NEXUS FORUM.EU

RESEARCH & INNOVATION ROADMAP

Cognitive Computing Continuum

WHITE PAPER

Final Version - June 2026

www.eucloudedgeiot.eu

Executive summary

The 2026 NexusForum.EU Research & Innovation Roadmap sets out a European capability agenda for the Cognitive Computing Continuum: a landscape in which AI, data, cloud, edge, HPC, telecommunications, hardware, software, energy systems and sectoral adoption increasingly interact. The roadmap is not based on a single infrastructure model. Instead, it recognises that Europe will need a combination of specialised AI and HPC infrastructures, open hardware-software ecosystems, cloud-edge and telco-edge capabilities, federated systems, sectoral platforms, testbeds and operational support structures.

STRATEGIC DESTINATIONS

The roadmap is guided by six Strategic Destinations that define the outcomes Europe should pursue through research and innovation.

First, Europe needs scalable and energy-efficient AI and data processing. This requires the ability to train, fine-tune, deploy and operate AI and data workloads across HPC, cloud, edge and specialised accelerator environments, while addressing energy use, cooling, grid integration and sustainability.

Second, Europe needs an AI stack built on an open European hardware and computing ecosystem. This includes processors, accelerators, compiler toolchains, runtimes, AI libraries, middleware and developer tools that strengthen European autonomy and competitiveness.

Third, Europe needs a competitive European AI and machine learning ecosystem. This requires capacity to build, evaluate, deploy and govern advanced AI systems, including frontier models, open-source (open-weight) models, agentic AI, AI Factories, sectoral AI platforms and AI operationalisation capabilities.

Fourth, Europe needs secure, sovereign and interoperable European computing capabilities. In some domains, this will require federation across providers, infrastructures and borders. In others, it will require secure hardware, trusted software supply chains, certifiable systems, open standards, resilience and auditability.

Fifth, Europe needs advanced digitalisation and AI adoption in industry and public sectors. This includes embedded and edge systems, industrial AI, OT/IT convergence, data spaces, digital twins, robotics, private clouds, sectoral platforms and large-scale validation environments.

Sixth, Europe needs leadership in disruptive and emerging computing paradigms. This includes neuromorphic computing, quantum technologies, hybrid quantum-classical systems, photonic/optical processing technologies, post-exascale computing, novel accelerators and other non-conventional architectures.

FROM PILLARS TO STRATEGIC DESTINATIONS

The roadmap is structured around five Pillars that translate these Strategic Destinations into concrete research and innovation priorities. The Strategic Destinations describe the desired European outcomes, while the Pillars organise the capability areas through which those outcomes can be achieved.

© 2024-2026 NexusForum.EU

Pillar I addresses the AI and machine learning capabilities needed for a competitive European AI ecosystem. It contributes especially to frontier AI capacity, agentic AI, distributed AI, open AI stacks and AI deployment across heterogeneous infrastructures.

Pillar II focuses on the infrastructure foundations of the Cognitive Computing Continuum. It connects energy-efficient AI processing, telco cloud-edge convergence, cloud-edge interconnection and federated infrastructure capabilities.

Pillar III addresses the hardware-software stack required for European autonomy and long-term competitiveness. It links semiconductor design, RISC-V, AI accelerators, inference hardware, memory, interconnects and emerging processor architectures.

Pillar IV focuses on the tooling needed to manage, optimise and operate heterogeneous and federated computing systems. It connects AI-native infrastructure management, developer productivity, orchestration, lifecycle automation and system-level optimisation.

Pillar V focuses on adoption, testbeds, benchmarks and sectoral implementation. It connects the roadmap to industrial and public-sector digitalisation, OT/IT convergence, operational AI, robotics support and large-scale validation.

Together, the Strategic Destinations and Pillars form a structured agenda: Europe must build the AI systems, infrastructure, hardware-software stacks, operational tooling and adoption mechanisms required to deploy advanced AI and computing capabilities at scale.

PILLAR I: FOUNDATIONAL AI AND ML TECHNOLOGIES

Agentic AI and neurosymbolic AI

Europe should strengthen its capacity to develop, operate and govern frontier AI systems while taking a pragmatic route to agentic AI, open-source (open-weight) large language models and neurosymbolic approaches. The priority is not only to advance model capability, but to create trustworthy AI systems that can reason, use tools, integrate knowledge, retain memory and execute workflows reliably across real-world environments.

Key priorities are to:

- Build European capacity to create and operate frontier AI models.
- Advance a pragmatic approach to Agentic AI and open-source Large Language Models.
- Develop dynamic agentic orchestration and workflow execution graphs.
- Strengthen harness engineering through memory, knowledge integration and verified tool use.
- Enable Agentic AI and LLMs for devices, edge environments and mobile phones.

AI and Data Workloads across Heterogeneous Computing Environments

AI and data workloads increasingly need to operate across heterogeneous infrastructures, including HPC, cloud, edge, telco edge and device environments. Europe should invest in software abstractions, portability mechanisms and federated AI capabilities that allow

© 2024-2026 NexusForum.EU

workloads, models and stateful applications to move, adapt and operate securely across this landscape.

Key priorities are to:

- Develop stateful continuous analytics and distributed state abstractions.
- Support portable AI applications and open AI technology stacks.
- Enable stateful AI mobility and model migration destinations.
- Advance federated and distributed AI.

PILLAR II: COGNITIVE COMPUTING CONTINUUM CONVERGENCE & INFRASTRUCTURE

Sustainable and Energy-efficient AI Infrastructure

As AI infrastructure expands, Europe must address energy consumption, cooling, grid integration and sustainability as core design requirements. Datacentres, AI Factories, edge infrastructures and continuum systems should be treated not only as compute assets, but also as flexible participants in Europe's energy system.

Key priorities are to:

- Develop energy and cooling solutions for high-density datacentres.
- Align renewable energy supply with datacentre workloads so datacentres can operate as flexible energy assets in the smart grid, and use their waste heat.
- Advance energy-grid-aware scheduling across the Computing Continuum.

Telco Cloud-Edge: Telco as One of the Main Tenants and Infrastructure Providers

Telecommunications infrastructure is a strategic part of the European computing continuum. Telco cloud-edge convergence can provide distributed and AI-native capabilities for sectors that depend on reliable connectivity and low latency.

Key priorities are to:

- Advance Open Radio Access Networks.
- Enable seamless data connectivity and predictive handover across networks.
- Develop AI-native telco cloud-edge capabilities.
- Build federated telco edge and Network-as-a-Service models.

Federations and Cloud-Edge AI Interconnect Framework

Europe needs interoperable cloud-edge frameworks that allow services, data and AI workloads to operate across providers and infrastructure domains where federation creates strategic value. This includes technical interconnect mechanisms as well as market-enabling structures.

Key priorities are to:

- Create a cloud-edge AI Interconnect Framework.

© 2024-2026 NexusForum.EU

- Develop decentralised cross-provider marketplaces.

PILLAR III: AN AI–ENABLING HARDWARE–SOFTWARE STACK

European Semiconductor Design

Europe's long-term AI and computing competitiveness depends on the ability to design, integrate and use competitive processor technologies. RISC-V and open hardware-software ecosystems are central to strengthening European autonomy, industrial capacity and innovation.

Key priorities are to:

- Develop competitive RISC-V processors and systems for European and global markets.
- Ensure availability of the necessary software stacks.

AI Inference Hardware: AI Accelerators, Memory, and Interconnects

AI inference is becoming a critical workload across datacentres, edge systems, industrial environments and devices. At the same time, LLM inference and agentic AI are becoming increasingly bound by memory and interconnects between memory, storage, accelerator, and CPU. This opens up new opportunities for new AI inference processing architectures that are different from GPUs. Europe should strengthen the full inference hardware stack, including accelerators, memory systems and interconnects, so that AI deployment can become more energy-efficient, scalable and suited to diverse application environments.

Key priorities are to:

- Develop processing and memory architectures optimised for LLM and agentic AI inference.
- Develop high-bandwidth, energy-efficient and reconfigurable interconnects for AI systems.
- Adopt a hardware-software-algorithm codesign approach across European processing initiatives.
- Advance emerging optical, or photonics-based, AI inference chips.

Emerging Processor Architectures

Europe should invest in disruptive and emerging computing paradigms that may shape future AI, simulation and optimisation capabilities. This includes both nearer-term hardware-software integration and longer-term research into fundamentally new computing approaches.

Key priority areas are to:

- Advance neuromorphic systems.
- Develop hybrid quantum and classical computing fusion.
- Integrate quantum computing infrastructure.

© 2024-2026 NexusForum.EU

PILLAR IV: AI–TOOLING FOR INTELLIGENT INFRASTRUCTURE MANAGEMENT AND DEVELOPER PRODUCTIVITY

Federation and System-level Optimisation for the Computing Continuum

Heterogeneous and federated computing systems require new optimisation, orchestration and trust mechanisms. Europe should develop tools that allow resources to be used efficiently across providers and infrastructure domains while supporting security, interoperability and decentralised coordination.

Key priorities are to:

- Advance continuum and cross-provider optimisation.
- Develop federated and decentralised continuum orchestration, trust and market mechanisms.

AI-native Management and Application Development for a Heterogeneous Computing Continuum

The complexity of the computing continuum will exceed what can be managed through manual operations alone. AI-native management should support deployment, lifecycle automation, maintenance, observability and incident response, while improving developer productivity across heterogeneous environments.

Key priorities are to:

- Enable “zero-touch” deployment and lifecycle automation.
- Develop AI-native operations, maintenance and incident response for continuum systems.

PILLAR V: SECTORAL ADOPTION, SUPPORT STRUCTURES, TESTBEDS & BENCHMARKS

Convergence of Operational Technologies and Information Technologies

Industrial and public-sector adoption will depend on the secure convergence of operational technology and information technology. Europe should support resilient industrial AI, secure data integration, data spaces, digital twins, lifecycle management and cybersecurity for cyber-physical environments.

Key priorities are to:

- Develop on-premises OT edge and resilient industrial AI.
- Enable secure OT/IT data integration and industrial data spaces.
- Advance industrial digital twins and AI-enabled operational optimisation.
- Strengthen cybersecurity and lifecycle management for converged OT/IT systems.

AI Operationalization

© 2024-2026 NexusForum.EU

Europe must move from AI experimentation to operational deployment at scale. This requires large-scale testbeds, validation environments, orchestration capabilities and support for autonomous AI agents, robotics and multi-agent systems that can be evaluated and adopted in real-world settings.

Key priorities are to:

- Establish large-scale testbeds.
- Support autonomous AI agents, multi-agent orchestration and robotics.
- Advance evaluation and monitoring of Human–AI collaboration in deployed systems

OVERALL DIRECTION

Across the five Pillars, the roadmap calls for a balanced European R&I agenda that

- strengthens frontier AI and agentic systems,
- builds sustainable and interoperable continuum infrastructure,
- secures control over critical hardware-software stacks,
- equips developers and operators with AI-native tooling,
- and accelerates adoption with ecosystem advances in industrial systems and with testbeds and operational support.

The overarching objective is to ensure that Europe can build, deploy and govern advanced AI and computing capabilities in ways that reinforce competitiveness, sovereignty, sustainability, trust and sectoral impact.

