x640 resolution to reduce computational overhead. While this allows real-time processing, critical details such as partially obscured objects (or even small weapons) could be lost. For instance, a compact firearm in a crowded baggage scan or a concealed blade on a person might go undetected at lower resolutions. Increasing resolution across the entire frame would require prohibitively large models or excessive compute power, making edge deployment impractical.

Metis AIPU's Multi-Model Approach: Precision Without Compromise

Axelera AI's Metis AIPU solves this problem through a layered inference strategy of computer vision workloads at the edge. In an object detection scenario, the system leverages three concurrent models running on a single Metis chip:

- **Base Object Detection (YOLOv8):** Performs object detection on the image to identify additional objects of interest (e.g., luggage). This model runs in parallel to the cascaded models doing the pose estimation and segmentation.
- **Pose Estimation (YoloV8-pose):** Detects human figures and tracks body posture to identify regions of interest (ROIs) in the high-resolution image.
- **High-Resolution Segmentation (YoloV8-seg):** Runs as a cascaded model, receives the ROIs from pose estimation to perform instance segmentation at the original 4K resolution to detect objects of specific classes (e.g., weapons) associated with human subjects.

By dynamically focusing computational resources on high-risk zones, the system maintains the efficiency of low-resolution processing while retaining the granularity

needed to detect small or obscured threats. The Metis AIPU's architecture—featuring four independent AI cores, each delivering at least 52 TOPS—enables parallel execution of these models. Two cores handle pose estimation, one runs segmentation, and the fourth manages base detection, ensuring real-time performance without bottlenecks.

Scalable Applications Beyond Security

The same methodology applies to other application areas such as personal protective equipment (PPE) compliance monitoring in industrial or laboratory settings. Here, the system could:

- Use pose estimation to identify workers in a frame.
- Apply high-resolution ROI crops to inspect gloves, goggles, or helmets at pixel-level detail.
- Flag non-compliance in real time, even in crowded or dynamic environments.

Traditional systems relying solely on low-resolution object detection might miss improperly worn PPE (e.g., a face mask worn below the nose). By combining pose data with localized high-resolution analysis, the Metis AIPU ensures precise verification without requiring oversized models or external cloud processing.

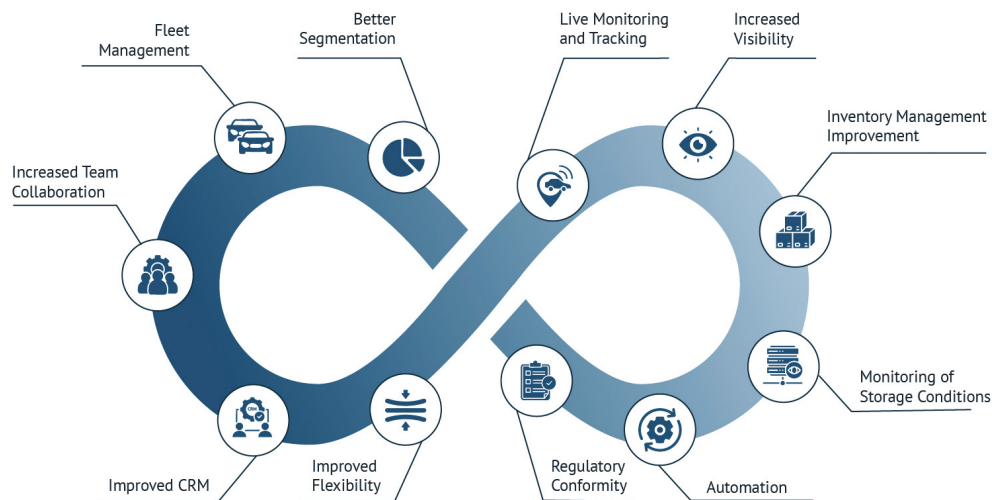
Why This Matters for Edge AI Engineers

Axelera AI's approach demonstrates how flexible hardware design can overcome edge computing's inherent limitations. The Metis AIPU eliminates the need to choose between resolution and efficiency, enabling engineers to:

1. **Maximize existing camera hardware without sacrificing performance:** Leverage high-definition 4K video feeds without down sampling
2. **Reduce latency:** Parallel model execution avoids sequential processing delays.
3. **Lower deployment costs:** A single chip replaces multiple devices, simplifying system integration.

For engineers designing edge AI solutions, this capability is particularly valuable in scenarios where false negatives carry high risks, such as security breaches or workplace safety violations. The Metis AIPU's ability to run diverse models concurrently also future-proofs deployments, allowing systems to adapt to evolving threats or regulatory requirements through software updates rather than hardware overhauls.

Axelera AI's Metis AIPU redefines edge AI efficiency by enabling intelligent resource allocation. By combining coarse-grained detection with targeted high-resolution analysis, the platform addresses a key pain point in object detection: preserving accuracy in complex, real-world environments. Whether applied to airport security, industrial safety, or similar high-stakes use cases, the Metis AIPU provides a scalable, cost-effective framework for engineers to deploy robust AI solutions at the edge.



Benefits of edge AI-powered IoT in scaling supply chain management (Image Credit: appinventiv)^[x1]

Scalability and Flexibility: Edge AI's Adaptive Framework

As industries evolve, AI systems must scale efficiently, handling more devices, processing greater data volumes, and expanding across new locations without compromising performance. Edge AI provides this adaptability by distributing intelligence across networks, ensuring seamless expansion while maintaining real-time insights. Whether managing thousands of IoT devices in supply chains or optimizing vast agricultural landscapes, Edge AI enables scalable, high-performance computing where it's needed most.

Scaling Intelligence Across Logistics Networks Through IoT

Modern supply chains depend on real-time visibility and rapid decision-making to mitigate disruptions and optimize resource allocation. Edge AI,

integrated with IoT sensors, enhances these capabilities by processing logistics data directly at distribution centers, warehouses, and transport hubs. Instead of transmitting vast amounts of information to centralized servers, smart sensors analyze temperature fluctuations, motion anomalies, and stock shortages on-site, triggering instant alerts when deviations occur. This localized intelligence helps logistics teams prevent costly bottlenecks, reduce downtime, and improve asset utilization.

Edge AI enables organizations to dynamically deploy and scale AI capabilities, from individual edge devices to enterprise-wide systems, while maintaining consistent performance and security standards. Edge AI architecture offers flexibility through horizontal scaling (adding more devices), vertical scaling (enhancing processing power), and geographical scaling (distributed deployments). This multi-dimensional approach allows organizations to adapt their edge AI implementations based on specific use

cases, regulatory requirements, and performance demands. It also supports standalone edge deployments and hybrid models that integrate with existing cloud infrastructure, providing a foundation for sustainable digital transformation initiatives.

For example, organizations like P&O Ferrymasters have optimized cargo capacity by 10% using AI-powered vessel loading procedures^[57], maintaining real-time visibility across their supply chain. Additionally, AI-driven forecasting has contributed to a 20% reduction in conversion expenses, with improved labor productivity accounting for 70% of these savings^[58]. By deploying multi-node edge AI frameworks, companies can expand their operations seamlessly, rolling out AI-driven logistics solutions across distribution centers without overburdening cloud networks. The ability to scale horizontally and vertically makes Edge AI an indispensable tool for supply chain resilience.

Edge AI in 2025: Scalability, Efficiency, and Real-World Impact

The evolution of edge AI stems from two parallel advancements: rising compute density at the edge and model miniaturization. Over the next few years, these trends will enable complex intelligent systems on resource-constrained devices. From advanced computer vision on microcontrollers (MCUs) to generative AI applications, edge-native models will increasingly operate in tandem, solving multifaceted problems without relying on cloud infrastructure. These systems will be further refined by advancements in data efficiency: automated labeling, active learning, and semi-supervised techniques powered by foundation models and generative AI will reduce reliance on manually annotated datasets. Synthetic data generation and AI-driven augmentation will enable robust training across edge environments, even with sparse real-world data. The future of edge AI lies in interconnected networks of specialized algorithms optimized for specific tasks. Enabled by in-context (zero-shot) learning, these on-device models can be reconfigured to adapt to new scenarios without retraining, even on devices with limited power and memory.

Edge Impulse offers access to the latest edge AI developments as they become available. Its end-to-end platform streamlines data collection, model training, and deployment across hardware from MCUs to powerful processors, empowering developers to build and deploy efficient AI models that run seamlessly on the edge.

Addressing Demand in Remote and Low-Power Applications

Industries such as energy, agriculture, and environmental monitoring require edge AI solutions that prioritize bandwidth, latency, energy efficiency, reliability, and privacy (abbreviated “BLERP”). Edge Impulse addresses these needs by enabling domain experts to deploy tailored solutions without requiring deep AI expertise.

- **Bandwidth:** Transmitting raw sensor data (e.g., high-resolution images from wildfire detection cameras) consumes significant network resources. Edge AI processes this data locally, sending only actionable insights, such as “fire detected at coordinates X, Y,” reducing bandwidth demands by orders of magnitude.
- **Latency:** Applications like road condition monitoring require real-time responses. Edge inference eliminates cloud round-trips, enabling immediate alerts (e.g., ice detection) without the 100+ millisecond delays inherent in cloud-based systems.
- **Energy efficiency:** Battery-powered devices like livestock health monitors rely on ultra-low-power operation. Edge AI minimizes energy consumption by avoiding resource-heavy cloud communication and optimizing models for local compute (e.g., quantized models on MCUs).

- **Reliability:** In remote environments with unstable connectivity, such as deforestation detection sensors in rainforests, edge devices operate autonomously. Data is processed on-device, ensuring functionality even during network outages.
- **Privacy:** Sensitive data, such as biometrics from wearable health monitors, never leaves the device. This mitigates the risks of interception or misuse, a critical requirement for consumer and industrial applications alike.

By processing data locally, edge AI reduces bandwidth usage, cuts latency, and extends device lifetimes through energy-efficient inference. Reliability and privacy are inherent: devices function offline, and sensitive data remains on-premises.

Optimizing Energy Efficiency in Wearables

While edge AI's energy efficiency is vital for remote applications, its importance extends to domains like wearable devices, such as smart rings, health monitors, and fitness trackers, where compact form factors and extended battery life define usability. Edge Impulse addresses these challenges through a holistic approach to model design that prioritizes efficiency at every stage:

1. **Input reduction:** Selecting optimal signal processing to shrink input data size.
2. **Model tuning:** Pruning and quantizing architectures for peak performance per watt.
3. **Compiler optimization:** Generating lean, hardware-specific code.

Furthermore, Edge Impulse integrates efficiency considerations directly into the design workflow. Developers receive real-time estimates of on-device performance, including latency, memory usage, and power draw, which enables iterative refinement before deployment. This proactive optimization is essential for battery-dependent wearables and industrial IoT sensors, where thermal limits and user experience hinge on balancing performance with power consumption.

Collaborations and Hardware-Agnostic Deployment

Edge Impulse's emphasis on efficiency and hardware flexibility is validated through partnerships that solve tangible, real-world challenges. By collaborating with domain experts across industries, the platform demonstrates how edge AI can adapt to diverse requirements. Some notable examples:

- **Healthcare:** Hyfe's cough detection algorithms analyze audio patterns to identify respiratory illnesses, enabling early diagnosis without compromising user privacy^[59].
- **Automotive:** Globalsense monitors sound data for automotive diagnostics and real-time crash detection, improving emergency response times in connected vehicles^[60].
- **Consumer electronics:** A major manufacturer deployed voice-controlled earbuds with sub-20ms latency, balancing accuracy and power efficiency for seamless user interaction^[61].
- **Aging-in-place:** A US-based company leverages motion sensors for fall detection in senior care, detecting incidents directly onboard to ensure rapid alerts^[27].

These collaborations are a result of Edge Impulse's hardware-agnostic philosophy. Developers can prototype models, simulate performance across devices, and deploy to production-grade hardware without vendor lock-in. This reduces time-to-market and ensures scalability. Integration with tools like NVIDIA Omniverse further streamlines workflows, enabling synthetic data generation to augment training datasets.

Edge AI's true potential lies in its ability to solve niche, impactful problems, which require more than raw computational power. They demand systems that respect the constraints of the physical world, like limited bandwidth, scarce energy, and the need for privacy. Edge Impulse equips engineers with advanced tools to optimize models holistically, collaborate across industries, and deploy flexibly across hardware. The result is smarter devices and systems that empower industries to innovate responsibly and scalably.

Smart Agriculture: Scaling Precision Farming for Global Food Demands

With the global population projected to reach 9.8 billion by 2050^[13], agriculture must scale intelligently to meet rising food demands while minimizing environmental impact. Edge AI allows farms to expand their technological footprint without increasing complexity, analyzing soil conditions, monitoring weather patterns, and automating irrigation systems in real time.

Rather than sending data to a distant server, advanced sensors and AI models evaluate factors like soil moisture or pest activity the moment they're detected, allowing swift intervention. Projects such as CrackSense exemplify

how real-time sensing can ensure fruit quality for crops like citrus, pomegranate, and table grapes, ultimately reducing spoilage and waste^[62]. This way, edge AI equips farmers with the knowledge to act on the spot through real-time monitoring and detection.

In terms of resource optimization, smart irrigation systems equipped with edge AI have demonstrated great effectiveness, dynamically adjusting water distribution based on localized soil moisture analysis and cutting water usage by 25%. Similarly, AI-powered pest detection can reduce pesticide application by up to 30%, ensuring precision farming with minimal waste^[62].

Autonomous farming is also scaling with edge AI, allowing agricultural

operations to deploy fleets of AI-enabled drones and robotic systems for real-time monitoring and automated harvesting. Solutions like NVIDIA's Jetson Orin Series facilitate on-site data processing, enabling predictive analytics for optimal planting and harvesting schedules. By embedding intelligence directly into farming equipment, edge AI transforms unpredictable natural conditions into manageable, data-driven decision-making factors.

Across industries, from logistics to agriculture, edge AI ensures scalability without sacrificing efficiency. By processing data at the source, businesses can seamlessly expand their AI capabilities, unlocking new levels of automation, cost savings, and operational resilience.



AI-generated image representing real-time crop monitoring and smart agriculture
(Image Credit: J. Renaz)^[xiii]

Chapter III:

The Technological Enablers of Edge AI

The deployment and operation of AI systems and models at the edge come with many benefits for industrial organizations, yet they still pose a host of challenges. For instance, challenges posed by the limited processing power of edge devices, compared to conventional centralized systems, still need to be addressed. Edge deployments also limit the data that are centrally aggregated, which results in the lack of adequate data points for certain applications. Also, there are still issues related to scaling edge AI solutions across diverse environments and challenges to ensuring seamless interoperability between heterogeneous devices.

To address these limitations, there is significant traction around edge AI technology, which leads to a continuous improvement of the main technological enablers of the different edge AI paradigms. These enablers

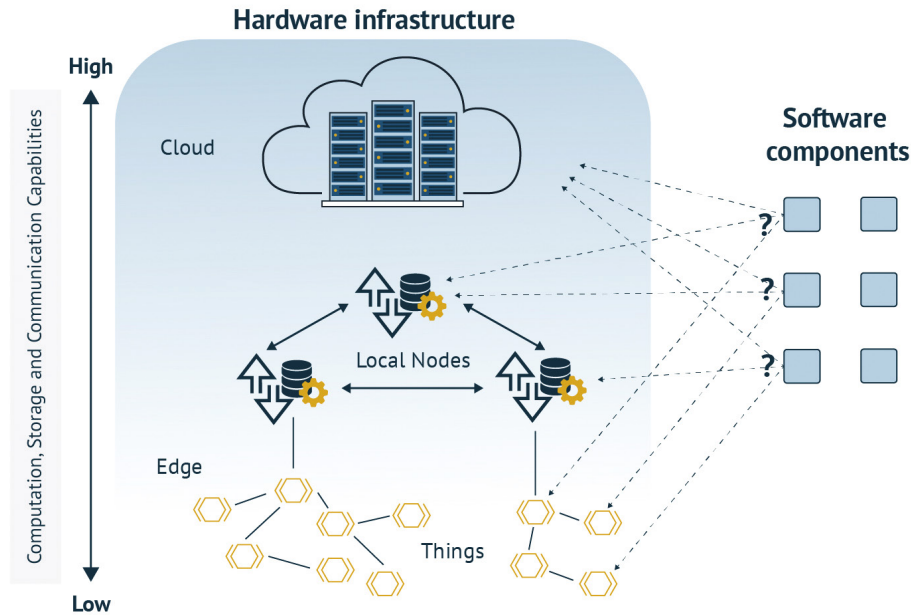
are advanced in directions that address the scaling, interoperability, and distribution challenges of edge AI deployments. Likewise, the development of these enablers spans many different aspects of AI systems, including hardware and software infrastructures, as well as novel AI models and paradigms.

Hybrid Edge-Cloud AI: Optimized Intelligence and Resource Management

The emergence of edge AI has been driven by the proclaimed limitations of traditional cloud-centric AI systems, such as latency, privacy concerns, and bandwidth constraints. Edge AI addresses these issues based on data processing locally close to the source of data, such as within devices at the

network's edge. This approach reduces latency through immediate data processing and decision-making while enhancing privacy as sensitive data are not shared to the cloud.

Furthermore, due to the reduced data transfers, edge AI decreases the attack surface and the bandwidth usage of AI applications at the edge. Nevertheless, there are still several AI-based use cases where cloud AI is needed. This is, for example, the case with applications that require many data points and are compute-intensive, such as training and using large language models with tens of billions of parameters. To address such use cases while retaining the benefits of edge AI for real-time applications with sensitive data, the preferred AI deployment model combines cloud and edge computing infrastructures.



Hybrid Edge-Cloud reference architecture for IoT systems
(Image credit: Wevolver, adapted from: M. Ashouri et al.)^[xiii]

The integration of edge and cloud computing has created a hybrid model that leverages the strengths of both approaches. Edge AI provides real-time processing capabilities, while cloud computing offers the computational power necessary for training complex models and handling very large-scale data analytics. This synergy allows businesses to optimize resource use: edge devices handle immediate, latency-sensitive tasks while the cloud manages more intensive computations and long-term data storage. Also, hybrid models enable continuous learning and model updates. Data processed at the edge can be aggregated in the cloud to refine AI models, which are then redeployed to edge devices for improved performance. This combination enhances scalability, flexibility, and efficiency for a wide range of use cases in different sectors.

During the early days of hybrid cloud/edge AI deployments, companies had to statically define the most appropriate placement of AI models, functions, and workloads considering the compute, energy efficiency, security, privacy, and latency requirements of their use cases. This placement was destined to optimally resolve performance and security trade-offs in the scope of heterogeneous cloud/edge infrastructure. Nowadays, these edge AI and cloud AI infrastructures come with intelligent resource management and AI services orchestration functions that optimize the placement of AI workloads between cloud and edge, considering parameters, application profiles, and use case requirements.

The state of the art (SOTA) in cloud/edge resource management for AI applications involves cross-layer orchestration of AI workflows, as well

as employment of edge functions based on stateless cloud/edge deployment paradigms like Function as a Service (FaaS). Emerging resource management approaches are also employing machine learning towards provisioning and placing AI workloads in dynamic and intelligent ways. The latter will be destined to deliver tangible benefits to both infrastructure providers (e.g., cloud providers) and operators/deployers of AI applications. Yet, to fully realize the benefits of edge computing, dedicated AI hardware is evolving to provide high-performance computing in compact, power-efficient form factors.

The Next Generation of Specialized Edge Hardware

In the years to come, edge AI will enable almost all organizations to access their unique layer of intelligence by leveraging their very own data. This will be empowered by the evolution of edge hardware, which will enable very scalable applications. It is, therefore, no accident that during 2024, an AI hardware enterprise, namely NVIDIA, was the company with the fastest growing market capitalization in the New York Stock Exchange (NYSE).

Specialized hardware will accelerate edge AI by providing the necessary computational power in a compact

form. As a prominent example, NVIDIA Jetson modules are delivering high-performance capabilities in edge applications like computer vision^[63]. To this end, they integrate Central Processing Units (CPUs), Graphics Processing Units (GPUs), memory, and interfaces into small form factors, which makes them ideal for deploying complex deep learning models on edge devices. As another example, Qualcomm's processors (e.g., chips used in Wi-Fi hubs) incorporate powerful Neural Processing Units (NPUs) that enable efficient on-device AI processing^[64].

Energy efficiency is another concern when it comes to edge AI deployments. The latter must be power efficient, which drives the development of Ultra-Low-Power Hardware. For

instance, Ambiq's Subthreshold Power Optimized Technology (SPOT) platform exemplifies ultra-low-power hardware designed for edge applications^[65]. SPOT enables devices to operate at significantly reduced voltage levels, which is key for enhancing battery life without sacrificing performance. This technology is, for example, important for digital health devices that require continuous operation without frequent recharging.

