

Enabling Agentic AI across the Cloud-Edge Computing Continuum: A Research Roadmap

Hui Song, Arda Goknil, Dumitru Roman
SINTEF AS, Norway

Maria Fazio, Massimo Villari
University of Messina, Italy

Massimo Mecella
Sapienza University of Rome, Italy

Maryam Doborjeh, Nikola K. Kasabov
Auckland University of Technology, New Zealand

JeongGil Ko
Yonsei University, South Korea

Hyunwhan Joe, Hong-Gee Kim
Seoul National University, South Korea

Thomas Fahringer, Zahra Najafabadi, Radu Prodan
University of Innsbruck, Austria

Abstract—Agentic AI is shifting AI applications from passive model inference to goal-driven, tool-using, and collaborative autonomous systems. Yet, current deployments remain concentrated in data centers or powerful personal devices. This paper provides a research roadmap for enabling large-scale, enterprise-oriented agentic AI across the computing continuum, from cloud to edge, IoT, and emerging hardware platforms, where heterogeneity, mobility, energy constraints, and governance requirements fundamentally reshape agent design and operation. We argue for a continuum-native agent paradigm that decomposes agents into intelligence, persona, and memory, allowing their independent placement, migration, and replication. Building on this abstraction, we structure the open research space into three pillars (agent operation, agent connectivity, and agent trustworthiness) and elaborate on potential research directions and challenges. Finally, we outline evaluation directions grounded in representative industrial use cases and benchmarks. The roadmap aims to guide research and development toward enabling the deployment of agentic AI as first-class workloads on the cloud-edge continuum.

I. INTRODUCTION

While generative AI (GenAI) is powerful in conversation, content creation, and reasoning, its future lies in agents that act, decide, and collaborate autonomously to solve real-world problems. While agent development frameworks such as AutoGen [1] and LangGraph [2] are maturing, and special-purpose agents such as Codex [3] (coding), Prism [4] (scientific writing), and OpenClaw [5] (personal assistance) are gaining attention, large-scale business-oriented agentic AI systems are largely beyond the horizon [6].

Infrastructure capacity is a decisive factor when adopting GenAI agents at scale: domains such as traffic management and telecoms operations may require thousands of concurrent agents interacting with various real-world entities [7], [8], [9]. Relying on managed GenAI services is prohibitively expensive, while self-hosting still depends on scarce, costly, and energy-intensive datacenter or on-premises GPU resources [10], [11], [12]; breaking this bottleneck requires exploiting the full potential of resources from cloud to

edge [13]. At the same time, GenAI is shifting from monolithic general-purpose LLMs toward compositions of many smaller models and multi-step (e.g., Mixture-of-Experts [14] or reinforcement-learning-based reasoning [15]) processes, which can outperform a single large model but still remain largely datacenter-bound. Extending these approaches across the continuum, where specialized models and agents collaborate under heterogeneous, constrained conditions, offers a path to scalable distributed intelligence [16], [17], [18], [19]. Finally, latency, reliability, privacy, and trust-critical applications require placing at least part of the agent population on edge devices close to the scene [20], [21], [22], [23].

This paper presents our vision for enabling large-scale agentic AI to operate effectively, energy-efficiently, and trustworthily across the cloud–edge continuum, allowing agents to collaborate on complex real-world problems while remaining aligned. This vision goes beyond today’s continuum support for software services and data pipelines: GenAI agents can be energy-intensive when tied to specific devices, difficult to coordinate when highly distributed, and hard to trust due to the inherent uncertainty in their decisions and outputs.

This paper outlines a research roadmap to fill these gaps and make agentic GenAI practical across the cloud-edge continuum. Our contributions are: 1) a continuum-native agent architecture connecting agent abstractions to infrastructure concerns; 2) a prioritized research agenda organized into three pillars—*agent operation*, *connectivity*, and *trustworthiness*—that distinguishes near-term engineering challenges from longer-term exploratory directions; 3) an interdisciplinary vision bringing insights from brain and social sciences to inform scalable collaboration, governance, and alignment; and 4) use cases to validate the performance, reliability, cost, and trustworthiness of agentic AI on the computing continuum.

II. STATE-OF-THE-ART AND RESEARCH GAPS

Recent advances in cloud–edge computing and agentic AI make large-scale continuum-native agents increasingly feasible, but key gaps remain. This section reviews the state of the art, open challenges, and relevant roadmaps across infrastructure, agentic AI, and human-stakeholder perspectives.

The European Commission funds this work through the Horizon Europe Project CoAgent. Prism, the scientific writing workspace powered by GPT-5.2, helped us structure the paper, polish some sections, and proofread it.

1) *Computing continuum as AI infrastructure*: The computing continuum extends resource virtualization and Infrastructure as Code from central cloud to edge resources. The growing complexity requires AI and ML approaches for service placement [24] and resource management [25] across the continuum, leading to a cognitive continuum. AI further enables intent-based network [26] and continuum [27] management, in which stakeholders describe their desired outcomes while delegating the selection of an appropriate solution to AI. While modern service computing has evolved to support elastic, scalable, and composable services across cloud and edge environments [24], [28], the emergence of AI agents introduces new opportunities and challenges for goal-driven, self-choreography of business processes to solve real-world tasks under uncertainty [29]. GenAI today runs predominantly in hyperscale data centers and HPC facilities built around high-end accelerators (GPUs/TPUs), using LLM serving stacks such as vLLM [30] for multi-GPU, multi-node inference. Despite rapidly emerging edge accelerators and memory-aware execution techniques [31], on-device and near-device GenAI remains resource-constrained. To enable edge inference, model compression methods such as quantization [32], pruning [33], distillation [34], and low-rank adaptation are increasingly important, while neuromorphic computing provides an extremely energy-efficient path for edge-native GenAI [35]. However, systematic and engineering support for agents on edge and unconventional devices remains largely underdeveloped.

