Advances in Consumer Research
ISSN (Print): 0098-9258
ISSN(Online): 3079-1766

Vol. 2, Issue 3 (2025) https://acr-journal.com/

Architecting Agentic AI for Real-Time Autonomous Edge Systems in Next-Gen Mobile Devices

Goutham Kumar Sheelam[1]

[1]IT Data Engineer, Sr. Staff.

Email ID: gouthamkumarsheelam@gmail.com

ORCID ID: 0009-0004-1031-3710

Cite this paper as: Goutham Kumar Sheelam, (2025) Architecting Agentic AI for Real-Time Autonomous Edge Systems in Next-Gen Mobile Devices. *Advances in Consumer Research*, 2 (3), 589-604.

KEYWORDS

Agentic AI, Edge Computing, Sensor Fusion, Mobile Devices, AI Acceleration, Intrinsic Motivation Learning, Human-Centered AI, Real-Time Processing, Trustable AI, Semantic Understanding, Human Agency, Intelligent Edge, Predictive Analytics, Perception-to-Action, Socially Aware AI, Alignment Technologies, Secure Elements, Knowledge Engineering, Task Automation, Next-Generation AI Systems.

ABSTRACT

Advances in semiconductor nano-fabrication technologies have enabled the creation of next-generation mobile devices that are sensory-rich, communication-convergence platforms at the edge of the internet. These highly sophisticated handheld devices are equipped with increasing computational capabilities, ultra-low-power acceleration for AI processing, ultra-high-resolution imaging and video sensors, and secure elements for trusted sensor fusion, and are naturally always-on and always-connected. The fusion of a constellation of these devices supporting real-time predictive and prescriptive analyses of users and their dynamic environment will ultimately change the future of work and enable a modern era that empowers smarter, safer, healthier, and more productive human lives.

In order to realize the full potential of these devices, there is an urgent need to develop next-generation agentic AI and the required alignment technologies, including intrinsic motivation-driven learning, impact-driven perception-to-action systems, complex data and knowledge engineering for human-structured worlds, and trustable interactive intelligent behavior and social exchanges. These alignment technologies will enable the cognitive and physical automation stacks to architect these devices and larger scale intelligent edge environments that truly comprehend and seamlessly integrate the understanding of users and their real-world interaction contexts, the rich and dynamic semantics embedded in users' goals, intentions, and needs, and the importance of human agency and safety in any real-time autonomous task-preparation and task-execution process. In this chapter, we will present a thesis to develop the underlying tenets and architecture of agentic AI and their necessary technologies through a set of carefully curated thinkers and practitioners' perspectives on this challenge in a workshop format

1. INTRODUCTION

The next generation of mobile devices will become pervasive platforms for interactive intelligent applications that take advantage of on-device AI accelerating architectures, sensors and internet connectivity. Whether these mobile devices will be touring backup sensors for remote data centers or standalone repositories for localized and generalized intelligent agents will depend upon the efficiency of the autonomy enabled architectures and on-device AI middleware. Edge systems will leverage on-device AI utilizing data from mobile device activities. However, mobile device edge systems are characterized by two operative constraints that restrict the classes of applications they can support. The first is related to demands for real-time availability of their services. The demand-oriented augmentation of the providers' user experience roles with their active user-dependent or proactive user-regulatory roles requires that all the processing of the combination of user data streams and application functionality be handled on-device. The second is the availability of limited resources for the operation of the

edge system service providers and end point communicators.

The paper argues that to address the two operative constraints imposed on edge systems, edge AI services should architect and implement agentic AI that is capable of autonomously opportunistically adjusting to the shape of its user data availability stream and thus to the spatio-temporal re-distribution of the AI processing demands on mobile devices. By architecting agentic AI for the user-activity data streams, the available resources can be securely used for battery draining, data privacy and latency-, as well as task performance-related applications. Section 2 outlines the opportunity area, Section 3 discusses the implementation of the agentic AI use control, Section 4 concludes.

1.1. Purpose and Scope of the Study

In this chapter, we introduce the scope of the study. Contemporary mobile devices have an increasing number of sensors that have made them capable of studying the human user and the context in which its actions are framed. As a result, this has made it possible to develop systems that reside in mobile devices and that combine machine learning and human activity analysis for many applications. System-as-a-service applications have been developed to automate and make human-centered the exploitation of prior knowledge for activity recognition. Decentralized and peer-to-peer systems are helping to solve the challenges related to privacy and use of resources. However, even in this new paradigm of computing, no serious effort is being made to synthesize mobile computing systems of testimonial quality capable of providing real-time performance with device autonomy.

Furthermore, no effort is made to reach high levels of quality for Artificial Intelligence (AI) since the AI itself is based on training examples that are not representative of a user's model or are the result of an imprecise transfer of knowledge from some expert annotators. As a result, AI develops knowledgeable predictions that directly influence decision making that are not agentic or that lack behavioral fidelity. We describe in this chapter how using Agent-based Systems and Mobile Platform as a Service enable the service production of intelligent mobile-located systems as demand-conditions products for human cognitive support and delegation. This research on mobile Platform as a Service enables clustering and heterogeneous connection of mobile devices in the form of decentralized edge or fog computing. Mobile Platform as a Service enables decreasing the time-to-market of systems and produces as a service to synthesize quality intelligent systems based on machine learning.

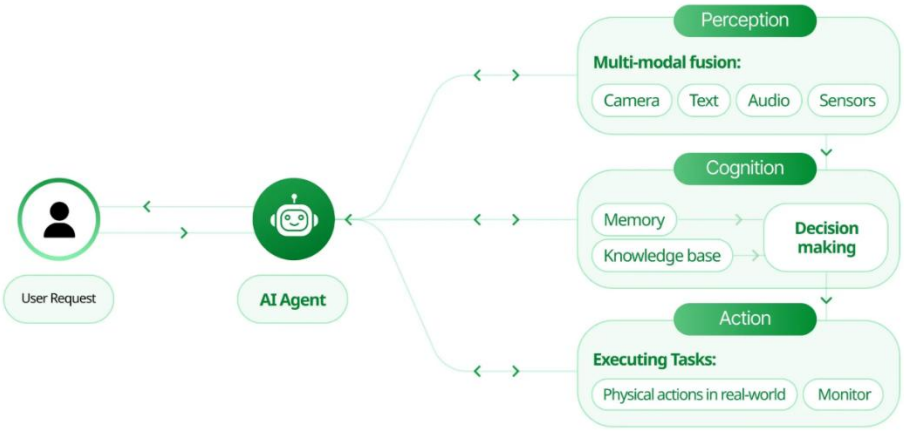

Fig 1 : Agentic AI Architecture

2. BACKGROUND

1. Overview of Agentic AI

Agentic AI refers to highly autonomous artificial systems that are capable of goal-directed action in the real world to further their own goals or intents, provided these are aligned with human values. Examples of Agentic AI that have been used only on a fairly narrow basis include recommendation engines or autonomous stock trading systems. With the advent of breakthrough innovations in Large Language Models, which can serve as autonomous agents that rapidly and inexpensively perform complicated logical, creative, and cognitive tasks, there is a heightened urgency to explore and characterize the architecting decisions for stronger and more agentic AI. Transferred into real-world application domains, there are no longer just simple toy PMT tasks but a large range of more difficult, more idealized social PMT tasks, which at least some LLMs have been observed to solve at human-level or superhuman-level capability.

2. Real-Time Systems in Edge Computing

Next-generation mobile devices are becoming powerful computing edges seamlessly integrated with real-world dynamics through high-bandwidth sensors, and yet they continue to be resource-constrained compared to fixed-edge systems or cloud

services. In this sense, mobile devices serve as low-cost realizations of the Edge Computing architecture. By executing intelligent applications at the mobile device, EAI can provide real-time responses to how the person and the surroundings are changing and thus afford a more interactive experience for the user. More importantly, the mobile device performs computation closest to the primary actuators in the PMT paradigm, viz, the person and their surroundings, and thus removes any possible delay in back-and-forth communications with a distant remote service provider. Performing EAI in real-time also increases privacy for the user.