While specialized hardware enables efficient computation at the edge, its full potential is only realized when paired with edge-native algorithms optimized for real-time inference, minimal data dependencies, and energy efficiency.



Ambiq's Apollo510 EVB
(Image Credit: Ambiq)^[xiv]

Scalable Edge NPU IP for SoC integration, from Embedded ML and Computer Vision up to Generative AI

The market for edge AI chips in multiple applications is rapidly expanding, driven by the increasing demand for power-efficient, low-latency AI processing directly on devices. Licensable NPU IP (Intellectual Property) is a crucial enabling technology for edge AI chip designers targeting consumer devices, industrial automation, and vehicle safety. Ceva is leading the way by developing scalable NPU IPs that accelerate the deployment of Smart Edge chips and devices.

Ceva-NeuPro-Nano: Highly Efficient, Self-Sufficient Edge NPU for Embedded ML Applications

With over 4 billion inference chips for Embedded ML (TinyML) devices forecasted to ship annually by 2029, this Edge NPU IP is the smallest of Ceva's NeuPro NPU product family^[66]. It delivers the optimal balance of ultra-low power and high performance in a small area to efficiently execute Embedded ML workloads across AIoT product categories, including hearables, wearables, home audio, smart home, smart factory, and more. Ranging from 10 GOPS up to 400 GOPS per core, Ceva-NeuPro-Nano enables energy-efficient, always-on audio, voice, vision, and sensing use cases in battery-operated devices across a wide array of end markets.

Ceva-NeuPro-Nano is a standalone, fully programmable NPU, not an AI accelerator, and therefore does not require a host CPU/DSP to operate. The IP core includes all the processing elements of a standalone NPU, including code execution and memory management. Its architecture is fully programmable and efficiently executes neural networks, feature extraction, control code, and DSP code. It also supports the most advanced machine-learning data types and operators, including native transformer computation, sparsity acceleration, and fast quantization, delivering a highly optimized solution with excellent performance.

Ceva-NeuPro-M: Scalable NPU Architecture for Transformers and Generative AI Applications

Ceva-NeuPro-M is a scalable NPU architecture with exceptional power efficiency of up to 3500 Tokens per Second/Watt for Llama 2 and 3.2 models^[67]. With 30% of generative AI inference predicted to be on-device in the next 2 years, the Ceva-NeuPro-M NPU IP delivers exceptional energy efficiency tailored for edge computing while offering scalable performance to handle AI models with over a billion parameters. Its award-winning architecture introduces significant advancements in power efficiency and area optimization, enabling it to support

massive machine-learning networks, advanced language and vision models, and multimodal generative AI.

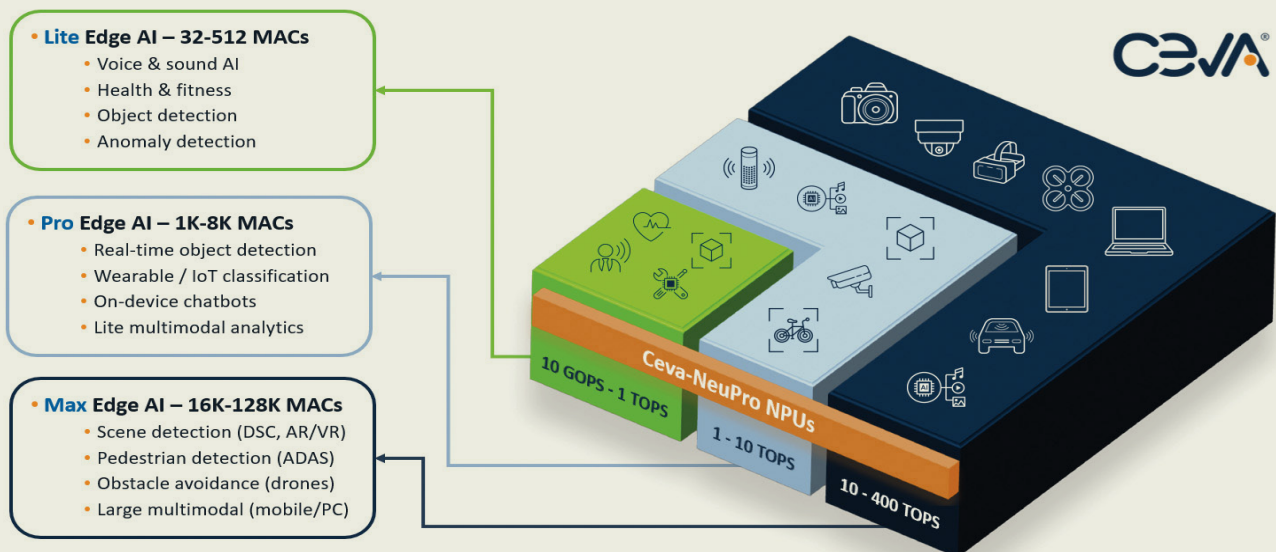
Even mid-range AI workloads, such as computer vision (object detection and classification), speech recognition, and small-scale NLP (keyword spotting), are becoming dominated by the use of transformers (e.g., ViT, BERT). Transformer support in edge NPUs is becoming mandatory for local text generation, context-aware AI assistants, and multimodal models for AR/VR, robotics, and advanced user-interface applications.

With a processing range of 400 GOPS to 200 TOPs per core, leading area efficiency, advanced transformer support, sparsity, and compression, the Ceva-NeuPro-M optimizes key AI models seamlessly. Thanks to its highly scalable design, it provides an ideal IP solution for embedding high-performance AI processing in SoCs across a wide range of edge AI applications.

AI SDK for Ceva-NeuPro NPUs

The Ceva-NeuPro Studio is a robust tool suite that complements the Ceva-NeuPro NPUs by streamlining the development and deployment of AI models. It includes tools for network optimization, graph compilation, simulation, and emulation, ensuring that developers can train, import, optimize, and deploy AI models with the highest efficiency and precision.

There are limitless possibilities to build Edge AI chips with diverse AI capabilities, from Embedded ML in consumer and industrial IoT to multimodal and edge generative AI in personal computing and automotive. Learn more about how Ceva's licensable NeuPro NPUs and wireless connectivity IPs are helping to build chips that power Smart Edge devices - www.ceva-ip.com



The Ceva-NeuPro NPU family
(Image Credit: Ceva)^[xv]

Edge-Native Models and Algorithms

Since the early days of edge AI, emphasis has been put on shrinking conventional machine learning models to fit and deploy them in edge devices. In recent years, there has been a surge of interest in edge-native AI algorithms, enabling real-time inference in the scope of applications like computer vision and speech recognition.

Edge-native algorithms are tailored for real-time inference on resource-constrained devices. They are optimized to balance accuracy with computational efficiency for edge. In this direction, they employ techniques such as model quantization and pruning, which reduce the size of AI models without any essential drop in AI performance. Edge-native algorithms are suitable for deployment on edge devices with limited resources (e.g., CPU and memory resources).

Some of the edge-native algorithms are also classified as “data-efficient” techniques, as they can perform well with smaller datasets, which makes them particularly useful in scenarios where data is limited or costly to

obtain. Data-efficient algorithms maintain high-performance levels without the need for large volumes of data that are typically used in traditional machine-learning methods.

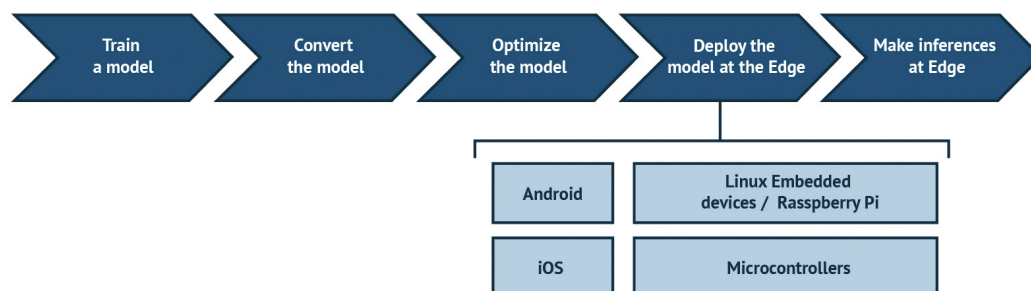
To support the scalable and resource-efficient deployments of edge native algorithms, technologies like Docker and Kubernetes are being adapted for edge deployments. These technologies simplify application management across diverse devices, which boosts scalability and resource optimization. Furthermore, there is a rise in DevEdgeOps techniques, i.e., DevOps practices adapted for Edge Environments. Specifically, such DevEdgeOps practices are adapted specifically for edge computing environments to address unique challenges like connectivity issues and diverse hardware requirements. This ensures efficient deployment and maintenance of edge-native applications.

Nowadays, there are also edge AI frameworks that facilitate the development and deployment of edge native AI models and algorithms. For example, tools like TensorFlow Lite and OpenVINO^[68] make it easier to deploy AI models optimized for edge

environments. Moreover, Edge AI supports hyper-personalized services, such as dynamic traffic management or smart retail experiences, based on the processing of data locally and the development of instant responses.

There are also opportunities for improving edge native algorithms based on their integration with state-of-the-art edge networking infrastructures like 5G networks. Specifically, the synergy between 5G and edge computing enhances the performance of edge algorithms by providing ultra-low latency and high-speed data transmission. This class of 5G-enabled, enhanced algorithms will play a considerable role in applications like autonomous vehicles, remote surgeries, and various immersive augmented reality applications.

Edge-native models will be increasingly deployed on a smaller scale in miniaturized devices, giving rise to the micro-edge and thin-edge AI paradigm. The latter refers to deploying lightweight AI models on minimal hardware resources, which is essential for extending AI capabilities to smaller devices like sensors or wearables.



TensorFlow Lite helps deploy AI models at the edge
(Image Credit: Wevolver, adapted from: [SeedStudio](#))^[xvii]

Moving LLMs and Generative AI to the Edge

For over two years following the emergence of OpenAI's ChatGPT, Large Language Models (LLMs) and Generative AI (GenAI) have been considered among the most promising developments of the AI community, especially with the recent surge of competing models like DeepSeek R1, Anthropic's Claude 3.5, Google's Gemini 1.5 and 2.0, and many more. Yet, as detailed in Wevolver's recent report titled "Edge AI Technology Report: Generative AI Edition," GenAI is increasingly finding its way away from cloud servers toward the edge^[69]. The local execution of generative models enables devices to provide personalized experiences without relying heavily on cloud resources. This local processing reduces latency and enhances privacy, which are among the key considerations in environments with intermittent connectivity or stringent security requirements.

Companies like Qualcomm and Arm are leading efforts to make GenAI models smaller and more efficient, in

order to make them usable in real-time applications like autonomous vehicles, smart homes, and industrial Internet of Things (IoT) applications.

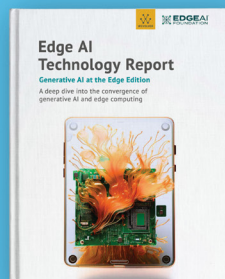
In the coming years, the integration of GenAI into edge devices will become more ubiquitous across consumer electronics and industrial systems. Furthermore, advances in hardware (e.g., specialized processors) and software frameworks (e.g., ExecuTorch^[70]) will continue to drive further improvements in AI model performance and accuracy. Most importantly, GenAI at the edge will enable autonomous AI agents that will be capable of solving problems, handling complex tasks, and collaborating with other agents. These agents will mimic human digital workers at the edge, which will be able to complete complex tasks close to the field in near real time.

Speaking of mimicking humans, a different paradigm is emerging: neuromorphic computing. Inspired by the brain's event-driven processing, neuromorphic chips aim to unlock ultra-low-power AI capabilities, addressing efficiency concerns that traditional architectures struggle with.

The Role of Neuromorphic Chips

Neuromorphic chips represent an emerging technology that is designed to mimic the human brain's neural architecture. These chips are inherently efficient at processing sensory data in real time due to their event-driven nature. Therefore, they hold promise to advance edge AI based on a new wave of low-power solutions that will be handling complex tasks like pattern recognition or anomaly detection.

In the next few years, neuromorphic chips will become embedded in smartphones, enabling real-time AI capabilities without relying on the cloud. This will allow tasks like speech recognition, image processing, and adaptive learning to be performed locally on these devices with minimal power consumption. Companies like Intel and IBM are advancing neuromorphic chip designs (e.g., Loihi 2^[71] and TrueNorth^[72], respectively) that consume 15–300 times less energy than traditional chips while at the same time delivering exceptional performance.



Have you seen our Generative AI at the Edge Report?

An In-Depth Guide for Engineers to See How Edge AI is Redefining Real-Time Processing, Personalization, and Security

Download from Wevolver now!



Emerging technologies like memristors and 3D architectures will improve the scalability and efficiency of neuromorphic chips. Memristors emulate synaptic behavior for more brain-like processing, while 3D integration reduces latency and enhances computational density. Furthermore, event-driven processing models (e.g., spiking neural networks) will be used to further optimize energy efficiency by mimicking the asynchronous nature of biological neurons. As AI models evolve toward brain-inspired architectures, interpretability becomes a key consideration. Ensuring that decisions made by both conventional and neuromorphic AI systems remain transparent and understandable is critical, especially in high-risk domains.

Explainability in Edge AI: Building Trust and Transparency

Edge AI models' simple and domain-specific nature renders them more interpretable than large cloud-based models. However, explainability remains a key requirement for regulatory compliance, trust, and real-world deployment, especially in industries like healthcare, finance, and industrial automation. In a recent book on "Advancing Edge Artificial Intelligence," the authors emphasized that edge AI must provide transparent, comprehensible decisions to gain adoption in safety-critical applications, where end-users require the reasoning behind a certain prediction^[73].

Explainable AI (xAI) at the edge refers to the ability of edge AI models to provide transparent, interpretable, and

justifiable decisions while operating under resource constraints. It ensures that AI-driven predictions can be understood, audited, and trusted by users, regulators, and stakeholders, particularly in high-risk applications.

Unlike cloud-based AI, which can rely on compute-intensive explainability methods like shapley additive explanations (SHAP) or local interpretable model-agnostic explanations (LIME), explainability in edge AI must balance interpretability with real-time performance. Hardware constraints, processing power, and latency requirements can impact the feasibility of such explainability methods at the edge. Lightweight variants of such techniques, including saliency maps (such as Grad-CAM), precomputed feature attribution (i.e., SHAP or LIME sent from cloud to edge), and context-aware explanations (i.e., rule-based, lightweight interpretable models) can help make decisions more transparent without compromising efficiency.

For example, the explainability technique Grad-CAM helped manufacturing engineers verify defect detection models by ensuring the AI focuses on actual product flaws rather than irrelevant background features. In healthcare, it assisted in medical imaging by confirming that models focus on relevant areas, such as lung regions in pneumonia detection, thereby enhancing trust among medical professionals^[74].

On the regulatory front, edge AI must meet regulatory requirements, particularly in healthcare, finance, and autonomous systems, where legal frameworks like the EU AI Act and GDPR mandate transparency. A 2025

paper on responsible AI highlights explainability as critical for high-stakes applications such as surgical planning and risk assessment, ensuring that AI-generated decisions remain auditable and justifiable^[75]. Similarly, Trustful AI underscores that traceability is just as vital as accuracy, enabling regulators and stakeholders to scrutinize AI outcomes^[76].

Beyond compliance, explainability is essential for bias detection and debugging. A study on edge security cameras revealed

Dataset biases in edge devices like security cameras can cause models to misidentify objects, such as individuals in wheelchairs, as these cameras were trained mostly on standing figures. Researchers used xAI techniques like D-RISE, thanks to its adaptability to diverse models, to identify feature dependencies, leading to targeted dataset augmentation and improved fairness^[77]. They also demonstrated how feature attribution allowed engineers to pinpoint which sensor readings influenced predictive maintenance models, making AI-driven insights more actionable^[78].

Autonomous vehicles and industrial robots are other applications for edge explainable AI (xEdgeAI), requiring transparency and explanations for their actions, particularly in failure scenarios where human oversight is necessary. However, achieving this without compromising performance remains a challenge. Emerging concept-based explanations and real-time saliency maps are improving interpretability while maintaining efficiency^[73]. As edge AI adoption grows in 2025, explainability will become an operational necessity. Organizations

must implement explainability frameworks that balance transparency, performance, and trust, ensuring AI-driven decisions remain both reliable and actionable.

Privacy-Preserving Distributed Learning Paradigms for Edge AI

Distributed learning paradigms like federated learning (FL) and swarm learning (SL) enable different actors to share data for AI model training and execution in a way that preserves privacy. As a prominent example, FL enables decentralized model training across distributed data sources, while preserving data privacy and security. The FL paradigm allows multiple entities (e.g., IoT devices and edge servers) to collaboratively learn a shared model without exchanging their local data.

Specifically, federated learning mitigates the issues related to data silos and residency requirements through collaborative learning that does not centralize sensitive information. Currently, FL deployments are in their infancy, as this distributed learning paradigm faces challenges such as data heterogeneity, communication overhead, and vulnerability to attacks. To alleviate these issues, ongoing research focuses on the development of robust federated learning frameworks and secure aggregation protocols.

Swarm learning is another innovative distributed learning paradigm, which is inspired by swarm intelligence. It employs a decentralized network of nodes, each with its own data and

AI model, to collaboratively learn from one another without exposing the underlying data. Many practical implementations of the SL approach leverage blockchain technology to ensure trust, security, and consensus among participating nodes.

Specifically, SL implementations allow nodes to exchange encrypted model updates through blockchain-based protocol, which enhances security and reduces the risk of a single point of failure. In the scope of an SL deployment, nodes can form dynamic swarms based on shared interests or goals, which allows them to benefit from collective intelligence without centralized control. SL is expected to offer numerous advantages for AI collaboration. However, it is also associated with various technical, social, and ethical challenges, which ask for further research prior to the widespread deployment and adoption of this technology.

Overall, the technological enablers of edge AI in 2025 are transforming how data is processed across networks. These technology enablers address the main limitations of cloud-centric approaches through local processing capabilities, specialized hardware, optimized algorithms, and innovative chip designs. In the years to come, they will enable Edge AI to play a pivotal role in offering a unique layer of intelligence to organizations in almost all different sectors.

Chapter IV:

Building an Edge

AI Ecosystem

The edge AI ecosystem today is at a stage where its long-term success depends on how hardware vendors, software developers, cloud providers, and industry stakeholders align their efforts. The push toward real-time AI inference, decentralized processing, and optimized edge computing architectures is creating a demand for more structured collaboration between industry, academia, and government. Without interoperability standards, scalable deployment models, and shared R&D efforts, edge AI risks fragmentation, limiting its potential impact across key sectors such as manufacturing, healthcare, and mobility.

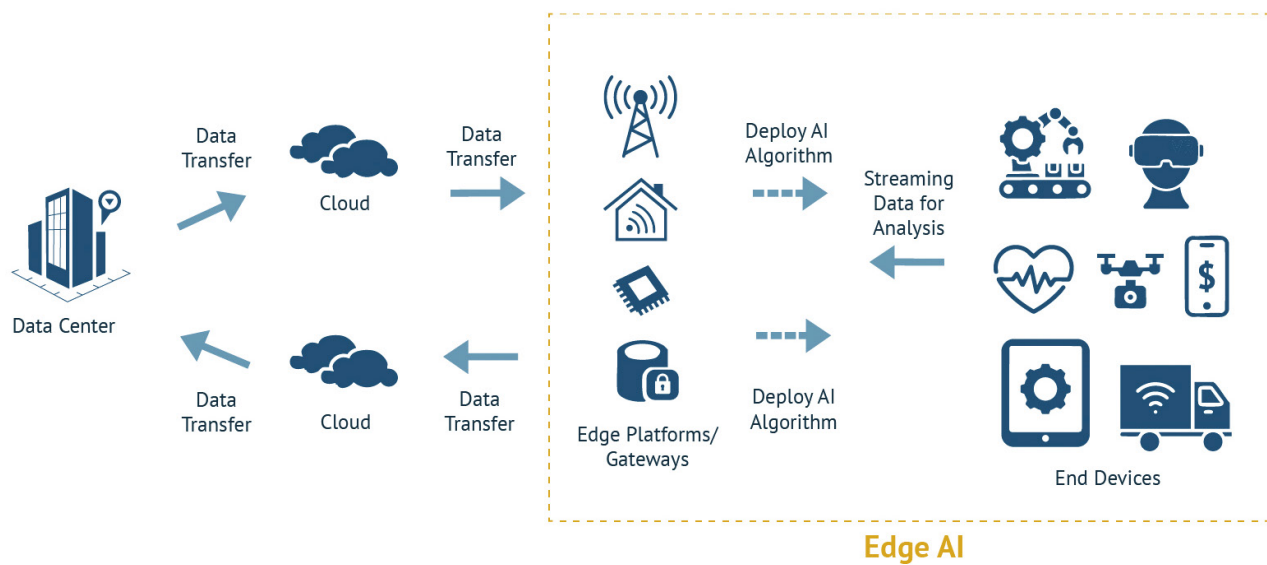
The industry is responding by forming alliances that address the challenges of edge deployment at scale. Semiconductor companies are working with AI model developers to ensure optimized inference at the

edge, cloud providers are integrating with edge-native platforms to enable hybrid architectures, and governments are funding initiatives to make edge AI more accessible and secure. Organizations like the Edge AI Foundation play a central role in this ecosystem, bringing together technology providers, researchers, and regulatory bodies to establish best practices, shared development frameworks, and certification programs that accelerate adoption while ensuring security, efficiency, and sustainability.