2) *Agentic AI, multi-agency, and distributed system frameworks*: As enterprise adoption of GenAI accelerates, attention is shifting from chat to tool-using, planning agents that execute multi-step tasks (e.g., via the Model Context Protocol [36]). At the same time, development is moving towards distributed approaches with maturing open frameworks for building agent teams, such as AutoGen [1] and the agentic AI foundation [37]. Meanwhile, reasoning and decision-making methods (e.g., chain and graph-of-thoughts [38]) enhance multi-step problem-solving, and process supervision and reinforcement learning for reasoning improve correctness and robustness [39]. Memory is a critical aspect in LLMs, where dynamic memory management, such as vAttention [40], improves serving efficiency. In agentic systems, long-term task-specific memory extends beyond raw context windows to hierarchical and external stores [41], with dedicated memory databases designed for agents [42]. In addition to agent-specific frameworks, actor-based platforms, such as Akka [43], provide a mature foundation for building scalable, message-driven applications composed of loosely coupled components, supporting distribution, elasticity, and fault tolerance. Higher-level coordination further introduces explicit behavioral models and lifecycle management; for example, collaborative state machines [44] provide a technical foundation for reliable collaboration among large-scale distributed agents. However, these frameworks assume stable execution environments and therefore do not address the challenges that arise when components become autonomous agents with uncertain behavior and evolving connections.

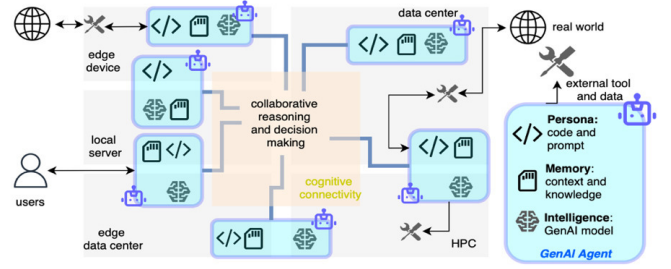


Fig. 1. Schematic of the agent-continuum architecture.

3) *Trustworthy and human-centric AI*: Trusted execution environments (TEEs) and secure communication channels underpin confidential computing on edge/IoT devices [45], while distributed ledgers enable auditable communication [46]; together they can provide trusted cloud-edge infrastructures for software systems. However, agentic systems introduce uncertainty, long-horizon behaviors, and new failure modes, requiring continuous observability, runtime risk governance and mitigation, and alignment mechanisms and guardrails for safety, bias, and fairness [47]. Despite rapid progress in agentic applications, trustworthiness, risk, and alignment remain largely underexplored. Research in this direction requires a shift from trusting human operators to trusting autonomous agents. Moreover, brain-inspired approaches remain a key inspiration for efficient and adaptive intelligence, including spiking neural networks [48], [49] and neuromorphic computing [50] for energy-efficient inference and Global Workspace Theory (GWT) for decentralized multi-agent decision-making [51], but adapted to large-scale agentic systems operating across the computing continuum.

a) *Related roadmaps and visions*: Recent roadmaps frame our work by describing the evolution of cloud-edge-IoT into federated computing continuum [52], the infusion of intelligence into that continuum [53], the orchestration and operation of agentic AI systems [54], and the use of agentic AI to augment software engineering [55]. Our roadmap builds on these perspectives, but places a unique focus on agentic AI as a first-class continuum workload and outlines the research space for using the continuum as infrastructure for efficient, collaborative, and trustworthy agentic systems.

III. CONTINUUM-ORIENTED AGENTIC AI ARCHITECTURE

Building on emerging practice in agentic systems and motivated by the lack of agent-oriented support in today's computing continuum, we propose a conceptual architecture for continuum-native GenAI agents (Figure 1). The grey boxes denote resource units across the continuum, from individual devices to aggregated service platforms (e.g., clusters or clouds), while the blue boxes represent agents deployed on these units. Each agent comprises three functionally distinct building elements. 1) *Intelligence* is the GenAI model that endows the agent with cognitive capabilities (e.g., reasoning and decision-making) with built-in knowledge. It is a shared capability, as multiple agents may use the same model, and a

single agent may switch models. 2) *Persona* is the prompt or code that activates and drives the GenAI model to perform specific tasks. It defines the agent’s mission and expected behavior and typically changes infrequently. 3) *Memory* is the context an agent maintains while performing tasks, including accumulated conversations with other agents and relevant domain- or task-specific knowledge the agent must retain.

This novelty does not replace existing agent frameworks, but complements them with a continuum-oriented deployment abstraction. While existing architectures typically emphasize reasoning and tool use, role-based multi-agent collaboration, stateful orchestration, or long-term skill and memory management, our decomposition instead exposes deployment units with different continuum requirements. *Intelligence* is compute-intensive and placed on cloud GPUs, edge accelerators, or specialized model-serving nodes. *Persona* is comparatively small, stable, and auditable, making it suitable for replication, versioning, and policy checking. *Memory* is stateful, privacy-sensitive, and latency-sensitive, and may therefore need to remain close to the user, device, or operational context. A practical implementation could map intelligence to a model-serving endpoint, persona to a versioned prompt, policy, or agent configuration, and memory to a local or distributed store combining conversation history, vector indexes, knowledge graphs, and task artefacts. For example, in predictive maintenance, a customer-site agent may keep sensitive maintenance memory locally, use an audited persona defined by the device provider, and invoke a local compressed or a cloud model depending on latency, privacy, and energy constraints.