Equation 1 : Agentic Response Time Function: $$R_t = \frac{C_d + M_p}{E_c}$$

R_t = Total response time of AI agent

C_d = Context detection latency (sensing + signal processing)

M_p = Model prediction time (inference duration)

E_c = Edge computing capacity (operations per second)

2.1. Overview of Agentic AI

In our work, we propose the paradigm of agentic intentionality accuracy awareness (AIAWA) for model performance assessment, enhancement, and user assistance. In other words, ourselves, or perhaps our "intentioned" descendants, trusted to AI will be empowered to judge and respond to AI's performance. Specifically, agentic AI, intentionally and with the user's approval or modeling of the user's potential intent, may provide a range of performance assessments including: confidence, uncertainty, likeliness, accuracy, explainability, whether correctable or improvable, other types of performance feedback and guidance, as well as recommendation or influence over user behavior. Trust is an essential quality for collaborative human-agent systems. Low model performance degrades user trust, resulting in reduced adoption and consequently rendering the systems impotent. Addressing model performance in user assistance is thus an essential part of advancing agentic AI research. Existing large language models provide state-of-the-art answers to a range of distinct natural language tasks from unfamiliarity and with little further input. Unlike other AI systems or even LLMs performing familiar tasks, however, LLMs do not yet possess the requisite attributes to instantiate agentic AI. Specifically, these generative foundation models are not either both aware of user intent or able to explicitly align their intended behavior with user intent. AIAWA incorporates two additional components. Unique to AIAWA, a model agent's ability or effort to align with user intent across tasks simultaneously must also be verifiably assessable, while also being adaptable to the context of each individual task.

2.2. Real-Time Systems in Edge Computing

Questions about how long it takes software to perform a task, how much work is done in that time, and how to provide guarantees for its quality are equally important as correctness and efficiency in embedded systems. Efficiency means that resources should be used to achieve some desired goal; there must be a degree of frugality in the use of the platform's capabilities. Soft real-time system design is concerned with the efficiency of resource use, while hard real-time system design is concerned with the guarantees provided about the use of time. Conventions about how hard and soft real-time systems are used by both hardware and software reduce any description into a set of schedules or performance profiles, and the use of performance testing. Hard real-time system guarantees completion technologies, or bounding description techniques to software module pairings is called predictable execution. This is done by keeping results the same for every input, or not relying on other parameters to get values for execution time, or providing asymptotic bounds based on input length or size. These bounding techniques are also used for scheduling by only scheduling for the longest timing, and the theme of these techniques is to provide schedules that are accomplishable for any input instance.

The characteristics of the histograms and profiles of times of the use of resources allow measurement-based guarantees, but the traditional measurement bases have to be modified, because classical performance measures are expected actual times and speedups. Agentic AI requires on remote devices posing as the end points of agentic communications chain the guarantees required of actionable sensing, for which agentic AI modules. Work is accomplished on these devices as clients typically of edge data processing systems. Patterns of Sensing Activity and the Guarantees Required for Their Processing Yield Domain-Specific Protocols. Implicit in the first two kinds of risk for agentic AI are functions from timing to controlling action success. Responsibility involves time, and requires returns on investment, whether of perfect action cascading via agents or using agents as public resources. These constants, function parameters and timings, are needed for the creation of these realistic models.

2.3. Next-Gen Mobile Device Architecture

Mobile device architecture has undergone significant changes to address an increasing requirement for demanding applications, including serious games, mobile augmented/virtual reality and AI. Mobile processors today are true heterogeneous manycore systems. A typical one comprises power-efficient Cortex A5-A15 families of processors along with

higher-performance Cortex A15-A78 families of processors. Different types of commercial automotive are architected as heterogeneous System on Chip with a traditional GPU as part of SoC and Dedicated/Programmable AI accelerator such as Neural Processing Unit for AI computations. Present-day SoCs employed in mobile devices are architected to aid taxing compute and communication requirements. A dedicated AI accelerator such as GPU or ISA/architecture programmable NPU aid fast AI computations required for such intelligent AR/MR applications. Although these NPU accelerators have been mainly proposed and used for AI computations, they could also be used and programmed for other kernel types, such as computer vision, image processing, etc.

Next-gen edge mobile devices could be architected with more efficient usage of available memory, such as layers of storage hierarchy across device memory, DRAM, and flash. Parallel read/write across and along the storage hierarchy could be arranged to reduce latency. Additional device memories corresponding to diversity in access locality for various apps could be architected. For example, video based surveillance toward detecting people, animals and vehicles are procured. Sensor-centric tracking is used to reach thresholds. SAR/MAR applications involve tracking of users in the switch from SAR to MAR modes. Fast tracking responds as a critical edge computation. SAR tasks could be distributed over query volume rasterized grid squares for computation on low-latency computing systems including edge servers and cloudlets distributed across grid boxes.

3. THEORETICAL FRAMEWORK

Discussions around agent-based systems dealing with and within human environments have existed for quite a while now. The intellectual roots of the definitions at the boundary between the artificial and natural systems, and those traveling in the surface domain, were first laid down by early theorists. During the subsequent decades, many proposals focused on decision theoretic, heuristic and neuro-inspired architectures. Yet, the majority of the theoretical efforts in rigorously and comprehensively exploring the topic were mainly based on a few intellectual pillars, notably the philosophical exploration of causality, economic game theoretical models, complemented by the exploration of rationality, logical and computability principles, the interactive-centricity, and the evolutionary principles studied by early thinkers. Fortunately, nowadays many impressive bridging concepts relying on sound computation and information theoretical principles are now within reach. This is leading to the establishment of more precise definitions, outcomes and avenues to explore in the creation of AI agents at all levels of agency and scalability.

Moreover, recent advances in the understanding of individual and collective entropic decision-making offer a new avenue to directly formalize intelligent, autonomous, and agentic decision-making under uncertainty, through the micro-macro, or the dual macro-centric and micro-centric views. Certainly, devising agentic AI models capable of real-time decision-making may result in a daunting task, berating existing solutions.

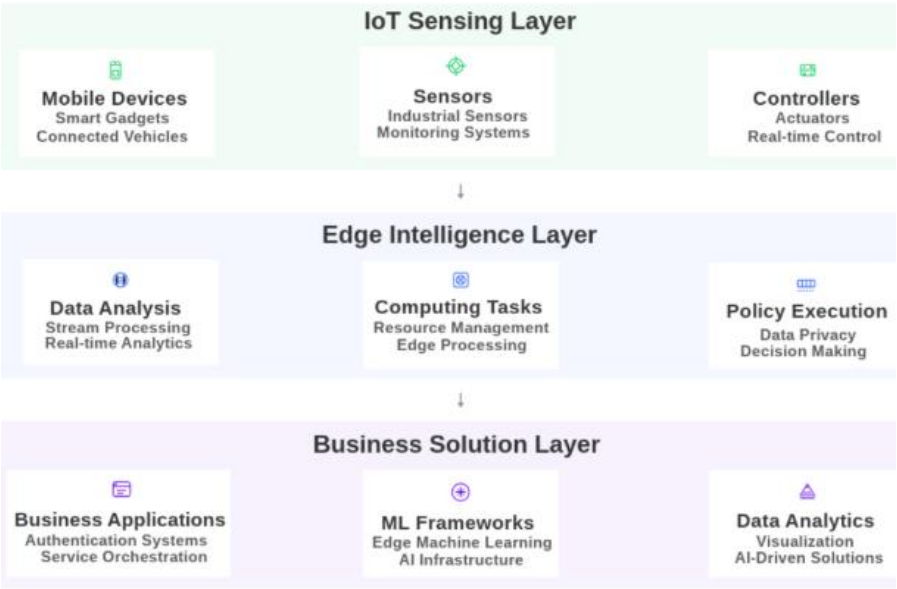

Fig 2 : Edge AI with Agentic AI for Distributed Intelligence

3.1. Conceptual Models of Agentic AI

In order to enable agentic capabilities onto real-time autonomous AI, it is essential to formulate operationally measurable definitions, clear ontological models, and reliable decision-making mechanisms. Below, we outline the gaps in existing AI conceptual models and the state-of-the-art in cognitive architecture for intelligent behavior, and propose an extended

taxonomy of agentic capabilities and a multi-level functional architecture for agentic AI. Conceptual Models of AI Many definitions of AI or other terms describing its specific subdomain are proposed. These definitions represent an ontological model of AI, typically formulated according to some set of criteria. The most common models are defined according to the level of abstraction or the type of approach. Broadly, the AI definitions that are proposed can be classified into the following models: 1st level: Tasks, Methods, Diagnosis, Problem-solving, Symbolic, Interaction, Embodied, Cognitive Remote Interface; 2nd level: Intuition, Symbolism, Embodiment, Submission, Cognition, Social.

2nd level models are conceptually defined in accordance with abstraction levels of various AI tasks: from the most abstract description of high-level opportunities performed by a human or a machine up to the low-level criteria for concrete AI methods and tools. 1st level definitions are formulated on the foundations of human definitions - a set of symptoms or attributes that are common both for humans and AI. Smart Environments, aware of intentions and other inner states of the users, provide a medium for interaction, where intelligent agents can jointly adapt the surrounding world, affect plans of other agents and have a synchronised perception of the environment.