At the heart of these developments is the need for standardized yet flexible architectures. A three-tier edge AI framework is evolving to accommodate heterogeneous hardware, distributed workloads, and real-time AI applications. Companies that rely on edge AI must navigate a landscape where vendor lock-in, security risks, and deployment

complexities can create significant barriers. The emergence of open-source platforms, industry consortiums, and cross-sector partnerships is critical to ensuring that edge AI can scale efficiently without introducing unmanageable operational overhead or excessive infrastructure costs.

This chapter explores the structure of the edge AI ecosystem, the key players driving its expansion, and the collaborative efforts shaping its future. It examines how semiconductor manufacturers, cloud providers, and AI developers are working together to create optimized hardware-software stacks, how industry consortia like the Edge AI Foundation are standardizing edge deployment, and how companies are overcoming the challenges of large-scale edge AI adoption through strategic partnerships and technological advancements.



Edge AI ecosystem includes edge devices, edge servers, and cloud platforms (Image Credit: ABI Research)^[xvii]

Edge AI Ecosystem & Architecture: A Multi-Layered Framework

Edge AI operates within a three-layered architecture that distributes computational workloads across edge devices, edge servers, and cloud platforms. This structure allows AI models to execute real-time inferencing at the edge while leveraging higher computing power when needed. Each layer plays a distinct role in processing, aggregating, and refining data for intelligent decision-making.

Edge Devices: Real-Time Inferencing at the Source

Edge devices are the first point of interaction with real-world data.

These include IoT sensors, industrial robots, smart cameras, and embedded computing systems deployed in manufacturing, healthcare, automotive, and retail environments. Their primary function is low-latency AI inferencing—processing data on-site without relying on continuous cloud connectivity.

To enable real-time decision-making, edge devices execute optimized AI models that use quantization, pruning, and model compression to function within power and memory constraints. NVIDIA Jetson, Intel Movidius, and ARM Cortex processors provide specialized architectures that enhance AI inference efficiency in constrained environments. However, device heterogeneity remains a challenge. Edge AI devices vary significantly in hardware architectures, AI frameworks, and connectivity protocols, making standardization efforts critical. Initiatives like the Open

Edge Computing Initiative (OEI) aim to establish interoperable frameworks that enable seamless integration of AI across diverse edge environments. Different hardware configurations require frameworks that support cross-platform compatibility, such as TensorFlow Lite, Open Neural Network Exchange (ONNX), and Apache TVM, ensuring that models can run on diverse edge devices without extensive rework.

Edge Servers: Local AI Execution & Aggregation

Edge servers act as computational intermediaries between edge devices and the cloud. These are often deployed in factories, hospitals, retail locations, and autonomous vehicle networks, aggregating data from multiple sources and executing more

complex AI workloads than individual edge devices can handle.

A key advantage of edge servers is localized AI inferencing: running heavier models without offloading data to a remote data center. This reduces latency, bandwidth costs, and security risks associated with cloud dependency. Industrial gateways, micro data centers, and high-performance edge nodes are used in this layer, leveraging AI accelerators like Intel Xeon D processors, NVIDIA EGX edge AI platform, and AWS Outposts.

Edge servers also manage dynamic model updates, ensuring that AI models deployed on edge devices remain optimized and retrained as conditions change. Rather than relying on cloud-based retraining, federated learning allows models to be updated locally before synchronizing with a central repository. This approach is critical in healthcare and industrial automation, where real-time adaptability is essential.

While edge AI encompasses a range of computing layers, not all edge deployments operate under the same constraints. Carlos Morales, Vice President of AI at Ambiq, emphasizes the need to differentiate between edge computing and endpoint devices: “Despite being lumped in with all of Edge, the market is partitioned into ‘edge’ and ‘endpoint.’ The same ecosystem shouldn’t really try to address both since the constraints are so different.”

Edge devices, such as embedded cameras or industrial sensors, are designed for low-power AI inferencing, while more powerful edge servers act as intermediaries that handle complex

AI workloads before relaying data to the cloud. The need for hardware specialization at both levels highlights why standardization efforts must account for this diversity rather than assuming a uniform approach to edge AI development.

Cloud Platforms: Centralized AI Coordination & Model Training

The cloud remains essential for model development, large-scale data analysis, and storage. It serves as the backbone for training deep learning models before they are optimized and deployed to the edge. AI models are typically developed and refined on Google Cloud TPU, Microsoft Azure Machine Learning, AWS SageMaker, and IBM Watson AI.

Once trained, models are deployed to edge devices and edge servers, where they execute inference tasks in production environments. The cloud also serves as the backbone for AI model monitoring, analytics, and centralized orchestration, ensuring that deployments remain efficient across thousands, or even millions, of edge endpoints. For large-scale edge deployments, cloud providers offer edge-specific services, such as AWS IoT Greengrass for device management and machine learning inference at the edge, Microsoft Azure IoT Edge for secure containerized AI workloads, and Google Cloud IoT for AI model deployment and integration with on-premise edge computing.

Despite its advantages, cloud reliance presents data privacy, security, and bandwidth challenges. This has driven hybrid AI approaches, where sensitive

data remains at the edge, while only selective insights are transmitted to the cloud for deeper analysis.

Data Flow & Processing in Edge AI: From Collection to Insight Generation

Edge AI systems rely on efficient data movement between layers to balance latency, security, and computational efficiency. The data pipeline follows a structured process:

1. **Data Capture at the Edge:** Sensors, cameras, and embedded systems collect raw data.
2. **On-Device Processing:** AI models at the edge filter, classify, and preprocess data before transmitting insights.
3. **Edge Server Aggregation:** Multiple edge devices send data to a local edge server, where it undergoes further analysis and refinement.
4. **Cloud Synchronization:** Only selected insights (such as aggregated predictions or anomaly detection results) are transmitted to the cloud, minimizing bandwidth usage.
5. **Model Updates & Feedback:** The cloud retrains models using large-scale data and then distributes optimized updates back to edge devices.

Efficient data transfer between layers depends on lightweight communication protocols. For example, Message Queuing Telemetry Transport (MQTT) is used in IoT environments for low-bandwidth data exchange. Advanced Message Queuing Protocol (AMQP) provides reliable messaging between edge and cloud. EdgeX Foundry is an open-source framework for secure data orchestration in heterogeneous edge AI environments^[79].

To further streamline operations, processes like containerization and virtualization are widely used in edge deployments. Docker and Kubernetes allow AI applications to run consistently across different hardware configurations, addressing the issue of edge device diversity. These containerized models enable scalability, ensuring that AI workloads remain adaptable to changing computational needs.

The Edge AI Foundation: Unifying the Industry for Scalable Deployment

The Edge AI Foundation has emerged as a key figure in aligning the efforts of semiconductor companies, cloud providers, AI software developers, and enterprises to create a cohesive, scalable, and interoperable edge AI ecosystem. While individual companies focus on proprietary hardware and software optimizations, the Edge AI Foundation operates as a coordinating body, ensuring that edge AI technologies evolve within an open, standardized, and

sustainable framework. This facilitates multi-stakeholder collaboration and establishes best practices. Its programs address common barriers such as proprietary toolchains, lack of cross-platform compatibility, and inconsistent deployment models, creating an environment where companies, startups, and researchers can work together to accelerate the adoption of edge AI across key sectors, including manufacturing, mobility, healthcare, and energy.

A major function of the Edge AI Foundation is fostering cross-sector collaboration, ensuring that corporations, research institutions, and emerging AI startups align their efforts toward common goals. To achieve this, the foundation has established a network of partner organizations, open-source initiatives, and structured talent development programs. Key areas of collaboration include:

- **Industry-Academia Partnerships:** Working with universities and research institutions to ensure that cutting-edge AI advancements are directly applicable to real-world edge AI challenges. The EDGE Academia-Industry Partnership (EDGE AIP) is helping accelerate research into ultra-low-power AI models optimized for edge devices^[80].
- **Startup Incubation & Acceleration:** Providing mentorship, funding, and networking opportunities for startups developing next-generation edge AI solutions. Early-stage companies often struggle with access to optimized hardware, high-performance AI models, and cloud-edge

integration frameworks. The foundation's programs aim to bridge these gaps.

- **Standards Development & Open-Source Projects:** Promoting the use of interoperable AI frameworks, ensuring that edge AI models are portable across different hardware and software environments. This reduces vendor lock-in and makes edge AI deployment more accessible across industries.

Accelerating The Edge AI Development Lifecycle

The journey from concept to deployment for edge AI products is riddled with challenges. embedUR, a company with over 20 years of embedded software expertise, is redefining this lifecycle with a robust approach that blends deep knowledge of embedded systems, AI model development and optimization, and strategic collaborations with silicon vendors.

At the core of embedUR's strategy is their ModelNova platform, a resource hub providing pre-trained AI models, curated datasets, and blueprints tailored for edge devices. A typical blueprint in ModelNova might combine a face recognition model, a curated dataset, and platform specifications for a microcontroller, providing a ready-to-implement guide for developers. Unlike generic model repositories, ModelNova focuses on edge-ready AI building blocks that significantly cut down the time required to create proofs of concept (PoCs). Instead of weeks, developers can get AI models running on hardware in minutes, thanks to embedUR's pre-optimized resources. This rapid prototyping capability allows product designers to validate ideas faster and iterate more efficiently.

Balancing Trade-Offs in Edge AI Design

embedUR's process starts with understanding the solution's Minimum Viable Product (MVP) requirements, long-term goals, and design constraints. This involves critical trade-offs in performance, cost, power, and features, all essential considerations for resource-constrained edge devices. Through their expertise in sensor data processing, feature

extraction, and model adaptation, embedUR ensures that AI models are functional and optimized for real-world performance.

They also tackle challenges related to sensor variation. Many public datasets use high-quality images that edge devices with low-cost sensors can't replicate. embedUR's curated datasets ensure models perform reliably in real-world edge environments. In addition to real datasets, these models can leverage synthetic data generated by Generative Adversarial Networks (GANs) as a way to augment real-world data. Whether it is selecting high-resolution camera images for detailed object detection or optimizing for low-latency applications with smaller AI models, embedUR helps navigate these decisions effectively.

Reducing Optimization Time: Weeks to Minutes

One of embedUR's key differentiators is the availability of pre-trained, optimized, and ready-to-run models on the ModelNova platform. These models are designed to eliminate the need for extensive optimization efforts, allowing users to quickly implement AI on their devices. Instead of spending weeks adapting, training, and optimizing models, users can simply select a use case, download a model tailored to their target platform, and launch it within minutes. This streamlined process eliminates the need for extensive optimization efforts, empowering developers to focus on their core product

ideas. While embedUR also offers services to curate, train, optimize, and test AI models, the true speed advantage lies in the accessibility of models that have already undergone this process, streamlining the transition from proof-of-concept to production-ready solutions.

Behind the scenes, embedUR employs internal ML Ops tools to streamline model development, training, and deployment. These tools ensure efficient workflows, benefiting both embedUR's team and their partners. By handling the complexities of model porting, embedUR lets developers focus on their core product ideas rather than wrestling with optimization challenges.

Collaborations & Partnerships for Seamless Integration

embedUR's ability to adapt AI models to embedded Linux and non-Linux microcontrollers and collaborate seamlessly with silicon vendors like Synaptics, STMicro, Infineon, Silicon Labs, and NXP gives it a distinct edge in the market. Its long-standing relationships with these vendors allow it to integrate AI capabilities efficiently into various hardware platforms, enabling cost-effective, low-power, and high-performance products. This collaborative approach accelerates the launch of new AI-enabled devices and ensures developers have access to the latest hardware innovations. In a nutshell, embedUR assists clients in selecting platforms that balance cost, performance, and power constraints, ensuring the chosen hardware aligns with the product's goals.

„Edge AI's growth will be driven by strong collaboration between silicon vendors, developers, and product designers. The industry needs platforms like ModelNova that provide AI-ready building blocks—pre-optimized models, datasets, and use-case blueprints—to simplify development and shorten the time to market. At embedUR systems, we're enabling partnerships that foster innovation, whether it's helping silicon vendors optimize new AI platforms or enabling developers to bring intelligent edge solutions to life faster.”

– Eric Smiley, VP Business Development

Furthermore, the launch of Edge AI Labs in partnership with the Edge AI Foundation (EAIF) exemplifies embedUR's commitment to fostering collaboration. Edge AI Labs, powered by ModelNova, serves as a dynamic platform where academics, developers, and product designers can exchange models, datasets, and insights. This initiative aims to bridge the gap between cutting-edge AI capabilities and practical product applications, enabling quicker innovation cycles and broader adoption of edge AI. Through community-driven features like model submissions, dataset sharing, and interactive discussions, Edge AI Labs is set to become a hub for inspiration and innovation.

Beyond AI, embedUR offers expertise across the entire IoT development lifecycle, including communications integration (such as Wi-Fi and BLE), user interfaces, cloud management, and firmware. This holistic approach ensures that clients receive not only cutting-edge AI solutions but also comprehensive, productized systems ready for deployment. Smiley put it together nicely in the following formula:

(Your Idea) + (ModelNova components) + (embedUR total embedded/IoT/AI solution expertise) = (Turnkey Development Lifecycle)

The Future of Edge AI: Smarter, Faster, and Scalable

embedUR's solutions have been applied in various domains, including image segmentation for people detection, face recognition for security, and audio denoising for smart devices. Looking ahead, embedUR anticipates a wave of new AI-specific chipsets featuring integrated neural engines optimized for low-power applications. As these advancements roll out, embedUR's mission with ModelNova is clear: to empower product developers with the tools and knowledge needed to unlock the full potential of edge AI^[81].

By offering ready-to-deploy models, curated datasets, and practical blueprints, embedUR is accelerating the development lifecycle and paving the way for faster, smarter, and more scalable Edge AI innovations^[82].

Strategic Industry Partnerships Driving Edge AI Adoption

Industry partnerships are accelerating the deployment of edge AI by optimizing AI workloads for real-world environments. Semiconductor companies are working with AI developers to improve model efficiency on specialized hardware, cloud providers are integrating edge-native computing solutions, and research institutions are collaborating with industry leaders to advance scalable architectures. These strategic alliances are addressing key challenges, including power constraints, model optimization, and interoperability, ensuring that Edge AI can operate reliably across industries.

Hardware and Cloud Collaborations

Intel is driving Edge AI adoption through its Edge AI Partner Enablement Package, which equips businesses with tools, frameworks, and technical resources to accelerate AI deployment at the edge^[83]. This initiative provides optimized AI inference solutions, reference architectures, and industry-specific implementation guides, helping companies integrate AI workloads on Intel hardware efficiently. OpenVINO remains a cornerstone of Intel's Edge AI strategy, enabling deep learning inference optimization across CPUs, GPUs, and AI accelerators. By supporting a broad ecosystem of developers and enterprise partners, Intel ensures that AI applications run efficiently on resource-constrained edge devices.

Another notable collaboration involves Qualcomm and Meta, which have worked to integrate Meta's Llama large language models (LLMs) directly onto Qualcomm's edge processors. This partnership reduces the dependence on cloud-based LLMs, allowing devices to execute generative AI workloads on-site. The result is improved response times and reduced operational costs, particularly for applications like voice assistants and automated customer support^[84].

Earlier this year, MemryX and Variscite announced a partnership aimed at enhancing edge AI efficiency^[85]. By combining MemryX's AI accelerators with Variscite's System on Module (SoM) solutions, this collaboration simplifies AI deployment on edge devices, particularly for industrial automation and healthcare applications. The integration allows developers to work with pre-optimized AI hardware, reducing latency and power consumption while ensuring faster time-to-market.

Google and Synaptics Collaborate on Edge AI for the IoT

AI at the IoT Edge has inherent advantages, such as low latency, efficiency, and privacy, which directly contribute to the user experience. This is particularly the case when advances in multimodal processing are applied to enable context-aware computing. However, implementing this in resource-constrained environments such as the IoT Edge requires a novel approach to hardware, software, tools, partnerships, and ecosystems. This has led to a collaboration between Google and Synaptics that will see Google's Kelvin MLIR-compliant machine-learning (ML) core integrated into the Synaptics Astra™ AI-Native compute platform for the IoT (Figure 1). Together, the two companies will work to define the optimal implementation of multimodal processing for context-aware computing at the IoT Edge for applications such as wearables, appliances, entertainment, embedded hubs, and monitoring.

Astra was designed from the ground up to meet the needs of Edge AI while simplifying the development process. It combines scalable, low-power silicon with open-source, easy-to-use software and tools, a strong partner ecosystem, and robust Veros™ intelligent wireless connectivity.

The platform builds upon Synaptics' foundation in the application of neural networks for pattern recognition and its field-hardened AI hardware and compiler design expertise for the IoT, as well as its in-house support of a broad base of modalities, such as vision, image, voice, and sound, all of which can be combined to provide context for seamless device interactivity. It was launched at Embedded

World 2024 with three embedded MPUs: the SL1680, SL1640, and SL1620. The new SR-Series high-performance MCUs are set to launch at EW2025.

Collaborating with Google: Open ML Meets Purpose-Built Hardware

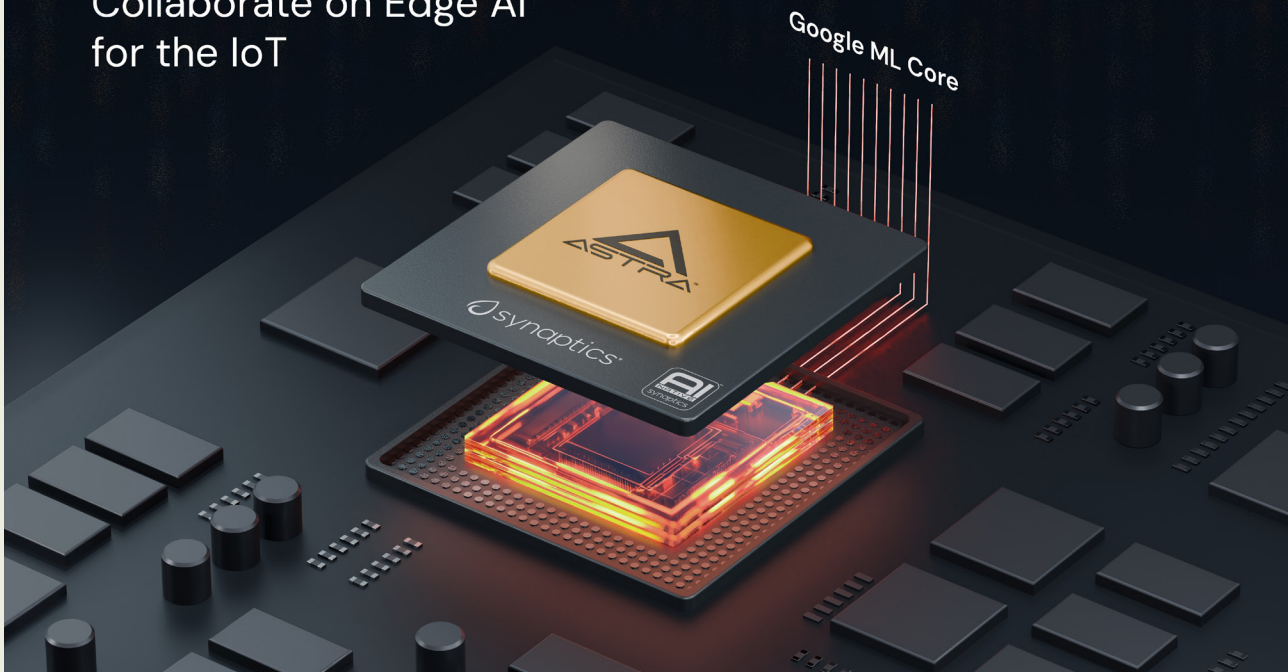
Google's partnership with Synaptics derives from their shared embrace of open-source approaches. Integrating Google's Kelvin MLIR-compliant ML core into Astra processors will allow developers to use standardized tools like TensorFlow Lite while optimizing models for Synaptics' neural processing units (NPU). This provides two advantages:

1. **Flexibility:** Developers can use familiar tools to build models, avoiding proprietary ecosystems.
2. **Optimization:** Synaptics' secure SyNAP compiler allows developers to fine-tune models for its NPUs, reducing latency and power consumption compared to generic edge chips. This is critical for real-time decision-making in resource-constrained edge AI applications.