IV. RESEARCH AGENDA

Building on the architecture, we organize the research landscape into three interdependent pillars (Figure 2) that address the key requirements for scalable agentic systems across the computing continuum, and directly target the gaps in Section II: agent operation, connectivity, and trustworthiness. Accordingly, each research challenge connects to the pillar that reflects its technical focus, while relevant dependencies on the other pillars are made explicit.

1) *Agent operation* Agent operation provides the execution substrate for efficient hosting, migrating, and managing intelligence, persona, and memory across heterogeneous continuum resources, enabling: a) *intent-driven model-as-a-service* for continuum-wide agent intelligence; b) *intent-driven orchestration* of agents with personas and memories; and c) *compression and hosting* of models on the edge.

2) *Agent connectivity* enables, on top of the operation, autonomously and spontaneously deployed collaborative agents, tools, and knowledge sources to discover one another, communicate, share context, and choreograph tasks through a layered stack of research foci: a) on-demand *networking* layer, to support b) *service-layer* agent-tool interaction, c) *knowledge-layer* with shared insights for distributed intelligence, and d) *choreography* layer for self-organizing business processes.

3) *Agent trustworthiness* provides four continuum-wide mechanisms that close the human-agent loop by constraining

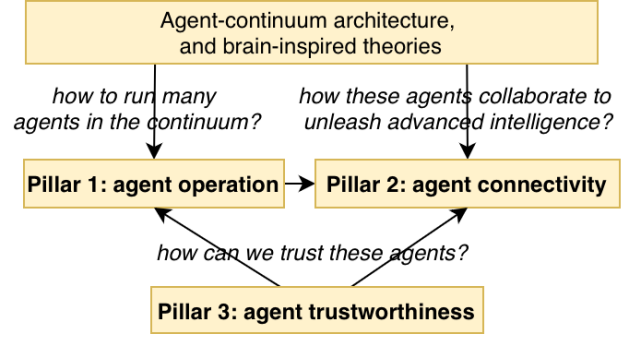


Fig. 2. Interdependent research pillars for continuum-native agentic AI.

and monitoring agents’ operation and connectivity, aligning, protecting, observing, and governing them when actions affect humans, organizations, or physical systems: a) capturing and aligning with human intents and values, b) providing trusted execution and communication for critical agents, c) providing observability of agent behavior with trusted evidence, and d) governing risks for uncertainty and world impact.

Table I summarizes these priorities as problem–outcome pairs and distinguishes near-term challenges from longer-term directions. We prioritize research problems by deployability, feasibility with current cloud–edge and agent technologies, and expected impact on scalable, energy-efficient, and trustworthy operation. The main near-term challenges are: 1) decomposed placement and lifecycle management of intelligence, persona, and memory, 2) reliable connectivity under dynamic cloud–edge conditions, and 3) runtime trust assurance through alignment, trusted execution, observability, and risk governance. Neuromorphic hosting, global-workspace-inspired collaboration, and metacognitive continuum management are longer-term exploratory themes.

A. Pillar 1: Agent Operation

Agent operation concerns the flexible, energy-efficient deployment and management of agents and their components across the computing continuum. The expected outcome is a continuum execution layer that determines where intelligence, persona, and memory should reside and how they should move as workload, privacy, and resource conditions change.

1) *Continuum-wide intent-based model-as-a-service*: Many AI and cloud providers offer managed LLM services [56], [57], [58]. Behind the managed APIs, open-source LLM serving stacks such as vLLM [30] (cloud/cluster) and Ollama [59] (edge/on-device) are increasingly used in practice, requiring extensions to support flexible deployment across cloud and edge and to leverage the many GenAI models available in Hugging Face. A continuum-wide model-as-a-service layer needs to advance the current pre-defined cloud-based LLM serving by allowing stakeholders to specify a model-using intent, such as intelligence level, specialized domain, latency, token throughput, and energy efficiency. Research should be focused on: 1) *modeling intents* for model serving inspired by intent-based networking, but using a more natural-language

TABLE I
PRIORITIZED RESEARCH PROBLEMS AND EXPECTED OUTCOMES FOR CONTINUUM-NATIVE AGENTIC AI.

Research Pillar	Research Problem	Expected Outcome
Operation	– Place and serve GenAI models across heterogeneous cloud, edge, and accelerator resources. – Manage persona and memory independently from model execution. – Execute agents on constrained or unconventional edge devices.	– Intent-driven model-serving mechanisms that optimize latency, cost, energy, and capability. – Lifecycle mechanisms for versioning, migration, checkpointing, cloning, and recovery of agent state. – Compression methods for constrained edge execution, with neuromorphic hosting treated as a longer-term option and evaluated through accuracy, latency, and energy trade-offs.
Connectivity	– Enable agents to dynamically discover tools, services, and other agents. – Coordinate distributed agents into goal-driven processes. – Maintain collaboration under unreliable or heterogeneous networks.	– Continuum-aware discovery and interaction protocols enriched with latency, trust, energy, and context metadata. – Declarative choreography and self-improvement mechanisms for composing agent capabilities into reliable workflows. – Intent-driven networking and routing mechanisms with measurable QoS, resilience, and recovery guarantees.
Trustworthiness	– Align agent behavior with user intent, organizational policy, and human values. – Protect critical agent execution and communication. – Detect and mitigate unsafe, uncertain, or non-compliant behavior.	– Intent-synthesis and value-calibration mechanisms with traceable updates and conflict handling. – Attestation, secure communication, and audit for sensitive workflows. – Observability and risk-governance mechanisms that produce evidence, explanations, and recovery actions.

style to fit agent-oriented specification; 2) *intent-driven deployment* of relevant models, with continuous optimization of performance and energy consumption via continual learning; and 3) *LLM serving stack integration*, from cloud/cluster stacks (e.g., vLLM) to edge/on-device stacks (e.g., Ollama), under a standard model-as-a-service interface.