© 2024-2026 NexusForum.EU

Table of contents

Strategic Destinations **2**

From Pillars to Strategic Destinations **2**

Pillar I: Foundational AI and ML Technologies **3**

Pillar II: Cognitive Computing Continuum Convergence & Infrastructure **4**

Pillar III: An AI-enabling Hardware-Software Stack **5**

Pillar IV: AI-tooling for Intelligent Infrastructure Management and Developer Productivity . **6**

Pillar V: Sectoral Adoption, Support Structures, Testbeds & Benchmarks **6**

Overall direction **7**

1 Introduction and the aim of this roadmap **13**

1.1 The aim and purpose of this roadmap 15

1.2 Roadmap development methodology and stakeholder engagement 16

1.3 Roadmap structure, technology pillars, and capability map 17

1.4 Strategic Destinations for Europe 20

A European perspective **22**

2 European policy and competitiveness **23**

3 Major relevant European initiatives **26**

3.1 Establishing a European ecosystem based on RISC-V 26

3.2 EuroHPC and the establishment of AI Factories 28

3.3 Access to European data, common data spaces, and data privacy 29

4 From Strategic Destinations to R&I Priorities: How to Read the Roadmap **32**

4.1 The five pillars of the roadmap 32

4.2 Strategic Destination pathways 34

4.3 High-level summary of recommendations 37

4.4 Using the detailed topic roadmaps 38

5 Agentic AI and neurosymbolic AI **40**

5.1 Priority: Build European capacity to create and operate frontier AI models 43

5.2 Priority: A pragmatic approach to Agentic AI and open-source Large Language Models 44

5.3 Priority: Dynamic agentic orchestration and workflow execution graphs 45

5.4 Priority: Harness engineering with memory, knowledge integration and verified tool use 46

5.5 Priority: Agentic AI and LLMs for devices, edge, and mobile phones 48

6 AI and Data Workloads across Heterogeneous Computing Environments **51**

6.1 Priority: Stateful continuous analytics and distributed state abstractions 54

© 2024-2026 NexusForum.EU

6.2 Priority: Portable AI applications and open AI technology stacks 55
6.3 Priority: Stateful AI mobility and model migration Destinations 58
6.4 Priority: Federated and distributed AI 60
7 Sustainable and Energy-efficient AI infrastructure 64
7.1 Priority: Energy and cooling solutions for high-density datacenters 66
7.2 Priority: Solving renewable energy supply and datacenter workloads as a flexible energy asset in the smart grid 67
7.3 Priority: Energy-grid-aware scheduling in the Computing Continuum 68
8 Telco Cloud-Edge: Telco as One of the Main Tenants and Infrastructure Providers ... 70
8.1 **Open Radio Access Networks (Open RAN)** 72
8.1.1 **Priority:** Secure and Governable Multi-Vendor Open RAN Architectures 73
8.1.2 **Priority:** AI-Native and Cross-Layer Optimisation in Open RAN 73
8.2 Seamless data connectivity and predictive handover across different networks 74
8.2.1 **Priority:** Predictive and Resilient Handover Across Heterogeneous Networks 75
8.2.2 **Priority:** Application-Aware Connectivity and Graceful Degradation Mechanisms 76
8.3 AI-Native Telco Cloud-Edge 77
8.3.1 **Priority:** AI-Driven Automation of Telco Operations 78
8.3.2 **Priority:** Domain-Specific and Trustworthy AI Models for Telecom 79
8.4 Federated Telco Edge and Network-as-a-Service (NaaS) 80
8.4.1 **Priority:** Pan-European Telco Edge Federation and Programmable Network Capabilities .. 81
9 Federations and Cloud-Edge AI Interconnect Framework 83
9.1 Priority: Create a cloud-edge AI Interconnect Framework 85
9.2 Priority: Decentralized cross-provider marketplace 86
10 European Semiconductor Design 88
10.1 Priority: Develop competitive RISC-V processors and systems for European and global markets 90
10.2 Priority: **Ensure availability of necessary Software Stacks** 91
11 AI inference hardware: AI Accelerators, Memory, and Interconnects 94
11.1.1 Priority: Processing and memory architectures for AI inference chips 97
11.1.2 Priority: Develop optical/photonics AI inference chips 99
12 Emerging Processor Architectures 101
12.1 Neuromorphic systems 101
12.1.1 Priority: Hardware–software co-design and interoperability for continuum integration 103
12.1.2 Priority: Standardised models, training methods, and software maturity for adoption 104
12.2 Hybrid quantum and classical computing fusion 105
12.2.1 Priority: Hybrid algorithm workflows and orchestration for quantum–classical fusion 107

© 2024-2026 NexusForum.EU

12.2.2 Priority: Cloud delivery model for quantum-enabled computing services 108
12.3 Integration of quantum computing infrastructure 109
12.3.1 Priority: Federated QC–HPC–cloud infrastructure and service readiness 111
12.3.2 Priority: Transition to fault-tolerant quantum computing through full-stack scale-up and procurement pull 112
13 Federation and system-level optimisation for the Computing Continuum 115
13.1 Priority: Continuum and cross-provider optimisation 117
13.2 Priority: Federated and decentralised continuum orchestration, trust, and market mechanisms 117
14 AI-native management and application development for a heterogeneous computing continuum 119
14.1 Priority: “Zero-touch” deployment and lifecycle automation 121
14.2 Priority: AI-native operations maintenance and incident response for continuum systems 122
15 Convergence of Operational Technologies and Information Technologies 125
15.1 Priority: On-premises OT edge and resilient industrial AI 128
15.2 Priority: Secure OT/IT data integration and industrial data spaces 129
15.3 Priority: Industrial digital twins and AI-enabled operational optimisation 131
15.4 Priority: Cybersecurity and lifecycle management for converged OT/IT systems 132
16 AI Operationalization 134
16.1 Priority: **Large-Scale Testbeds** 137
16.2 Priority: Autonomous AI Agents, Multi-Agent Orchestration, and Robotics Support 139
16.3 Priority: Evaluation and Monitoring of Human–AI Collaboration in Deployed Systems 141
17 Conclusions 144
18 List of Contributors 146

© 2024-2026 NexusForum.EU

List of figures

Figure 1: Overview of the relationships between the Strategic Destinations and Pillars 14

Figure 2 : An overview of the capability landscape for the Cognitive Computing Continuum. 18

Figure 3: Overview of the five Pillars and the technology areas under each Pillar. 33

Figure 4: Illustration of the prefill and decode stages in LLM inference 95

© 2024-2026 NexusForum.EU

NEXUS FORUM.EU

PART A:

Introduction and Background

1 INTRODUCTION AND THE AIM OF THIS ROADMAP

Computing, networking and data infrastructures have historically evolved as separate vertical stacks, each with its own hardware, software, operational practices, procurement models and governance requirements. Industrial automation and operational technology systems were designed for deterministic behaviour, safety, long lifecycles and high availability. Cloud systems, by contrast, emerged around elasticity, software-defined infrastructure, rapid provisioning, multi-tenancy and continuous service evolution. High-performance computing developed around large-scale scientific simulations, and specialised performance requirements, while embedded and edge systems evolved around locality, reliability, energy constraints and real-time interaction with the physical world.

These worlds are now converging. Advances in IoT have led to a sharp increase in operational data being captured and are placing advanced compute capabilities closer to the devices. Cloud-native and SaaS ecosystems have changed how digital capabilities are delivered and operated. 5G networks and future 6G capabilities are advancing how connectivity and compute are integrated. Modern AI systems require increasingly complex data pipelines, model lifecycle management, compute capacity for training and inference, evaluation mechanisms, deployment environments and safe update processes.

Emerging applications in areas such as connected and autonomous vehicles, cooperative mobility systems, smart logistics, remote industrial operations, AR/VR applications, smart energy systems and emergency response services, demonstrate the necessity of having computing capabilities distributed across some combination of devices, edge nodes, telco infrastructure, cloud platforms and, where relevant, HPC resources.

The result is a shift from isolated vertical stacks towards a broader computing continuum: a capability landscape in which data processing, AI workloads and digital services can be developed, trained, deployed, operated and governed across device edge, industrial edge, telco edge, cloud, HPC and emerging accelerator infrastructures. Workloads and services may be placed in different environments according to requirements for latency, performance, cost, energy, trust, security, resilience, data governance and regulatory compliance. This requires orchestrating resources across such heterogeneous platform capabilities, and new tools to make it easy for developers to deploy applications across them.

More recently, the rise of the Agentic AI paradigm is changing not only how humans interact with computers, but also reinventing what software is and how it is created. These new AI systems connect large language models to tools, memory, workflows, applications, and software systems. Natural language is becoming a programming language, and an important part of the software engineering toolset.

In this roadmap, the **European Cognitive Computing Continuum** refers to this broader capability landscape. It does not imply that every workload must move dynamically across all infrastructure domains, or that every system must be federated across multiple providers. Rather, it captures the need for Europe to build strong capabilities across multiple computing environments, and to make those environments interoperable, portable or federated where this creates value.

© 2024-2026 NexusForum.EU

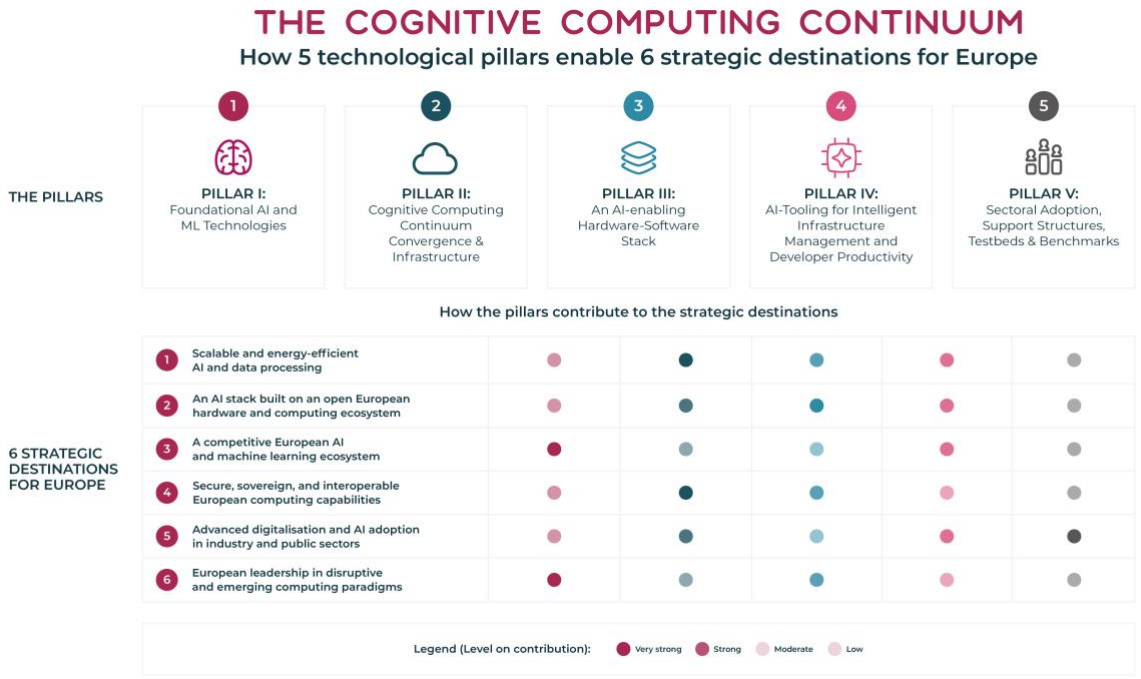


Figure 1: Overview of the relationships between the Strategic Destinations and Pillars

The term **cognitive** reflects the growing role of AI in this landscape. AI is not only a workload that requires compute, data and deployment environments. It is also becoming part of the operational and coordination layer of digital infrastructure itself. AI can support workload placement, predictive operations, service management, incident response, energy optimisation, software engineering, compliance monitoring and agentic workflows. This creates new opportunities, but also new requirements for reliability, explainability, safety, cybersecurity and governance.

In some domains, Europe will need federated and multi-provider architectures, particularly to avoid lock-in, enable cross-border services, support data spaces, connect public and private infrastructures, improve resilience or create interoperable markets. In other domains, progress will depend primarily on specialised hardware, vertically integrated systems, AI Factories, open software stacks, sector-specific platforms, single-provider infrastructures, or tightly optimised hardware-software co-design.

Purpose of this roadmap and how it is structured

The purpose of the roadmap is to identify research and innovation capabilities needed for Europe to build, deploy and govern advanced AI, data and computing systems across this diverse landscape. It connects AI, data, cloud, edge, HPC, telco infrastructure, semiconductors, software, cybersecurity, sustainability and sectoral adoption into a coherent European capability agenda.

The roadmap therefore provides a structured view of the capability areas in which European research and innovation can strengthen competitiveness, digital sovereignty, sustainability and AI adoption, while recognising that different domains require different architectural and deployment models.

© 2024-2026 NexusForum.EU

Part A of this document provides the European context and background, while **Part B** (Sections 4-16) introduces the research & innovation roadmap, outlining concrete priorities and recommendations. Part B is organized according to five technology pillars, see Section 1.2, with different sections for different technology areas. These pillars are connected to six Strategic Destinations introduced in Section 1.3.

1.1 The aim and purpose of this roadmap

This Research & Innovation Roadmap for the Cognitive Computing Continuum is developed as part of the EU-funded project NexusForum.EU. The purpose of the project, and this roadmap, is to support European computing ecosystems and provide support and recommendations to the European Commission in the development of their research & innovation agenda and funding calls.

The primary audience is the European research and innovation ecosystem, including the European Commission, Member States, research organisations, technology providers, infrastructure operators, industrial users, SMEs, startups and standardisation communities.

This is the third iteration of the roadmap, following earlier iterations completed in 2024. When the NexusForum.EU project was conceived in early 2023, the main strategic priorities for the computing continuum were centred on multi-provider cloud-edge-IoT infrastructures, telco edge, data federations, the expansion of edge nodes in Europe, and the creation of a European single market for data through Common European Data Spaces.

Since then, the technological and policy landscape has changed significantly. Rapid advances in generative AI, agentic AI, AI accelerators, quantum technologies, neuromorphic systems, open hardware, data-centre investment, energy constraints and geopolitical competition have broadened the scope of what a European computing roadmap needs to address. At the same time, earlier priorities around cloud-edge federation, interoperability, data spaces, telco-edge convergence and digital sovereignty remain important.

The aim of this roadmap is to provide a vision and set of R&I priorities for Europe's future computing capability base. It preserves continuity with earlier work on cloud-edge-IoT, telco edge and data federations, but expands the scope to reflect the rapid development of AI, hardware, software, quantum and neuromorphic computing, sustainability requirements and sectoral digitalisation.

The roadmap is not based on a single architectural model. Instead, it recognises that different technologies and use cases require different deployment models. Some priorities concern specialised AI and HPC infrastructures. Some concern vertically integrated hardware-software stacks. Some concern open-source and open-standard ecosystems. Some concern sector-specific platforms, industrial edge systems or private cloud environments. Others concern federated multi-provider environments, cross-border services and shared European infrastructure.