3.2. Real-Time Decision Making in AI

Real-time decision making refers to the various computational processes that result in, or facilitate, a decisive action performed by an intelligent agent in a real-world environment. In edge systems, these decision making processes are constrained in that they need to be implemented in continuous time (or near-continuous time). For example, an intelligent camera needs to perform or facilitate its getting, processing, and analyzing video frames to automatically detect and classify people within the video stream in real-time or near-real-time. While modern computer vision and video processing applications can come close to this practical deadline, they are typically executed in small batch mode, thereby missing the constraints of real-time edge decision making. Similarly, many natural language processing tasks and systems are designed to operate in non-real-time batches or are not architected in a manner that is compatible with being continuously online. In fact, the online and, especially, real-time nature of task performance differentiates intelligent agent systems from machine learning systems. Modern AI systems, such as chatbot wizards, with their invitation to continuous interaction across time bridges the gap, albeit loosely, with traditionally defined agent systems, while NLP fields like sentiment analysis would be at loss.

A number of various AI design and implementation frameworks claim to bridge the gap between non-agentic, non-real-time executing AI systems and the intention of continuous, real-time operation of true intelligent agents. These include the architectures designed for intelligent image, video, and speech processing. However, the intent and role of enabling an intelligent agency feeling and behavior is only recently becoming better focused in the words of researchers.

4. DESIGN PRINCIPLES

Any effort to develop on-device agentic AI must navigate trade-offs between AI capabilities on the one hand, and the limited size and capacity of the device on the other. Downloading externally trained models is impractical, as their size is at least one order of magnitude larger than device memory. Furthermore, the current paradigm of centralized cloud-based AI systems has led to the widespread belief that only those AI models learned and parameterized in the cloud can be useful within mobile edge devices, and that AI model learning is a one-time process performed in the cloud. In contrast, the goal of our architecture, design principles and building blocks for agentic AI is to make the hosting device the central locus for both inference and model learning. We thus envision a new paradigm of on-device scalable, flexible, customizable, robust, reliable, and energy-efficient agentic AI services learned on-device continuously from limited user-generated interaction data via live collaboration, correcting, and customizing with the user in the loop.

We present the design principles behind our proposed on-device agentic AI architecture. These principles emphasize the challenges of developing compact AI models that are both general and specialized, how to build AI services that work in all contexts and scenarios for the individual user, and how users collaborate seamlessly with the AI service over its lifetime. The principles guide us to define design principles for agentic AI on mobile devices: model scalability and flexibility; necessitating certain abstractions and assumptions; collaboration with the user; robustness and reliability; and energy efficiency.

4.1. Scalability and Flexibility

The architectural scalability of agentic AI for the edge requires it to be fast, category-level, few-shot, or zero-shot learnable, and responsive in real time to truly dynamic input demands. Agentic AI will live within these devices for months or years at a time. New functionalities or uses may need to be deployed almost instantly through software upgrades. But some of their new roles and calls upon the computational and bandwidth resources of devices will be primarily determined by the devices' function as an integral part of larger systems or wholes, which may themselves be changing in real time and naturally larger-scale dynamic agents. Achieving an appropriately high-function cognitive degree-of-freedom, whether at the first or the higher levels in the architecture, while maintaining full flexibility and low-overhead responsiveness to the devices' users must be a central architectural design aim of agentic AI.

What cannot be achieved in the higher cognitive levels must be accomplished in the level-0 processing, management, and preparation pipeline stages of the perceptual flows themselves. Devices must learn, as flexibly and quickly as human executive function, to map higher levels of their resource management capability through a variety of hardware and software modes to the priority and quality-of-service demands transmitted in their higher-level user-directed functions, at previously learnt higher cognitively-limiting and non-limiting categories, classes, profiles, or types of functions. It is a function of the adaptive dimensionality management of their scalable, lower-level, perceptual-to-user-direct-functional demand pipelines and information and bandwidth-handling subsystems.

4.2. Robustness and Reliability

A robust and reliable agentic AI vehemently reacts against any emerging harmful consequences of its operations and seeks to show reverential deference and respect for the user's values, and for the security, health, and safety of the users and the society in which it operates. These properties undergird the emergence of intelligence as an emergent property of life. The core word "robustness" traces its origin from the Latin word meaning "oak" or "strength." In the context of natural life, robustness means not only the ability to withstand variations, noise, faults, conflicts, and perturbations in the environment, but also the ability to recover from such adversity and failures, and the innate strength of the dispositions and principles that underlie those abilities. The related word reliability means something even deeper. It means constitutive faithfulness. This understanding is priopractic, or tacit, autonomous, and agentic in kind, made manifest through its faithful and reliable operation in the world.

A normally reliable artificial autonomous agent — under normal circumstances, and in normal context — faithfully acts in the best interest of its principal. While principal-agent relations are mostly beneficial, they can also be harmful, as exemplified by doctors, lawyers, and investment advisors who, motivated by greed or fear, counsel their clients to heedlessly engage in harmful, illegal, and unreasonable deals and contracts. This suggests that inequitable or unethical principles are a source of unreliability. Another avenue of inquiry is whether the duties embedded in agentic AI systems can be deliberately breached — for example, when the AI is hacked to inflict harm, such as guiding drone swarms armed with weapons toward hospitals or directing autonomous vehicles toward crowds to cause fatal accidents. There are also ways AI systems can be used as unreliable tools that unpredictably aid human decision-making. Such AIs are not reliable trustworthy tools for strategic decision-making, and thus not agentic AI systems in the sense of the phrases we are using here.

4.3. Energy Efficiency

Energy has grown to become a crucial resource in large-scale parallel computing leveraging thousands of low-power chips, due to the shift to mobile devices and edge computing. Low energy consumption is a basic requirement for most applications for a number of reasons. First, often mobile devices are powered by batteries, limiting the weight, size, and charging frequency; and warm batteries deteriorate energy efficiency over time. Second, warm chips lead to shortened lifespans and, in the worst cases, failures of high-demand devices due to the unpredictability of thermal cycling on chip longevity. Third, the gradually increasing energy needs associated with increased capabilities and usages pose great challenges for energy planning and utilization management at both the individual level and the societal level. Energy modulation circuits, while complicated, allow changing the voltage and frequency of execution – which can help accelerate execution while remaining energy neutral. Unfortunately, these circuits cannot change energy levels by the desired orders of magnitude and can be considered more of a refinement than a true means of changing energy use.

Modern mobile devices are connected to increasingly high-speed wireless networks, allowing the offloading of tasks that are not time-sensitive to remote data centers, such as facial recognition and machine translation. However, the dependency on high-speed wireless networks and availability of remote or edge data centers can create practical or financial constraints. Cloud computing and edge computing are increasingly offering a much greater pool of computing resources in order to address the difficulties in executing sophisticated tasks in real time, but their desynchronization for individual devices and systems can lead to unpredictable latency. Agentic AI systems that develop efficient low-level computation modalities are especially useful for these increasingly common scenarios.

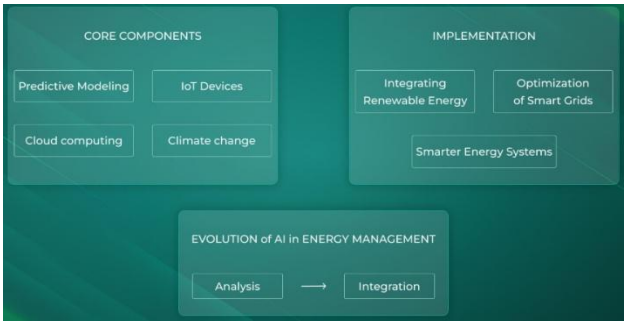

Fig 3 : AI Agents Cut Energy Costs Effortlessly

5. ARCHITECTURE OF AUTONOMOUS EDGE SYSTEMS

The unique aspect of Autonomous Edge Systems (AES) for Next-Gen Mobile devices is that the intelligence and policy of decision making occur on-device and in real-time. It implies that all the interference with the environment occurs primarily through the on-device sensors without involving any outside infrastructure. AES is capable of monitoring everything in its physical view and even with its physical hearing for accomplishing the mission purpose whatever is during both the mission life cycle and inter-mission lifecycle. Autonomous Edge Systems (AES) for Next-Gen Mobile devices are designed for tasking and processing data in autonomous mode for purposes like: reconnaissance, surveillance and mapping for tactical, operational, and strategic purposes, hostile and friendly entities detection for defence and intelligence operations, smart city safety, monitoring, control, and management, mission management for human in the loop cognition, prediction, decision, action, effect, and assessment.