2) *Intent-driven orchestration of agent persona and memory*: The agent orchestration handles storing the agent’s persona and memory to leverage intelligence from the model-as-a-service layer. Research should investigate a comprehensive, multimodal database solution, combined with file systems, to store source code, plain text, knowledge graphs, and multimedia data, providing the context (both persona and memory) for agent conversation. Potential research directions include: 1) graph representation to store the essential knowledge and accumulated memories along the agent lifecycle; 2) knowledge embedding (e.g., vector databases) and on-demand searching for efficient retrieval of relevant memories; 3) “forgetting mechanism” to gradually remove irrelevant and outdated data; 4) smart placement, deployment, and resource allocation of the agent storage modules in the computing continuum; and 5) novel agent lifecycle actions, like cloning and teleporting, based on memory management.

3) *Edge-native models and agents*: GenAI models trained with large datasets require significant resources for inference. Compression reduces model size and computational requirements via quantization and distillation, but does not guarantee application-level performance. Research needs to focus on new mechanisms: 1) *pruning the neural network* for specific purposes and knowledge (e.g., “an agent that manages traffic planning does not need to know much about ancient Rome”) to achieve performance on mobile edge devices; 2) *contrastive learning* for selective dimensional reduction and testing environments to validate the performance and device suitability of the slimmed models for the user needs; and 3) *hosting on*

unconventional edge devices such as converting transfer-based models into spiking neural network (SNN) models [49] to run on neuromorphic computing accelerators, or embedding knowledge into SNN to store the persona and memory of agents, while using remote models as the intelligence.

B. Pillar 2: Agent Connectivity

Building on the operational substrate established by the agent operation, agent connectivity addresses how deployed agents discover one another, exchange information, invoke tools, and coordinate tasks across heterogeneous cloud-edge resources. The expected outcome is a collaboration layer that lets agents form, adapt, and recover goal-driven workflows despite distribution, heterogeneity, and network instability.

1) *Choreography of agent processes*: Solving real-world problems often requires multi-step processes involving many specialized agents. Self-organizing autonomous agents should choreograph complex business processes and workflows among them, driven by high-level goals and requirements, exploiting individual capabilities and semantics, and considering their continuum-oriented features, such as locations, latency, energy consumption, and access to external systems. Research needs to address: 1) *declarative and agent-oriented representation* of goals, requirements, and agent skills for automating efficient agentic discovery and composability of capabilities; 2) *reinforcement learning* to reward agents in forming the right processes towards achieving the high-level goals; and 3) *self-improving agent choreography* to keep assessing the effectiveness and performance of established agent processes and seek opportunities for potential improvements.

2) *Collaborative reasoning and decision-making*: Reasoning and decision-making for complex real-world problems are among the most advanced uses of GenAI. Recent multi-model and multi-step methods (e.g., chain and graph-of-thought) naturally align with multi-agent architectures, motivating self-reinforced multi-agent reasoning with a distributed knowledge

base for insight sharing. Key challenges include 1) *distributed knowledge sharing* algorithms; 2) generating and interpreting *knowledge embeddings*; and 3) *propagating embeddings* for decentralized system adaptation. Brain-inspired mechanisms, particularly *GWT* [51]—which explains collaborative decision-making via insight sharing, attention, and competition—agreement among decentralized cortical units—offer promising design principles and warrant focused study.

3) *Real-world engagement*: Agentic systems need connections with external AI agents and systems available worldwide. While the *Model Context Protocol (MCP)* and agent skills [60] emerged as de facto standards, the continuum-based agent-world interaction requires extensions to support real-time situational awareness, ensuring agents receive an up-to-date world view and multi-view world modelling, enabling agents to select among raw sensor feeds, analytic summaries, predictive insights, and digital-twin simulations, according to task demands and confidence thresholds. This includes the following directions: 1) *meta-information extension* to MCP and agent skills, with real-time performance metrics, time-constraint annotations, sensor and simulation profiles, resource preferences, and power consumption data; 2) *intent modelling* that captures an agent’s expectation on real-world engagement; and 3) patterns to wrap and link to *external tools* on different infrastructures, from IoT sensors to HPC-based simulation and analytics tasks.

4) *Reliable and low-latency agent networks*: Continuum-native agents rely on heterogeneous, geographically dispersed networks, from the public Internet to private WANs and radio access networks (RAN). Even small delays or packet losses can disrupt workflows, knowledge exchange, real-world engagement, and collaborative decision-making. Agent overlay networks therefore need low-latency paths and end-to-end QoS guarantees, supported by SDN/NFV orchestration, traffic prioritization such as network slicing, and dynamic path optimization such as SD-WAN. Key research needs include *intent-based network management* for capturing agent communication requirements and learning-based monitoring, and *rerouting of critical agent traffic* around congestion while satisfying service-level agreements (SLA).

C. Pillar 3: Agent Trustworthiness

The third pillar cuts across operation and connectivity: as agents are deployed, migrated, connected, and coordinated across the continuum, their execution, communication, decisions, and real-world actions must remain trustworthy, transparent, and aligned with human values. The expected outcome is an assurance layer that continuously aligns, protects, observes, and governs agent behavior across digital and physical contexts by capturing intents and values, supporting trusted execution, monitoring behavior, and managing risks.

1) *Human-agent intent synthesis for AI alignment*: A human-interface layer is needed to keep agents goal-directed and value-aligned across the computing continuum. Building on conversational AI (e.g., intent/value extraction) and

cognitive-science theories of shared intentionality, with potential directions including: 1) *intent synthesis* that combines passive NLU with active, heuristic questioning to derive consistent intents from incomplete or evolving conversations; 2) *multi-stakeholder value calibration* that reconciles user goals with organizational policies, societal norms, and ethical constraints; 3) *context-adaptive intent lifecycle management* for maintaining, versioning, and propagating intents.