The roadmap therefore organizes the vision into complementary capability areas that are related but not necessarily mutually dependent. This makes it possible to support progress in AI, hardware, software, cloud, edge, HPC, data, connectivity, cybersecurity and sustainability

© 2024-2026 NexusForum.EU

without assuming that all of these domains must converge into one uniform infrastructure model.

1.2 Roadmap development methodology and stakeholder engagement

The Research and Innovation Roadmap was developed through an iterative methodology combining evidence gathering, expert synthesis, stakeholder consultation and progressive refinement. The process built on technological capabilities mapping and analysis of the state of the art across AI, data, cloud, edge, IoT, HPC, telecommunications, semiconductors, software, cybersecurity, sustainability and sectoral adoption. It was also informed by the analysis of 418 Horizon Europe projects funded between 2021 and 2025, which helped identify areas of European strength, fragmentation and potential capability gaps.

The roadmap was progressively developed across three iterations. Earlier versions established the initial scope, thematic structure and basis for consultation, while subsequent revisions incorporated new evidence and responded to changes in the technological, industrial and policy landscape. These updates reflected developments including generative and agentic AI, AI Factories and accelerators, quantum and neuromorphic computing, RISC-V and open hardware ecosystems, data-centre capacity and energy constraints, and growing concerns around competitiveness, resilience and digital sovereignty.

A central methodological step was the creation of a common framework linking six Strategic Destinations, five technology Pillars and detailed research and innovation priorities. The Strategic Destinations describe the long-term outcomes that European research and innovation should support, while the Pillars organise the technological, operational and adoption-related capabilities required to achieve them. Individual roadmap topics were developed by considering capability gaps, technology trends, European policy priorities, industrial needs, infrastructure requirements and barriers to adoption. Where relevant, the roadmap also identifies enabling measures such as standardisation, benchmarking, testbeds, open-source development, ecosystem coordination, procurement pathways and deployment support.

Stakeholder engagement and validation took place continuously throughout the roadmap process. Feedback was gathered through the NexusForum.EU community platform, thematic Working Groups, webinars, workshops, roundtables, consultation sessions, expert panels and NexusForum Summits. The Working Groups brought together participants from research, industry and policy to consider areas including AI, cloud-edge-IoT convergence, interoperability and standardisation, cybersecurity and resilience, digital sovereignty, sustainability and emerging computing infrastructures. These activities were used to test proposed priorities, identify missing topics, refine recommendations and ensure that the roadmap reflected practical challenges and opportunities across the Cognitive Computing Continuum.

The roadmap was also validated through international engagement, particularly with stakeholders from Japan and the Republic of Korea. International workshops and cooperation activities enabled comparison of European priorities with global developments in AI, cloud-edge infrastructure, cybersecurity, semiconductors, 6G, quantum technologies, interoperability and digital sovereignty. Taken together, the methodology combined analytical

© 2024-2026 NexusForum.EU

mapping and expert judgement with participatory, stakeholder-driven validation, helping to ensure that the final roadmap was grounded in evidence, responsive to technological change and relevant to the wider European research and innovation ecosystem.

1.3 Roadmap structure, technology pillars, and capability map

The European Cognitive Computing Continuum is not a single technology, platform, infrastructure or market segment. It is better understood as a capability landscape spanning computing, networking, data, software, hardware, services and operations. Some parts of this landscape will need to interoperate across providers, borders and governance domains. Other parts will advance through specialised systems, open ecosystems, sectoral platforms, private infrastructures or vertically integrated hardware-software stacks.

The figure below is introduced as a simplified capability map to illustrate this point. It is not complete, but it provides another view of the computing continuum landscape. The purpose of the map is not to describe or prescribe a specific architecture, but to show how different capability areas relate across the computing landscape.

© 2024-2026 NexusForum.EU

Figure 2 : An overview of the capability landscape for the Cognitive Computing Continuum.

The roadmap covers multiple deployment models. Some priorities concern specialised AI or HPC infrastructures, such as AI Factories, EuroHPC systems and post-exascale platforms. Some concern vertically integrated hardware-software stacks, such as RISC-V systems, AI accelerators, embedded AI and automotive platforms. Some concern open ecosystem stacks, including open-source middleware, compiler toolchains, runtimes and software libraries. Others concern single-provider or private environments, sectoral platforms, federated infrastructures, multi-provider market environments or hybrid continuum deployments.

Across these deployment models, the capability map identifies four main infrastructure domains: HPC, cloud, edge servers and the device edge. These domains differ in location, ownership, performance envelope, operational model and regulatory context. HPC provides large-scale capability for AI training, simulation, scientific computing and high-performance

© 2024-2026 NexusForum.EU

analytics. Cloud provides elastic service environments, data management and scalable deployment models. Edge servers support low-latency, local, sector-specific and privacy-sensitive workloads. The device edge connects embedded intelligence, sensing, actuation, robotics, industrial systems and cyber-physical environments to the wider computing landscape.

The map also identifies three recurring technology layers: hardware, runtimes and middleware, and application services. The hardware layer covers processors, accelerators, memory, interconnects, storage, energy systems, embedded platforms and data-centre infrastructure. This is where Europe's ambitions for open hardware, RISC-V, AI accelerators, semiconductor design, HPC technologies and energy-efficient computing become concrete.

The runtimes and middleware layer provides the software capabilities needed to make heterogeneous systems usable and efficient. It includes technologies for hardware abstraction, workload placement, portability, distributed state, orchestration, security, trust, observability and developer productivity. In some cases, these capabilities support cross-provider federation; in others they support specialised or vertically integrated systems.

The application services layer is where infrastructure becomes useful to sectors, public administrations, researchers, AI developers and industrial users. It includes AI services, data services, digital twins, robotics support, analytics, model lifecycle services, data-space integration and sector-specific digitalisation capabilities.

The map also highlights two enabling toolchains: hardware design tooling and software engineering tools. These should be treated as strategic capabilities in their own right. Hardware design tooling is necessary if Europe wants to strengthen its processor, accelerator and semiconductor ecosystem. Software engineering tools are equally important because advanced computing capabilities will only be adopted if developers can build, test, deploy, monitor, secure and evolve applications across heterogeneous environments.

Finally, the map identifies horizontal foundations: system architecture, infrastructure and operation domains, and service management and operations. Some of these capabilities are federation-critical; others are useful within specialised, private or vertically integrated systems.

The roadmap is organised into five pillars.

Pillar I focuses on foundational AI and machine learning technologies.

Pillar II addresses continuum convergence and infrastructure, including telco cloud-edge and IT/OT convergence.

Pillar III covers the hardware-software stack, including processors, accelerators, software stacks and disruptive computing paradigms.

Pillar IV addresses AI-enabled infrastructure management and developer productivity.

Pillar V addresses sectoral adoption, support structures, testbeds and benchmarks.

Sustainability, **cybersecurity**, **interoperability**, **open source** and **digital sovereignty** are treated as transversal topics.

In summary, the six strategic destinations provide the direction of travel, while the Pillars and the capability map provides the organising logic for the roadmap. Together, they connect

© 2024-2026 NexusForum.EU

Europe's ambitions for competitiveness, digital sovereignty, sustainability and AI leadership to concrete R&I priorities across AI, data, hardware, software, cloud, edge, HPC, telco infrastructure and emerging computing paradigms.

1.4 Strategic Destinations for Europe

In developing this roadmap, **six strategic destinations** were used as working hypotheses to describe the outcomes that European research and innovation needs to be strong in. Each topic mentioned in the roadmap (Part B) refers to the relevant strategic destinations to which it contributes.

1. **Scalable and energy-efficient AI and data processing**

   Europe needs scalable and energy-efficient capabilities for training, fine-tuning, deploying and operating AI and data workloads across HPC, cloud, edge and specialised accelerator environments. Some workloads will benefit from federated placement across infrastructures, while others will be best served by specialised, tightly optimised systems.

2. **An AI stack built on an open European hardware and computing ecosystem**

   Europe should strengthen open and competitive hardware-software stacks, including processors, accelerators, compiler toolchains, runtime environments, AI libraries, middleware and developer tools. These stacks should support European autonomy and competitiveness whether deployed in specialised systems, single-provider platforms or federated environments.

3. **A competitive European AI and machine learning ecosystem**

   Europe needs the capacity to train, fine-tune, evaluate, deploy and govern advanced AI systems. This includes AI Factories, sectoral AI platforms, industrial adoption, AI in science, trusted datasets, model evaluation, agentic AI, and operational tooling. Federation may improve access and portability, but many AI advances will occur independently of multi-provider infrastructure.

4. **Secure, sovereign, and interoperable European computing capabilities**

   Europe needs secure, sovereign and interoperable computing capabilities. In some cases this requires federation across providers, borders and infrastructures. In others it requires certifiable single-provider systems, secure hardware, trusted software supply chains, open standards, auditability and resilience.

5. **Advanced digitalisation and AI adoption in industry and public sectors**

   Industrial and public-sector adoption will involve different architectural models: embedded and edge systems, private clouds, AI Factories, sectoral platforms, data spaces, digital twins, robotics environments and federated cross-border services. The roadmap should support this diversity rather than assume a single continuum architecture.

6. **European leadership in disruptive and emerging computing paradigms**

   Europe should build leadership in emerging computing paradigms such as neuromorphic computing, quantum technologies, post-exascale systems, novel accelerators and non-

© 2024-2026 NexusForum.EU

conventional architectures. These domains require deep research, hardware-software co-design, testbeds and ecosystem development, but do not inherently depend on federated multi-provider infrastructures.

© 2024-2026 NexusForum.EU

A European perspective

Despite significant initiatives at EU level to harmonize regulatory standards,[1] the digital market in Europe remains fragmented, with local markets individually lacking the critical mass for players to scale and compete with their global counterparts. This fragmentation can hinder the development of unified technological solutions and the emergence of a fully integrated European Digital Single Market. Therefore, harmonizing standards and creating interoperable systems across borders is essential for technological sovereignty.

The report *"The future of European competitiveness"*[2] (the *Draghi report*) assesses the current state of European competitiveness, with a particular focus on industrial policy, its caveats and hopes for its future. Due to geopolitical instability, the urge of not depending on other countries is higher, because the EU has realized that dependency easily engenders instability. According to this report, there has been a shift in paradigm, "The era of rapid world trade growth looks to have passed, with EU companies facing both greater competition from abroad and lower access to overseas markets". According to this report, Europe must radically change for digitalisation and decarbonisation to take place in the European economy, and investments need to rise to 5 % of GDP.

The above report further identifies three key areas for growth to take place and to close the innovation gap with the US and China, i) activate a joint plan aimed at strengthening competitiveness ii) decarbonising the economy; iii) take action to enhance security, while reducing dependencies on third parties. In particular, three main barriers prevent Europe from growth.

- Market *fragmentation* is high, which in return drives innovative companies away to more profitable continents. Administrative and regulatory burdens within the European Internal Market prevent innovative companies from thriving.
- The EU is not taking *full advantage* of its common resources. In the field of innovation, collaboration is weak, despite the known spillover benefits of investing in cutting-edge technologies. Here Draghi gives an interesting figure: the public spending on research and innovation in Europe roughly equals the US in terms of GDP share, however, only one tenth of this spending takes place at the EU level.
- Europe does not coordinate on its *industrial strategy*, and its slow and fragmented law-making process hinders its effectiveness in keeping pace with the rest of the evolving world. Overall, the report aims at delivering action points for a new European industrial strategy that can relieve the Union from these barriers.

[1] *Relevant examples include the GDPR and the Free Flow of Non-Personal Data Regulation.*

[2] *https://commission.europa.eu/topics/strengthening-european-competitiveness/eu-competitiveness-looking-ahead_en*

© 2024-2026 NexusForum.EU

2 European policy and competitiveness

Cloud, edge and connectivity technologies are strategic enablers for Europe's digital transformation. Together, they provide the backbone of the distributed computing infrastructure that allows data to be processed where it is generated (at the edge), where it can be aggregated and managed efficiently (in cloud environments), and where extremely large-scale workloads – particularly AI training and high-performance analytics – can be executed (in HPC/supercomputing environments). This distributed reality is increasingly fundamental to modern AI systems, IoT deployments and advanced connectivity (5G/6G).

Europe's direction can be summarised as a reinforcing triangle:

- **Compute infrastructure at scale:** ensuring Europe has accessible, high-performance compute capacity for AI training and deployment (including through EuroHPC-based initiatives such as AI Factories).
- **Data availability and trusted sharing:** enabling data to flow across sectors and borders under European rules, and creating practical mechanisms such as data spaces to support reuse and innovation.
- **Trustworthy AI governance:** creating legal certainty and safeguards through the AI Act while supporting innovation pathways and adoption.

The European Union has set a clear 2030 direction for digital transformation through the **Digital Decade Policy Programme 2030**, which sets targets across digital skills, digital infrastructures, business digitalisation and digital public services.[3] The policy programme frames the urgency to scale secure, high-performing and sustainable digital infrastructure across Europe, while ensuring interoperability and the capacity to support strategic technologies.