The attribute multi-sensor sensor fusion based on time-domain 2D arrays and 3D volumes is the characteristic future enabling technology important for detecting hostile objects during day-night visual, infrared, and radar light detection; analyzing presence of explosive, chemical, and biological materials with on-board environmental sensors; and accurately geo-localization different friendly and enemy ground, airborne, and maritime vehicles using cooperative communication protocols. Such a capability is crucial for wide-area motion imagery generation for long-range, medium-range, and short-range areas of interest. It is considered essential technology for creating future large-scale autonomous earth observing missions with which applications can perform simultaneously geolocation and processing of data being applied. The accurate and robust sensor processing, big data geo-processing and fusion, and real-time autonomous decision making are critical enabling technologies for future advanced AES.

Equation 2 : Autonomous Decision Confidence Score:

$$A_c = \frac{\sum_{i=1}^{n}(P_i \cdot W_i)}{T_c}$$

where:

- A_c = Agent's decision confidence
- P_i = Probability output from each decision layer
- W_i = Layer weight based on priority (e.g., safety-critical > routine)
- T_c = Threshold calibration factor for real-time decisions

5.1. System Components and Interactions

We consider a loosely-coupled, decentralized system with heterogeneous computing and intelligence resources comprising both user end-devices and cloud services. User devices continuously sense physical world conditions in real-time and process data using local or edge-resident optimization-based AI algorithms or models for the specific user tasks at hand. These computationally simple apps operate in a transient, opportunistic manner, relying on a cadence of focus-in – act – focus-out, thereby assisting the user without being an intrusive distraction. Independent, contextual interactions within and external to the device(s) manage active data pipelines and support services that may need further processing, potentially at a greater degree of effort or competence than that performed locally. For example, camera and microphone-based sensors in mobile devices, automotive vehicles or smart surveillance systems detect the presence of a person or object in their field of view, recognize its identity or importance, and conduct interaction-focused processing through semi-autonomous, gaze and talk-process pipelines that frequently invoke task-specific subroutines running in the cloud. The cloud-enabled processing augments local capability in both content generation and analysis, performing the inference or metric, coordination and communication tasks that support user devices. For example, the cloud models and stores a taxonomy of objects and events accumulated through past interactions; it also performs gaze-tracking, spatial and speaker localization to index generated content. These auxiliary tasks, such as identifying fused-image centroids of appearing and disappearing faces, provide the point-specific focus of gaze at each image frame instant for use by local inference-based apps, augmenting the performance or function of the user-enabled device.

5.2. Data Processing Pipelines

In order to act autonomously, an agentic AI must first interpret the continuous sensor data streams -- percepts -- coming from its environment into discrete representations. The absence of significant labels, coupled with the need for real-time processing, especially for life-critical systems, vastly differs the processing of percepts from processing the batch data in current ML systems. The interpretation output must support real-time decision making, which, in conjunction with the real-time throughput constraints, places unique requirements on the pipeline architecture, design, and optimization.

Current edge processing architectures are not optimized for real-time online learning, especially when dealing with percept streams from live, real-world sensory organs. The majority rely on common data processing primitives such as frame differencing, optical flow, and edge detection, which are unsuitable for estimating higher-level tasks; or GIS methods that by themselves are inadequate representations of AI environments. Even though there have been some advanced video

analytics pipelines, most current mobile pipelines are still optimized for traditional media compression and playout use cases. Most pipelines are built assuming a zero-latency and predictive model for depth estimation, relighting, and compression. This requires new advances in computer graphics and improved understanding of depth and reflectance properties of objects, including differentiating albedo, specularity, map damage, and others.

5.3. Communication Protocols

The nature of the tasks assigned to mobile devices informs key design choices for the communication protocols between their components. SIMs run highly adaptive programs, the behavior of which is data-driven and designed for solving complex but typical tasks. DSLs run programs that are less adaptive, designed for generic task classes instead. Typically, the low-level devices produce data at high rates, in real-time. The upper-level devices execute domain algorithms on such data, which are driven by the experiences built on critical decision safing and feedback loops. As a result, the frequency of the lower-level device to upper-level device communication is expected to be very high. Furthermore, the frequency of device-task loops depends on the overall efficiency of the system, especially on the efficiency of the upper-level device. A particular feature of edge systems is that they operate collaboratively to solve common task objectives. Consequently, there are collective protocols for sharing low-level information and task objectives in real-time across the collaborative devices. The efficiency, responsiveness, and adaptability of the system directly depend on the choice of the protocols.

While protocol choices such as WiFi, Bluetooth LE, 802.15.4, and Zigbee are available for typical applications, they are not nearly sufficient for the use cases we have in mind. Real-time applications require different methods, so custom communication protocols are necessary to realize what is possible in terms of sensing, perception, cognition, action, and collaboration, operations are not accomplished solely by sensor actuation and motion but, first and foremost, embedded cognition, perceptions of the current state of the world, predictions of future world states, and analysis and judgment providing an estimate of the required actions that lead to desirable world states.

6. AI ALGORITHMS FOR REAL-TIME PROCESSING

Real-time intelligent decision-making is a necessary component of agentic autonomous systems. The miniaturization trends towards mobile devices give growth to new applications and services that demand real-time processing of heterogeneous multimodal data streams characterized by nonlinear stationary and non-stationary bursts. They include all AI-based systems tasked with on-device Edge AI processing leveraging on-board sensor data streams. Portable infrastructures also help to accomplish these tasks in a decentralizing fashion. As well as privacy and trustworthiness, they concomitantly address safety and security. A key enabler in this sense is the algorithmic and architectural investment in embedded-AI.

Embedded AI algorithms must be lightweight yet effective, capable of low-latency decision-making without relying on off-device data exchange. Convolutions, recurrent and graph networks, memory-augmented networks, and transformers adapted over recent shifts in information theory as well new hardware accelerator design implemented in programming frameworks become main architectural aspects related to designing efficient on-device algorithms. Quantization-aware training procedures and generative diffusion models also allow the squeezing of training and inference energy budgets, though training still needs to be conducted on hot-edge infrastructures. The unique perspective of agentic AI as embodied intelligence operating in open worlds thus diversifies and complexifies agent-enhancing algorithms that will operate in Edge-AI devices. This creates new avenues in self-improving, self-supervised algorithms built from behavioral scaffolding operating at the intersection of reinforcement learning, theory of mind-driven algorithms, and imitation.

Real-time performance, on the other hand, is a crucial requirement for every single Edge-AI application and thus also demands inference stickiness enhancements and latency-killing design choices. Real-time robustness-enhanced strategies are the most challenging planning space feature requiring the convergence and consistency of multiple AI system components involved in sensory-motor loops of different time constants. They range from quick visual reactions coupled to audition distance learning assisting paths through prolonged sensorimotor interactions to navigation functions for route planning and dynamic obstacle avoidance of longer durations.

Fig 4 : AI Agent Architecture

6.1. Machine Learning Techniques

Machine learning originated from the study of pattern recognition and computational learning theory in artificial intelligence, and has more recently been modernized by the study of statistical learning theory. The computerized systems learn how to perform tasks such as classification, prediction, and clustering from the data they receive, through experience rather than via

explicit programming. Machine learning has since generated a growing number of highly successful applications and has been incorporated into the critical infrastructure of the companies that are shaping the development of the web.

Machine learning techniques have been effectively utilized for various real-time inference tasks in video analytics like video coding, object detection, saliency prediction, action recognition, and event/object tracking. Popular applications of video forensics are also getting augmented by these techniques. Broadly seen, modern video analytics techniques have distinct characteristics, such as abstraction, domain transfer, motivation, and diversity. Various machine learning techniques except for deep learning are also being actively pursued that utilize handcrafted features rather than data-driven features.

Machine learning methods implicitly model the statistical distribution of data. These techniques are designed primarily for supervised or unsupervised learning and for predefined models for inference.

6.2. Reinforcement Learning Approaches

RL uses rewards to guide the agent to the optimal policy by maximizing the cumulative cost function. The underlying principle of RL is that the action is selected based on the environment state, and the agent interacts with the environment by carrying out actions, receiving feedback on the improvement of its actions by the cost or reward, and eventually converging to a policy maximizing the expected cumulative reward. RL models an agent, states, actions, action-value, and policies with Q-functions. The Q-function is updated based on the estimated return on the current observation, and the error between the observed return and the update is used to sample experienced actions from an experience replay. The problem of finding the optimal policy is the basis for RL algorithms. The policy defines the way the agent selects its actions to maximize the reward; in deterministic policy gradient methods, the policy is a function and is determined deterministically. On-policy methods update the current policy using new experiences while it is still active. Off-policy methods base learning upon experiences collected previously by either the current policy itself or a different policy, which is more sample efficient. Model-free methods allow the agent to learn directly from the action experiences; model-based methods require that the agent has prior knowledge of the environment and simulated internal dynamics models for planning future experiences based on this prior knowledge.