2) *Trusted execution and communication environment*: Hardware-based confidential computing (e.g., on-chip TEEs) can protect decision and guardrail logic (persona), sensitive knowledge (memory), and code/credentials for accessing critical real-world systems, using verified or certified control code at load time and enforcing a “policy firewall” that sandboxes unexpected actions. Runtime-switchable enclaves and co-TEEs can mutually cross-check to detect side-channel leakage. For inter-agent coordination, lightweight permissioned distributed ledgers can trace requests and action commitments, ensuring that only attested agents join sensitive workflows and that all messages are tamper-evident and auditable. Together, these mechanisms limit errant or malicious actions and protect honest agents from compromised peers or operators.

3) *Attention-driven observability for responsible AI agents*: Rich observability is necessary for streaming fine-grained telemetry, from energy use, decision traces, or data access footprints, into a pluggable plane inspired by *5G network data analytics functions*. Runtime adaptation includes sharpening its view when stakeholder intent or context demands deeper inspection, and relaxing when risk is low. Without predefined rules, a self-supervised “attention engine” must continuously scan for out-of-distribution behavior, like wasteful energy spikes, socially harmful decisions, or emergent biases.

4) *Uncertainty-aware risk governance*: Built on a novel theory of *AI risk*, based on the agent behavior uncertainty (or *creativity* related to model “temperature”) and its real-world impact, uncertainty-aware risk governance detects, forecasts, and mitigates adverse events for agentic workflows. Potential research directions include: 1) *automated risk-model extraction* fusing agent processes, historic bow-ties, logs, and manuals into consistent hazard-barrier graphs; 2) *process analytics* across the continuum, learning and composing per-activity probability laws, drift, anomaly, and root-cause; 3) *risk-aware decision-making* using safe/distributional reinforcement learning with conditional value-at-risk or expected-cost objectives, action shielding, and simulation-in-the-loop; and 4) *assurance by design* with continuous conformance checks versus intents, SLAs, and automatic generation of risk packages.

D. Brain-inspired research and metacognition

Brain-inspired theories and techniques are long-term design inspirations rather than immediate implementation requirements. Neuromorphic and SNN-based low-power inference are pioneers inspired by the sparse, event-driven signaling of biological neurons, and supporting LLMs and agents on such unconventional hardware requires returning to the theoretical basis of neuroinformatics, comprising: 1) *GWT*

for distributed decision-making modeling the decentralized brain thought; 2) *collaboration by shared intentionality* for agent alignment, a hallmark of human social cognition; and 3) *adaptive neurocognitive models* of risk and reward to enhance agents' ability to assess the consequences of actions and calibrate decision thresholds. In the long term, such mechanisms may inspire new forms of adaptive connectivity, collective decision-making, and self-monitoring in computing continuum-native agent systems, but their feasibility requires substantial theoretical and empirical investigation. Realizing continuum management tools guided by the research directions discussed creates opportunities to interact with application-plane agents, learn from their behavior, and retrospectively improve one's own performance in orchestration, connections, resource optimization, and more.

V. EMPIRICAL GROUNDING: USE CASE AND BENCHMARK

To validate continuum-native agentic approaches, we propose four representative use cases that stress large-scale cloud-edge agent deployments. They anchor the roadmap by exercising key capabilities such as code generation, goal decomposition, constrained decision-making, intent negotiation, safe tool use, system monitoring and remediation, and cyber-physical control, and providing a basis for comparing cloud-only, edge-only, and continuum-native agent architectures.

A. Agentic smart mobility

We envisage a multi-layer agentic architecture spanning roadside/edge devices, mobile edge servers, control centers, and national coordination hubs for real-time traffic control, incident response, and road maintenance. Agents process local sensor streams, interpret regional events, and generate coordinated response and maintenance plans, optionally supported by HPC-based simulation and prediction. Beyond real-time control, they can support long-term mobility and urban planning by simulating what-if scenarios in transportation, urban design, population growth, and climate change, thereby enabling citizen engagement and democratic decision-making.

1) *Key challenges include:* 1) extreme distribution of agents on mobile edge devices far away from each other; 2) real-time reaction, decision-making, and planning; 3) critical decisions with potential impacts on traffic efficiency and safety; and 4) complex multi-layer collaboration with many agents, human operators, and roles.

2) *Roadmap relevance involves:* 1) a continuum-wide model hosting and agent orchestration in a hierarchical road infrastructure; 2) distributed multi-agent reasoning across geographically dispersed infrastructure; 3) real-time situational awareness and traffic control via agent-tool engagement with road management equipment, connected vehicles, and operator dashboards; 4) function-oriented model compression for agents on mobile edge devices; 5) risk governance for critical traffic decisions; and 6) human-agent alignment for real-time engagement with traffic operators and long-term engagement with citizen stakeholders.

B. Predictive device maintenance

Device providers, from healthcare equipment to machine tools, need a global service network for device maintenance. They face increasing complexity due to rapidly advancing device functionality, the high cost of downtime, and regulatory requirements. An agentic system significantly supports service engineers. 1) *Knowledge agents* extract and maintain internal knowledge from the device provider and the customer; 2) *troubleshooting agents* identify the cause of an issue by analyzing system log data, medical device context, hospital routines, and using data analytics code generated on the fly; and 3) *proactive maintenance planners* monitor device performance and environmental conditions in real time, predict failures, and recommend actions to the maintenance team. This pilot requires agent capabilities, such as data analytics and coding, to handle diverse data sources and to plan actions.