A more comprehensive background about the relevant European policy landscape can be found in the Digital Policy report produced by the NexusForum.EU project. The rest of this section will give a brief introduction to the policy landscape relevant to the roadmap.

Digital technologies and EU Competitiveness

Since then, the Draghi report was released, which again stressed the critical role of computing technologies for the future of European competitiveness and digitalisation.[4] The report identifies and recommends three strategic priorities for *digitalisation and advanced technologies*:

- **High-speed/capacity broadband networks and related equipment and software** (i.e. fixed, wireless, and satellite/hybrid networks) to enable connectivity and distribute secure, ubiquitous and sustainable digital services essential to EU citizens and businesses.

- **Computing and AI**, i.e. infrastructure, platforms and advanced technologies needed to autonomously develop and scale up digital services, enabling companies to

[3] *https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/europes-digital-decade-digital-targets-2030_en*

[4] *https://commission.europa.eu/topics/strengthening-european-competitiveness/eu-competitiveness-looking-ahead_en#paragraph_47059*

© 2024-2026 NexusForum.EU

innovate, boost their productivity and upscale, notably concerning cloud, high-performance computing and quantum, as well as AI and its industrial applications.

- **Semiconductors**, a key driver and enabler for the electronics value chain, and a strategic element of Europe's security and industrial strength across sectors.

Building on the Draghi report, the Commission presented the Competitiveness Compass in January 2025.[5] The compass introduces a number of actions for closing the innovation gaps in advanced technologies, including AI, robotics, quantum, and space technologies. It further highlights a need to diversify and strengthen European supply chains.

The European Data Strategy and Common European Data Spaces

The need to invest in cloud technologies is again mentioned in the European Data Strategy, which outlines the critical role of data and aims at creating a single market for data.[6] The strategy emphasises the need to make data available and invest in data-sharing tools and infrastructures to store and process data. The Data Act is an important component in this strategy, by giving users greater control of the data generated by their connected devices.[7] A major implementation pillar of this strategy is the development of **Common European Data Spaces**, intended to overcome technical and organisational barriers to data sharing and to enable data-driven innovation at scale across sectors and Member States.[8]

The Simpl initiative is an open source middleware platform meant to support data access and interoperability between European data spaces. It supports technical implementation of secure and interoperable cloud-edge federations as a foundation of the common European data spaces.[9]

The AI Continent Action Plan and European approach to AI

The EU's AI policy landscape has also matured substantially.[10] Following the regulatory framework established in the **AI Act**,[11] the EU has placed strong emphasis on accelerating AI innovation and adoption. The Commission's **AI Innovation Package** supports startups and SMEs and includes measures such as privileged access to supercomputing resources and the establishment of the **AI Office**.[12]

More recently, the Commission's **AI Continent Action Plan** has reinforced the strategic link between AI ambition and computing infrastructure, explicitly highlighting **computing infrastructure** as one of its strategic areas and emphasising the deployment of **AI Factories** as hubs to train and fine-tune AI models and provide services to the wider ecosystem. The Action Plan's framing, together with ongoing EuroHPC initiatives, underscores that Europe's competitiveness in AI depends not only on rules and talent, but also on scalable, accessible and sustainable compute capacity.

[5] *https://commission.europa.eu/topics/competitiveness/competitiveness-compass_en*

[6] *https://digital-strategy.ec.europa.eu/en/policies/strategy-data*

[7] *https://digital-strategy.ec.europa.eu/en/policies/data-act*

[8] *https://digital-strategy.ec.europa.eu/en/policies/data-spaces*

[9] *https://digital-strategy.ec.europa.eu/en/policies/simpl*

[10] *https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence*

[11] *https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai*

[12] *https://digital-strategy.ec.europa.eu/en/factpages/ai-innovation-package*

© 2024-2026 NexusForum.EU

The recently announced **Cloud and AI Development Act** aims to further strengthen European data centre capacity to benefit AI innovation. It focuses on three main areas: research and innovation, infrastructure capacity, and autonomy. On R&D&I, it aims to support next-generation cloud and AI technologies for frontier AI, industrial AI and physical AI, including through grand challenges, national cloud and AI strategies, and Experience and Acceleration Centres for AI. On capacity, it seeks to at least triple EU data centre capacity within five to seven years, simplify permitting, improve access to energy, land, water and financing, and ensure sufficient compute for AI, cloud services and data-intensive applications.

The Act also has a strong sovereignty dimension. It proposes an EU-wide framework for assessing cloud and AI sovereignty, supports deployment in critical sectors, promotes EU-based innovation and supply-chain resilience, and introduces common procurement mechanisms for public administrations. It also encourages open-source solutions as a way to improve resilience. Overall, the initiative is designed to improve conditions for businesses, cloud providers, investors, researchers and public administrations, while reducing Europe's strategic dependencies in critical digital infrastructure.

© 2024-2026 NexusForum.EU

3 Major relevant European initiatives

Several important European initiatives are helping to shape the Cognitive Computing Continuum. These initiatives address complementary layers of the European digital ecosystem, from the underlying computing architectures and infrastructure needed to run advanced AI workloads to the data-sharing frameworks required to enable trustworthy and scalable innovation. The following examples illustrate how Europe is building capabilities across these interconnected domains. Together, they strengthen the technological foundations, shared resources and governance mechanisms needed to support European AI innovation, digital sovereignty and competitive deployment across cloud, edge, HPC and sectoral environments.

3.1 Establishing a European ecosystem based on RISC-V

Over the past decade and a half there has been significant investment in implementing European processors, developing and maintaining their roadmaps, targeting both HPC and embedded and IoT applications. The EU goal is the production of cutting-edge and sustainable semiconductors in Europe, with at least 20% of world production in value by 2030. By that time, Europe should have manufacturing capacities below 5nm nodes, aiming at 2nm, and improve energy efficiency in the domain by a factor ten. To achieve this the RISC-V open standard Instruction Set Architecture (ISA)[13] plays a central role in EUs strategy. RISC-V is royalty-free and open-specification allowing any market actor to implement microprocessors or accelerators using it. This lowers the barriers to innovation and market entry, and through collaborative efforts lessens the overhead for individual actors to develop and maintain non-value-adding supporting hardware and software needed for intellectual properties (IP) being developed and brought to market. For this reason, RISC-V has been called "the Linux of hardware"[14]. For Europe RISC-V offers the opportunity to develop fully European processors, in contrast to today's reliance on processors and IPs owned by and developed on other continents.

EuroHPC JU (see section 3.2) aims to build a diverse IP portfolio of processors, accelerators, quantum chips, and AI accelerators, the pilots and demonstrators needed to realize them, as well as the necessary ecosystem to scale them to exascale including the associated software stack and applications. In the short term (2024-2026) this builds on the efforts in the European Processor Initiative (EPI) to build ARM-based processors, but in the medium term (2026-2028) this shall be complemented with competitive RISC-V general purpose processors and GPUs, with EuroHPC JU post-exascale systems being one of the first customers to drive demand in Europe and fulfil the EU goal of autonomy in strategic processing technologies[15].

The European Processor Initiative, EPI[16], is a Framework Partnership Agreement under EuroHPC JU (150 MEUR) to develop European supercomputing technologies, including

[13] *RISC-V International, "About RISC-V," https://riscv.org/about/*

[14] *https://www.eetimes.com/european-union-seeks-chip-sovereignty-using-risc-v/*

[15] *Alexandra Kourfali (EuroHPC JU), EPI Forum 2024, https://www.european-processor-initiative.eu/dissemination-material/epi-forum-in-barcelona/*

[16] *https://www.european-processor-initiative.eu*

© 2024-2026 NexusForum.EU

European microprocessor and accelerator technologies, to improve performance and power ratios. EPI develops general purpose processors based on the ARM ISA, and accelerators based on the RISC-V ISA. The EPI project is complemented by EUPILOT[17] (30 MEUR) aiming to deliver all-European open-source and open-standard based software and hardware HPC systems, EUPEX[18] (40 MEUR) to produce a European pilot for exascale computing, and the eProcessor[19] (8 MEUR) to develop a European out-of-order high-performance RISC-V CPU.

European processor investment has been targeting both ARM and RISC-V ISAs and while the ARM ISA continues to play an important role in the coming years (2024-2029), there has been a shift towards more investment in European RISC-V based technologies, up to 2030 and beyond, with the establishment of the Digital Autonomy with RISC-V in Europe[20] (DARE, 240 MEUR) Framework Partnership Agreement to develop large-scale high-performance computing ecosystem based on RISC-V. DARE is expected to run between 2025 and 2030 to deploy RISC-V pilot systems with a path towards post-exascale RISC-V development and procurement in the next decade. These projects do not only leverage technology, IP, and lessons learned from previous European efforts but constitute a coherent and strategic approach to co-develop European RISC-V processors, their software stacks, but also the porting of applications of key European value to the new platforms.

The European Partnership Chips Joint Undertaking, Chips JU, was established following the European Chips Act to succeed the Key Digital Technologies JU with the goal to reinforce EU strategic autonomy and address European challenges in electronic components and systems to establish scientific excellence and innovation leadership through the establishment of pilot lines, design platforms, and competence centres.[21,22] Chips JU and its predecessors has a strong involvement in European RISC-V which with renewed EU strategic significance today is guided by the 2022 Recommendations and Roadmap for European Sovereignty in Open-Source Hardware, Software, and RISC-V Technologies[23] and the 2023 Roadmap towards a High-Performance Automotive RISC-V Reference Platform[24]. In 2024 Chips JU had contracted two projects in this area; TRISTAN[25] (54 MEUR, 2022-2025) and ISOLDE[26] (39 MEUR, 2023-2026). TRISTAN develops a repository of European RISC-V quality building blocks for SoC designs in key European application domains such as automobile, industrial, as well as the necessary Electronic Design Automation (EDA) tools and the full software stack. The follow-up ISOLDE project expands the European RISC-V ecosystem by maturing the CVA6 and NOEL-V European RISC-V superscalar processors for applications including safety- and security-critical systems, system-level hardware components and their software stacks to contribute to the European RISC-V strategy in embedded high-performance computing and industrial IoT.

---

[17] *https://eupilot.eu*

[18] *https://eupex.eu*

[19] *https://eprocessor.eu/*

[20] *https://eurohpc-ju.europa.eu/specific-grant-agreement-sga-development-large-scale-european-initiative-hpc-ecosystem-based-risc-v_en*

[21] *https://www.chips-ju.europa.eu/Our-vision/*

[22] *https://digital-strategy.ec.europa.eu/en/news/chips-competence-centres-strengthen-semiconductor-expertise-across-europe-about-kick*

[23] *https://digital-strategy.ec.europa.eu/en/library/recommendations-and-roadmap-european-sovereignty-open-source-hardware-software-and-risc-v*

[24] *Jariu Kinaret (Chips JU), https://eurohpc-ju.europa.eu/document/download/7e21bf39-bb19-43b8-88b6-dd51650d348e_en?filename=pdf%20chips%20parrt%202.pdf*

[25] *https://tristan-project.eu/*

[26] *https://www.isolde-project.eu*

© 2024-2026 NexusForum.EU

Taken together these European efforts, spanning the Compute Continuum, expose a strong European commitment to a joint European ecosystem based on RISC-V.

3.2 EuroHPC and the establishment of AI Factories

The European High Performance Computing Joint Undertaking (EuroHPC JU) was created in 2018 to make Europe a world leader in supercomputing, with a budget of EUR 8.2 billion for the period 2021-2027. It is a legal and funding entity that allows the EU and participating countries to coordinate their efforts and pool their resources to develop a European supercomputer ecosystem based on European technologies.[27]

EuroHPC aims to boost scientific excellence and industrial strength in Europe and support the digital transformation of its economy. EuroHPC aims to develop and maintain a *"world-leading federated, secure and hyper-connected supercomputing, quantum computing, service and data infrastructure ecosystem,"* and increase the use of HPC in both public and private sector.[27]

The *AI Innovation Package*, proposed in January 2024, proposed the establishment of AI Factories in Europe, leveraging the EuroHPC supercomputer ecosystem to develop trustworthy AI in Europe.[28] In July 2024 an amendment to the EuroHPC Regulations came into force,[29] expanding its mission to include the development and operation of *"AI Factories located around EuroHPC supercomputing facilities to support the growth of a highly competitive and innovative AI ecosystem in Europe."*

This amendment allows EuroHPC to acquire and operate dedicated AI-optimised supercomputers, in particular for training generative AI and general-purpose AI models that require significant computing resources. The purpose of these AI factories is to offer a one-stop shop for startups, the scientific community, and other innovative AI users to develop powerful AI models for a variety of emerging AI applications.

The EuroHPC AI Factories initiative represents a collaborative effort from across 17 European countries, pooling together EU and national resources to create a robust and interconnected network of AI hubs. This initiative aims to position Europe as a leader in AI innovation and development.

The first seven AI factories will be deployed in 2025 across Europe, specifically in Finland, Germany, Greece, Italy, Luxembourg, Spain, and Sweden.[30] These sites were selected to host the new AI factories, which will include brand new AI-optimised supercomputers in five of these countries. Notably, some of these will leverage cloud technologies to provide cloud-style interfaces and services that are more familiar to the AI community.