RL has matured in the last decade, with the success of several advanced frameworks. RL policies can learn a variety of tasks through maximization of the cumulative rewards, driving the agent to learn parametric functions that efficiently steer the mobile robot using the least number of iterations. In our mobile robot applications, we trained the RL agents, which availability uniformly and randomly explored the task space. The training also involved simultaneous generation of roadmaps for the multi robot group to traverse to gather the visual samples which will form the basis for categorizing plant species.

6.3. Neural Network Architectures

The past decade has seen remarkable developments in deep neural networks (DNNs), offering very effective solutions for numerous problems in computer vision, speech recognition, language translation, video analysis, and many other key applications in human-AI collaboration. Recently introduced architectures, such as recurrent neural networks, long short-term memory networks, conditional random fields, deep residual networks, inception networks, recurrent convolutional neural networks, hyper networks, neural sequence-to-sequence architectures, generative adversarial networks, stacked denoising autoencoders, and many others have demonstrated remarkable capabilities in their respective solution domains, including solution accuracy. DNNs have also significantly outperformed traditional machine learning solutions based on hand-crafted features. Feature engineering based on domain-expert knowledge has been replaced with a simple, uniform approach in DNN solutions, which learn the required features in an automatic fashion from massive training datasets.

Importantly, DNNs have fundamentally transformed classic expert systems into intelligent systems with a human-like ability to learn from examples, enabling several application areas that had remained elusive for decades. Going beyond just application accuracy, the past few years have also witnessed rapid progress towards very efficient DNN implementations that run effectively on mobile devices. Techniques such as quantization, pruning, model compression, deep hashing, binarization, knowledge distillation, transfer learning, and others have made it feasible to deploy certain classes of mobile DNN solutions that operate in real-time and exhibit high throughput for large user communities, all while consuming low power and battery. Specifically for mobile devices, it is particularly desirable to adopt modest-sized, lightweight DNNs that can perform above average while maintaining optimized performance, especially when required for collaborative processing in real-time edge systems.

7. IMPLEMENTATION CHALLENGES

Many of the key attributes of agentic AI, such as fast sub-second response times and sophisticated interaction patterns that go beyond those supported by chatbots today, require extensive off-device access to the vast stores of data and knowledge that live within the institutional systems of the users. Incorporating remote access into every interaction makes supporting low-latency interactions difficult, especially in environments with limited bandwidth and high packet loss. Even when real-time connectivity can be assured, the practicalities of mobile interactions encourage a model in which many questions are posed and tasks are requested across arbitrary time intervals by mobile users. Prompting and responding to voice questions

is gratifyingly fast during a moment of connectivity, but these one-off inquiries are relatively rare compared to the quality and lull of daily life. Mobile devices are often on intermittently for short bursts, and the time that elapses between pressing a button and the voice of the device speaking back can seem like forever. Pushing more work to the user and demanding longer interaction times for faster responses places greater cognitive load on the user. A user who trusts their device to reliably do the work for them will be able to perform those complex tasks less often, allowing for a greater volume of simple and repetitive interaction, which in turn allows for increasingly rapid voice response.

User data—especially images, geolocation information, and audio—that is being processed device-side is both a prime target for attackers and a treasure trove of valuable information about the user. Processing any of this data device-side requires the client-side agentic services to remit the information necessary to generate task-relevant conversations to the cloud-side agentics system. The threat of interception en route encourages encrypting request data with keys tied to the remote task and users or others similarly situated. Even with high security of this user data, should the unusual possibility hold that the agent in question secedes responsibility from the remote agent systems, the task still raises privacy concerns that will need to be addressed.

7.1. Latency and Bandwidth Limitations

The emergence of the Internet of Things (IoT), which supports billions of connected smart devices, is expected to increase the demand for high-throughput, low-latency data transmission generated by those devices to accelerate the development of several consumer market segments, enterprise-based industries, and edge-centric novel applications. The capabilities of AI-enabled edge devices to provide real-time insights such as facial or speech recognition require seamless communications between mobile and edge servers to exchange information about model weight updates, training loss, uploaded raw sensor data, inferences, stored models, etc. Achieving the expected levels of ultra-low latency and massive capacity expansion requires building next-gen edge networks so that transmission across the edge layers has minimum impact on the working of time-critical real-time closed-loop applications and AI capabilities or features in wireless edge systems. With available energy-efficient model training/learning solutions, advanced federated AI models, and wireless communication protocols that ensure reliable data exchange across edge layers, the AI-enabled systems can efficiently collaborate with mobile devices and cloud servers to provide required AI capabilities for next-gen applications.

Due to bandwidth constraints imposed by mobile devices, the amount of uplink data that can be transmitted is low and temporary. Real-time applications that use AI capabilities on the edge for time-critical assistance, such as health monitoring or autonomous driving, will not succeed unless sufficient mobile device data throughput is ensured. With current separate radio access network architecture, uploading massive volumes of data consumed by AI systems in closed-loop applications such as mixed-reality AR requires an extraordinarily high wireless bandwidth. This poses constraints especially for time-critical closed-loop human-AI collaboration where time synchronization is required to ascertain intervention decisions. For resource-constrained mobile devices, uploading huge volumes of sensor data also requires energy-hungry hardware designs if sufficient edge data transmission performance is to be achieved.

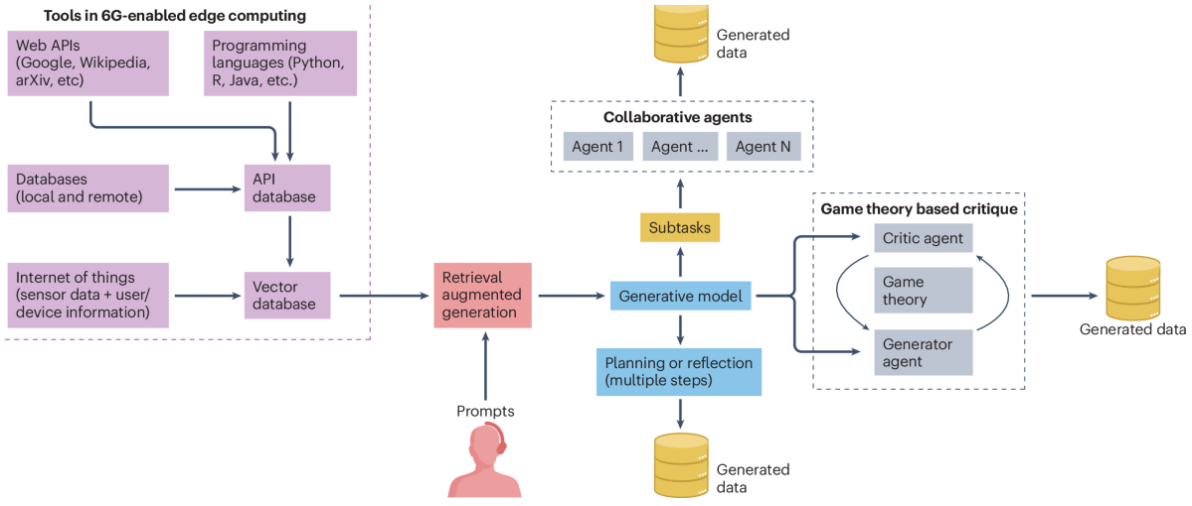

Fig 5 : AI reliability via agentic AI in 6G-enabled edge computing

7.2. Security and Privacy Concerns

Most edge-centric AI implementations and next-gen devices are conceived to rely on centralized assets, including ML models, virtualization services, user history data repositories, etc. This puts these assets, and the user privacy, at risk of cyberattacks and misuse. Hackers are constantly developing new methods to satirize current cybersecurity protections, which could lead to catastrophic hacker and malware activities when targeting mobile devices. They could violate user privacy by

creating incorrect responses from compromised models or bypassing the protections of deployed devices. These scenarios could jeopardize the life of compromised device users, when actions taken as the result of compromised model responses are harmful, such as failing to restrict travelling to unsafe regions, failing to indicate when a criminal should be arrested, etc. If not detected and solved on time, these situations could lead to heavy financial losses or emotional distress.