1) *Key challenges include:* 1) high-volume, heterogeneous data (e.g., real-time streams, machine logs, and hospital records); 2) stringent multi-stakeholder regulatory requirements for medical-device use and maintenance; 3) complex yet customer-specific analytics and maintenance workflows; and 4) strong security and data-privacy requirements between device providers and users.

2) *Roadmap relevance motivates:* 1) secure hosting and trusted execution for agents deployed on customers' sites; 2) long-term observability that produces evidence for regulatory compliance; and 3) collaborative choreography and distributed decision-making for cross-device, cross-customer troubleshooting under prescribed yet customizable workflows.

C. Drone mission control

Today's unmanned aircraft systems still rely on human operators to make and refine mission plans for the drone swarms. An agentic system enables drones to adjust their missions autonomously in response to rapid context changes, making prompt decisions in fast-changing environments. Embedded edge agents onboard the drones have situational awareness and handle navigation (re-)planning, while orchestration agents in ground support stations manage the swarm's route planning and maintain a global overview. During the swarm operation, the orchestration and connectivity of agents must adapt to changing routes, plans, and tasks, with advanced lifecycle management, for example, to "teleport" an agent across drones to cover a task from a different location or perspective, or clone an agent into multiple drones to duplicate a task.

1) *Key challenges comprise:* 1) resource-constrained computing environment for onboard agents with limited energy supply (batteries prioritized for drone motion); 2) distributed collaboration under unstable connectivity and operational environment; 3) open world operation exposed to security threats; 4) complex management of different and potentially conflicting plans, routes, and mission objectives simultaneously; and 5) autonomous decisions should remain aligned with ethical rules and human values, for example by prioritizing human safety over task completion and drone preservation.

2) *Roadmap relevance motivates*: 1) pruning small multi-modal models on heterogeneous edge devices, including neuromorphic SoCs for ultra-low-power inference; 2) distributed multi-agent collaboration with efficient memory and insight sharing across the swarm; 3) low-latency, reliable connectivity under intermittent networking, supported by TEEs and secure communication channels; 4) mission-aware risk identification and mitigation; and 5) continuous alignment with operator intent, organizational objectives, and broader human values.

D. Agentic telecom operation

Managing the RAN and its attached computing resources is mission-critical and requires continuous optimization in response to changing conditions. Intent-based networking must translate high-level stakeholder intents into configuration actions that balance cost, energy, latency, throughput, and reliability. Future intent managers could act as autonomous agents, using real-time data and closed-loop automation to configure the RAN, refine longer-term adaptation policies, and collaborate across the infrastructure to adapt to near-real-time conditions and user demand.

1) *Key challenges include*: 1) large-scale agents distributed across the telecommunication infrastructure; 2) long-lived agents with constant decision making; 3) fast-changing network conditions requiring agents to adapt their collaboration; 4) mission-critical tasks involving agents deviating from expected behavior or disrupting critical infrastructure; and 5) multi-objective optimization of agents overseeing different areas, competing and constantly negotiating with each other.

2) *Roadmap relevance motivates*: 1) human-agent intent synthesis and alignment for intent-based network management; 2) advanced memory mechanisms to maintain accumulated context over long-lived agent lifecycles; 3) safe and effective use over complex network functions and configurations; and 4) semantic choreography for distributed multi-agent reasoning with configurable observability, generating vendor- and customer-facing reports that explain agent actions in the RAN.

E. Benchmarks

Benchmarks for continuum-native agentic systems must combine task-level outcomes with deployment-level constraints, instantiated by the four use cases: 1) resolving a traffic incident through distributed sensing and response planning; 2) diagnosing a device fault from logs, manuals, and operational context; 3) replanning a drone mission under intermittent connectivity and limited onboard energy; and 4) translating a telecoms intent into safe configuration actions. For each task, competing approaches should be evaluated under the same workload and infrastructure conditions, comparing at least cloud-only monolithic agents, edge-only agents, and decomposed continuum-native agents. Core metrics include task success, solution quality, time-to-action, end-to-end latency, energy use, communication overhead, model-serving cost, resource utilization, policy violations, explanation quality, audit completeness, recovery time, and robustness under failures, mobility, and changing resource availability.

Evaluation protocols should vary the number of agents, sites, tools, network disruptions, and resource budgets, and should report both absolute performance and Pareto trade-offs among quality, latency, cost, energy, and trustworthiness. The intelligence-persona-memory model can further be assessed through ablations that disable model switching, memory locality, persona replication, or state migration.

VI. CONCLUSIONS

This paper presented a research roadmap for enabling agentic AI across the cloud-edge computing continuum. We argued that large-scale agentic systems should be first-class continuum workloads and proposed separating agent intelligence, persona, and memory as independently placeable and governable resources. Building on this abstraction, we identified three interdependent research pillars: 1) agent operation for flexible and energy-efficient deployment; 2) agent connectivity for distributed collaboration; and 3) agent trustworthiness for alignment, protection, observability, and risk governance. We further grounded the roadmap in four use cases and outlined benchmark tasks, metrics, baselines, and ablation protocols for evaluating continuum-native agentic systems. In the longer term, this research may support metacognitive agentic ecosystems that monitor their internal state, recognize knowledge limits, self-correct, optimize resources, and adapt strategies under real-world uncertainty. We will explore this vision through the CoAgent [61] Horizon Europe project.

ACKNOWLEDGMENT

The CoAgent project receives funding from the European Union's Horizon Europe research and innovation program under grant agreement 101297191.