These AI Factories will act as dynamic ecosystems that foster innovation by providing comprehensive support, including access to AI-optimised HPC resources, training, and technical expertise. They will serve as hubs for AI startups, SMEs, and researchers, promoting

[27] *https://eurohpc-ju.europa.eu/about/discover-eurohpc-ju_en*

[28] *https://digital-strategy.ec.europa.eu/en/factpages/ai-innovation-package*

[29] *https://digital-strategy.ec.europa.eu/en/news/setting-ai-factories-now-possible-after-eurohpc-regulation-amendment*

[30] *https://eurohpc-ju.europa.eu/selection-first-seven-ai-factories-drive-europes-leadership-ai-2024-12-10_en*

© 2024-2026 NexusForum.EU

collaboration across Europe and driving advancements in areas like healthcare, energy, and climate.

3.3 Access to European data, common data spaces, and data privacy

There is an abundance of national and international (EU-level) legislation and initiatives meant to stimulate data sharing across the economy. In the EU, the EU Data Act[31] mandates equipment manufacturers to make the data collected and produced by the equipment available to users and third parties, supporting ecosystems such as the Common European Data Spaces[32]. Other ongoing initiatives like the International Data Spaces[33] and Gaia-X[34] aim to bridge the gap between policy, legislation and technical implementation of workable solutions. In Singapore the Infocomm Media Development Authority offers a Trusted Data Sharing Framework and regulatory sandbox to allow businesses and public authorities to experiment with secure data sharing solutions and accelerate innovation[35]. Similarly, in Japan the Ministry of Economy, Trade and Industry and the Information Processing Promotion Agency launched an initiative for interoperable cross-border data infrastructures called the "Ouranos Ecosystem"[36]. The European Commission funds a wide range of projects and initiatives to implement the European data strategy[37]. For example, European Research initiatives such as the European Open Science Cloud[38] are working on developing a cloud infrastructure for sharing data for scientific and public services purposes.

In February 2020, the European Commission recognized the need to establish the foundations of a "European Strategy for Data"[39]. Its aim is for the EU to take a leading role in empowering society through data, enabling better decision-making in both business and the public sector. This initiative was followed in May by the Recovery and Resilience Facility's "20% Digital EU Flagship Scale Up", in October by the members' declaration for a "European Cloud," and in March 2021 by the Digital Decade Strategy, which set targets for Edge and Cloud by 2030.

The data space concept was created to facilitate the sharing of data between European entities. Several initiatives have been launched in Europe to this end:

- The Data Spaces Support Centre (DSSC), funded by European Commission as part of the Digital Europe Program, establishes a network of relevant organizations and initiatives involved in the development of Data Spaces.
- Gaia-X was established in 2021 as a privately funded not-for-profit organization with the support of BMWK. GAIA-X has funded 11 consortia until 2024, with a total budget of 122

[31] *EU Data Act https://digital-strategy.ec.europa.eu/en/policies/data-act*

[32] *Common European Data Spaces https://digital-strategy.ec.europa.eu/en/policies/data-spaces*

[33] *International Data Spaces Association: https://internationaldataspaces.org/*

[34] *Gaia-X Federated Data Sharing Infrastructure: https://gaia-x.eu/*

[35] *Singapore Infocomm Media Development Authority: https://www.imda.gov.sg/how-we-can-help/data-innovation*

[36] *Japan's Initiatives for Interoperable Data Infrastructures Officially Named "Ouranos Ecosystem" https://www.meti.go.jp/english/press/2023/0429_001.html*

[37] *A European strategy for data https://digital-strategy.ec.europa.eu/en/policies/strategy-data*

[38] *European Open Science Cloud: https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science/european-open-science-cloud-eosc_en*

[39] *https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0066*

© 2024-2026 NexusForum.EU

million Euros. While initially focused on building the European Cloud, Gaia-X later opened participation to Cloud Hyperscalers, ensuring their compliance with EU values through a Label Framework.

- The International Data Spaces Association (IDSA), one of the first non-profit organizations to contribute to the European strategy for data, formed in 2016 and incorporated under German law. The IDSA evolved from the Fraunhofer Society project called the Industrial Data Space in 2014, which was renamed to International Data Spaces (IDS) in 2015. IDSA is working on the "Dataspace Protocol," focusing on the Data Exchange layer of Data Spaces, with the aim to make it an ISO standard. Expanding beyond Europe, the IDSA now has a network of hubs globally, including in Japan, Malaysia, and a competence centre in China.
- FIWARE, originally funded under the European Union's Seventh Framework Programme for Research, was a non-profit association of industry, academia and SMEs worldwide providing standards and technology development around Data Exchange and Data Sharing and a full framework of open-source components to build smart solutions, digital twins and data spaces.

© 2024-2026 NexusForum.EU

NEXUS FORUM.EU

PART B:

Research & Innovation Roadmap and Priorities

www.eucloudedgeiot.eu

4 From Strategic Destinations to R&I Priorities: How to Read the Roadmap

The European Cognitive Computing Continuum is not a single infrastructure, architecture or market. It is a capability landscape that spans AI, data, cloud, edge, high-performance computing, telecommunications, hardware, software, energy systems and sectoral adoption. Across this landscape, Europe will need to strengthen several complementary technology pathways: agentic AI and emerging AI technologies, together with the specialised hardware and deployment infrastructures needed to support them; large-scale HPC, AI Factory and testbed environments required to evaluate, train, validate and scale AI technologies before deployment; open and competitive hardware-software ecosystems; federated and interoperable infrastructures where these create value; secure industrial and public-sector deployment environments; and emerging computing paradigms with long-term strategic potential.

The six Strategic Destinations defined in Part A express important outcomes that European research and innovation should help achieve. The five pillars in Part B organise the technological and operational capabilities required to move towards these outcomes. The detailed topic roadmaps then identify the R&I priorities, support actions, standardisation needs, testbeds and longer-term investments that can build these capabilities.

These relationships are not one-to-one. A topic may contribute to several Strategic Destinations, and a Strategic Destination may require capabilities from several pillars. Equally, not every capability is required for every pathway. For example, federation is essential where Europe seeks interoperable multi-provider services, cross-border infrastructures, common data-space participation or distributed markets. It is less central to pathways based primarily on the development and deployment of leading AI systems, private industrial environments or the introduction of novel processor and hardware architectures that rely on hardware-algorithm-software co-design within a dedicated technology stack. Likewise, neuromorphic and quantum technologies are important strategic options for Europe's future position, but they are not dependencies for near-term AI deployment or industrial digitalisation.

This section provides a concise guide to the structure and logic of the roadmap. It shows how the pillars relate to the Strategic Destinations, identifies the most important capability pathways, and clarifies where R&I priorities are foundational, enabling, conditional on deployment choices, or oriented towards longer-term leadership.

4.1 The five pillars of the roadmap

The detailed roadmap is organised into five pillars, each addressing a distinct but connected part of Europe's future computing capability base.

Pillar I: Foundational AI and ML Technologies addresses Europe's capacity to create, deploy and operate advanced AI systems. It covers frontier and agentic AI, neuro-symbolic approaches, open-source model strategies, AI workflows and orchestration, memory and verified tool use, AI at the device edge, and AI and data workloads across heterogeneous computing environments. Its purpose is to strengthen Europe's AI capability not only at the

© 2024-2026 NexusForum.EU

level of models, but also in the systems, software and operational mechanisms that make AI usable and trustworthy.

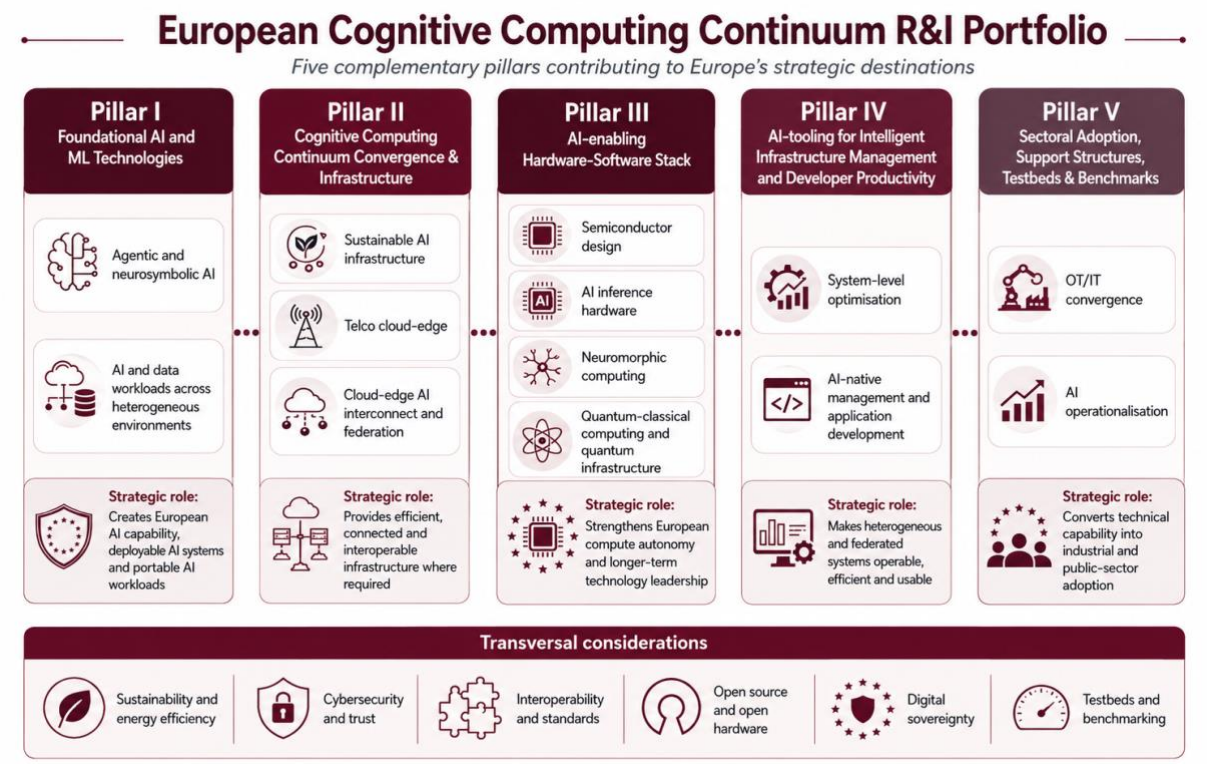


Figure 3: Overview of the five Pillars and the technology areas under each Pillar.

Pillar II: Cognitive Computing Continuum Convergence & Infrastructure addresses the infrastructure required to support increasingly distributed and data-intensive applications. It covers sustainable and energy-efficient AI infrastructure, telco cloud-edge convergence, Open RAN, seamless connectivity and predictive handover, AI-native telecommunications infrastructures, federated telco edge services, and cloud-edge AI interconnect frameworks. Its purpose is to support efficient, connected and interoperable infrastructure where distributed operation, low latency, energy management or federation creates strategic value.

Pillar III: An AI-enabling Hardware-Software Stack addresses Europe's capacity to develop and control critical compute technologies. It covers European semiconductor design, RISC-V processors and associated software stacks, AI inference accelerators, memory and interconnects, neuromorphic systems, hybrid quantum-classical computing and quantum infrastructure integration. Its purpose is to build open, usable and competitive European hardware-algorithm-software ecosystems, while developing longer-term options in emerging computing paradigms.

Pillar IV: AI-tooling for Intelligent Infrastructure Management and Developer Productivity addresses the operational and developer capabilities needed to make heterogeneous infrastructures practical. It covers federation and system-level optimisation, trusted and decentralised orchestration, AI-native infrastructure management, zero-touch deployment, lifecycle automation, observability, maintenance and incident response. Its purpose is to ensure that increasingly complex computing systems can be efficiently deployed, managed, secured and optimised.

© 2024-2026 NexusForum.EU

Pillar V: Sectoral Adoption, Support Structures, Testbeds & Benchmarks addresses how European capabilities are brought into operational use. It covers the convergence of operational technologies and information technologies, resilient industrial AI, secure OT/IT data integration, industrial digital twins, cybersecurity and lifecycle management, large-scale testbeds, AI operationalisation, autonomous agents, multi-agent orchestration and robotics support. Its purpose is to connect R&I investment to industrial and public-sector deployment, experimentation, validation and adoption.

Across all five pillars, several considerations are transversal rather than confined to an individual topic. Sustainability and energy efficiency influence data-centre design, AI hardware, workload placement and infrastructure operation. Cybersecurity, trust and auditability are important for agentic AI, industrial systems, supply chains and federated services. Interoperability, open standards and open-source ecosystems support adoption, portability and European autonomy. Testbeds and benchmarks provide the means to evaluate whether research advances can become operational capabilities.

4.2 Strategic Destination pathways

The following pathway summaries identify the principal topic areas contributing to each Strategic Destination. They are intended to guide readers towards the relevant detailed topic roadmaps in Part B and to highlight the most relevant priorities and dependencies for each Strategic Destination.