The collaboration reflects a broader industry shift toward collaborative approaches to hardware-software integration, which aims to address the real-world challenges of context-aware Edge AI computing.



Collaborate on Edge AI for the IoT



Google's ML core on Synaptics Astra platform (Image Credit: Synaptics)^[xviii]

Astra's AI-Native Architecture: Built for Context, Not Just Compute

Unlike retrofitted edge AI solutions, Astra processors are engineered from the ground up for multimodal AI workloads. The architecture combines CPUs, GPUs, and DSPs with dedicated neural processing units (NPUs) tasked exclusively with ML inference, all with a unified memory structure that minimizes data movement between vision, audio, and sensor processing blocks. The Veros wireless connectivity portfolio includes Wi-Fi, Bluetooth, Zigbee/Thread, UWB, and GPS/GNSS to ensure reliable, robust, interoperable, and efficient communication in congested RF environments.

Applications for context-aware computing, where devices analyze real-time sensor data to make dynamic decisions, include a smart thermostat. Using Astra, a device could

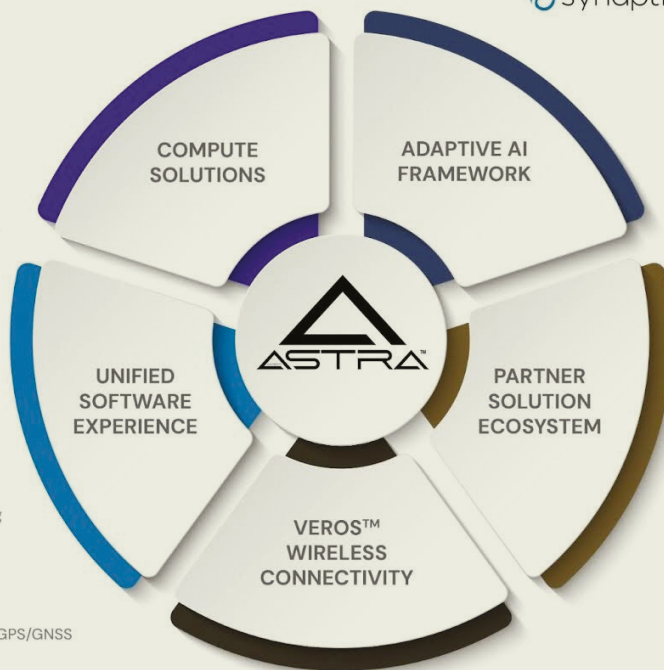
efficiently correlate motion, humidity, and ambient light to adjust temperature settings. This demonstrates how local processing optimizes energy use and the user experience without relying on cloud-based computing and compromising privacy.

The Road Ahead: Engineering Scalable Edge AI Solutions for the IoT

Edge AI's value lies not only in moving compute away from the cloud but also in redefining how devices interact with their humans and the environment. The open-source approach taken by Synaptics and Google will define and accelerate the deployment of solutions that will make these interactions intuitive and seamless.



- 
COMPUTE SOLUTIONS
 Build differentiation with our power-efficient, multi-modal, AI-enabled MPUs and high-performance AI and connectivity MCUs.
- 
ADAPTIVE AI FRAMEWORK
 Move from “making AI work” to “working with AI” with our open, cross-platform tooling, modeling, & optimization infrastructure.
- 
UNIFIED SOFTWARE EXPERIENCE
 Leverage an open, consistent cross-platform experience, supporting Linux, Android, & the leading RTOS offerings.
- 
PARTNER SOLUTION ECOSYSTEM
 Build market-ready systems & solution stacks with industry-leading ODMs, ISVs, service providers, and system integrators.
- 
VEROS™ WIRELESS CONNECTIVITY
 Pair Astra with Synaptics’ Veros, the most robust, industry-proven connectivity portfolio, spanning Wi-Fi® Bluetooth®, BLE, 802.15.4, & GPS/GNSS



ASTRA: The AI-Native IoT compute platform from Synaptics
(Image Credit: Synaptics)^[xviii]

“We are on the brink of a transformative era in Edge AI devices, where innovation in hardware and software is unlocking context-aware computing experiences that redefine user engagement,” said Vikram Gupta, Senior Vice President and General Manager of IoT Processors, Chief Product Officer at Synaptics. “Our partnership with Google reflects a shared vision to leverage open frameworks as a catalyst for disruption in the Edge IoT space. This collaboration underscores our commitment to delivering exceptional experiences while validating Synaptics’ silicon strategy and roadmap for next-generation device deployment.”

“Synaptics’ embrace of open software and tools and proven AI hardware makes the Astra portfolio a natural fit for our ML core as we ramp to meet the uniquely challenging power, performance, cost, and space requirements of

Edge AI devices,” said Billy Rutledge, Director of Systems Research in Google Research. “We look forward to working together to bring our capabilities to the broad market.”

Academic and Government Initiatives Supporting Edge AI

Industry-academic collaborations are playing a crucial role in advancing edge AI research and deployment. Amazon's Scholars and Visiting Academics program provides AI researchers with flexible opportunities to engage in real-world AI challenges while maintaining their academic roles. Similarly, Amazon's University Hubs program supports faculty-led research in AI optimization, benefiting both industry and academia^[86].

In Europe, the PREVAIL initiative brings together research institutions such as CEA-Leti, Fraunhofer, imec, and VTT to develop next-generation edge AI chips. By creating a multi-hub prototyping platform, this initiative allows companies to test AI hardware designs in real-world applications before scaling production^[87].

In the UK, the National Edge AI Hub serves as a collaborative platform uniting academia, industry, and the public sector to advance edge AI technologies. Led by Newcastle University, the Hub brings together a multidisciplinary team from institutions across the UK. The Hub's mission focuses on enhancing data quality and decision accuracy in time-critical applications such as healthcare and autonomous electric vehicles. Its activities encompass five main areas: cyber-disturbance modeling and simulation for edge computing, edge computing for AI, AI-driven edge cyber-resilience, an academic-industry technology incubator in edge AI, and industry-based, user-inspired application-driven validation. By

fostering a collaborative research community and leveraging existing UK investments, the National Edge AI Hub aims to amplify the impact of edge AI research and facilitate technology transfer and commercialization^[88].

Similarly, the U.S. National Science Foundation's NAIRR Pilot is a large-scale initiative aimed at democratizing AI access. Intel, NVIDIA, Microsoft, Meta, OpenAI, and IBM are among the industry participants contributing compute power and AI tools to researchers developing secure and energy-efficient AI applications. By creating a shared infrastructure, NAIRR helps accelerate innovation while ensuring AI resources are available beyond large tech companies^[89].

Challenges and Future Considerations in Edge AI Deployment

Energy Efficiency and Sustainability

The shift toward edge AI reduces dependence on cloud computing, but efficient on-device processing remains a major challenge. AI inference at the edge requires optimized hardware capable of balancing computational power with low energy consumption. Initiatives like PREVAIL are developing next-generation edge AI chips and advancing hardware that supports efficient AI processing in resource-constrained environments. Beyond hardware innovation, software-level optimizations, such as model quantization and sparsity techniques, are being leveraged to extend battery life in edge devices while maintaining inference accuracy.

Microcontroller Unit (MCU)-based AI inference is another critical bottleneck in energy efficiency. Many edge devices rely on MCUs with constrained power budgets, but the lack of efficient MCU AI runtimes makes it difficult to deploy advanced AI models without excessive energy drain. Companies like Ambiq are tackling this issue by focusing on ultra-low-power AI processing solutions, ensuring that AI workloads can run effectively on battery-operated and energy-sensitive applications such as wearables, smart sensors, and industrial IoT.

Security and Data Privacy

The distributed nature of edge AI deployments creates multiple security risks, including model tampering, unauthorized access, and data interception. Unlike centralized AI models stored in secured data centers, edge AI inference occurs across a fragmented network of devices, each of which must be safeguarded against cyber threats.

Federated learning presents one solution to mitigate privacy concerns by enabling decentralized AI training without transmitting raw data. Instead of centralizing sensitive information, devices process data locally and share only encrypted model updates, reducing exposure to potential breaches. Meanwhile, zero-trust security frameworks are being adopted to strengthen authentication mechanisms in edge AI systems, ensuring that every data exchange is validated at the hardware, software, and network levels. Companies implementing zero-trust architectures are focusing on hardware-based security enclaves, trusted execution

environments (TEEs), and secure boot mechanisms to prevent unauthorized modifications to edge AI models.

Regulatory compliance is also a growing concern, particularly as industries such as healthcare and finance adopt edge AI. Strict data sovereignty laws require companies to implement edge-native encryption standards and on-device data processing strategies to meet GDPR, HIPAA, and other compliance frameworks. The push for standardized edge AI security certifications is gaining traction, with organizations working toward defining best practices for secure AI inference at the edge.

Scalability and Infrastructure Management

Scaling edge AI from pilot projects to full-scale deployments presents logistical and infrastructural hurdles. Unlike cloud-based AI, where centralized servers handle model execution, edge AI requires distributed orchestration across thousands or even millions of devices. Managing model updates, optimizing resource allocation, and ensuring seamless communication across heterogeneous hardware architectures are among the top challenges companies face.

5G and next-generation connectivity solutions play a crucial role in unlocking large-scale edge AI adoption. With ultra-low latency and high-bandwidth capabilities, 5G enhances real-time AI processing by enabling rapid data exchange between edge nodes. This is particularly beneficial in autonomous systems, smart cities, and industrial IoT, where immediate AI-driven responses are essential.

According to IDC's forecast, global investment in edge IT infrastructure is expected to grow by 60% by 2028 as enterprises prioritize AI-driven edge computing in their digital transformation strategies^[90]. To support large-scale edge AI adoption, lightweight AI orchestration frameworks are being integrated into edge deployments. Tools like KubeEdge extend Kubernetes' capabilities to edge environments, enabling distributed AI workload management across cloud and on-premise edge servers. Meanwhile, Eve-OS, an LF Edge initiative, provides a bare-metal virtualization framework optimized for constrained devices.

However, a fragmented software ecosystem remains a challenge. Many edge AI deployments are hindered by proprietary toolchains and vendor-specific solutions that make cross-platform deployment difficult. The push for open-source AI frameworks and standardized edge inference APIs is helping address this issue by enabling interoperability across different hardware and software stacks. Initiatives such as EdgeX Foundry provide an open framework for integrating AI and IoT applications at the edge, while ONNX facilitates model portability across various AI hardware accelerators. Meanwhile, LF Edge, a Linux Foundation project, works toward creating a unified edge computing ecosystem by standardizing critical edge infrastructure components. These open-source efforts are reducing the risk of vendor lock-in and allowing enterprises to adopt scalable, hardware-agnostic AI solutions.

The Path Forward

For edge AI to reach full-scale adoption, industry leaders must continue addressing the pain points of energy efficiency, security, and scalability. Innovations in MCU AI runtimes, regulatory-compliant zero-trust architectures, and containerized edge AI workloads will be critical in ensuring seamless, cost-effective deployment across industries. Strategic partnerships between semiconductor companies, AI software developers, and connectivity providers will further drive advancements, creating a future where AI-powered edge devices can operate reliably, securely, and efficiently across diverse real-world applications.

With growing investment in standardized architectures, optimized software stacks, and public-private collaborations, edge AI is poised to become a dominant paradigm for AI processing. Initiatives like the Edge AI Foundation are central to this transformation, ensuring that the industry moves toward an interoperable, scalable, and collaborative future. Organizations that engage early with these initiatives will be best positioned to leverage the full potential of edge AI.

Chapter V: The Future of Edge AI

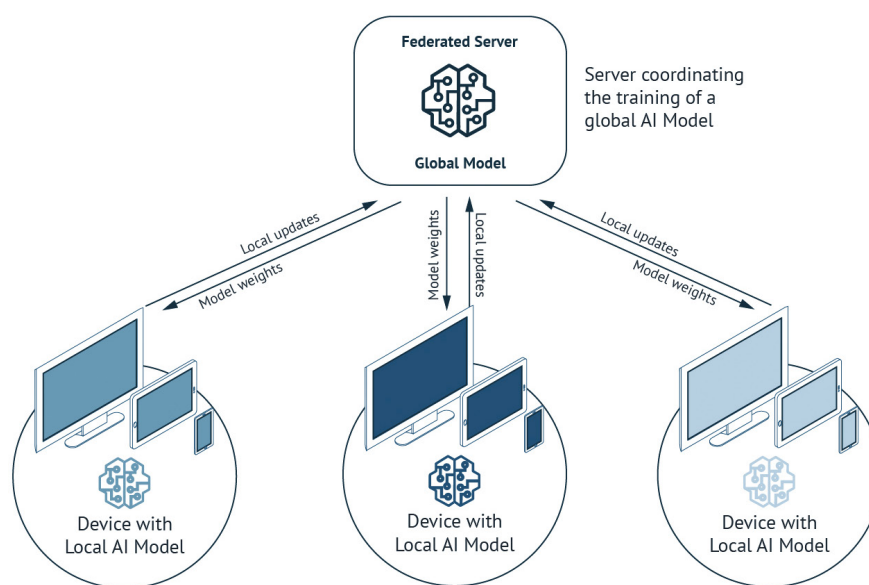
By 2030, intelligence will no longer be confined to centralized data centers. AI will operate at the source—on every device, sensor, and autonomous system—powering industries, cities, and everyday life. Machines will no longer wait for cloud responses to make critical decisions. Instead, edge AI will be the primary driver of real-time, autonomous intelligence, shaping a world where devices think, learn, and adapt locally.

Our 2024 State of Edge AI report uncovered how the demand for real-time AI, low-latency processing, and data privacy is accelerating edge AI adoption^[91]. Researchers, open-source communities, and enterprise leaders are driving innovations that are making edge AI more precise, energy-efficient, and scalable. However, its future is deeply interconnected with the progress of supporting

technologies: advancements in silicon, next-generation AI models, and communication networks like 6G. The continuous evolution of these foundational elements will determine the speed and scale of Edge AI's adoption.

This year's future outlook builds on these insights, revealing how federated learning, edge-native AI models, quantum-enhanced intelligence, and generative AI at the edge are converging to create self-learning, privacy-first AI systems. Autonomous vehicles will train each other without relying on centralized datasets. Hospitals will deploy AI models that evolve in real time based on patient data, ensuring hyper-personalized treatment. Industrial robots will operate with predictive intelligence, detecting and fixing errors before they happen.

Emerging innovations in neuromorphic computing, multi-agent reinforcement learning, and post-quantum cryptography are also redefining what's possible, enabling AI systems that are faster, more secure, and vastly more efficient. In a nutshell, tomorrow's AI will be self-sustaining, privacy-preserving, and infinitely scalable. This chapter explores these breakthroughs and how they will shape the next decade of edge AI, where intelligence is deployed and continuously optimized at the edge.



Federated learning is a robust strategy for collaborative model training across edge devices without the need to transfer raw data back to a central location
(Image Credit: Wevolver)

5 Emerging Trends in Edge AI

1. Federated Learning: Decentralized Intelligence at the Edge

Federated learning (FL) is evolving beyond privacy preservation into a cornerstone of decentralized intelligence. Over the next five years, federated frameworks are expected to actively enhance model adaptability, autonomy, and cross-industry collaboration. AI systems will learn dynamically at the edge, improving their intelligence across autonomous vehicles, decentralized medical AI, industrial automation, and global IoT networks without relying on centralized training. Market forecasts support this, as federated learning is poised to deliver nearly \$300 million

in market value by 2030, with a projected CAGR of 12.7%^[92].

However, ensuring AI models trained on decentralized devices can still generalize well across different environments remains a key challenge. In response, researchers are developing multi-prototype FL, an approach where multiple specialized models are trained instead of a single global one^[93]. This is particularly relevant in healthcare and industrial automation, where data can vary significantly across locations. For instance, hospitals in different regions may have distinct patient demographics, and factories using AI-driven quality control may have unique manufacturing conditions. Multi-prototype FL enables each environment to retain models that best fit its specific needs while still benefiting from global insights.

Another major driver of FL's evolution is the integration with next-generation networks such as 6G. As edge deployments scale, ultra-low-latency networks will allow AI models to synchronize across distributed devices more efficiently, reducing the time it takes to refine and deploy updates. The emergence of quantum federated learning (QFL) is also being explored to reduce the communication burden between devices, making the process more efficient for large-scale IoT networks^[94].

Security remains a critical focus for federated learning at the edge. The growing concern over adversarial attacks and model vulnerabilities has led to the exploration of post-quantum cryptography (PQC) for FL deployments. With quantum computing on the horizon, traditional encryption methods may no longer

be sufficient to protect decentralized AI models. Post-quantum encryption techniques are being developed to future-proof federated learning against emerging cybersecurity threats^[95]. Furthermore, AI-driven optimization of PQC algorithms, including reducing computational overhead, enhances their feasibility for deployment in resource-constrained edge environments^[96].

Federated Learning Across Industries

On the application front, FL is becoming a key enabler of autonomous, real-time AI across various industries. With advancements in model personalization, network efficiency, and security, FL is expected to drive the next phase of scalable, intelligent edge AI deployments across sectors ranging from healthcare to automotive and beyond.

- In autonomous mobility, FL can enable fleet-wide intelligence sharing, where self-driving vehicles refine collision avoidance, route optimization, and environmental adaptation by exchanging compressed model updates rather than raw sensor data. As V2X communication expands, AI-driven traffic systems will continuously refine predictions using real-world, distributed data.
- In healthcare, federated models can improve diagnostics by allowing hospitals, research centers, and wearable medical devices to collaborate without sharing patient data. AI models detecting rare diseases will adapt to diverse regional datasets

while preserving privacy through homomorphic encryption and post-quantum cryptography.

- In manufacturing, industrial AI will enable robots across different production lines to exchange model refinements securely, optimizing predictive maintenance and defect reduction. This will drive the rise of self-optimizing factories, where AI autonomously improves efficiency without external oversight.
- FL is also converging with multi-agent reinforcement learning (MARL), enabling swarm intelligence in robotics, logistics, and smart city infrastructure. By coordinating learning across distributed AI agents, industries could see decision-making speeds improve by up to 50% while reducing cloud dependency^[95].

By the end of the decade, federated learning will underpin a decentralized AI ecosystem, where edge devices refine models autonomously, ensuring continuous adaptation, enhanced security, and compliance with global data regulations^[96]. Federated learning's decentralized design aligns perfectly with emerging data localization mandates, which now affect over 75% of global markets^[97]. Rather than moving sensitive information across borders, FL transmits only model parameters. This will accelerate adoption among manufacturers wary of intellectual property leakage and healthcare institutions handling confidential patient records.