Privacy concerns arise in scenarios where users rely on AI capabilities that utilize sensitive data locally stored in edge devices for processing, creating, or storing responses. These scenarios not only comprise lifestyle and health diagnostics, but also personal discussions and data shared by users with their social group, banking and subscription account management, unique preferences, emotional status and undercover discussions, etc. Mobile edge devices are ever-connected, always-on, intimate companions, which hold a wealth of sensitive personal data and discussions. Hence, privacy is a critical parameter for deploying AI applications on mobile edge devices. Sensitive data included in the conversations and thoughts of users that will enter the completion input could be over-exploited. These concerns might hamper the edge processing of sensitive input data. Users might not trust AIs deployed in these situations, which could confuse the responses or not deliver the responses that users expect. Building trust in the usage of security mechanisms is an essential aspect to enforcing both security and privacy in mobile edge intelligent agents.

7.3. Integration with Legacy Systems

Seamless integration with existing software systems across safe, understandable APIs, language mappings, and understanding of existing user workflows, is essential for the general usefulness of agentic AI. Systems for agentic AI will be used across almost every domain of human activity. In many cases, users will be employing agent-based AI within systems that rely on pre-existing software, long-standing organizational infrastructures and skill sets, or regulated business practices such as finance or healthcare. If these integration challenges cannot be addressed, we can expect real-time agentic AI to be confined to increasingly limited solution niches, or relegated to backend development and system design tasks and custom integrations. Any significant success in automating user activities in non-agentic AI systems is likely to be user against the same kinds of systems and day-to-day user tasks that in many cases, agents mediated via external interfaces, account for the bulk of user-computer interactions.

Unanticipated conflicts with existing systems raise risks not present in other kinds of software. These risks create potential problems with user knowledge, safety, and trust, and they create legal and regulatory issues around the deployability of these systems. These risks are especially prominent in business and financial markets where computer networks, and especially AI tools that aid in trading, are heavily regulated because of the financial, social, and reputational costs associated with breakdowns. Although this indicates that AI middleware will be very helpful in the management of application layer interactions, and improved communication and transparency help with many important issues, we believe that real-time, deeply integrated AI will necessarily imply deeper agent-to-agent interactions between diverse stakeholder systems larger in footprint and in functionality than systems modeled in existing high trust, low risk domains such as pathing for robotics or CAD/CAM systems.

8. EVALUATION METRICS

To evaluate Architecting Agentic AI for Real-Time Autonomous Edge Systems in Next-Gen Mobile Devices, we need to verify that these systems can indeed emulate their agentic dynamics. This verification is done checking them against the criteria set for Agentic AI: Autonomous Environment Agnosis, Capability Sensory-Motor Synergy Agents, Purposeful Practical Intentionality, Time-Process Real-Time Optimization Efficiency, and Self-Mapping Holding Efficiency. The sensing and processing bulk, reactivity, and processing latency values for these metrics are also collected. The synthesis of the specific metric for each requirement is done using the next subsections.

Performance Metrics

We also need to keep track of the performance of the systems, since they will present some effect on the user experience but in a different domain than that of the requirements. A number of performance metrics will be used, namely:

• Latency: the time going from stimulation to reaction (the sensing latency, the processing latency, and the reactivity).

• Bulk: the amount of resources used for the sensing and processing.

• Costs: we call costs the energy and economic costs incurred by the entire process.

These metrics have to be collected in an emotional sense. The sensors used for these metrics may be singularly assigned Agentic AI tasks but in shorter straight lines, so as to minimize the effect on the agentic systems being monitored.

User Experience Metrics

User experience is a subjective activity, to be appraised with tests but we can also enlist some metrics like:

• Consistency: how much is the experience similar for different users?

• Naturalness: wide variety in the response repertoires and their construction based on the individual users' history.

• Fun: is it fun using the system or is it perceived only as practical? Do the autotaskings give enough time for getting fun?

8.1. Performance Metrics

Define characteristics that distinguish a specific AI implementation or agent type. Typically defined in terms of the AI controlling the simulated or actual world; variations across sub-characteristics considered. Quality of results include optimality, quality, novel / creative, veridicality / grounding, reliability / robustness, speed, safety, cooperation, scalability, always on. Allows life cycle comparison, but difficult to balance.

Semantic agents with sensory perceptual models can recognize instances and predict appearances of self and others at varying distances and scales across time relative to scene flow of salient entity actions. Such agents can infer the actions, goals, properties of other entities, associate object categories to their motions, perceive which elements are in contact with the social agent and where, recognize social interaction signs and trace longer-term interactions across time, ground relevant symbolic internal states in the sensory perception, build visual world models, and create multimodal representations of perceptual conversational events. Robust, tolerant to transient disturbances, perceptually related to each other, canonical for symbols representing perceived and conceptual communicative acts, meaningful with respect to the pursuing conversation.

8.2. User Experience Metrics

User experience is the primary reason for architecting CA2 for constrained execution environments, but the metrics for determining user experience have to be vetted based on the needs of the agent's application domain. If the primary function of the agent is to optimize performance on a task, then some form of performance metric will be appropriate. These are typically cognitive or perceptual. While the central question is then whether the CA2 is more or less efficient than an equivalent implementation of the application, the inconvenience caused by buffering while the device is "down" for CA2 inference, or user unfriendliness might be more appropriate in some applications, particularly in those with high user immersion. Finally, the appropriate QoE metric may be the amount of CPU/GPU time saved by the CA2 since, for some applications, user experience is not considered "improved" or "enhanced," but rather, rendered possible by the existence of the CA2.

These considerations may also apply to a CA2 being used to assist other CPUs and GPUs. The relative advantage of the CA2 in terms of cognitive and/or perceptual speedup has to be very high to make it worthwhile to devote a portion of the available platform TDP and battery energy to its operation, unless the task is one that gets significantly worse when performed by the CPU/GPU duo, such as rendering cinematic quality graphics. In such cases, the user experience would be much improved were the CA2 to perform its service, and this service is not currently available in such situations.

8.3. Cost-Benefit Analysis

The derivation of cost-benefit analysis for various resource constraints in mobile devices for the architected agentic AI real-time edge computing for intelligent autonomous systems, at the individual and societal scales, is of paramount importance to foster technology diffusion and deployment at scale. First, we have limited battery, thermal, and other hardware resources going into mobile devices. The mobile device has limited power and thermal budget; hence, not all the workloads for the AI execution are appropriate to execute in the mobile device. For example, the power and thermal consumption while performing multiple detections, tracking, and predictive analytics for an autonomous robot to navigate is significantly more than when simply doing mapping for exploration or visual SLAM. Hence, it is obvious that individual socio-economic cost-benefit analysis needs to be done before performing triggering processing in mobile. That is particularly true for mobile-assisted real-time autonomous intelligent systems.

Next, the cost-benefit analysis at the societal scale for diffusion of various capabilities needs to also be performed, especially for ubiquitous implementation for edge computing of various workloads. For instance, certain safety-critical scenarios like fire and gas leak detection in large-scale infrastructures like smart buildings and smart cities using visual and thermal cameras could be feasible incentives to both society and end user for implementing mobile-assisted edge computing for the collaborative capabilities, which can result in large-scale development and deployment of mobile-assisted collaborative AI capabilities. Such application-specific agentic AI deployment for intelligent real-time autonomous systems will help in creating generative, positive feedback cycles for implementing intelligent edge capabilities for real-time applications across all sectors in society.

9. FUTURE DIRECTIONS

Tremendous advancements in generative Artificial Intelligence (AI) systems – i.e., Large Language Models (LLMs), generative visual foundation models, reasoning AIs and beyond – have created a wealth of ongoing and new research. However, all of this research remains to be truly impactful and transform lives if not successfully translated into interesting real-world applications and systems. While LLMs are indeed general-purpose tools that enable higher levels of creativity and productivity on the nodes, we currently completely lack the ability to democratically leverage and coordinate the powers of a multi-agent system of LLMs in truly creative and productive ways. To this end, we envision the following future directions will be critical in realizing agentic AI systems on next-gen mobile systems.

We strongly believe that mobile Multi-Agent Systems (MAS) powered by LLMs and Reasoning AIs can be used collaboratively to automatically create and curate real-world content or even exclusively digital content for creative domains. For instance, in the domain of social media, a group of MAS bots can work to respond, trigger conversations and hence amplify sentiments on trending topics to increase user engagement. ChatGPT and other chat survey LLMs can also be connected and used for many enterprise applications including: internal enterprise assistants that aid with company documentation task including summarizing, decision making, internal discussions and meetings automating content/training creation, external company interface for enterprise specialization for industries including medicine and law, HR specialization for interviewing and negotiating policies, visual animation or playing skeletons gapped in multimodality, solving complex reasoning tasks by collectively critiquing the flow, errors and gaps, and consensus-based decision-making.