REFERENCES

- [1] "Autogen: A programming framework for agentic ai," 2026. [Online]. Available: <https://github.com/microsoft/autogen>
- [2] "Langgraph: Build resilient language agents as graphs," 2026. [Online]. Available: <https://github.com/langchain-ai/langgraph>
- [3] "Codex." [Online]. Available: <https://openai.com/blog/openai-codex/>
- [4] "Prism: A free workspace for scientific writing and collaboration, with gpt," 2026. [Online]. Available: <https://prism.openai.com/>
- [5] "Openclaw personal assistant agent," 2026. [Online]. Available: <https://github.com/openclaw/openclaw>
- [6] McKinsey. The state of AI in 2025: Agents, innovation, and transformation. [Online]. Available: <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai/>
- [7] F.-X. Devailly, D. Larocque, and L. Charlin, "IG-RL: Inductive graph reinforcement learning for massive-scale traffic signal control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 7496–7507, 2022.
- [8] W. Jia and M. Ji, "Multi-agent deep reinforcement learning for large-scale traffic signal control with spatio-temporal attention mechanism," *Applied Sciences*, vol. 15, no. 15, p. 8605, 2025.
- [9] Y. Liu, G. Luo, Q. Yuan, J. Li, L. Jin, B. Chen, and R. Pan, "GPLight: Grouped multi-agent reinforcement learning for large-scale traffic signal control," in *IJCAI*, 2023, pp. 199–207.
- [10] P. Patel, E. Choukse, C. Zhang, Í. Goiri, B. Warriar, N. Mahalingam, and R. Bianchini, "Polca: Power oversubscription in llm cloud providers," *arXiv preprint arXiv:2308.12908*, 2023.
- [11] Q. Liu, D. Huang, M. Zapater, and D. Atienza, "GreenLLM: Slow-aware dynamic frequency scaling for energy-efficient llm serving," *arXiv preprint arXiv:2508.16449*, 2025.