2. Edge Quantum Computing and Quantum Neural Networks

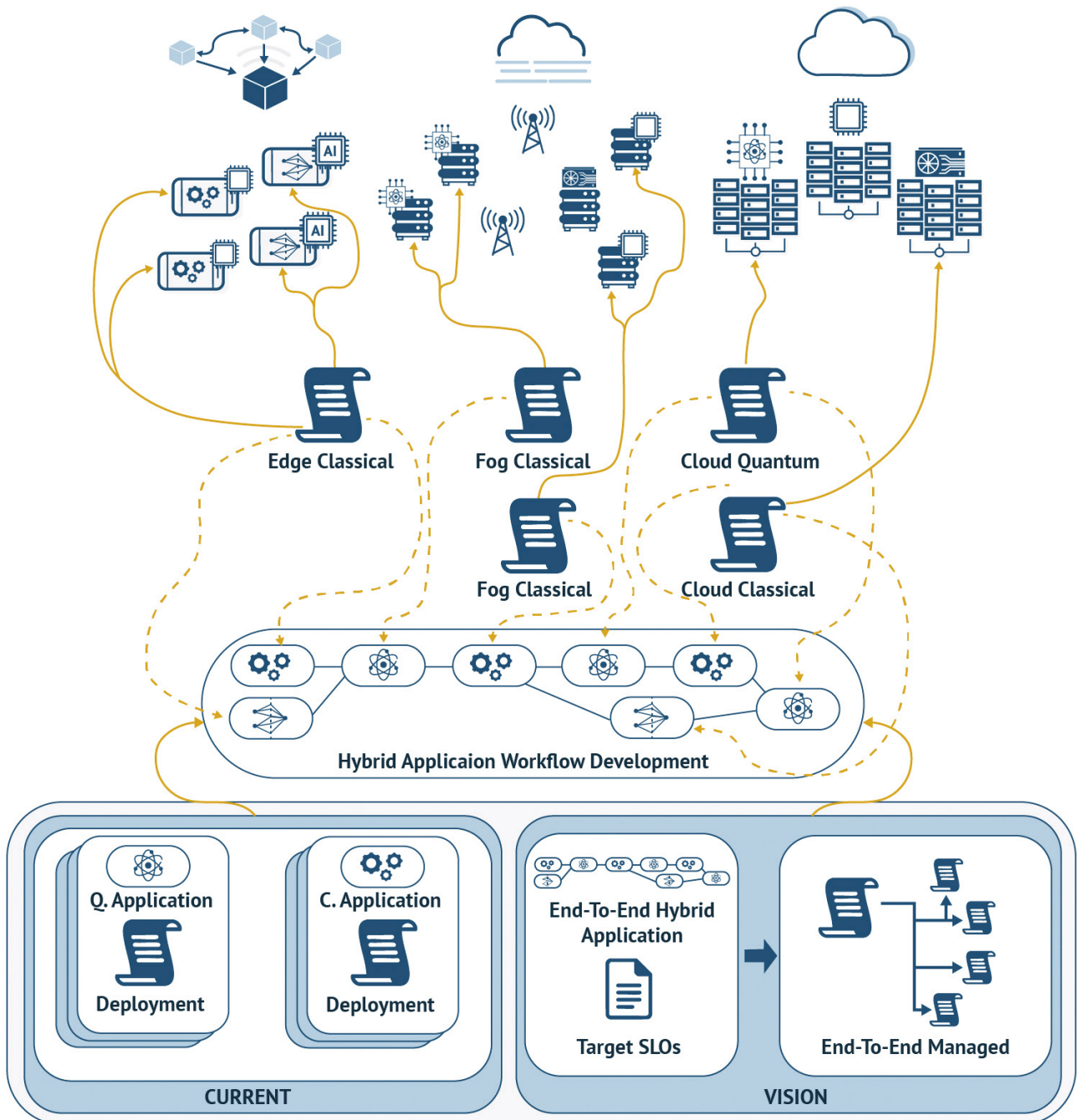
Quantum computing is set to redefine the capabilities of edge AI. While today's AI at the edge relies on optimized deep learning models and low-power hardware accelerators, quantum computing introduces a fundamentally different approach: leveraging quantum states to process exponentially larger datasets and optimize decision-making at speeds unattainable by traditional methods. As quantum processing units (QPUs) move beyond cloud-based infrastructure, hybrid quantum-classical AI will emerge at the edge, enhancing real-time decision-making across industries such as finance, healthcare, energy, and industrial automation.

Quantum Neural Networks (QNNs): Smarter AI for the Edge

Quantum neural networks (QNNs) are a new class of AI models that leverage quantum properties to detect patterns and relationships in data that "classical AI" might struggle with. Unlike existing neural networks, which require increasing amounts of power and memory to improve performance, QNNs can process information in more compact and efficient ways.

Early research integrating QNNs into classical deep learning has shown up to 2-3x faster training speeds for optimization-heavy AI tasks and a 50-70% reduction in computational costs^[98]. This means AI models at the edge, whether in autonomous vehicles, industrial robots, or financial fraud detection systems, can learn and adapt

EDGE-CLOUD CONTINUUM



A Distributed Classical-Quantum Hybrid Platform
(Image Credit: A. Furutanpey et al.)^[xix]

much faster without needing constant retraining in the cloud.

To make QNNs viable for edge AI, hybrid computing architectures are being developed. These use a combination of classical deep learning for preprocessing and quantum layers for complex optimization. Another breakthrough is quantum circuit cutting, which breaks down large quantum computations into smaller, manageable tasks that can run efficiently on edge QPUs. These methods allow small-scale quantum devices to contribute to AI inference at the edge without relying on large, cloud-based quantum systems^[99].

Quantum Speedups for Edge AI

Edge AI applications typically require rapid analysis and real-time decision-making. Quantum computing brings exponential speed improvements to several key areas:

- **Financial Fraud Detection:** Quantum-enhanced AI can process over 10,000 transactions per second at the edge, identifying suspicious activity in real time^[100].
- **Smart Grid Optimization:** Quantum AI dynamically adjusts power distribution based on real-time energy demand, reducing waste and improving efficiency. Research has shown that using quantum-assisted algorithms can lead to 10% in energy consumption savings^[101].
- **Drug Discovery & Healthcare AI:** Quantum-powered AI can cut drug discovery timelines significantly. Quantum computing in computer-

assisted drug discovery (CADD) could expand the range of biological mechanisms that can be modeled, accelerate screening processes, and reduce trial-and-error iterations in drug development. Identifying viable candidates earlier and avoiding research pathways likely to fail minimizes costly dead ends, streamlining the discovery phase^[102].

These gains come from quantum computing's ability to solve optimization and probabilistic problems far more efficiently than classical systems, making it a natural fit for edge applications that demand speed and accuracy.

Bringing Quantum Computing to Edge Devices

Until now, quantum computing has been confined to cloud-based data centers due to its hardware requirements, including extreme cooling. However, new advancements in mobile QPUs, based on diamond-based processors, will likely make it possible to run quantum algorithms at room temperature. In the coming years, quantum computing will not be limited to the cloud but could be embedded in autonomous systems, industrial robots, and IoT devices at the edge^[99].

Hybrid quantum-classical split computing architectures are developed to integrate quantum AI into edge environments. In this setup, classical AI handles initial processing, while quantum computing refines and optimizes results. This reduces reliance on centralized quantum resources, making edge AI systems faster and more self-sufficient.

While mobile QPUs bring quantum computing closer to real-world edge deployments, challenges remain. Quantum error correction, hardware stability, and computational overhead in constrained environments must be addressed before quantum computing can be widely deployed at scale. These hurdles, however, are actively being tackled through advancements in error-resilient quantum algorithms and quantum architectures, paving the way for truly decentralized quantum intelligence at the edge.

3. Edge AI for Autonomous Humanoid Robots

Back in 2016, at a conference of the UK government's Robotics and Autonomous Systems Network, robotics luminary Hiroshi Ishiguro criticized the use of the cloud to offload robotic intelligence to a remote data center, pointing out the issue of latency. „Accessing a cloud computer takes too long. The half-second time delay is too noticeable to a human. In real life, you never wait half a second for someone to respond. People answer much quicker than that.” That delay would push you into a trough of creepiness known as the „uncanny valley,” where the robot seems human-like but is subconsciously perceived as an unnatural imitation rather than a genuine human presence^[103].

Edge AI solves this by enabling humanoid robots to process information locally, ensuring real-time decision-making, natural interactions, and uninterrupted autonomy. This shift is accelerating as humanoid robots enter industries such as manufacturing, healthcare, and retail.

For humanoid robots operating in unstructured environments, low-latency AI inference is essential. In other words, every millisecond counts. The ability to interpret human actions in real time is an active requirement for industrial and service robots. That is where edge AI is crucial, allowing robots to:

- Adapt instantly to human movement in warehouses and factories
- Make split-second navigation decisions in dynamic environments
- Recognize and predict actions with human action recognition (HAR) models

Recent research on edge-AI-powered HAR has demonstrated 97.58%

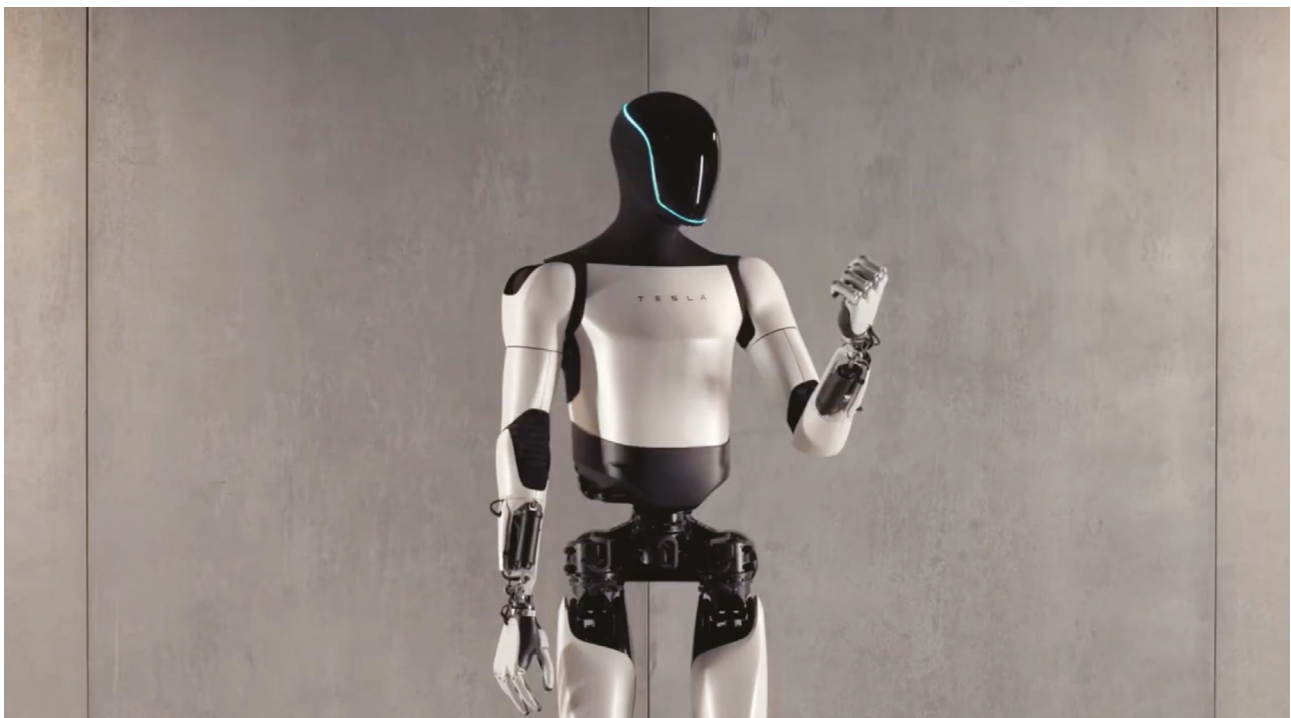
precision in single-action recognition and 86% accuracy in continuous action tracking, enabling robots to interpret gestures, predict movement, and respond accordingly^[104]. In a factory setting, this means robots can adapt instantly to assembly line changes or human coworkers' movements, avoiding accidents and improving efficiency. In warehouses, humanoid robots using edge AI can autonomously navigate unpredictable environments, optimizing logistics without waiting for cloud-based instructions.

Embodied AI for Interactive Robots

The next phase of humanoid robotics will be defined by embodied AI, where AI models become more adaptive, responsive, and capable of self-improvement. Recent advancements in

reinforcement learning and AI world models are pushing humanoid robots beyond pre-programmed behaviors, enabling them to learn from real-world interactions and refine their decision-making processes^[105]. Here, both edge and cloud computing are needed for such robots to function autonomously and effectively. Cloud AI enables large-scale training, long-term knowledge accumulation, and global model updates, while edge AI ensures real-time perception, decision-making, and interaction with the environment^[106].

A leading example of this hybrid approach is NVIDIA's Project GR00T, which is accelerating the development of general-purpose humanoid robots by integrating edge AI for real-time inference and cloud-based simulation for large-scale training. Through GR00T-Mobility and GR00T-Perception,



Humanoid robot by Tesla (Image credit: Teslarati)^[xx]

robots gain enhanced dexterity, perception, and full-body control, allowing them to perform increasingly complex tasks with human-like coordination^[107].

The future of humanoid robots also hinges on multimodal AI, where speech, vision, and motion are processed simultaneously at the edge, enabling seamless, natural interactions in real-world settings. Multimodal AI fuses multiple data sources in real time, allowing robots to make relatively accurate, context-aware decisions. A humanoid robot in a retail store, for instance, doesn't just respond to spoken questions. It also analyzes a customer's body language, facial expressions, and surrounding environment to infer intent and deliver more helpful responses.

Alongside the emergence of multimodal AI is the rise of small language models (SLMs). While large language models (LLMs) require significant cloud resources to function, SLMs are optimized for on-device processing, enabling robots to generate conversational responses, understand commands, and provide assistance without relying on an internet connection. This shift is critical for applications in customer service, healthcare, and industrial automation, where reliable, real-time communication is essential. SLMs are particularly effective in low-power, resource-constrained environments, making them ideal for humanoid robots that operate autonomously.

In retail environments, humanoid robots like 1X Robotics' EVE assist customers by responding to verbal inquiries, analyzing facial expressions, and navigating store layouts, all

without requiring continuous internet access^[108]. Meanwhile, in hospitals and elder care settings, AI-powered robots monitor patients, assist with mobility, and detect subtle changes in behavior that could indicate medical emergencies, all with on-device processing that ensures data privacy and security.

As SLMs become more efficient and multimodal AI continues to improve, humanoid robots will move beyond scripted, predefined interactions toward genuinely intelligent, adaptive engagement with humans. This will enable robots to handle more complex tasks, integrate into more industries, and function as proactive assistants.

4. AI-Driven AR/VR: The Next Evolution

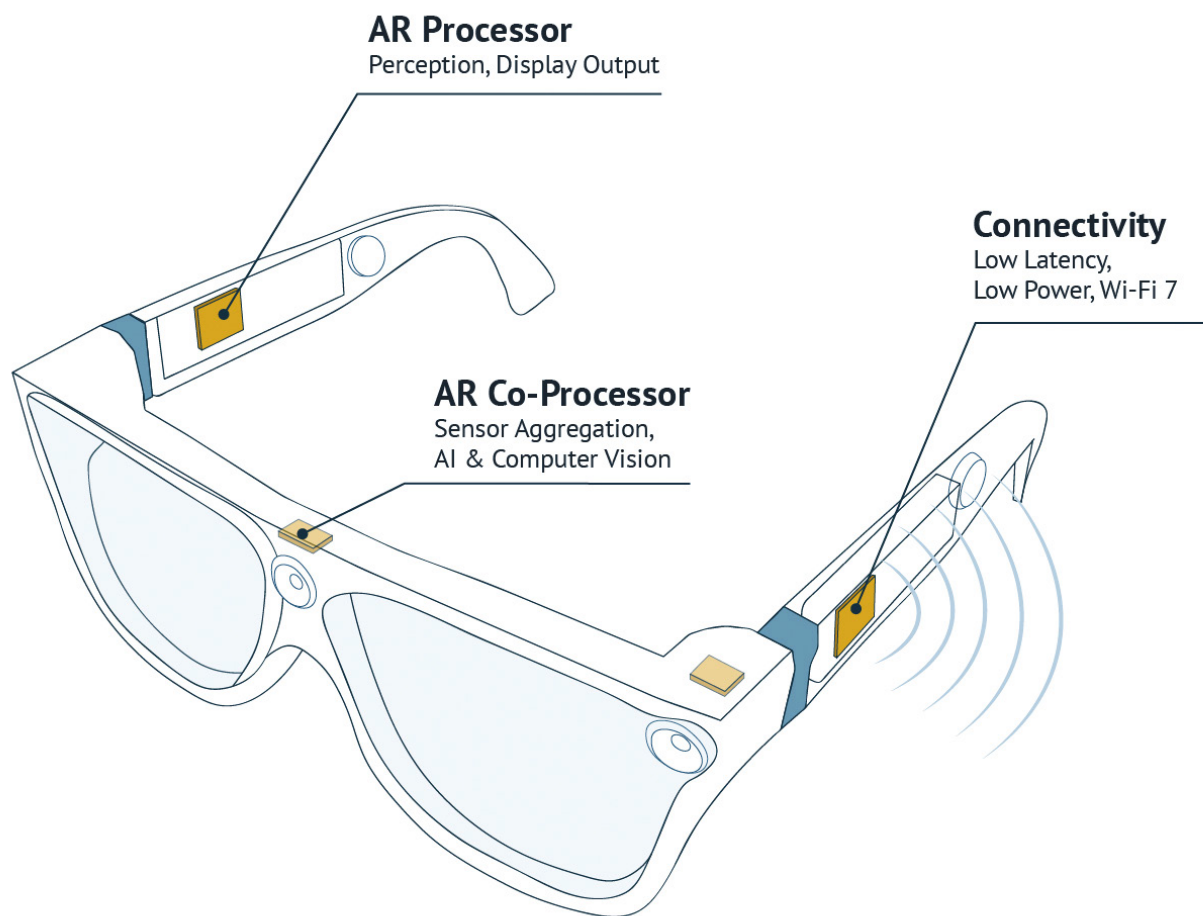
Augmented Reality (AR) and Virtual Reality (VR) are not confined to gaming and entertainment anymore. As AI capabilities expand, these technologies are becoming deeply embedded in industry, healthcare, and enterprise applications, shifting from isolated experiences to fully integrated, intelligent ecosystems. The future of AR/VR will be defined by adaptive, AI-powered environments that respond to users in real time, breaking the barriers between the digital and physical worlds.

Edge AI is a critical enabler of this evolution. Next-generation AR/VR devices will process information locally, allowing for real-time responsiveness and increased energy efficiency. AI-driven spatial computing will allow AR glasses and VR headsets to dynamically adjust overlays, depth perception, and environmental interactions based on

context. In industrial settings, this means AR-powered workspaces will provide engineers with hands-free, AI-generated instructions that adapt to real-world conditions in real time. In healthcare, AR-assisted surgery will integrate AI-powered overlays that enhance precision, updating in milliseconds based on the surgeon's movements without cloud-induced delays.

A major breakthrough shaping the next phase of AR/VR is the rise of AI-driven avatars and virtual beings^[109]. Unlike current virtual assistants, which follow predefined scripts, the next generation of AI avatars will interact with users in a more natural and context-aware way. By leveraging edge AI, these virtual entities will process voice, gestures, and facial expressions instantly, creating a seamless conversational experience. In retail, AI-powered digital assistants will transform customer interactions, offering real-time, personalized assistance in virtual storefronts. In corporate settings, AI-driven virtual coworkers will become part of hybrid workplaces, integrating into AR-enhanced collaborative spaces where meetings feel as natural as in-person interactions.

As AR/VR devices become more intelligent, they are also becoming smaller and more power-efficient. The future of AR lies in compact, AI-native wearables that eliminate the need for bulky hardware or constant cloud connectivity. Companies like Qualcomm and Meta are developing ultra-low-power AI chips designed for lightweight AR glasses, making standalone, high-performance AR possible. For example, Qualcomm's QCC112 chip is designed for ultra-low-power operation, making it suitable



Overview of the 3-chip/module Qualcomm Snapdragon AR2 Gen 1 platform
(Image Credit: Qualcomm)^[xxx]

for small form-factor devices like AR glasses^[42]. Such advancements create lightweight, comfortable, and unobtrusive AR/VR wearables, enhancing user adoption and experience. 2025 will bring major advancements in the AR/VR market from industry leaders like Apple, Google, Samsung, and Meta^[110]. By 2030, AI-powered AR glasses are projected to be almost as ubiquitous as smartphones, enabling real-time language translation, AI-enhanced navigation, and digital overlays that merge seamlessly with physical environments^[111].

The convergence of AI, AR, and the metaverse will further push the boundaries of human-computer interaction^[112]. Future AR/VR systems will not just display information but generate adaptive, AI-driven digital environments that respond intelligently to users. Interfaces will evolve dynamically, tailoring experiences to individual preferences, work habits, and social interactions. The next era of mixed reality will not be a passive experience but an AI-powered, hyper-personalized digital layer that continuously adapts to the world around us.

As AI and AR/VR technologies mature, the gap between digital and physical reality will continue to shrink. In the coming decade, intelligent AR/VR will redefine how we work, learn, and interact, with AI not just powering virtual experiences but making them feel as natural and intuitive as the real world.

5. Neuromorphic Computing: The Future of Energy-Efficient AI

Neuromorphic computing is poised to become increasingly prevalent in the edge AI space by introducing brain-inspired architectures that offer significant advantages in energy efficiency and processing capabilities. Unlike traditional computing systems that separate memory and processing units, neuromorphic systems integrate these functions, mimicking the parallel and event-driven nature of the human brain. This design enables them to handle complex, real-time data processing tasks with minimal energy consumption, making them ideal for edge applications. For example, the NeuRRAM chip, introduced in a Nature study in 2022, has an analog computing architecture twice as energy-efficient as state-of-

the-art „compute-in-memory” chips, enabling sophisticated cognitive tasks on edge devices without cloud connectivity^[113,114]. This leap mirrors the shift from desktop PCs to smartphones, unlocking portable applications once deemed impossible.