Strategic Destination 1. Scalable and energy-efficient AI and data processing	
Capabilities and role in the pathway	**Foundational:** efficient inference hardware, memory and interconnects; energy-efficient data-centre and continuum infrastructure; deployment and operation of AI and data workloads across heterogeneous environments. **Enabling:** system-level optimisation and AI-native management of infrastructure. **Conditional / longer-term:** federation where optimisation spans providers or infrastructure domains; neuromorphic systems for future low-power AI applications.
Especially relevant roadmap sections	**Section 6:** *AI and Data Workloads across Heterogeneous Computing Environments*, especially 6.1–6.4. **Section 7:** *Sustainable and Energy-efficient AI infrastructure*, especially 7.1–7.3. **Section 11:** *AI inference hardware: AI Accelerators, Memory, and Interconnects*. **Section 13:** *Federation and system-level optimisation for the Computing Continuum*, especially 13.1. **Section 14:** *AI-native management and application development for a heterogeneous computing continuum*. **Section 12.1:** *Neuromorphic systems*.
Strategic Destination 2. An AI stack built on an open European hardware and computing ecosystem	

© 2024-2026 NexusForum.EU

Capabilities and role in the pathway	**Foundational:** European processor and semiconductor design; RISC-V systems; AI inference accelerators; software stacks and portable AI execution environments. **Enabling:** compilers, runtimes, libraries, developer tools, benchmarks and application porting; European infrastructures and pilots that create demand for European technologies. **Longer-term:** neuromorphic and quantum-related hardware-software ecosystems.
Especially relevant roadmap sections	**Section 4.1:** *Establishing a European ecosystem based on RISC-V*. **Section 4.2:** *EuroHPC and the establishment of AI Factories*. **Section 6.2:** *Portable AI applications and open AI technology stacks*. **Section 10:** *European Semiconductor Design*, especially 10.1–10.2. **Section 11:** *AI inference hardware: AI Accelerators, Memory, and Interconnects*. **Section 12:** *Emerging Processor Architectures*, especially 12.1–12.3.
Strategic Destination 3. A competitive European AI and machine learning ecosystem	
Capabilities and role in the pathway	**Foundational:** frontier, agentic and neuro-symbolic AI; deployable AI workloads across heterogeneous computing environments; operationalisation of AI through testbeds, lifecycle practices and agent-based systems. **Enabling:** AI Factories; model evaluation; AI-native developer and operations tooling; small models and edge AI; trusted knowledge, memory and tool-use mechanisms. **Conditional:** federated AI where shared access, distributed execution, model mobility or portability across infrastructures creates value.
Especially relevant roadmap sections	**Section 4.2:** *EuroHPC and the establishment of AI Factories*. **Section 5:** *Agentic AI and neurosymbolic AI*, especially 5.1–5.5. **Section 6:** *AI and Data Workloads across Heterogeneous Computing Environments*, especially 6.2–6.4. **Section 14:** *AI-native management and application development for a heterogeneous computing continuum*. **Section 16:** *AI Operationalization*, especially 16.1–16.2.
Strategic Destination 4. Secure, sovereign and interoperable European computing capabilities	
Capabilities and role in the pathway	**Foundational:** interoperable cloud-edge AI frameworks; federated telco-edge services; cross-provider optimisation, orchestration and trust mechanisms; cybersecurity and lifecycle management for connected industrial systems. **Enabling:** open hardware-software stacks; AI-native operations; secure data integration and European data-space participation; portable AI applications and open AI technology stacks that reduce lock-in and support interoperability across heterogeneous infrastructures. **Conditional:** federation is central where services, data or resources operate across providers, borders or governance domains; other environments may depend primarily on secure specialised or private infrastructures.

© 2024-2026 NexusForum.EU

Especially relevant roadmap sections	**Section 4.1:** *Establishing a European ecosystem based on RISC-V*. **Section 4.3:** *Access to European data, common data spaces, and data privacy*. **Section 6.2:** *Portable AI applications and open AI technology stacks*. **Section 8:** *Telco Cloud-Edge*, especially 8.1 *Open Radio Access Networks (Open RAN)*, 8.2 *Seamless data connectivity and predictive handover across different networks*, and 8.4 *Federated Telco Edge and Network-as-a-Service (NaaS)*. **Section 9:** *Federations and Cloud-Edge AI Interconnect Framework*, especially 9.1–9.2. **Section 13:** *Federation and system-level optimisation for the Computing Continuum*, especially 13.1–13.2. **Section 14:** *AI-native management and application development for a heterogeneous computing continuum*. **Section 15.2:** *Secure OT/IT data integration and industrial data spaces*. **Section 15.4:** *Cybersecurity and lifecycle management for converged OT/IT systems*.
Strategic Destination 5. Advanced digitalisation and AI adoption in industry and public sectors	
Capabilities and role in the pathway	**Foundational:** resilient industrial-edge AI; secure OT/IT integration; industrial digital twins; AI operationalisation, testbeds and robotics support. **Enabling:** agentic and neuro-symbolic AI for applied systems; heterogeneous and edge AI deployment; data spaces; connectivity and telco-edge capabilities for mobile or latency-sensitive applications. **Conditional:** federation where industrial value chains, public services or sectoral ecosystems require shared data or distributed service provision.
Especially relevant roadmap sections	**Section 5.4:** *Harness engineering with memory, knowledge integration and verified tool use*. **Section 5.5:** *Agentic AI and LLMs for devices, edge, and mobile phones*. **Section 6:** *AI and Data Workloads across Heterogeneous Computing Environments*. **Section 8.2:** *Seamless data connectivity and predictive handover across different networks*. **Section 8.3:** *AI-Native Telco Cloud-Edge*. **Section 15:** *Convergence of Operational Technologies and Information Technologies*, especially 15.1–15.4. **Section 16:** *AI Operationalization*, especially 16.1–16.2.
Strategic Destination 6. European leadership in disruptive and emerging computing paradigms	
Capabilities and role in the pathway	**Foundational:** neuromorphic systems; hybrid quantum-classical computing; integration of quantum infrastructure with HPC and cloud services; novel accelerator and post-exascale pathways. **Enabling:** semiconductor design; software stacks, compilers and runtimes; orchestration; shared benchmarks; testbeds; early application pilots and procurement pathways. **Conditional:** federated access models where specialised emerging-computing resources are offered as shared European services.

© 2024-2026 NexusForum.EU

Especially relevant roadmap sections	**Section 4.1:** *Establishing a European ecosystem based on RISC-V*. **Section 4.2:** *EuroHPC and the establishment of AI Factories*. **Section 10:** *European Semiconductor Design*. **Section 11:** *AI inference hardware: AI Accelerators, Memory, and Interconnects*. **Section 12:** *Emerging Processor Architectures*, especially 12.1 *Neuromorphic systems*, 12.2 *Hybrid quantum and classical computing fusion*, and 12.3 *Integration of quantum computing infrastructure*.

4.3 High-level summary of recommendations

Hardware capability depends on software maturity and demand creation. European investments in RISC-V processors, AI accelerators, memory architectures, neuromorphic systems and quantum technologies will not translate into strategic capability through hardware development alone. These investments need to be matched by investments to create demand, and supporting ecosystems with compilers, runtime systems, libraries, model-serving frameworks, developer tools, benchmarks, application-porting programmes and routes to early use through European infrastructures and sectoral pilots.

AI infrastructure depends on operational and developer capability. Scalable compute capacity, including AI Factories, HPC resources, cloud platforms and edge infrastructures, is necessary but not enough. European users must also be able to develop, evaluate, deploy, monitor, secure and update AI systems across different environments. This places particular importance on portable AI software stacks, lifecycle management, observability, AI-native operations and practical developer workflows.

Federation is central where European interoperability and shared capability require it. In multi-provider, cross-border and data-space settings, Europe needs mechanisms for interoperability, identity, trust, service discovery, data exchange, workload placement, lifecycle management and system-level optimisation. These capabilities are critical where Europe aims to create interoperable markets, shared services or resilient infrastructures across organizational boundaries. However, federation should be pursued where it creates identifiable value; it should not be assumed to be the organizing requirement for all AI, hardware or sectoral systems.

Industrial and public-sector adoption depends on operational validation. Capabilities for industrial AI, cyber-physical systems, robotics, digital twins, public-sector services and critical infrastructures must be validated under realistic operating conditions. This requires testbeds, benchmarks, secure data integration, cybersecurity and lifecycle management, as well as attention to latency, resilience, safety, privacy, energy use and long equipment lifecycles.

Emerging computing paradigms require staged and sustained investment. Neuromorphic, photonic/optical, and quantum technologies have the potential to strengthen Europe's future position in low-power intelligence, simulation, optimisation and advanced computational services. Their development should be linked to software stacks, hybrid configurations with classical computing, testbeds, skills and early application pilots. They should be advanced as strategic options with long-term value, rather than presented as immediate dependencies for the wider computing continuum.

© 2024-2026 NexusForum.EU

4.4 Using the detailed topic roadmaps

The remainder of Part B provides detailed topic roadmaps within each of the five pillars, including the specific R&I priorities and recommendations relevant to different time horizons and implementation instruments. Readers concerned with a particular Strategic Destination can use the pathway summaries above to identify the most relevant topic areas and then consult the corresponding detailed roadmaps.

For research and innovation programme design, the topic roadmaps should be read in combination rather than in isolation. In particular, infrastructure investment should be considered alongside software and ecosystem maturity and energy requirements; hardware development alongside toolchains, benchmarks and adoption mechanisms; federation alongside trust, interoperability and viable use cases; and sectoral adoption alongside testbeds, cybersecurity and operational validation.

Taken together, the Strategic Destinations, pillars and topic roadmaps express a portfolio approach to European research and innovation. Europe's objective is not to create a single uniform computing platform, but to build a coherent, interoperable and competitive capability base across AI, data, hardware, software, connectivity, sustainability, security and adoption. This capability base should enable Europe to deploy federation where it creates value, strengthen specialised and sectoral systems where these are most appropriate, and invest consistently in the emerging technologies that may shape future European leadership.

© 2024-2026 NexusForum.EU

NEXUS FORUM.EU

PART B

PILLAR I:

Foundational AI and ML Technologies

5 Agentic AI and neurosymbolic AI

Primary destinations
3. A competitive European AI and machine learning ecosystem
Secondary destinations
2. An AI stack built on an open European hardware/computing ecosystem
5. Advanced digitalisation and AI adoption in industry and public sectors

Background and driving factors

Agentic AI is emerging as a new software and systems paradigm in which large language models are composed with tools, memory, workflows, retrieval systems, software services, other agents and human approval mechanisms to plan and execute complex tasks. Agentic AI marks a shift from using large language models primarily as standalone conversational systems or one-shot generative AI tools, towards using them as components in larger, goal-directed systems and neurosymbolic AI systems. The research and innovation challenge is not just to improve the underlying language models, but also how to engineer the system around the language models and their interaction.

The state of the art shows that useful agentic behaviour depends not only on model capability, but also on the surrounding execution architecture. Production-oriented agent systems increasingly rely on explicit workflow graphs, memory and retrieval layers, tool contracts, verification and auditability, runtime traces, human-in-the-loop and approval gates, and fallback logic. In other words, the practical frontier lies in the harness around the model: the engineering layer that determines how context is assembled, how tools are invoked, how state is preserved, how behaviour is observed, and how outputs are verified before action is taken. Emerging standards and frameworks reinforce this direction, including open protocols for connecting agents to tools and data sources, and observability conventions for tracing model calls, tool use, and agent steps across systems.

Europe in relation to the state of the art

Agentic AI is advancing rapidly, and enterprises are experimenting with integrating agentic AI in their own systems. Agentic AI has also created new opportunities for companies and startups to build services using existing LLMs (commercial and open-source), without having to train new ones from scratch. On the other hand, the enterprise LLM market is highly dominated by a few vendors, and we have seen a decline in the market share of open-source[40]

[40] *"Open-source LLM" is a loosely used concept, and in many cases refers to "open-weight LLM". That is, the AI model parameters are public and the model can be hosted and used freely, but the data and methods used to train the model might not be public. In this roadmap, we will use "open-source" LLM synonymously with "open-weight" LLM.*

© 2024-2026 NexusForum.EU

LLMs. Companies increasingly turn to frontier LLMs in favour of developing internal AI models or using open-source models, according to recent reports from Menlo Ventures.[41,42]

These reports from Menlo Ventures estimate that the top three AI frontier labs (Anthropic, OpenAI, and Google) account for 88 % of the enterprise segment in the LLM API market (the top four, including Meta, account for 96 % of the market). This report also shows a relative decline in the use of open-source LLMs in enterprise markets, and a growing adoption of Chinese open-source LLMs in builder ecosystems.

This means that, in addition to the AI frontier labs providing their own agentic AI solutions (with their own harnesses and ecosystems), other companies are increasingly building their own agentic AI systems on top of the LLMs of AI frontier labs, with non-EU open-source LLMs as alternatives.