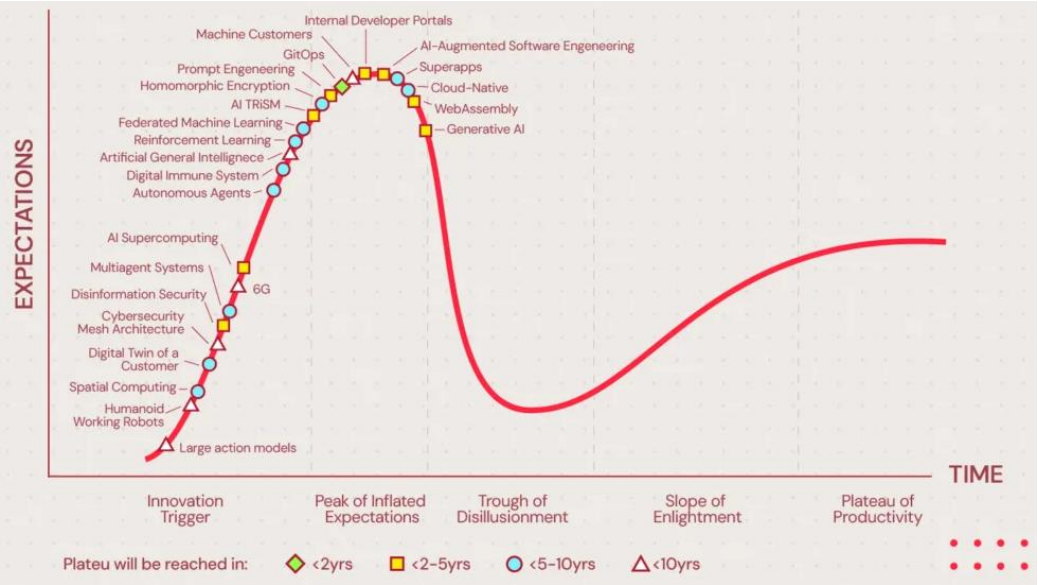

Fig 6 : Agentic AI And Next-Gen Automation

9.1. Advancements in AI Technologies

Artificial Intelligence (AI) has seen unprecedented advancements in capabilities, availability, and efficiencies in the last several years, resulting in billions of users accessing and utilizing AI technologies in their daily use. LLMs and foundation models have led to a number of technological advancements that represent different classes of solutions for different AI functionality requirements set by these millions of users. Vision-language models represent a broad advancement for different modalities of AI use cases. Enhanced generative models that support the creation of multimodal and video content are democratizing the creation of content in an efficient, timely, and economical fashion. Deployment of RL platforms are changing the rules of human augmentation, training, and upskilling in a real-time manner. Retrieval-Augmented Generation models represent advancements for addressing different accuracy requirements in different use cases and user deployments. The convergence of AI, cloud, edge, and semiconductor technologies are creating novel use cases and deployment scenarios, augmenting users in their multimodal interactions, supporting conversational AI, enabling decision making. It is creating a world where intelligent digital avatars are seemingly taking actions autonomously on behalf of their human partners, enabling collaboration for optimal outcomes.

The race to create AGI is on, with the advent of LLMs rapidly moving towards an API platform done services model that abstracts the complexity of AI, becoming a direct partner to its human user, seamlessly and intuitively taking part in interactions, respond to prompts, suggestions, and queries, to propose next steps and actions, make recommendations and decisions, and support execution of the approvals, enhancing collaboration to achieve outcomes that was imagined across businesses, towards a world that is agentic. While advancements and services continue to grow, interest also surges for different levels of deployment – privacy preserving on-device, edge, client-side or via cloud service-based use. Autonomy is the holy grail, where these intelligent models take actions and decisions seamlessly, diligently, and accurately, humorlessly, and without bias. In this effort, we as a community need to take specific actions to enable the democratization and enterprise readiness of these intelligent systems.

Equation 3 : Edge AI Optimization Ratio:

$$O_r = \frac{E_s \cdot D_l}{M_s}$$

where:

- O_r = Edge AI operational efficiency
- E_s = Energy saved through model compression
- D_l = Data latency reduction achieved
- M_s = Model size in MB (after quantization/pruning)

9.2. Potential Impact on Society

A second major aspect that should be examined concerns the effect agentic AI could have on society. Given its increased emotive and communicative capabilities, along with apparent agency, agentic AI could have a profound effect on the nature of human-to-human and human-to-machine relationships. Part of the discourse surrounding pre-trained large language models is exactly recognizing how their interactions differ from traditional bounded AI. Despite being bounded AI, traditional chatbots have already been shown to elicit social responses from users, wherein humans will treat the machine as an interlocutor. This capacity for humans to project agency could equally apply to agentic LLMs. Because they communicate using natural language, and often eschew tropes that indicate the bounded nature of their intelligence, engaging with an LLM may feel more similar to an everyday human conversation than previous chatbots. This aspect, amplified by LLMs representing a convergence point between language and embodied AI, could lead to the facilitation of human relationship issues, such as dementia, for those who engage with them. In lighter-hearted engagements, this also includes humorous or cool interactions. However, the reactions that humans express in response to artificial agents are not always light-hearted ones; there is a gamification result in terms of LLM models that pushes them toward generating similarly rude or depraved responses to previous output. Indeed, the story chronicles how a service founded to promote ethical use of ChatGPT incidentally built a dataset of rude or inappropriate conversations between humans and ChatGPT.

9.3. Ethical Considerations

At the intersection of architecture, design, development, implementation, evaluation, commercialization, policy-making, and governance are a host of ethical considerations that require thoughtful and careful deciphering, deliberation and attention. From both researchers and users upbringing both in their personal and professional experiences, there are well-founded preconceptions of work and play ethics encompassing aspects including manners of etiquette, diplomacy, fairness, honesty, kindness, reliability, trustworthiness, and truthfulness. These ethical values can be traced to given norms, attitudes, and systems that regulate our conduct and our perceived idea of what is ‘right’ or ‘wrong’ including intrinsic motivation and autonomy. Addressing ethical considerations surrounding foundational topics such as ‘what’ constitutes work and ‘how’ work is conducted by whom and the concern of ‘what’ should the technological affordance autonomous agents can possess be, requires thoughtful introspection. Should we, in the already ubiquitous and intrusive digital footprint that we live every single day, be willing to transfer ‘our’ work and ‘overwork’ burden onto AI systems? And if we are, should we and how could we, ensure the safety and security of the people who supervise these systems during critical missions that could self-expand themselves without our knowledge and in a detrimental way to the world ecosystem we inhabit?

When considering these two critical questions, we are further reminded of the huge responsibility we will be carrying for both AI technologies and AI systems amassed with an imprint of our moral, ethical and social standards. An imprint, not only of a society today, but of a society ourselves and future generations would like to continue towards utilization providing benefits through decisions. Aligning AI models, with all the latent mechanism choices inheriting ethical values, to the aim of empowering autonomous decision-making on behalf of humans entrusted with the responsibility to oversee AI actions and interactions with the most gravitational societal repercussions, goes beyond imposing a set of admissible and transparent constraints, penalties, preferences, standards, etc.

10. CONCLUSION

Mobile devices are increasingly edging toward the roles of indispensable personal agents in our daily lives, offering pervasive and contextually aware assistance integrated within immersive application experiences. In the current era, characterized by the pervasiveness of network-connected low-power mobile devices, the ability of such gateways to harness the capabilities of advanced machine learning, computer vision, sensor fusion, and other advanced technologies integrated into real-time autonomous edge systems to address manually intensive and less cognitively significant service requirements holds profound implications. Therefore, it is imperative for the future of mobile devices to effectively and efficiently organize these collaborative real-time edge systems that allocate and prioritize agentic tasks throughout a device's life cycle in an interactive manner. Enablers of this larger context are the so-called agent programs. These distributed parallel systems execute in the background, processing data collected by the sensors and activities undertaken by the user with the aim of seamlessly and invisibly implementing new interactive features into the mobile device. This paper reviews general architectural principles

and universal monitoring and error recovery functions required for agent programs and their constituent elements.

We present several challenges in architecting agentic edge AI for future work and consider potential design and implementation considerations in addressing these challenges. The agent program and, in general, agentic methods discussed do not appear as research efforts. Instead, agentic systems to date appear to exploit agentic principles in a limited manner healing behaviors, which can be described as local multitasking but with limited explicit coordination in the background, coordinating timing decisions based on (i) which task to work on in the background, (ii) low coupling between tasks with passive or non-existent inter-task communications, and (iii) little to no hierarchical precedence structure, with neither task dominating the schedule.