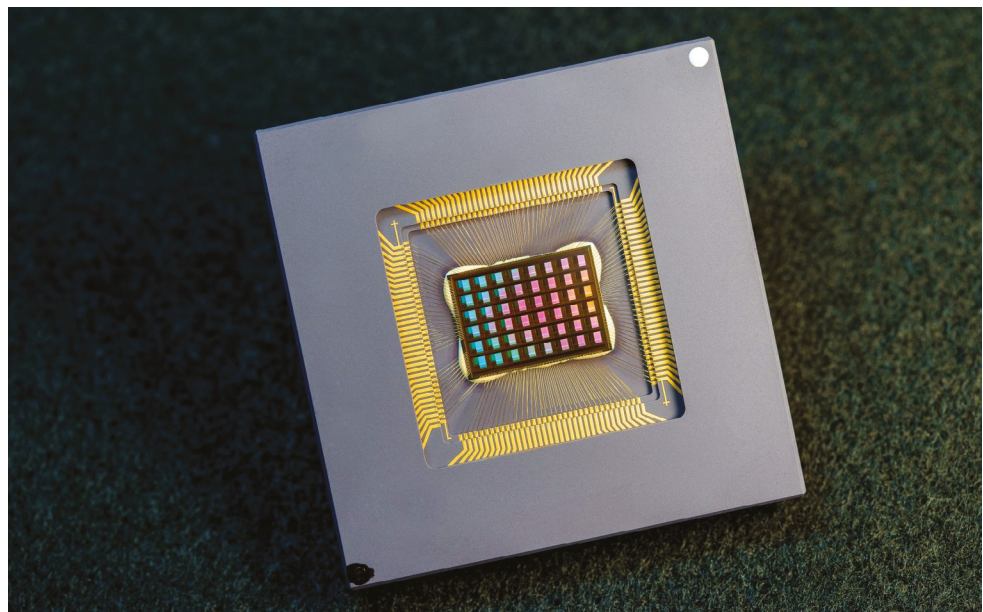
Research and early commercial deployments indicate that neuromorphic chips can redefine how intelligence is deployed at the edge. The focus is shifting from proof-of-concept systems to scalable, real-world applications. The next stage of neuromorphic AI will be shaped by its ability to scale, integrate with hybrid architectures, and enable new forms of decentralized, real-time intelligence.

Scaling Neuromorphic Computing for Complex AI Tasks

Recent research suggests that neuromorphic systems need to scale

significantly to handle the growing complexity of AI workloads. A study in Nature highlights the shift from small-scale neuromorphic experiments to large-scale architectures^[115], such as Intel's Hala Point, which now incorporates over 1.15 billion neurons^[116]. The next step is increasing the number of neurons and synapses while maintaining energy efficiency, enabling real-world tasks beyond simple pattern recognition.

Scaling neuromorphic AI requires integrating sparsity, a key biological principle where unnecessary neural connections are pruned. This optimization could drastically improve energy efficiency and processing density. If successful, large-scale neuromorphic systems could match deep learning accelerators in performance while consuming a fraction of the power.



The NeuRRAM chip can run computations within its memory, storing data in an analog spectrum (Image credit: University of California San Diego)^[box1]

Scaling neuromorphic AI requires integrating and optimizing several key principles:

- **Sparsity** enables efficient processing by pruning unnecessary connections, reducing computational overhead without loss of accuracy.
- **Neuronal scalability** allows systems to scale from compact edge devices to large, multi-chip architectures capable of solving complex tasks.
- **Asynchronous communication** eliminates bottlenecks by ensuring event-driven data flow and improving responsiveness.
- **Dynamic reconfigurability** enables adaptable AI that can modify its neural pathways in real time for greater flexibility.
- **Redundancy and correlation** enhance fault tolerance and improve computational stability by leveraging neural redundancy.
- **Sensor and compute interfaces** will need standardization to support seamless integration with external devices, particularly in vision, audio, and biomedical applications.
- **Resource awareness** ensures systems dynamically manage energy, memory, and compute resources based on real-time demands^[115].

If implemented successfully, these principles will drive neuromorphic AI toward large-scale, high-efficiency deployments at the edge.

Hybrid Systems: The Path to Widespread Adoption

For the foreseeable future, hybrid architectures will bridge the gap between conventional AI and neuromorphic computing. Research indicates that neuromorphic chips will be deployed alongside GPUs, TPUs, and analog AI processors, each handling different computational tasks^[117]. This approach will enable neuromorphic cores to manage low-power, real-time processing, such as sensor fusion in smart cities or autonomous vehicles, while traditional chips handle larger-scale computations.

Memristor-based neuromorphic chips are also gaining attention. The startup Techifab is developing memristor-based components that integrate memory and processing into a single unit, reducing data transfer energy loss^[118]. „Our goal is to use the brain as a model to create a technology that makes complex decisions logically and traceably with minimal energy consumption,” says Heidemarie Krüger, physicist and founder of Techifab. These advances could further optimize neuromorphic chips for high-performance edge AI systems.

Real-Time Learning and Adaptive Intelligence

Unlike static deep learning models, future neuromorphic AI will continuously learn and adapt at the edge. Current AI deployments require cloud retraining, but neuromorphic processors will enable local adaptation based on real-time data. This will improve robotics, industrial automation, and autonomous systems, where real-time decision-making is critical.

For example, wearable devices using neuromorphic computing are already demonstrating fast, energy-efficient processing for gesture recognition and biomedical applications. With latency reduced to as low as 5.7 milliseconds and power consumption at just 41 mW, neuromorphic chips are proving capable of real-time edge intelligence^[119].

Neuromorphic Chips + 6G + Quantum Computing

The long-term trajectory of neuromorphic computing extends beyond existing edge AI systems. The integration of neuromorphic AI with 6G networks and quantum computing is expected to enable ultra-low-latency, massively parallel processing at the edge. BrainChip's Akida processor and state-space models, such as Temporal Event-Based Neural Networks (TENNs), are early indicators of this direction, demonstrating the feasibility of lightweight, event-driven AI architectures for high-speed, real-time applications^[120].

As scaling challenges are addressed, neuromorphic chips will move from niche applications to mainstream adoption, powering the next generation of autonomous machines, decentralized AI, and real-time adaptive systems. The future of edge AI will depend on how efficiently intelligence is deployed. Neuromorphic computing is positioned to make that shift possible.

New Approaches for GenAI Innovation at the Edge

Recent advances in LLM training algorithms, such as Deepseek's release of V3, have rattled the traditional AI marketplace and challenged the belief that large language models require high computational investments to achieve performant results. This demonstrates that trying a new approach can have a big impact on a key challenge the AI market faces: high computational power and resultant costs to train and execute LLM models. New approaches must be considered to address similar challenges at the edge, such as the large compute and memory bandwidth requirements of transformer-based GenAI models that result in on costly and power-hungry edge AI devices.

State Space Models: More Efficient than Transformers

State Space Models (SSMs), with high-performance models like Mamba, have emerged in the last two years in cloud-based LLM applications to address the high computational complexity and power required to execute transformers in data centers. Now, there is growing interest in using SSMs to implement LLMs at the edge and replace transformers, as they can achieve comparable performance with fewer parameters and less overall complexity.

Like transformers, SSMs can process long sequences (context windows). However, their complexity is on the order of the sequence length $O(L)$, compared to the order of the square of the sequence length $O(L^2)$ for transformers—and with $1/3$ as many parameters. Not to mention that SSM implementations are less costly and require less energy.

These models leverage efficient algorithms that deliver comparable or even superior performance. Brainchip's unique approach is to constrain an SSM model to better fit physical time series or streaming data and achieve higher model accuracy and efficiency. BrainChip innovation

of SSM models constrained to streaming data are called Temporal Enabled Neural Networks, or TENNs. Combined with optimized LLM training, they pave the way for a new category of price and performance LLM and VLM solutions at the edge.

Deploying LLMs at the Edge: Efficiency and Scalability

BrainChip addresses the challenge of deploying LLMs at the edge by using SSMs that minimize computations, model size, and memory bandwidth while producing state-of-the-art (SOTA) accuracy and performance results to support applications like real-time translation, contextual voice commands, and complete LLM models with RAG extensions. Brainchip can condense the software model and the implementation into a tiny hardware design. A specialized LLM design can execute the edge LLM execution in under a watt and for a few dollars using a dedicated IP core that can be integrated into the customer's SoCs. This enables a whole new class of consumer products that do not require costly cloud connectivity and services.

This ultra low power execution makes edge LLMs viable for always-on devices like smart assistants and wearables. Cloud LLM services are neither private nor personalized. A completely local edge AI design enables real-time GenAI capabilities without compromising privacy, ensuring users have greater control over their data and enabling a new class of personalization you can bring wherever you go.

Emerging designs like BrainChip's Akida core offer a scalable and efficient solution for engineers and product developers who want to integrate advanced AI capabilities into private, personalized consumer products, including home, mobile, and wearable products.

Final Thoughts on Preparing for the Next Wave

As organizations anticipate the next wave of technological advancement, strategic preparation in edge AI needs to happen today. Key focus areas include infrastructure investment, data privacy and security, and cross-industry collaboration.

In infrastructure, organizations should prioritize the deployment of microdata centers and next-generation IoT devices to process data closer to its source, reducing latency and enhancing real-time decision-making. Optimizing AI models for performance and efficiency is essential, ensuring that these models operate effectively within diverse device constraints. This balance between computational demands and available resources is key to aligning AI infrastructure with business objectives, ultimately enhancing operational efficiency and maintaining a competitive edge.

Organizations must also implement robust encryption protocols and access controls to safeguard information. Emerging techniques, such as confidential computing and multi-party computation by IBM Research^[121], add an extra layer of protection during data processing. Furthermore, compliance with evolving privacy regulations like GDPR and CCPA remains critical. Advanced methods such as federated learning enable decentralized data processing, minimizing breach risks while enhancing AI performance at the edge.

Yet, arguably, the most impactful factor in the success of edge AI lies

in collaboration across sectors to establish industry standards and interoperability. Partnerships among hardware vendors, software developers, and regulatory bodies foster innovation and standardization, ensuring cohesive ecosystems. Collaborative initiatives, such as those led by the edge AI Foundation, are essential to address global challenges and drive adoption^[122].

Organizations that proactively invest in infrastructure, prioritize security, and embrace collaborative ecosystems will position themselves as leaders in the edge AI space in 2025 and beyond.

References

1. S. Kinney, "RCR Wireless News," RCR Wireless News. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.rcrwireless.com/20250210/ai-infra-structure/convergence-of-test-time-inference-scaling-and-edge-ai>
2. J. Davis, "Looking ahead: 2025 will be the year of edge AI," Edge Industry Review. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.edgeir.com/looking-ahead-2025-will-be-the-year-of-edge-ai-20250210>
3. National Highway Traffic Safety Administration (NHTSA), Department of Transportation (DOT), "Federal Register Volume 87, Issue 46 (March 9, 2022)." Office of the Federal Register, National Archives and Records Administration, Mar. 09, 2022. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.govinfo.gov/content/pkg/FR-2022-03-09/pdf/2022-04894.pdf>
4. D. J. Jeffs, D. X. He, and J. Li, "Autonomous Vehicles Market 2025-2045: Robotaxis, Autonomous Cars, Sensors," Oct. 2024. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.idtechex.com/en/research-report/autonomous-vehicles-markets-2025-2045/1045>
5. R. Dhall, "Hyundai Motor Group Embraces NVIDIA AI and Omniverse for Next-Gen Mobility," NVIDIA Blog. Accessed: Feb. 18, 2025. [Online]. Available: <https://blogs.nvidia.com/blog/hyundai-motor-group-ces/>
6. Reuters, "Costs from supply chain disruptions drop by over 50% but headwinds remain -survey," *Reuters*, Aug. 09, 2023. [Online]. Available: <https://www.reuters.com/markets/costs-supply-chain-disruptions-drop-by-over-50-headwinds-remain-survey-2023-08-09/>
7. S. Hippold, "Gartner Predicts 25% of Supply Chain Decisions Will Be Made Across Intelligent Edge Ecosystems Through 2025," *Gartner*, Jan. 19, 2022. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.gartner.com/en/newsroom/press-releases/2022-01-19-gartner-predicts-25-percent-of-supply-chain-decisions-will-be-made-across-intelligent-edge-ecosystems-through-2025>
8. S. Saha, "IoT in Supply Chain Market Size, Demand & Trends 2023-2033," Future Market Insights. [Online]. Available: <https://www.futuremarketinsights.com/reports/iot-in-supply-chain-market>
9. Edge Impulse, "Industrial Asset Tracking with NVIDIA Omniverse and TAO." Accessed: Feb. 18, 2025. [Online]. Available: <https://edgeimpulse.com/case-studies/industrial-asset-tracking-with-nvidia-omniverse-and-tao>

10. Infosys BPM, "The Power of Predictive Maintenance in Manufacturing." Accessed: Feb. 18, 2025. [Online]. Available: <https://www.infosysbpm.com/blogs/manufacturing/predictive-maintenance-in-manufacturing.html>
11. Deloitte, "Deloitte Survey on AI Adoption in Manufacturing," Deloitte China. Accessed: Feb. 18, 2025. [Online]. Available: <https://www2.deloitte.com/cn/en/pages/consumer-industrial-products/articles/ai-manufacturing-application-survey.html>
12. J. Soldatos, S. Jaber, D. Pike, R. Rao, and J. Hertz, "Building the Factory of Tomorrow: A Comprehensive Guide to Manufacturing Automation," Wevolver. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.wevolver.com/article/building-the-factory-of-tomorrow>
13. U. Nations, "World population projected to reach 9.8 billion in 2050, and 11.2 billion in 2100," United Nations. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.un.org/en/desa/world-population-projected-reach-98-billion-2050-and-112-billion-2100>
14. FAO, "Emissions due to agriculture: Global, regional and country trends 2000–2018," 2020. Accessed: Feb. 18, 2025. [Online]. Available: <https://openknowledge.fao.org/server/api/core/bitstreams/cc09fbbc-eb1d-436b-a88a-bed42a1f12f3/content>
15. M. E. Jarroudi *et al.*, "Leveraging edge artificial intelligence for sustainable agriculture," *Nature Sustainability*, vol. 7, no. 7, pp. 846–854, Jun. 2024, doi: 10.1038/s41893-024-01352-4.
16. M. Padhiary, D. Saha, R. Kumar, L. N. Sethi, and A. Kumar, "Enhancing precision agriculture: A comprehensive review of machine learning and AI vision applications in all-terrain vehicle for farm automation," *Smart Agricultural Technology*, vol. 8, p. 100483, Aug. 2024, doi: 10.1016/j.atech.2024.100483.
17. M. Alam, M. S. Alam, M. Roman, M. Tufail, M. U. Khan, and M. T. Khan, "Real-Time Machine-Learning Based Crop/Weed Detection and Classification for Variable-Rate Spraying in Precision Agriculture," in *IEEE Xplore*, May 2020. Accessed: Feb. 18, 2025. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/9102505>
18. K. Heyl, F. Ekdardt, P. Roos, and B. Garske, "Frontiers," *Frontiers in Sustainable Food Systems*, vol. 7, Jan. 2023, doi: 10.3389/fsufs.2023.1088640.
19. M. Henriques, "The ageing crisis threatening farming," BBC. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.bbc.com/future/bespoke/follow-the-food/the-ageing-crisis-threatening-farming/>
20. J. Halvorson, "2022 Census of Agriculture Impacts the Next Generations of Farmers." Accessed: Feb. 18, 2025. [Online]. Available: <https://www.usda.gov/about-usda/news/blog/2023/02/22/2022-census-agriculture-impacts-next-generations-farmers>
21. P. Periyasamy, "Smart Agriculture with Computer Vision," embedUR. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.embedur.ai/smart-agriculture-with-computer-vision/>
22. O. Dewangan and P. Vij, "Self-Adaptive Edge Computing Architecture for Livestock Management: Leveraging IoT, AI, and a Dynamic Software Ecosystem," in *BIO Web of Conferences*, Jan. 2024. [Online]. Available: https://www.bio-conferences.org/articles/bioconf/abs/2024/01/bioconf_msn-bas2024_05010/bioconf_msn-bas2024_05010.html
23. World Health Organization: WHO, "Ageing and health," World Health Organization: WHO. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/ageing-and-health>
24. World Health Organization: WHO, "Patient safety," World Health Organization: WHO. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.who.int/news-room/facts-in-pictures/detail/patient-safety>
25. H. R. Holman, "The Relation of the Chronic Disease Epidemic to the Health Care Crisis - PMC," *ACR Open Rheumatology*, vol. 2, no. 3, Feb. 2020, doi: 10.1002/acr2.11114.
26. C. Wan, "Biometrics in Healthcare: The Future is Now," Ambiq. Accessed: Feb. 18, 2025. [Online]. Available: <https://ambiq.com/blog/biometrics-in-healthcare-the-future-is-now/>
27. Edge Impulse, "Revolutionizing Fall Detection with Edge AI." Accessed: Feb. 18, 2025. [Online]. Available: <https://edgeimpulse.com/case-studies/revolutionizing-fall-detection-with-edge-ai>

28. H. Singh, A. N. D. Meyer, and E. J. Thomas, "The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations," *BMI Quality & Safety*, vol. 23, no. 9, pp. 727–731, Sep. 2014, doi: 10.1136/bmjqs-2013-002627.
29. S. P. Shashikumar, G. Wardi, A. Malhotra, and S. Nemati, "Artificial intelligence sepsis prediction algorithm learns to say 'I don't know,'" *npj Digital Medicine*, vol. 4, no. 1, pp. 1–9, Sep. 2021, doi: 10.1038/s41746-021-00504-6.
30. R. van der Meulen, "What Edge Computing Means For Infrastructure And Operations Leaders," Gartner. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.gartner.com/smarter-withgartner/what-edge-computing-means-for-infrastructure-and-operations-leaders>
31. J. Roese, "Tech's Big Bang In 2025: AI Is The Spark Igniting A New Era," Forbes. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.forbes.com/sites/delltechnologies/2024/12/04/techs-big-bang-in-2025-ai-is-the-spark-igniting-a-new-era/>
32. S. Sinclair, "Unleashing the Edge: Use Cases, Challenges, and Requirements in Edge Infrastructure and Environments," Enterprise Strategy Group. [Online]. Available: <https://www.techtarget.com/esg-global/research-report/unleashing-the-edge-use-cases-challenges-and-requirements-in-edge-infrastructure-and-environments/>
33. IMARC Group, "Edge AI Market Size, Share, Industry Growth & Report 2033," IMARC. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.imarcgroup.com/edge-ai-market>
34. P. Research, "Edge AI Market Size to Surpass USD 143.06 Billion by 2034," *Precedence Research*, Sep. 18, 2024. [Online]. Available: <https://www.precedenceresearch.com/edge-ai-market>
35. E. Hofstetter, "Autonomous Vehicle Trends Taking Shape in 2024," Innoviz Technologies Ltd - HQ. Accessed: Feb. 18, 2025. [Online]. Available: <https://innoviz.tech/blog/autonomous-vehicle-trends-2024>
36. A. Berkovich, "How Edge Case Detection Enhances AI Safety in Autonomous Vehicles," Akridata • Edge Data Platform for Data-Centric AI. Accessed: Feb. 18, 2025. [Online]. Available: <https://akridata.ai/blog/edge-case-detection-safer-ai-autonomous-vehicles/>
37. C. Wang, Ed., "Examining Smart Driving: How Li Auto is Leading the Future Mobility Revolution," BitAuto Australia. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.bitauto.com/au/news/100196561302.html>
38. S. Rossi, "Autonomous and ADAS test cars generate hundreds of TB of data per day," Tuxera. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.tuxera.com/blog/autonomous-and-ad-as-test-cars-produce-over-11-tb-of-data-per-day/>
39. F. Sideco, "Qualcomm Expecting To 'Flex' Its Automotive Muscles In 2024," Forbes. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.forbes.com/sites/tiriasresearch/2024/01/06/qualcomm-expecting-to-flex-its-automotive-muscles-in-2024/>
40. 5GAA, "5GAA publishes updated Roadmap for C-V2X." Accessed: Feb. 18, 2025. [Online]. Available: <https://5gaa.org/5gaa-publishes-updated-roadmap-for-c-v2x/>
41. E. Hofstetter, "Autonomous Vehicle Trends Taking Shape in 2024," Innoviz Technologies Ltd - HQ. Accessed: Feb. 18, 2025. [Online]. Available: <https://innoviz.tech/blog/autonomous-vehicle-trends-2024>
42. K. Vinoth and P. Sasikumar, "Multi-sensor fusion and segmentation for autonomous vehicle multi-object tracking using deep Q networks," *Scientific Reports*, vol. 14, no. 1, pp. 1–32, Dec. 2024, doi: 10.1038/s41598-024-82356-0.
43. C. E. Staff, "Leveraging edge computing's power in Industry 4.0," Control Engineering. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.controleng.com/articles/leveraging-edge-computings-power-in-industry-4-0/>
44. A. Bala *et al.*, "Artificial intelligence and edge computing for machine maintenance-review," *Artificial Intelligence Review*, vol. 57, no. 5, pp. 1–33, Apr. 2024, doi: 10.1007/s10462-024-10748-9.
45. D. Dhinakaran, S. E. Raja, R. Velselvi, and N. Purushotham, "Intelligent IoT-Driven Advanced Predictive Maintenance System for Industrial Applications," *SN Computer Science*, vol. 6, no. 2, pp. 1–24, Feb. 2025, doi: 10.1007/s42979-025-03695-x.
46. M. Patel, M. Chui, and M. Collins, "The Internet of Things: Catching up to an accelerating opportunity," McKinsey & Company, Nov. 2021.