While there are many European companies developing leading AI models for other applications, and there's a strong basis of AI expertise in Europe, Europe is lagging behind when it comes to the absolute frontier LLMs, with Mistral as the one notable European exception. Elsewhere, South Korea has recently launched a sovereign LLM challenge to train LLMs for the Korean language that can compete with frontier AI models, and the country is emerging as a strong AI nation.[43]

There are also many notable startups and scaleups in the EU that have successfully built AI services on top of third-party LLM APIs from the AI frontier labs. This over-reliance on APIs from AI frontier labs is risky from both a market and geopolitical perspective. However, the rapid progress and market consolidation in Agentic AI means it's also risky to condition a European AI strategy on the development of EU-based LLMs. On the other hand, leading open-weight LLMs have been competitive with those developed by AI frontier labs, when integrated with the right model harness.

There are still market opportunities for advancing agentic AI in Europe and adopting European agentic AI solutions in enterprise workflows and public sector. Here, the EU could look to adopt a pragmatic strategy: promote the use of open-source (possibly of non-EU origin) LLMs hosted by EU providers and evaluate where they can be used, while at the same time building capacity in Europe to train state-of-the-art LLMs.

One major challenge is that the turnaround time for current funding frameworks for most research and innovation projects in the EU (call publication -> submission deadline -> approval and project start -> project end) is not aligned with the rapid rate of AI innovation globally.

Cross-cutting principles and dependencies

Advances in Agentic AI can be pursued independently of the rest of the roadmap, but there are further synergies with the topic *"AI and Data Workloads across Heterogeneous Computing Environments."* Programming models and middleware could be developed to make it easier to run agentic AI applications and distribute workflows in the continuum.

---

[41] *https://menlovc.com/perspective/2025-mid-year-llm-market-update/*

[42] *https://menlovc.com/perspective/2025-the-state-of-generative-ai-in-the-enterprise/*

[43] *https://www.koreaherald.com/article/10566046*

© 2024-2026 NexusForum.EU

Overview of priorities and recommendations

	Short-term	Medium-term	Long-term
Priority groups	**Agentic AI and neurosymbolic AI**		
Build European capacity to create and operate frontier AI models	● Build shared datasets and synthetic-data methods Create expert networks monitoring AI innovations anchored to AI Factories Reduce timelines between submission to project start for AI R&I routes Expand AI education and competence programmes	● R&I in alternative LLM architectures and training methods Support emerging AI paradigms, such as world models and physical AI	
A pragmatic approach to Agentic AI and open-source Large Language Models	● Develop criteria and policy support for using non-EU open weight LLMs hosted in the EU Develop methods for testing and verifying compliance with these criteria	● Adoption of EU-based LLMs as they become available	
Dynamic agentic orchestration and workflow execution graphs	● Develop concepts and tooling for workflow graph abstractions for agent execution and orchestration SLA- and constraint-aware agentic orchestration Security layers and access control for agentic AI	● Self-adaptive and self-optimizing agentic orchestration and workflows	
Harness engineering with memory, knowledge integration and verified tool use	Develop memory management techniques for LLMs and agentic AI systems Develop structured knowledge retrieval	Develop graph-grounded agentic workflows Develop hybrid symbolic-neural reasoning for digital	Develop tools that compile logic rules and domain constraints into AI systems

© 2024-2026 NexusForum.EU

		and knowledge graph integration Develop neuro-symbolic harnesses for safety-constrained AI Build shared reference datasets and knowledge resources	twins and industrial systems Define interfaces between LLMs, knowledge graphs and reasoning engines Establish benchmarks for grounded AI	Develop certifiable neuro-symbolic AI systems Develop self-maintaining knowledge-grounded AI systems Establish European knowledge-grounded AI platforms Establish European standards for AI memory and knowledge governance
	Agentic AI and LLMs for devices, edge, and mobile phones	Develop distillation and quantization methods for developing small language models Specify multi-tier knowledge graph architectures Develop knowledge graph delta synchronisation Link neuro-symbolic execution to hardware and runtime targets	Create standardised edge ontologies for priority verticals Develop distributed semantic memory for cloud-edge AI systems Develop hybrid device-edge-cloud agent architectures Develop energy-aware model selection and task routing Establish benchmarks for on-device agentic AI Define formats for small models, local memories and device-level agent capabilities	Develop self-optimising on-device AI systems Develop certifiable on-device and mobile agentic AI Develop cooperative device intelligence Establish European open platforms for mobile and embedded AI Establish best practices for on-device AI memory, updates and user control
		← Timeline →		

5.1 Priority: Build European capacity to create and operate frontier AI models

It is essential to build capacity in Europe to train and operate frontier LLMs, while maintaining a pragmatic AI strategy towards the use of open-weight models. In the longer term, Europe should also aim to develop capacity to develop the next generation of AI models, beyond large language models, including world models and physical AI.

Impact

Support the growth of European AI ecosystems and capacity in training the most advanced AI models, and quickly disseminate advances in AI through educational and skills development initiatives. Nurture and secure domestic AI expertise in Europe as well as research and

© 2024-2026 NexusForum.EU

innovation capacity in training and fine-tuning frontier AI models, to advance state-of-the-art AI expertise in general in Europe.

Recommendations

Short-term
[Ecosystem] Create a pan-European initiative for building common datasets and methods for generating reliable synthetic data for training frontier LLMs, and anchor it with the AI factory ecosystem.
[Ecosystem] Establish expert networks that monitor and spread early AI innovations and trends in Europe. Anchor them in the AI factory ecosystem to support rapid and broad dissemination.
[Support] Develop new mechanisms and procedures that reduce time between application submission to approval and project start for research and innovation project funding applications targeting AI innovations, for example by establishing pre-approval procedures or introducing a more flexible project form. Anchor these in the AI factories and reserve prioritized resources to support such initiatives and for developing and disseminating training material to the broader public.
[Support] Advance European education programs and competence development of expertise in LLM training and post-training methods, and agentic AI in universities and research institutes. Integrate AI transversally across subjects and education programs, and invest in inter-disciplinary skill development of teachers and researchers.
Medium-term
[Research & Innovation] Support development of alternative LLM architectures and training methods.
[Research & Innovation] Support the development of next-generation frontier AI models based on emerging paradigms, including world models and physical AI.

5.2 Priority: A pragmatic approach to Agentic AI and open-source Large Language Models

While Europe develops its own large language models, there will be a capability gap in the EU before they reach the level of the global frontier models. On the other hand, agentic AI is advancing rapidly, and enterprises increasingly rely on the LLM APIs of US AI frontier labs, while developers increasingly adopt Chinese open-source models. The EU should find ways of leveraging available open-weight LLMs hosted in the EU, for example evaluating their use for different applications (based on criteria of trustworthiness, security, etc.).

Impact

For Europe, agentic AI provides an opportunity to leverage existing open-source LLMs to establish leadership in this new paradigm. The rapid evolution of Agentic AI also risks vendor

© 2024-2026 NexusForum.EU

lock-in of a new generation of enterprise AI systems. As the field advances, emerging AI tooling and service ecosystems risk building dependencies on the AI APIs and services offerings of top US hyperscalers. Europe cannot risk falling behind in agentic AI, which represents the next AI evolution of large language models.

Recommendations

Short-term
[Research & Innovation] Develop criteria for evaluating whether a non-EU open-source LLM can be used safely, and make these criteria differentiated based on domain and end-use application.[44] Develop new methods and tooling for testing and verifying a non-EU open-source LLM against these criteria, and to align them to the criteria (ensuring compliance) using post-training or harness methods.
[Regulatory; Ecosystem] Reconcile the EU policy stance with the use of non-EU open-source LLMs, adopting a differentiated approach based on domain and end-use application.
Medium-term
[Research & Innovation; Ecosystem] As EU-based large language models become increasingly capable, promote their integration in agentic AI systems alongside non-EU open source models.

5.3 Priority: Dynamic agentic orchestration and workflow execution graphs

Agentic AI uses large language models to plan and execute workflows. There is a move towards using an intermediate layer with more formal execution graphs and workflow descriptions, rather than ad-hoc, language-based orchestration. This is one way in which the software engineering toolbox is maturing for agentic AI.

Impact

Single-shot Large Language Models alone are not sufficient to make AI useful in complex tasks and enterprise contexts. Recent performance gains of LLM-based systems and agentic AI are largely due to improvements in harness engineering and agent orchestration. Formulating agentic orchestration as execution graphs makes them more deterministic, provides tools to increase observability and transparency, and opportunities to dynamically adapt and optimise agentic workflows.

Recommendations

Short-term

[44] *For example, LLMs used for office productivity applications may not require the strictest ethical alignment; ensuring that LLMs used for coding agents haven't been trained to introduce exploits and security vulnerabilities.*

© 2024-2026 NexusForum.EU

[Research & Innovation] Develop concepts and tooling for workflow graph abstractions for agent execution and orchestration. Develop explicit workflow and execution-graph models for composing agentic systems spanning reasoning steps, tool calls, verification steps, and fallback paths.

[Research & Innovation] SLA- and constraint-aware agentic orchestration. Develop orchestrators that can decide where agent steps should execute based on latency, bandwidth, trust domain, privacy requirements, and resource constraints. Develop methods for agentic workflows with strict SLA requirements, for example latency.

[Research & Innovation] Security layers and access control for agentic AI. Develop methods and frameworks for security enforcement and observability in agentic AI, treating agentic AI as workflows of multiple users/systems with different access control.

Medium-term

[Research & Innovation] Self-adaptive and self-optimizing agentic orchestration and workflows. Develop agentic workflow engines that can re-plan at runtime in response to failures, mobility, asynchronous events, and changing workload conditions while preserving policy compliance and auditability. Develop self-optimisation methods that dynamically improve orchestration and workflows over time.

5.4 Priority: Harness engineering with memory, knowledge integration and verified tool use

The performance of LLMs in agentic systems depend on the harness around the model, such as memory systems, retrieval layers, knowledge graphs, symbolic reasoning components, tool interfaces, policy constraints, verification steps. Structured retrieval, knowledge graphs, and neuro-symbolic AI can help ground AI outputs in verified data structures, domain ontologies and operational context. In cloud-edge and 6G environments, these knowledge systems may be distributed across cloud, edge and device tiers; in other settings, they may operate within a single organisation, industrial edge platform or sectoral AI system.

Impact

Pure deep learning systems such as LLMs remain vulnerable to hallucination, weak logical consistency, poor long-term memory and limited grounding in verified facts. For many European application domains, including digital twins, manufacturing, mobility, smart cities, healthcare, energy and public-sector systems, factual accuracy, traceability and constraint adherence are non-negotiable.

Recommendations

Short-term

Research & Innovation: Develop memory management techniques for LLMs and agentic AI systems. Develop methods for managing short-term context, long-term memory, user memory, organisational memory and domain memory. Research should

© 2024-2026 NexusForum.EU

address memory relevance, forgetting, provenance, consent, update conflicts, privacy, security and auditability. Memory systems should distinguish between temporary conversational context, persistent factual knowledge, sensitive user-specific information and verified domain knowledge.

Research & Innovation: Develop structured knowledge retrieval and knowledge graph integration. Research lightweight retrieval-augmented generation where the retrieval source is a structured knowledge graph or ontology rather than only unstructured text. This should include methods for graph-based retrieval, entity linking, semantic query planning, provenance-aware retrieval, confidence scoring and explanation of retrieved evidence.

Research & Innovation: Develop neuro-symbolic harnesses for safety-constrained AI. Create methods for combining neural models with symbolic rules, domain ontologies, constraint solvers, knowledge graphs and verification steps. The goal is to prevent AI systems from producing outputs or actions that violate known domain constraints, safety rules or operational policies.

Ecosystem: Build shared reference datasets and knowledge resources. Support the creation of open or controlled-access domain knowledge resources for European priority sectors such as manufacturing, mobility, energy, smart cities, public administration and healthcare. These resources should include ontologies, knowledge graphs, benchmark tasks, provenance metadata and licensing models.

Medium-term

Research & Innovation: Develop graph-grounded agentic workflows. Integrate knowledge graphs and ontologies into agentic execution graphs so that agents can plan, act and verify outputs against domain facts and constraints. This should include methods for checking tool calls, workflow steps and generated outputs against structured knowledge before action is taken.

Research & Innovation: Develop hybrid symbolic-neural reasoning for digital twins and industrial systems. Support neuro-symbolic methods that combine learned models with engineering constraints, physical models, causal graphs, process models and digital-twin representations. This is especially important for industrial 6G applications, manufacturing, robotics, mobility and energy systems.

Standardisation: Define interfaces between LLMs, knowledge graphs and reasoning engines. Support common APIs and interchange formats for connecting AI models to knowledge graphs, ontologies, rule engines, constraint solvers and provenance systems. This should avoid lock-in to proprietary agent or RAG platforms.

Testing and Benchmarking: Establish benchmarks for grounded AI. Create benchmarks that measure factual accuracy, constraint adherence, explanation quality, robustness to outdated knowledge, provenance quality, memory consistency and hallucination reduction in knowledge-grounded AI systems.

Long-term

© 2024-2026 NexusForum.EU

Research & Innovation: Develop tools that compile logic rules and domain constraints into AI systems. Develop methods that automatically translate logical rules, safety constraints, engineering requirements or domain policies into model constraints, runtime guards, verification steps or neural representations. The goal is to ensure that AI systems cannot easily violate safety-critical constraints defined by knowledge graphs, ontologies or formal rules.