- [12] G. Wilkins, S. Keshav, and R. Mortier, "Hybrid heterogeneous clusters can lower the energy consumption of llm inference workloads," in *e-Energy*'24, 2024, pp. 506–513.
- [13] E. J. Husom, A. Goknil, L. K. Shar, and S. Sen, "The price of prompting: Profiling energy use in large language models inference," *arXiv preprint arXiv:2407.16893*, 2024.
- [14] M. Artetxe, S. Bhosale, N. Goyal, T. Mihaylov, M. Ott, S. Shleifer, X. V. Lin, J. Du, S. Iyer *et al.*, "Efficient large scale language modeling with mixtures of experts," *arXiv preprint arXiv:2112.10684*, 2021.
- [15] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi *et al.*, "Deepseek-r1 incentivizes reasoning in llms through reinforcement learning," *Nature*, vol. 645, no. 8081, pp. 633–638, 2025.
- [16] V. C. Pujol, B. Sedlak, T. Salvatori, K. Friston, and S. Dustdar, "Distributed intelligence in the computing continuum with active inference," *arXiv preprint arXiv:2505.24618*, 2025.
- [17] W. Chen, Z. You, R. Li, C. Qian, C. Zhao, C. Yang, R. Xie *et al.*, "Internet of agents: Weaving a web of heterogeneous agents for collaborative intelligence," in *ICLR*, 2025, pp. 36 374–36 411.
- [18] W. Chen, Y. Su, J. Zuo, C. Yang, C. Yuan, C.-M. Chan, H. Yu, Y. Lu, Y.-H. Hung *et al.*, "Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors," in *ICLR*, 2024, pp. 20 094–20 136.
- [19] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," *arXiv preprint arXiv:2305.14325*, 2023.
- [20] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1738–1762, 2019.
- [21] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE communications surveys & tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.
- [22] S. Bagchi, M.-B. Siddiqui, P. Wood, and H. Zhang, "Dependability in edge computing," *Communications of the ACM*, vol. 63, no. 1, pp. 58–66, 2019.
- [23] C. He, G. Liu, S. Guo, and Y. Yang, "Privacy-preserving and low-latency federated learning in edge computing," *IEEE Internet of Things Journal*, vol. 9, no. 20, pp. 20 149–20 159, 2022.
- [24] F. A. Salaht, F. Desprez, and A. Lebre, "An overview of service placement problem in fog and edge computing," *ACM Computing Surveys*, vol. 53, no. 3, pp. 1–35, 2020.
- [25] Z. Zhong, M. Xu, M. A. Rodriguez, C. Xu, and R. Buyya, "Machine learning-based orchestration of containers: A taxonomy and future directions," *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1–35, 2022.
- [26] A. Leivadreas and M. Falkner, "A survey on intent-based networking," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 625–655, 2022.
- [27] D. Firmani, F. Leotta, J. G. Mathew, J. Rossi, L. Balzotti, H. Song, D. Roman, R. Dautov, E. J. Husom, S. Sen *et al.*, "Intend: Intent-based data operation in the computing continuum," in *CAISe*'24, vol. 3692, 2024, pp. 43–50.
- [28] B. Sedlak, V. C. Pujol, I. M. de Abril, P. K. Donta, A. N. Toosi, and S. Dustdar, "Service orchestration in the computing continuum: Structural challenges and vision," *IEEE Internet Computing*, vol. 30, no. 2, pp. 88–98, 2026.
- [29] T. Masterman, S. Besen, M. Sawtell, and A. Chao, "The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey," *arXiv preprint arXiv:2404.11584*, 2024.
- [30] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with pagedattention," in *SOSP*'23, 2023, pp. 611–626.
- [31] M. Navardi, R. Aalishah, Y. Fu, Y. Lin, H. Li, Y. Chen, and T. Mohsenin, "Genai at the edge: Comprehensive survey on empowering edge devices," in *AAAI*'25, vol. 5, no. 1, 2025, pp. 180–187.
- [32] E. J. Husom, A. Goknil, M. Astekin, L. K. Shar, A. K. Sen, S. Sen, B. A. Mithassel, and A. Soylu, "Sustainable llm inference for edge ai: Evaluating quantized llms for energy efficiency, output accuracy, and inference latency," *ACM Transactions on Internet of Things*, vol. 6, no. 4, pp. 1–35, 2025.
- [33] K. Bhardwaj, "Pqv-mobile: A combined pruning and quantization toolkit to optimize vision transformers for mobile applications," *arXiv preprint arXiv:2408.08437*, 2024.
- [34] R. Agrawal, H. Kumar, and S. R. Lnu, "Efficient llms for edge devices: Pruning, quantization, and distillation techniques," in *ICMLAS*, 2025, pp. 1413–1418.
- [35] S. Al Abdul Wahid, A. Asad, and F. Mohammadi, "A survey on neuromorphic architectures for running artificial intelligence algorithms," *Electronics*, vol. 13, no. 15, p. 2963, 2024.
- [36] Model Context Protocol, "What is the Model Context Protocol (MCP)?" <https://modelcontextprotocol.io/docs/getting-started/intro>.
- [37] Agentic AI foundation, <https://aaif.io>, accessed: 2026-02-12.
- [38] Y. Yao, Z. Li, and H. Zhao, "Got: Effective graph-of-thought reasoning in language models," in *NAACL*'24, 2024, pp. 2901–2921.
- [39] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi *et al.*, "Deepseek-r1 incentivizes reasoning in llms through reinforcement learning," *Nature*, vol. 645, no. 8081, pp. 633–638, 2025.
- [40] R. Prabhu, A. Nayak, J. Mohan, R. Ramjee, and A. Panwar, "vattention: Dynamic memory management for serving llms without pagedattention," in *ASPLOS*'25, 2025, pp. 1133–1150.
- [41] P. Letswalo, "MemGPT: Engineering semantic memory through adaptive retention and context summarization," *Information Matters*, vol. 5, no. 10, 2025.
- [42] Zep, "Agent memory at enterprise scale," <https://www.getzep.com/>.
- [43] Lightbend, Inc., "Akka: Build powerful reactive, distributed, and resilient message-driven applications," <https://akka.io>.
- [44] M. Etheredge, T. Fahringer, F. Erlacher, E. Kohler, S. Pedratscher, J. Aznar-Poveda, N. Saurabh, and A. Lebre, "Collaborative state machines: A better programming model for the cloud-edge-iiot continuum," *arXiv preprint arXiv:2507.21685*, 2025.
- [45] Q. Wang and D. Oswald, "Confidential computing on heterogeneous cpu-gpu systems: Survey and future directions," *ACM Computing Surveys*, 2026.
- [46] Q. Zhu, S. W. Loke, R. Trujillo-Rasua, F. Jiang, and Y. Xiang, "Applications of distributed ledger technologies to the internet of things: A survey," *ACM computing surveys*, vol. 52, no. 6, pp. 1–34, 2019.
- [47] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Deroncourt, T. Yu, R. Zhang, and N. K. Ahmed, "Bias and fairness in large language models: A survey," *Computational linguistics*, vol. 50, no. 3, pp. 1097–1179, 2024.
- [48] S. Ghosh-Dastidar and H. Adeli, "Spiking neural networks," *International journal of neural systems*, vol. 19, no. 04, pp. 295–308, 2009.
- [49] N. Kasabov, K. Dhoble, N. Nuntalid, and G. Indiveri, "Dynamic evolving spiking neural networks for on-line spatio-and spectro-temporal pattern recognition," *Neural Networks*, vol. 41, pp. 188–201, 2013.
- [50] D. Marković, A. Mizrahi, D. Querlioz, and J. Grollier, "Physics for neuromorphic computing," *Nature Reviews Physics*, vol. 2, no. 9, pp. 499–510, 2020.
- [51] S. Goldstein and C. D. Kirk-Giannini, "A case for ai consciousness: Language agents and global workspace theory," *arXiv preprint arXiv:2410.11407*, 2024.
- [52] EUCloudEdgeIoT, "Research & Innovation Roadmap," <https://eucloudedgeiot.eu/research-innovation-roadmap/>.
- [53] H. Kokkonen, L. Lovén, N. H. Motlagh, A. Kumar, J. Partala, T. Nguyen, V. C. Pujol, P. Kostakos *et al.*, "Autonomy and intelligence in the computing continuum: Challenges, enablers, and future directions for orchestration," *arXiv preprint arXiv:2205.01423*, 2022.
- [54] R. Sapkota, K. I. Roumeliotis, and M. Karkee, "Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges," *Information Fusion*, p. 103599, 2025.
- [55] D. Amalfitano, A. Metzger, M. Autili, T. Fulcini, T. Hey, J. Keim *et al.*, "A research roadmap for augmenting software engineering processes and software products with generative ai," *ACM TOSEM*, 2026.
- [56] "Amazon bedrock," Amazon Web Services product page. [Online]. Available: <https://aws.amazon.com/bedrock/>
- [57] "Azure openai service." [Online]. Available: <https://learn.microsoft.com/en-us/azure/ai-services/openai/overview>
- [58] "Vertex ai generative ai," Google Cloud documentation. [Online]. Available: <https://cloud.google.com/vertex-ai/generative-ai/docs/overview>
- [59] "Ollama." [Online]. Available: <https://github.com/ollama/ollama>
- [60] Agent skills. [Online]. Available: <https://agentskills.io/home>
- [61] European Commission, "Cognitive Computing Continuum for Large-Scale Distributed GenAI Agents (CoAgent)," CORDIS: EU Research Results, 2026, horizon Europe project, Grant agreement ID: 101297191. [Online]. Available: <https://cordis.europa.eu/project/id/101297191>