47. A. Locke and K. Olikara, "Edge Computing's Critical Role in Industrial AI," Rockwell Automation. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.rockwellautomation.com/en-us/company/news/the-journal/ai-at-edge-provides-greater-autonomy.html>
48. Stream Analyze, "Enhancing Manufacturing Efficiency with Edge AI: Stream Analyze Case Study." Accessed: Feb. 18, 2025. [Online]. Available: <https://www.streamanalyze.com/solutions/manufacturing-automated-quality-control>
49. IBM, "Cost of a data breach 2024," IBM. [Online]. Available: <https://www.ibm.com/reports/data-breach>
50. GE Healthcare, "Home Is Where the Heart Is: ECG Device Helps Patients Monitor Cardiac Conditions From Anywhere," GE HealthCare (United States). Accessed: Feb. 18, 2025. [Online]. Available: <https://www.gehealthcare.com/insights/article/home-is-where-the-heart-is-ecg-device-helps-patients-monitor-cardiac-conditions-from-anywhere>
51. "Biobeat Medical - smart vital signs monitoring," biobeat. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.bio-beat.com/>
52. J. de Groot, "What is HIPAA Compliance?," Digital Guardian. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.digitalguardian.com/blog/what-hipaa-compliance>
53. Market Pay, "The Benefits of Self-Checkout: Efficient Customer Journeys, Shorter Queues, and More." Accessed: Feb. 18, 2025. [Online]. Available: <https://market-pay.com/en/blog/the-benefits-of-self-checkout-efficient-customer-journeys-shorter-queues-and-more-2>
54. J. Anglen, "AI-Powered Dynamic Pricing in Retail and E-Commerce," Rapid Innovation. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.rapidinnovation.io/post/ai-powered-dynamic-pricing-in-e-commerce>
55. S. Blog, "How Artificial Intelligence Can Benefit Retail Security," Scylla Resources. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.scylla.ai/how-artificial-intelligence-can-benefit-retail-security/>
56. J. Thomason, "Inside Amazon's new 'Just Walk Out': AI transformers meets edge computing," VentureBeat. Accessed: Feb. 18, 2025. [Online]. Available: <https://venturebeat.com/ai/inside-amazons-new-just-walk-out-ai-transformers-meets-edge-computing/>
57. P. Christiano, "Top 15 Use Cases And Applications Of AI Transforming Logistics In 2025 - ExpertBeacon," Expertbeacon. Accessed: Feb. 18, 2025. [Online]. Available: <https://expertbeacon.com/logistics-ai/>
58. V. Struk, "AI in Logistics: Revolutionizing Supply Chain and Operations," Relevant Software. Accessed: Feb. 18, 2025. [Online]. Available: <https://relevant.software/blog/ai-in-logistics-key-ways-by-which-ai-boosts-the-logistics-industry/>
59. Edge Impulse, "Hyfe: Translating Coughs to Actionable Insights." Accessed: Feb. 18, 2025. [Online]. Available: <https://edgeimpulse.com/case-studies/hyfe-translating-coughs-to-actionable-insights>
60. Edge Impulse, "GlobalSense Drives New Standards in Automotive Diagnostics with Edge AI." Accessed: Feb. 18, 2025. [Online]. Available: <https://edgeimpulse.com/case-studies/global-sense-drives-new-standards-in-automotive-diagnostics-with-edge-ai>
61. Edge Impulse, "Bringing Voice Control to Earbuds & Headsets." Accessed: Feb. 18, 2025. [Online]. Available: <https://edgeimpulse.com/case-studies/bringing-voice-control-to-earbuds-headsets>
62. Aetina, "How Edge AI is Transforming Farming: Future of Smart Farming." Accessed: Feb. 18, 2025. [Online]. Available: <https://www.aetina.com/about-news-detail.php?i=985>
63. "Jetson Modules, Support, Ecosystem, and Lineup," NVIDIA Developer. Accessed: Feb. 18, 2025. [Online]. Available: <https://developer.nvidia.com/embedded/jetson-modules>
64. "QCS8250 5G, Wi-Fi 6 & AI-enabled processor for enterprise IoT and healthcare applications," Qualcomm. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.qualcomm.com/products/internet-of-things/consumer/cameras/qcs8250>
65. "SPOT," Ambiq. Accessed: Feb. 18, 2025. [Online]. Available: <https://ambiq.com/technology/spot/>

66. "Ceva-NeuPro-Nano," Ceva. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.ceva-ip.com/product/ceva-neupro-nano/>
67. "Ceva-NeuPro-M," Ceva. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.ceva-ip.com/product/ceva-neupro-m/>
68. "OpenVINO 2024.6 – OpenVINOTM documentation." Accessed: Feb. 18, 2025. [Online]. Available: <https://docs.openvino.ai/2024/index.html>
69. S. Jaber and J. Soldatos, "Edge AI Technology Report: Generative AI Edition," Wevolver. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.wevolver.com/article/the-guide-to-generative-ai-at-the-edge>
70. "PyTorch ExecuTorch," PyTorch. Accessed: Feb. 18, 2025. [Online]. Available: <https://pytorch.org/executor-torch-overview>
71. Intel, "Next-Level Neuromorphic Computing: Intel Lab's Loihi 2 Chip," Intel. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.intel.com/content/www/us/en/research/neuromorphic-computing-loihi-2-technology-brief.html>
72. F. Ottati, "TrueNorth: A Deep Dive into IBM's Neuromorphic Chip Design," Open Neuromorphic. Accessed: Feb. 18, 2025. [Online]. Available: <https://open-neuromorphic.org/blog/true-north-deep-dive-ibm-neuromorphic-chip-design/>
73. O. Vermesan, V. Piuri, F. Scotti, A. Genovese, R. D. Labati, and P. Coscia, "Explainability and Interpretability Concepts for Edge AI Systems," in *Advancing Edge Artificial Intelligence*, New York: River Publishers, 2024, pp. 197–227. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.taylorfrancis.com/chapters/oa-edit/10.1201/9781003478713-9/explainability-interpretability-concepts-edge-ai-systems-ovidiu-vermesan-vincenzo-piuri-fabio-scotti-angelogenovese-ruggero-donida-labati-pasquale-coscia>
74. L. Moreau, "AI Explainability with Grad-CAM: Visualizing Neural Network Decisions," *Edge Impulse*, Jan. 08, 2025. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.edgeimpulse.com/blog/ai-explainability-with-grad-cam-visualizing-neural-network-decisions/>
75. R. Shamsuddin, H. B. Tabrizi, and P. R. Gottimukkula, "Towards responsible AI: an implementable blueprint for integrating explainability and social-cognitive frameworks in AI systems," *AI Perspectives & Advances*, vol. 7, no. 1, pp. 1–23, Jan. 2025, doi: 10.1186/s42467-024-00016-5.
76. COGNI, "Trustful AI: Transparency, Traceability, and Explainability in Focus - EdgeAI," EdgeAI -. Accessed: Feb. 18, 2025. [Online]. Available: <https://edge-ai-tech.eu/trustful-ai-transparency-traceability-and-explainability-in-focus/>
77. N. T. T. Hung, N. V. T. Khang, C. Q. Hung, T. V. Binh, N. Q. Khanh, and C. Hung, "Enhancing the Fairness and Performance of Edge Cameras with Explainable AI," in *IEEE Xplore*, IEEE, Feb. 2024, pp. 1–4. [Online]. Available: doi: 10.1109/ICCE59016.2024.10444383
78. H. T. T. Nguyen, L. P. T. Nguyen, and H. Cao, "XEdgeAI: A human-centered industrial inspection framework with data-centric Explainable Edge AI approach," *Information Fusion*, vol. 116, p. 102782, Apr. 2025, doi: 10.1016/j.inffus.2024.102782.
79. fiveable, "1.4 Edge AI Ecosystem and Architecture Overview," fiveable. Accessed: Feb. 18, 2025. [Online]. Available: <https://library.fiveable.me/edge-ai-and-computing/unit-1/edge-ai-ecosystem-architecture-overview/study-guide/BhLdeqVzzUPTC4Ee>
80. "Educating the world on Edge AI," Edge AI Foundation. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.edgeaifoundation.org/posts/announcing-the-tinyml-foundation-industry-academia-partnership>
81. "ModelNova - Pre-Trained Edge AI Models:Your Comprehensive AI Guide." Accessed: Feb. 18, 2025. [Online]. Available: <http://modelnova.ai>
82. "Embedded Systems & Edge AI Experts," embedUR. Accessed: Feb. 18, 2025. [Online]. Available: <http://embedur.ai>
83. "Edge AI Partner Enablement Package," Intel. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.intel.com/content/www/us/en/content-details/845320/edge-ai-partner-enablement-package.html>

84. V. Sukumar and R. C. N. do Amaral, "Qualcomm partners with Meta to support Llama 3.2. Why this is a big deal for on-device AI," Qualcomm. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.qualcomm.com/news/onq/2024/09/qualcomm-partners-with-meta-to-support-llama-3-point-2-big-deal-for-on-device-ai>
85. MemryX, "MemryX and Variscite Announce a Partnership to Enhance Edge AI Solutions," *Cision PR News-wire*, Jan. 06, 2025. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.prnewswire.com/news-releases/memryx-and-variscite-announce-a-partnership-to-enhance-edge-ai-solutions-302342491.html>
86. P. Natarajan and A. Szeto, "A new paradigm for partnership between industry and academia," *Amazon Science*, Mar. 23, 2023. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.amazon.science/news-and-features/a-new-paradigm-for-partnership-between-industry-and-academia-in-the-age-of-ai>
87. O. Pyper and A. Zaludaite, "EU Consortium Developing Next-Gen Edge-AI Technologies Is Accepting Design Proposals," *Fraunhofer-Verbund Mikro-elektronik in Kooperation mit den Leibniz-Instituten FBH und IHP*. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.forschungsfabrik-mikroelektronik.de/de/presse--und-medien/Presse/eu-consortium-accepts-design-proposals.html>
88. "National Edge AI Hub - EdgeAI solutions for real industry challenges," EPSRC National Edge Artificial Intelligence Hub. Accessed: Feb. 18, 2025. [Online]. Available: <https://edgeaihub.co.uk/>
89. National Science Foundation, "Democratizing the future of AI R&D: NSF to launch National AI Research Resource pilot," NSF - National Science Foundation. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.nsf.gov/news/democratizing-future-ai-rd-nsf-launch-national-ai>
90. IDC, "IDC FutureScape: Worldwide Digital Infrastructure 2025 Predictions," IDC. Accessed: Feb. 18, 2025. [Online]. Available: https://www.idc.com/research/viewtoc.jsp?containerId=US51665124&_gl=1*gaa3wn*_up*_MQ*_ga*ODY2MTY4NTY1LjE3Mz-k4NzQ2NTk*_ga_541ENG1F9X-*MTczOTg3NDY1OS4xLjAuMTczOTg3NDY1OS4wLjAuMA.*_ga_Y7C-NRMFF6J*MTczOTg3NDY1OS4xLjAuMTczOTg3NDY1OS4wLjAuMA..
91. S. Jaber, J. Soldatos, and R. Rao, "2024 State of Edge AI Report - Chapter 11: The Future of Edge AI," *Wevolver*, Apr. 08, 2024. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.wevolver.com/article/2024-state-of-edge-ai-report/the-future-of-edge-ai>
92. Grand View Research, "Federated Learning Market Size And Share Report, 2030," Grand View Research. [Online]. Available: <https://www.grandviewresearch.com/industry-analysis/federated-learning-market-report>
93. Y. Qiao, Md. S. Munir, A. Adhikary, H. Q. Le, A. D. Raha, and C. Zhang, "MP-Fed-CL: Multiprototype Federated Contrastive Learning for Edge Intelligence," *IEEE Xplore*. [Online]. Available: DOI: 10.1109/JIOT.2023.3320250
94. N. Innan, M. A.-Z. Khan, A. Marchisio, M. Shafique, and M. Bennai, "FedQNN: Federated Learning using Quantum Neural Networks," *IEEE Xplore*. [Online]. Available: DOI: 10.1109/IJCNN60899.2024.10651123
95. S. S. Gill *et al.*, "Edge AI: A Taxonomy, Systematic Review and Future Directions," *Cluster Computing*, vol. 28, no. 1, pp. 1–53, Oct. 2024, doi: 10.1007/s10586-024-04686-y.
96. M. Cole, "Have your cake and eat it, too: Federated learning and edge computing for safe AI innovation," *IAPP*, Jun. 05, 2024. [Online]. Available: <https://iapp.org/news/a/have-your-cake-and-eat-it-too-federated-learning-and-edge-computing-for-safe-ai-innovation>
97. S. P. C., K. Jain, and P. Krishnan, "Analysis of Post-Quantum Cryptography for Internet of Things," *IEEE Xplore*. [Online]. Available: DOI: 10.1109/ICICCS53718.2022.9787987
98. B. Tran, "Quantum Computing's Impact on AI: Training Speeds and Model Efficiency Stats," *PatentPC*, Feb. 10, 2025. Accessed: Feb. 18, 2025. [Online]. Available: <https://patentpc.com/blog/quantum-computings-impact-on-ai-training-speeds-and-model-efficiency-stats>
99. A. Furutanpey, J. Barzen, M. Bechtold, S. Dustdar, F. Leymann, and P. Raith, "Architectural Vision for Quantum Computing in the Edge-Cloud Continuum," *IEEE Xplore*. [Online]. Available: DOI: 10.1109/QSW59989.2023.00021

100. D. Boruga, D. Bolintineanu, and G. I. Racates, "Quantum-enhanced algorithms for real-time processing in cryptographic systems: A path towards post-quantum security," *World Journal of Advanced Engineering Technology and Sciences*, vol. 13, no. 2, pp. 193–204, Jan. 2024, doi: 10.30574/wjaets.2024.13.2.0561.
101. Q. News, "Quantum Computing In Energy: Quantum-assisted Grid Optimization," *Quantum Zeitgeist*. [Online]. Available: <https://quantumzeitgeist.com/quantum-computing-in-energy-quantum-assisted-grid-optimization/>
102. M. Evers, A. Heid, and I. Ostojic, "Matthias Evers," *McKinsey & Company*, Jun. 18, 2021. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.mckinsey.com/industries/life-sciences/our-insights/pharmas-digital-rx-quantum-computing-in-drug-research-and-development>
103. P. Marks, "Humanoid robots can't outsource their brains to the cloud due to network latency," *Ars Technica*, Mar. 02, 2016. Accessed: Feb. 18, 2025. [Online]. Available: <https://arstechnica.com/gadgets/2016/03/network-delays-rule-out-the-cloud-as-an-out-sourced-brain-for-humanoid-robots/>
104. S.-T. Wang, I.-H. Li, and W.-Y. Wang, "Human Action Recognition of Autonomous Mobile Robot Using Edge-AI," *IEEE Xplore*. [Online]. Available: <https://ieeexplore.ieee.org/document/9969627>
105. Y. Liu *et al.*, "Aligning Cyber Space with Physical World: A Comprehensive Survey on Embodied AI," *arXiv.org*. [Online]. Available: <https://arxiv.org/abs/2407.06886>
106. A. K. Ramasubramanian, R. Mathew, I. Preet, and N. Papakostas, "Review and application of Edge AI solutions for mobile collaborative robotic platforms," *Procedia CIRP*, vol. 107, pp. 1083–1088, May 2022, doi: 10.1016/j.procir.2022.05.112.
107. B. Dipert, "NVIDIA Advances Robot Learning and Humanoid Development With New AI and Simulation Tools," *Edge AI and Vision Alliance*. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.edge-ai-vision.com/2024/12/nvidia-advances-robot-learning-and-humanoid-development-with-new-ai-and-simulation-tools/>
108. G. Ombach, "Council Post: Where AI Meets The Physical World: The Rise Of Humanoid Robots," *Forbes*. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.forbes.com/councils/forbestechcouncil/2024/12/30/where-ai-meets-the-physical-world-the-rise-of-humanoid-robots/>
109. FXMedia Team, "The Future of AR/VR: Emerging Trends to Watch in 2025," *FXMedia: Solutions for Metaverse*. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.fxmweb.com/insights/the-future-of-arvr-emerging-trends-to-watch-in-2025.html>
110. P. Pathak, "Happy New Year 2025: Top 5 technology trends to watch, from AR/VR to GTA 6 and smarter AI," *Financial Express*, Dec. 31, 2024. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.financialexpress.com/life/technology-happy-new-year-2025-top-5-technology-trends-to-watch-from-arvr-to-gta-6-and-smarter-ai-3704534/>
111. R. Kumar, "Will AR Glasses Replace Smartphones by 2030," *Industry Wired*, Nov. 02, 2024. Accessed: Feb. 18, 2025. [Online]. Available: <https://industry-wired.com/will-ar-glasses-replace-smartphones-by-2030/>
112. A. Bosworth, "Accelerating the Future: AI, Mixed Reality and the Metaverse," *Meta*. Accessed: Feb. 18, 2025. [Online]. Available: <https://about.fb.com/news/2024/12/accelerating-the-future-ai-mixed-reality-and-the-metaverse/>
113. I. Patringenaru, "A new neuromorphic chip for AI on the edge, at a small fraction of the energy and size of today's compute platforms," *UC San Diego - Jacobs School of Engineering*. Accessed: Feb. 18, 2025. [Online]. Available: <https://jacobsschool.ucsd.edu/news/release/3499?id=3499>
114. W. Wan *et al.*, "A compute-in-memory chip based on resistive random-access memory," *Nature*, vol. 608, no. 7923, pp. 504–512, Aug. 2022, doi: 10.1038/s41586-022-04992-8.
115. D. Kudithipudi *et al.*, "Neuromorphic computing at scale," *Nature*, vol. 637, no. 8047, pp. 801–812, Jan. 2025, doi: 10.1038/s41586-024-08253-8.
116. Intel, "Intel Builds World's Largest Neuromorphic System to Enable More Sustainable AI," *Newsroom*. [Online]. Available: <https://newsroom.intel.com/artificial-intelligence/intel-builds-worlds-largest-neuromorphic-system-to-enable-more-sustainable-ai>

117. E. Insights, "Is Neuromorphic Computing the Future of AI?" Exoswan Insights. [Online]. Available: <https://exoswan.com/is-neuromorphic-computing-the-future-of-ai#h-so-is-neuromorphic-computing-the-future-of-ai>
118. L. M.-E. E. translation by G. Chupakhin, "How A Neuromorphic Chip Might Shape the Future of Industry," *Abbe Center of Photonics*, Feb. 18, 2025. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.acp.uni-jena.de/5225/how-a-neuromorphic-chip-might-shape-the-future-of-industry>
119. A. Vitale, E. Donati, R. Germann, and M. Magno, "Neuromorphic Edge Computing for Biomedical Applications: Gesture Classification Using EMG Signals," *IEEE Xplore*. [Online]. Available: doi: 10.1109/JSEN.2022.3194678
120. H. Akbar, "BrainChip Podcast: Neuromorphic Computing Shaping the Future of AI With Dr. Jason K. Eshraghian," *Wevolver*, Jul. 16, 2024. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.wevolver.com/article/brainchip-podcast-epi-33-neuromorphic-computing-shaping-the-future-of-ai-with-dr-jason-k-eshraghian>
121. "Confidential Computing," IBM. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.ibm.com/confidential-computing>
122. "EDGE AI FOUNDATION," EDGE AI FOUNDATION. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.edgeaifoundation.org/events>

Image Sources

- [i] P. Research, "Edge AI Market Size to Surpass USD 143.06 Billion by 2034," *Precedence Research*, Sep. 18, 2024. [Online]. Available: <https://www.precedenceresearch.com/edge-ai-market>
- [ii] "Siemens erweitert mit Industrial Edge die Feldebene um die Vorteile der Cloud," Siemens. Accessed: Feb. 19, 2025. [Online]. Available: <https://press.siemens.com/global/de/pressemitteilung/siemens-erweitert-mit-industrial-edge-die-feldebene-um-die-vorteile-der-cloud>
- [iii] "Smart Farming Revolution: How Edge AI is Reshaping Agriculture," Aetina Corporation. Accessed: Feb. 19, 2025. [Online]. Available: <https://www.aetina.com/application.php?t=64>
- [iv] H. Zhao, "An edge streaming data processing framework for autonomous driving," *Connection Science*, vol. 33, no. 2, Jun. 2020, doi: 10.1080/09540091.2020.1782840.
- [v] F. Sideco, "Qualcomm Expecting To 'Flex' It's Automotive Muscles In 2024," *Forbes*. Accessed: Feb. 18, 2025. [Online]. Available: <https://www.forbes.com/sites/tiriasresearch/2024/01/06/qualcomm-expecting-to-flex-its-automotive-muscles-in-2024/>
- [vi] J. KÖİVOGUL, "Artificial Intelligence (AI) and Predictive Maintenance," Oct. 19, 2023. Accessed: Feb. 19, 2025. [Online]. Available: <https://www.linkedin.com/pulse/artificial-intelligence-ai-predictive-maintenance-jean-ko%C3%AFvogui/>
- [vii] M. Tactic, "Future proofing smart manufacturing with edge AI," *Latent AI*. Accessed: Feb. 19, 2025. [Online]. Available: <https://latentai.com/blog/future-proofing-smart-manufacturing-with-edge-ai/>
- [viii] Stream Analyze, "Enhancing Manufacturing Efficiency with Edge AI: Stream Analyze Case Study." Accessed: Feb. 18, 2025. [Online]. Available: <https://www.streamanalyze.com/solutions/manufacturing-automated-quality-control>
- [ix] M. Saifuzzaman, T. N. Ananna, M. J. M. Chowdhury, M. S. Ferdous, and F. Chowdhury, "A systematic literature review on wearable health data publishing under differential privacy," *International Journal of Information Security*, vol. 21, no. 4, pp. 847–872, Jan. 2022, doi: 10.1007/s10207-021-00576-1.