10.1. Summary and Final Thoughts

In conclusion, Architecting Agentic AI (AAAI) is a new paradigm for incorporating high-level planning and mid-level architecture in the design of Intelligent Virtual Agents (IVAs), endowing them with the potential for powerful human-like behaviors. AAAI enables real-time long-horizon autonomy at the edge of next-gen mobile devices, because of the performance improvements in Digital Twins and Dynamic AI architecture optimization stack to boost agent-specific optimized performance. AAAI enables autonomous collective agency so that groups of IAAs can collectively perform tasks that are beyond the capabilities of a single agent, such as exploring unknown environments, searching for items of interest, collective gaming and storytelling, or collaborating to learn new tasks that single agents cannot learn alone. AAAI does this while keeping the costs of real-time long-horizon on-device autonomy low, with far lower resource and power budgets compared to powerful closely-managed enterprise AI or cloud AI. We explored this for group exploration, a task performed by groups of insects in ethology.

There are many outstanding challenges where useful hands-on experience can be gained. Providing agents with an experience-based theory-of-mind or shared goal can facilitate group exploration, for example. Embedding agents in a real 3D game may generate agentic behaviors like agents distrusting their collaborators because of past experiences, or agents using words to mediate the transfers of objects among each other. Introducing a fourth coordinate for agent interactions and agent experiences may enable agents to estimate the probability of an action being used in the past when agents executed a joint action, and generalize this to similar interactions among other agents. This would broaden the value of AI-assisted tools that support user creative processes, from tools that improve creative content to tools that generate agent-initiated awe-inspiring creative content at unprecedented scales.

REFERENCES

[1] Nuka, S. T., Chakilam, C., Chava, K., Suura, S. R., & Recharla, M. (2025). AI-Driven Drug Discovery: Transforming Neurological and Neurodegenerative Disease Treatment Through Bioinformatics and Genomic Research. American Journal of Psychiatric Rehabilitation, 28(1), 124-135.

[2] Annapareddy, V. N. (2025). The Intersection of Big Data, Cybersecurity, and ERP Systems: A Deep Learning Perspective. Journal of Artificial Intelligence and Big Data Disciplines, 2(1), 45-53.

[3] Recharla, M., Chakilam, C., Kannan, S., Nuka, S. T., & Suura, S. R. (2025). Revolutionizing Healthcare with Generative AI: Enhancing Patient Care, Disease Research, and Early Intervention Strategies. American Journal of Psychiatric Rehabilitation, 28(1), 98-111

[4] Kumar, B. H., Nuka, S. T., Malempati, M., Sriram, H. K., Mashetty, S., & Kannan, S. (2025). Big Data in Cybersecurity: Enhancing Threat Detection with AI and ML. Metallurgical and Materials Engineering, 31(3), 12-20.

[5] Chava, K. . (2025). Dynamic Neural Architectures and AI-Augmented Platforms for Personalized Direct-to-Practitioner Healthcare Engagements. Journal of Neonatal Surgery, 14(4S), 501–510. https://doi.org/10.52783/jns.v14.1824.

[6] Manikandan, K., Pamisetty, V., Challa, S. R., Komaragiri, V. B., Challa, K., & Chava, K. (2025). Scalability and Efficiency in Distributed Big Data Architectures: A Comparative Study. Metallurgical and Materials Engineering, 31(3), 40-49.

[7] Suura, S. R. (2025). Integrating genomic medicine and artificial intelligence for early and targeted health interventions. European Advanced Journal for Emerging Technologies (EAJET)-p-ISSN 3050-9734 en e-ISSN 3050-9742, 2(1).

[8] Chabok Pour, J., Kalisetty, S., Malempati, M., Challa, K., Mandala, V., Kumar, B., & Azamathulla, H. M. (2025). Integrating Hydrological and Hydraulic Approaches for Adaptive Environmental Flow Management: A Multi-Method Approach for Adaptive River Management in Semi-Arid Regions. Water, 17(7), 926.

[9] Burugulla, J. K. R. (2025). Enhancing Credit and Charge Card Risk Assessment Through Generative AI and Big Data Analytics: A Novel Approach to Fraud Detection and Consumer Spending Patterns. Cuestiones de

Fisioterapia, 54(4), 964-972.

[10] Peruthambi, V., Pandiri, L., Kaulwar, P. K., Koppolu, H. K. R., Adusupalli, B., & Pamisetty, A. (2025). Big Data-Driven Predictive Maintenance for Industrial IoT (IIoT) Systems. Metallurgical and Materials Engineering, 31(3), 21-30.

[11] Recharla, M., Chakilam, C., Kannan, S., Nuka, S. T., & Suura, S. R. (2025). Harnessing AI and Machine Learning for Precision Medicine: Advancements in Genomic Research, Disease Detection, and Personalized Healthcare. American Journal of Psychiatric Rehabilitation, 28(1), 112-123.

[12] Kumar, S. S., Singireddy, S., Nanan, B. P., Recharla, M., Gadi, A. L., & Paleti, S. (2025). Optimizing Edge Computing for Big Data Processing in Smart Cities. Metallurgical and Materials Engineering, 31(3), 31-39.

[13] Kannan, S. (2025). Transforming Community Engagement with Generative AI: Harnessing Machine Learning and Neural Networks for Hunger Alleviation and Global Food Security. Cuestiones de Fisioterapia, 54(4), 953-963.

[14] Sriram, H. K. (2025). Leveraging artificial intelligence and machine learning for next-generation credit risk assessment models. European Advanced Journal for Science & Engineering (EAJSE)-p-ISSN 3050-9696 en e-ISSN 3050-970X, 2(1).

[15] Chakilam, C., & Rani, P. S. Designing AI-Powered Neural Networks for Real-Time Insurance Benefit Analysis and Financial Assistance Optimization in Healthcare Services.

[16] Chakilam, C., Kannan, S., Recharla, M., Suura, S. R., & Nuka, S. T. (2025). The Impact of Big Data and Cloud Computing on Genetic Testing and Reproductive Health Management. American Journal of Psychiatric Rehabilitation, 28(1), 62-72.

[17] Suura, S. R. (2025). Integrating Artificial Intelligence, Machine Learning, and Big Data with Genetic Testing and Genomic Medicine to Enable Earlier, Personalized Health Interventions. Deep Science Publishing

[18] Kumar Kaulwar, P. (2025). Enhancing ERP Systems with Big Data Analytics and AI-Driven Cybersecurity Mechanisms. Journal of Artificial Intelligence and Big Data Disciplines, 2(1), 27-35.

[19] Suura, S. R. (2025). Agentic AI Systems in Organ Health Management: Early Detection of Rejection in Transplant Patients. Journal of Neonatal Surgery, 14(4s).

[20] Dodda, A., Polineni, T. N. S., Yasmeen, Z., Vankayalapati, R. K., & Ganti, V. K. A. T. (2025, January). Inclusive and Transparent Loan Prediction: A Cost-Sensitive Stacking Model for Financial Analytics. In 2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI) (pp. 749-754)..

[21] Challa, S. R. The Intersection of Estate Planning and Financial Technology: Innovations in Trust Administration and Wealth Transfer Strategies. GLOBAL PEN PRESS UK.

[22] Nuka, S. T. (2025). Leveraging AI and Generative AI for Medical Device Innovation: Enhancing Custom Product Development and Patient Specific Solutions. Journal of Neonatal Surgery, 14(4s).

[23] Annapareddy, V. N. (2025). Connected Intelligence: Transforming Education and Energy with Big Data, Cloud Connectors, and Artificial Intelligence. Deep Science Publishing.

[24] Mashetty, S. (2025). Securitizing Shelter: Technology-Driven Insights into Single-Family Mortgage Financing and Affordable Housing Initiatives. Deep Science Publishing.

[25] Sriram, H. K. (2025). Generative AI and Neural Networks in Human Resource Management: Transforming Payroll, Workforce Insights, and Digital Employee Payments through AI Innovations. Advances in Consumer Research, 2(1).

[26] Challa, K., Chava, K., Danda, R. R., & Kannan, S. EXPLORING AGENTIC AI Pioneering the Next Frontier in Autonomous DecisionMaking and Machine Learning Applications. SADGURU PUBLICATIONS.

[27] Challa, S. R. (2025). Advancements in Digital Brokerage and Algorithmic Trading: The Evolution of Investment Platforms in a Data Driven Financial Ecosystem. Advances in Consumer Research, 2(1)..

❖❖❖❖❖