- [x] J. Jenkins, "Amazon's Just Walk Out technology just got smarter—here's what's new," *US About Amazon*, Jul. 31, 2024. Accessed: Feb. 19, 2025. [Online]. Available: <https://www.aboutamazon.com/news/retail/amazon-just-walk-out-improves-accuracy>
- [xi] S. Srivastava, "IoT in Supply Chain and Logistics – Benefits, Use Cases & Challenges," Appinventiv. Accessed: Feb. 19, 2025. [Online]. Available: <https://appinventiv.com/blog/iot-in-logistics-and-supply-chain-management/>
- [xii] J. Renaz, "Harvesting Innovation: The Evolution of Smart Agriculture through Technology Integration," Jan. 06, 2024. Accessed: Feb. 19, 2025. [Online]. Available: <https://www.linkedin.com/pulse/harvesting-innovation-evolution-smart-agriculture-through-joel-renaz-kgovc/>
- [xiii] M. Ashouri, P. Davidsson, and R. Spalazzese, "Quality attributes in edge computing for the Internet of Things: A systematic mapping study," *Internet of Things*, vol. 13, p. 100346, Mar. 2021, doi: 10.1016/j.iot.2020.100346.
- [xiv] "Ambiq: Super-low Power Semiconductor for IoT with AI," Ambiq. Accessed: Feb. 19, 2025. [Online]. Available: <https://ambiq.com/>
- [xv] "Ceva - Leading Licensor of Innovative Silicon and Software IP Solutions," Ceva. Accessed: Feb. 19, 2025. [Online]. Available: <https://www.ceva-ip.com/>
- [xvi] A. Ming, "Everything about TensorFlow Lite and start deploying your machine learning model," Latest Open Tech From Seeed. Accessed: Feb. 19, 2025. [Online]. Available: <https://www.seeedstudio.com/blog/2022/05/08/everything-about-tensorflow-lite-and-start-deploying-your-machine-learning-model/>
- [xvii] Y.-K. Wong, "How the Edge Enables Groundbreaking AI Applications," ABI Research. Accessed: Feb. 19, 2025. [Online]. Available: <https://www.abiresearch.com/blog/edge-ai-applications>
- [xviii] "Press Release," Synaptics. Accessed: Feb. 19, 2025. [Online]. Available: <https://www.synaptics.com/company/news/synaptics-and-google-collaborate-on-edge-ai-for-the-iot>
- [xix] A. Furutanpey, J. Barzen, M. Bechtold, S. Dustdar, F. Leymann, and P. Raith, "Architectural Vision for Quantum Computing in the Edge-Cloud Continuum," *IEEE Xplore*. [Online]. Available: DOI: 10.1109/QSW59989.2023.00021
- [xx] S. Alvarez, "Tesla shows off impressive Optimus Gen 2 humanoid robot," TESLARATI. Accessed: Feb. 19, 2025. [Online]. Available: <https://www.teslarati.com/tesla-shows-off-optimus-gen-2-humanoid-robot-video/>
- [xxi] "Qualcomm Launches Snapdragon AR2 Designed to Revolutionize AR Glasses," Qualcomm. Accessed: Feb. 19, 2025. [Online]. Available: <https://www.qualcomm.com/news/releases/2022/11/qualcomm-launches-snapdragon-ar2-designed-to-revolutionize-ar-gl>
- [xxii] I. Patringenaru, "A new neuromorphic chip for AI on the edge, at a small fraction of the energy and size of today's compute platforms," UC San Diego - Jacobs School of Engineering. Accessed: Feb. 18, 2025. [Online]. Available: <https://jacobsschool.ucsd.edu/news/release/3499?id=3499>

About the Authors

Samir Jaber, Editor-in-Chief

Leipzig, Germany

Samir Jaber is an editor, writer, and industry expert on topics of technology, science, and engineering and is the editor-in-chief of the Edge AI report series with Wevolver. Samir is the Chief Editor and Founder of Wryters, a writing and consulting agency. He has comprehensive experience working with Fortune 500 companies and industry leaders as a writer, editor, and content manager, leveraging his technical background in mechanical engineering, nanotechnology, and scientific research. Samir is also a featured author in 30+ industrial magazines with a focus on Artificial Intelligence (AI), the Internet of Things (IoT), 3D printing, Autonomous Vehicles (AV), nanotechnology, materials science, and sustainability. His experience includes award-winning engineering research and patented engineering design in the fields of nanofabrication and microfluidics.

John Soldatos, Co-author

Athens, Greece

Honorary Research Fellow at the University of Glasgow

John Soldatos holds a Ph.D. in Electrical & Computer Engineering from the National Technical University of Athens (2000) and is currently an Honorary Research Fellow at the University of Glasgow, UK (2014-present). He was Associate Professor and Head of the Internet of Things (IoT) Group at the Athens Information Technology (AIT), Greece (2006–2019), and Adjunct Professor at the Carnegie Mellon University, Pittsburgh, PA (2007–2010). He has significant experience working closely with large multi-national industries (e.g., IBM, INTRACOM, INTRASOFT International) as an R&D consultant and delivery specialist while being a scientific advisor to various high-tech startup enterprises. Dr. Soldatos is an expert in Internet-of-Things (IoT) and Artificial Intelligence (AI) technologies and applications, including IoT/AI applications in smart cities, finance (Finance 4.0), and industry (Industry 4.0).

Deval Shah, Co-author

Adelaide, Australia

Deval Shah is a Senior Machine Learning Engineer at the Australian Institute for Machine Learning (AIML), where he specializes in designing Retrieval Augmented Generation (RAG) systems for enterprise search and policymaking applications. His focus on scalability and secure data handling underscores his expertise in production-ready AI solutions. Prior to AIML, Deval advanced from Machine Learning Engineer to Senior Software Engineer at Uncanny Vision, spearheading projects that enhanced license plate recognition models and vehicle re-identification microservices. Deval has collaborated with over 15 AI companies, published more than 100 articles, and successfully built substantial organic inbound channels for AI startups through his content-driven initiatives. His skill set spans full-stack development (Next.js, FastAPI), containerization (Docker), and cloud infrastructure (AWS, Azure DevOps).

About the Partner

Edge AI Foundation

Los Altos, CA

EDGE AI FOUNDATION, the global place of innovation, collaboration, advocacy, and education for energy-efficient, affordable, and scalable edge AI technologies, from tinyML to the edge of AI. We unite diverse industry leaders, researchers, and practitioners to drive collective progress and achieve breakthroughs that solve the world's biggest challenges. We're here to bring together researchers, developers, business leaders, and policymakers to tackle the big challenges in AI, from low-power machine learning to advanced edge computing. We believe in a future where trillions of intelligent devices work together to create a healthier, more sustainable world.

www.edgeaifoundation.org



About the Sponsors

embedUR systems

Software development
San Ramon, CA

An expert in AI, Edge Computing, IoT, Networking and Cloud, embedUR has leveraged its time-honed, bleeding-edge and unique capabilities to develop complex products and solutions across verticals. Our solutions have been deployed in millions of devices across the world and benefit a wide enterprise including a global Fortune 500 customer base.

Our vision is to create truly intelligent edge devices that can adapt to dynamic environments with precision and make real-time decisions efficiently, thereby bringing transformative changes in how edge devices perceive and interact with the world.

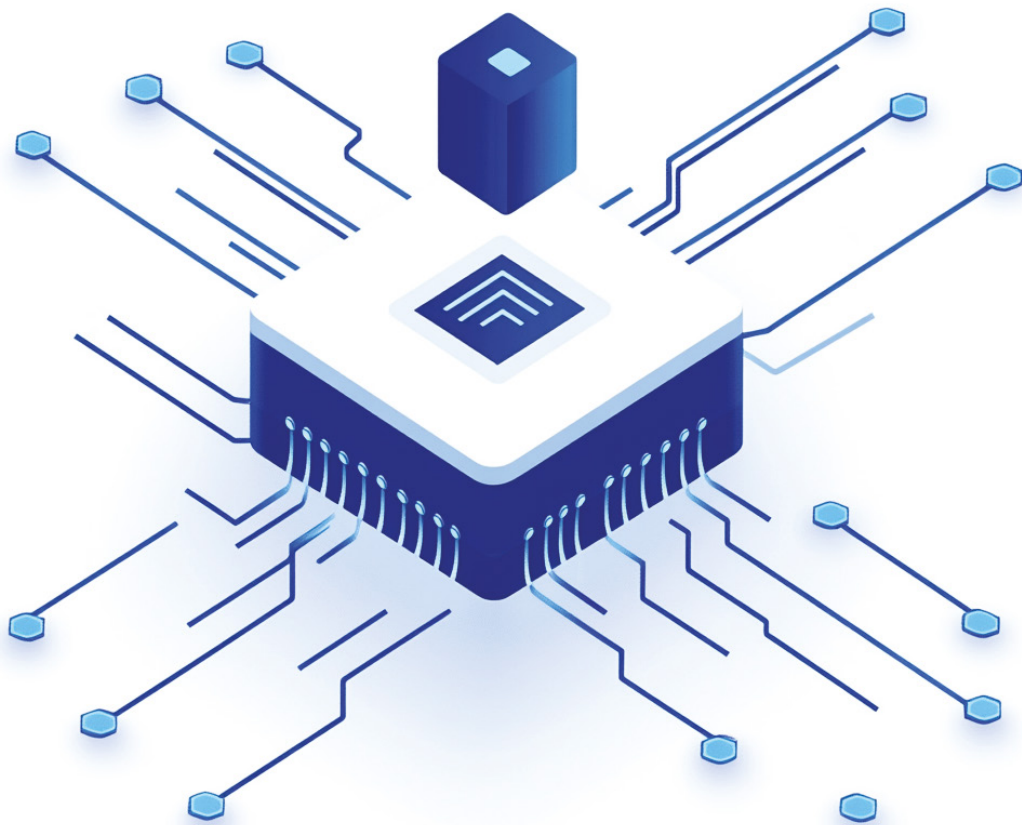
ModelNova is a groundbreaking model zoo for Tiny Devices with pre-trained AI models, optimized for different software frameworks and a variety of low-power hardware platforms with and without native AI acceleration. This innovative platform streamlines Edge AI product development, enabling rapid prototyping and deployment of intelligent applications on edge devices, in a fraction of the time it used to take to develop Edge AI solutions from scratch.

ModelNova addresses a significant challenge faced by engineers and developers: the complexity of selecting, creating, training, porting and optimizing AI models for different hardware platforms, especially low-

power IoT devices. This is a critical process, upon which the success of a project largely depends. It requires extensive expertise in low-level firmware, plus a deep understanding of the silicon they are trying to optimize for, which is a rare skill set. ModelNova bridges this gap by offering a diverse collection of pre-trained models for AI vision, speech, sound and more, which have already been fine-tuned to run on multiple combinations of hardware and software architectures.

www.embedur.ai
www.modelnova.ai





Ambiq

Semiconductors
Austin, TX

Our mission is to enable intelligence (artificial intelligence (AI) and beyond) everywhere by delivering the lowest power semiconductor solutions. We enable our customers to deliver artificial intelligence compute at the edge where power consumption challenges are the most profound. Our technology innovations, built on the patented and proprietary sub-threshold power optimized technology (SPOT), fundamentally deliver a multi-fold improvement in power consumption over traditional semiconductor designs. We've powered over 260 million devices today.

Ambiq is the only semiconductor company to successfully commercialize and scale subthreshold power in their integrated circuit designs, resulting in an unprecedented 10X-13X reduction in energy consumption.

During design, testing, and manufacturing, semiconductor companies avoid subthreshold power operations due to highly volatile and unpredictable variables such as temperature, process, voltage, etc. Ambiq has overcome all these challenges by continuously perfecting SPOT to maximize energy efficiency in the last 14 years, resulting in unmatched power savings.

Wearable manufacturers, for instance, can use Ambiq's microprocessors to add more complex features, such as enriched graphic display, data sensing and real-time analysis, advanced inferencing, always-on listening or speech detection, and more without draining the battery. This creates instant value for smart device manufacturers and their consumers, which is why Ambiq is in more than 70% of the top 10 major smartwatches and fitness bands.

Outside of wearables, Ambiq's SPOT platform can be used in industrial, medical, smart homes, gaming, and AR/VR applications, to name a few. As the world moves towards using AI to enhance our everyday lives, moving this intelligence to the edge is vital to transforming our future into a more data-driven and intelligent world.

www.ambiq.com



Advanced A



Edge Impulse

Machine Learning
San Jose, CA

Edge Impulse streamlines the creation of AI and machine learning models for edge hardware, allowing devices to make decisions and offer insight where data is gathered. Edge Impulse's technology empowers developers to bring more AI products to market, and helps enterprise teams rapidly develop production-ready solutions in weeks instead of years. Powerful automations make it easier to build valuable datasets and develop advanced AI for edge devices from MCUs to CPUs to GPUs. Used by health and wearable organizations like Ultrahuman, industrial organizations like Halma as well as top silicon vendors and over 150,000 developers, Edge Impulse has become the trusted ML platform for enterprises and developers alike.

edgeimpulse.com



Axelera AI

Semiconductor Manufacturing
Eindhoven, Netherlands

Axelera AI is providing the world's most powerful and advanced solutions for AI at the edge. Our game-changing Metis™ AI platform – a holistic hardware and software solution for AI inference at the edge – enables computer vision and generative AI applications to become more accessible, powerful and user friendly than ever before.

The core of our platform is our Metis AI Processing Unit (AIPU), which is based on proprietary digital in-memory computing technology (D-IMC) and RISC-V controlled dataflow technology. The AIPU offers industry-leading performance, usability, and efficiency at a fraction of the cost of existing solutions. Our technology is scalable and outperforms any other startup or incumbent.

Developed from the ground together with our Metis AIPU, our click-and-run Voyager SDK software stack allows developers to easily integrate inference acceleration on Metis into their AI pipeline. It also comes with a model zoo of pre-compiled, industry standard models to help developers get started quickly, straight out of the box.

The Voyager SDK automatically quantizes and compiles neural networks that have been trained on different frameworks – so you don't need to retrain – generating code that runs on the Metis AI platform with industry-leading accuracy. Optimized networks running on Metis AIPU are indistinguishable from those running on systems with floating-point units.

Our technology is integrated into AI acceleration cards (PCIe and M.2

form factors), boards, and inference-ready systems. This enables small to medium-sized enterprises to speed up adoption and streamline field installation. Our Metis systems are available today and already deployed at a number of customers.

www.axelera.ai

AXELERA
ARTIFICIAL INTELLIGENCE

Brainchip

Computer Hardware Manufacturing
Laguna Hills, California

BrainChip is the worldwide leader in edge AI neuromorphic processing and learning. The company's first-to-market neuromorphic processor, Akida™, mimics the human brain to analyze only essential sensor inputs at the point of acquisition, processing data with unparalleled efficiency, precision, and economy of energy. Keeping machine learning local to the chip, without the need to access the cloud, dramatically reduces latency while improving privacy and data security. In enabling effective edge compute to be universally deployable across real-world applications, such as connected cars, consumer electronics, and industrial IoT, BrainChip is proving that on-chip AI close to the sensor is the future for customers' products as well as the planet.

www.brainchip.com

brainchip
Essential AI

Synaptics

Semiconductor Manufacturing
San Jose, California

Synaptics is driving innovation in AI at the Edge, bringing AI closer to end users and transforming how we engage with intelligent connected devices, whether at home, at work, or on the move. As a go-to partner for forward-thinking product innovators, Synaptics powers the future with its cutting-edge Synaptics Astra™ AI-Native embedded computer, Veros™ wireless connectivity, and multimodal sensing solutions. We're making the digital experience smarter, faster, more intuitive, secure, and seamless. From touch, display, and biometrics to AI-driven wireless connectivity, video, vision, audio, speech, and security processing, Synaptics is the force behind the next generation of technology enhancing how we live, work, and play.

www.synaptics.com

 **synaptics**®

Ceva

Semiconductor and Software IP
Rockville, Maryland

For more than three decades Ceva has been a trusted provider of technology that enables people and electronics systems to interact in seamless, secure, and increasingly more intuitive and predictive ways. Ceva's semiconductor and software IP offerings are used by the world's top semiconductor and electronics companies to develop extraordinary and differentiated products that connect, sense, and infer - the three critical pillars of the rapidly evolving era of AI-enabled Smart Edge. Ceva's solutions enable a new generation of connected and distributed intelligence to make our lives safer, enjoyable, and more efficient.

Powering more than 19 billion devices Ceva has established leadership in reliable and secure wireless connectivity for use in both infrastructure and end points; low-power and highly efficient audio and vision sensing and interface solutions; and scalable neural-network-based AI processing. Ceva is committed to its customers' success and continues to drive innovation in smart connected systems in the AI era.

www.ceva-ip.com

 **CEVA**®

Ambient Scientific

Semiconductor Manufacturing
Santa Clara, California

Ambient Scientific is a fabless semiconductor company pioneering AI computing technologies to unlock and power the next generation of AI possibilities from the edge to the cloud. Our Analog In-Memory Compute technology, called DigAn®, revolutionizes AI compute by eliminating the need for a tradeoff between performance, efficiency and flexibility to deliver higher than ever AI performance at orders of magnitude lower power consumption. The flexibility of DigAn® ensures these advantages scale from cloud level infrastructure efficiencies to portable micro-edge AI possibilities.

GPX10, our first processor built on the DigAn® architecture, enables AI applications once deemed impossible in battery-powered devices, such as always-on voice detection and FaceID, all while consuming minimal power, small enough to run on a coin cell battery.

With a full-stack SDK supporting standard AI frameworks, a custom neural network compiler and training toolchains to enable hassle free data collection for training AI models, we empower developers to seamlessly deploy AI at the edge, unlocking new possibilities in ultra-low-power computing.

www.ambientscientific.ai



About Wevolver

Wevolver is a global platform and community that provides engineers with the knowledge and connections to develop better technology.

We bring a professional audience of engineers informative and inspiring content, such as articles, videos, podcasts, and reports, about state-of-the-art technologies.

The knowledge on Wevolver is published by various sources: universities, tech companies, and individual community members. Next to that, we manage a network of over 50 technical writers who create content for our customers and publish that on Wevolver.com

Millions of engineers leverage Wevolver to stay up to date, find knowledge when they are developing products, and leverage the platform to make meaningful connections.

Wevolver has won the SXSW Innovation Award, the Accenture Innovation Award, and the Top Most Innovative Web Platforms by Fast Company. Wevolver is how today's engineers stay cutting edge.

wevolver.com