Research & Innovation: Develop certifiable neuro-symbolic AI systems. Create methods for certifying AI systems that combine foundation models, knowledge graphs, symbolic reasoning, memory and runtime assurance. Certification should cover provenance, traceability, constraint adherence, update management, audit evidence and behaviour under uncertainty.

Research & Innovation: Develop self-maintaining knowledge-grounded AI systems. Develop AI systems that can detect missing, inconsistent or outdated knowledge; propose updates; request human validation; and propagate approved changes across cloud, edge and device tiers. Such systems should include safeguards against knowledge poisoning and unauthorised semantic updates.

Ecosystem: Establish European knowledge-grounded AI platforms. Create open, reusable European platforms for knowledge-grounded AI in key sectors. These platforms should combine domain ontologies, knowledge graphs, RAG infrastructure, symbolic reasoning, model interfaces, benchmarks, governance tools and deployment patterns for edge, cloud and hybrid environments.

Standardisation and Governance: Establish European standards for AI memory and knowledge governance. Develop standards and best practices for persistent AI memory, knowledge graph provenance, semantic interoperability, consent-aware memory, knowledge lifecycle management and auditability. These standards should support trustworthy AI deployment in regulated and safety-critical environments.

5.5 Priority: Agentic AI and LLMs for devices, edge, and mobile phones

Developing agentic AI and LLM capabilities for devices and mobile phones would enable AI systems that are more private, responsive, and resilient. By keeping more inference, context, memory and knowledge close to the user or physical environment, these systems can reduce latency, lower bandwidth requirements, operate under intermittent connectivity and limit the transfer of sensitive data to centralised cloud services.

Impact

This would support a wide range of European use cases, including personal assistants, connected vehicles, industrial maintenance, robotics, smart-city services, healthcare support, energy management, emergency response and AR/VR applications. In these settings, on-device AI can provide local reasoning, context awareness, natural-language interaction, anomaly detection, decision support and coordination with nearby edge services.

© 2024-2026 NexusForum.EU

The priority also has sovereignty and competitiveness implications. Europe should not depend only on remote proprietary AI APIs for everyday AI functionality in devices, vehicles, industrial systems and public-sector applications. Strong on-device and mobile AI capabilities can create demand for European processors, accelerators, software stacks, runtimes, model-compression tools and open AI ecosystems. This is particularly relevant for RISC-V, specialised AI accelerators, neuromorphic hardware and European software toolchains.

Recommendations

Short-term

Research & Innovation: Develop distillation and quantization methods for developing small language models. Distil and quantize large language models for mobile phones and edge with limited memory and processing power. Use these methods to also develop domain-specific LLM-driven AI agents that can be combined in agentic workflows and have a smaller footprint.

Research & Innovation: Specify multi-tier knowledge graph architectures. Develop architectures for partitioning knowledge graphs across infrastructure tiers, for example a global or sectoral knowledge graph in cloud/HPC environments, domain-specific knowledge graphs at the edge, and local knowledge graphs on devices. Research should define which semantic facts need to be local, which can remain centralised, and how consistency, latency, privacy and resilience trade-offs are managed.

Research & Innovation: Develop knowledge graph delta synchronisation. Research methods for synchronising only changed semantic facts, relationships, constraints or embeddings between cloud, edge and device tiers. This should include versioning, conflict resolution, provenance tracking, access control and rollback mechanisms.

Research & Innovation: Link neuro-symbolic execution to hardware and runtime targets. Explore how knowledge-graph reasoning, symbolic constraints and neuro-symbolic inference can be compiled or optimised for different hardware targets, including RISC-V processors, AI accelerators and neuromorphic. This should be treated as an enabler of accurate and efficient edge inference, not only as a high-level AI technique.

Medium-term

Ecosystem and Standardisation: Create standardised edge ontologies for priority verticals. Develop standardised “edge ontologies” for domains such as automotive, smart cities, manufacturing, energy, healthcare and robotics. These ontologies should support shared semantic understanding between AI models, sensors, digital twins, edge nodes, cloud services and human operators.

Research & Innovation: Develop distributed semantic memory for cloud-edge AI systems. Create memory architectures that allow AI systems to maintain consistent and context-aware behaviour across devices, edge nodes and cloud services. This includes mechanisms for semantic cache invalidation, memory summarisation, privacy-preserving memory sharing and policy-aware memory access.

Research & Innovation: Develop hybrid device-edge-cloud agent architectures. Develop agentic AI architectures where local device agents handle immediate, private or

© 2024-2026 NexusForum.EU

low-latency tasks, while edge or cloud agents provide larger-scale reasoning, coordination, retrieval, model updates and cross-device learning. Research should define how tasks are decomposed across tiers and how failures, handovers and offline modes are handled.

Research & Innovation: Develop energy-aware model selection and task routing. Create methods that decide whether a task should be handled by a small local model, an edge model or a cloud/HPC model based on energy, latency, privacy, connectivity, cost, confidence and safety requirements.

Testing and Benchmarking: Establish benchmarks for on-device agentic AI. Develop benchmarks for latency, energy consumption, memory footprint, factual grounding, offline capability, privacy preservation, robustness, tool-use reliability and user-perceived quality. Benchmarks should include realistic European use cases such as mobility, industrial maintenance, robotics, healthcare support and smart-city services.

Standardisation: Define formats for small models, local memories and device-level agent capabilities. Support interoperable formats for compressed models, local semantic memory, tool permissions, device capabilities and agent-to-edge communication. This should avoid lock-in to specific mobile operating systems, chip vendors or proprietary AI assistant ecosystems.

Long-term

Research & Innovation: Develop self-optimising on-device AI systems. Develop AI systems that can adapt model choice, memory use, retrieval, sensor processing and communication patterns over time while respecting energy, privacy, safety and user-control constraints.

Research & Innovation: Develop certifiable on-device and mobile agentic AI. Create methods for certifying AI systems that operate on devices and interact with sensors, actuators, user data, vehicles, robots or industrial equipment. Certification should address model behaviour, memory governance, tool permissions, security, update mechanisms, auditability and fail-safe operation.

Research & Innovation: Develop cooperative device intelligence. Explore methods for groups of nearby devices, vehicles, robots or sensors to cooperate through local AI agents while preserving privacy, security and energy efficiency. This includes swarm intelligence, local consensus, collaborative perception, distributed anomaly detection and edge-assisted coordination.

Ecosystem: Establish European open platforms for mobile and embedded AI. Create open European platforms, reference implementations and developer tools for deploying small language models, local agents, semantic memory and knowledge-grounded AI on mobile, embedded and industrial devices. These platforms should support European hardware targets, open runtimes and sector-specific requirements.

Standardisation and Governance: Establish best practices for on-device AI memory, updates and user control. Develop standards and governance practices for local AI memory, model updates, consent, deletion, provenance, security patches, user override, audit logs and responsible fallback to edge or cloud services.

© 2024-2026 NexusForum.EU

6 AI and Data Workloads across Heterogeneous Computing Environments

Primary destinations
2. An AI stack built on an open European hardware and computing ecosystem
4. Secure, sovereign, and interoperable European computing capabilities
Secondary destinations
1. Scalable and energy-efficient AI and data processing

Background and driving factors

Modern AI and data workloads increasingly need to operate across heterogeneous computing environments, including devices, industrial edge systems, telco edge, cloud platforms, HPC systems, AI Factories and specialised accelerator infrastructures. Different stages of the AI lifecycle place different demands on this infrastructure. Large-scale model training may require HPC or AI-optimised supercomputing. Fine-tuning and evaluation may use AI Factories, cloud platforms or sectoral infrastructures. Inference may run in cloud environments, at the edge, on devices, or across hybrid deployments depending on latency, privacy, cost, energy and reliability requirements.

This diversity creates a major software and systems challenge. AI applications are increasingly tied to specific hardware architectures, cloud service ecosystems, model-serving platforms, data-management systems and operational toolchains. Many organisations adopt hyperscaler AI services because they are convenient and integrated, but this can also create vendor lock-in and make later migration costly. At the same time, European AI infrastructures such as EuroHPC systems and AI Factories often differ from cloud environments in their software stacks, scheduling systems, data movement processes and developer workflows. Bridging these differences is essential if European users are to move from access to compute towards practical AI deployment.

The computing landscape is also becoming more heterogeneous at the hardware level. AI workloads may need to run on CPUs, GPUs, NPUs, RISC-V processors, ARM systems, specialised AI accelerators, neuromorphic systems or future quantum-classical infrastructures. This requires portable AI technology stacks, retargetable compilers, open model formats, common runtime abstractions and mature developer tooling. Without such software layers, new European hardware and infrastructure investments risk remaining difficult to use or isolated from mainstream AI development workflows.

Beyond portability, many AI and data applications are becoming stateful. They depend on sessions, rolling context, user state, digital-twin state, model memory, key-value caches, streaming data, distributed knowledge, and continuous analytics. This is especially important for industrial monitoring, connected mobility, smart energy systems, robotics, AR/VR, emergency response and edge AI. In these settings, it is not enough to move stateless containers or batch jobs. The system must manage where state is placed, how it is

© 2024-2026 NexusForum.EU

synchronised, how it is migrated, and how consistency, integrity, privacy and latency are preserved.

This section focuses on the software, data and runtime capabilities needed to make AI and data workloads usable across heterogeneous environments. This includes stateful continuous analytics, portable AI technology stacks, stateful model migration and federated or distributed AI. Some of these capabilities are federation-critical, especially where data or workloads span organisations and providers. Others are federation-enabling or federation-optional, and may be equally relevant within single-provider, industrial, cloud, edge, HPC or vertically integrated systems.

Europe in relation to the state of the art

Europe has strong assets for AI and data workloads across heterogeneous infrastructure: EuroHPC, AI Factories, industrial edge environments, telecom networks, data spaces, research infrastructures, embedded systems, robotics, automotive platforms and emerging European hardware initiatives. However, these assets are not yet experienced by many developers and users as a coherent AI deployment environment.

The global state of the art is currently shaped by highly integrated hyperscaler ecosystems, where compute, storage, data services, model APIs, development tools, deployment pipelines, monitoring and security are offered as unified platforms. These ecosystems reduce friction for developers, but also create dependency on proprietary services, hardware stacks and operational models. For European SMEs, public-sector users and industrial actors, the convenience of these platforms can make it difficult to adopt alternative infrastructures, even where European compute or data resources are available.

Europe's challenge is therefore not only to provide compute capacity, but to make European and European-accessible infrastructure user-friendly and part of the larger business ecosystems. AI developers should not need deep expertise in every underlying environment to deploy models across cloud, HPC, edge or device settings. They need portable model formats, open runtimes, common APIs, workflow tools, data movement services, observability, reproducible deployment patterns and guidance on when to use cloud, edge, HPC, AI Factories or specialised hardware.

Europe also has a particular interest in reducing dependency on narrow hardware and software ecosystems. Many AI workloads remain closely tied to CUDA-based GPU stacks or proprietary cloud AI services, although the situation is improving for LLM-specific frameworks, such as vLLM and SGLang. European investments in RISC-V, AI accelerators, open hardware, EuroHPC, AI Factories and edge AI will only succeed if they are matched by software maturity: compilers, runtime systems, ML libraries, model-serving frameworks, performance-portability tools and developer communities.

At the same time, Europe is well positioned to lead in trustworthy and operational AI for real-world environments. Industrial systems, mobility, energy, public services, healthcare, robotics and smart cities require AI systems that are not only powerful, but also portable, auditable, explainable, secure, energy-aware and capable of operating under local constraints. This creates an opportunity for Europe to differentiate itself through open, trustworthy and heterogeneous AI deployment stacks rather than trying only to replicate hyperscaler ecosystems.

© 2024-2026 NexusForum.EU

The priorities in this section therefore address a practical gap between infrastructure availability and AI adoption. They focus on making AI and data workloads portable, stateful, efficient, trustworthy and deployable across the diverse environments where European industry, research and public services actually operate.

Overview of priorities and recommendations

Priority groups	Short-term	Medium-term	Long-term
AI and Data Applications in the Computing Continuum			
Stateful continuous analytics and distributed state abstractions	Cloud-edge state synchronization mechanisms Programming models for stateful serverless at the edge	Developer platform tooling for stateful serverless in the continuum. Include AI integrations Run-time optimisation considering state sync/migration costs.	Multi-provider state management ensuring consistency and integrity.
Portable AI applications and stateful model migration	Develop portable AI runtime and model-serving abstractions. Improve support for heterogeneous AI execution Support retargetable ML compilers and portable ML libraries. Create reference AI deployment recipes for European infrastructures Promote open model and workload formats.	Develop cloud-HPC-edge workflow portability Develop performance-portable AI software stacks Support open-source maintenance for strategic AI software components. Establish portability benchmarks. Create application-porting programmes.	Develop technology-neutral AI execution environments. Establish European profiles for portable AI workloads. Build a sustainable European open AI software stack Enable automatic workload adaptation across hardware targets.
Stateful AI mobility and model migration Destinations	Define stateful AI workload and state models. Develop stateful inference handover mechanisms. Separate static model assets from dynamic runtime state. Study failure modes in stateful model migration.	Migrateable AI Container Standards Develop migrateable AI runtimes for edge and mobile environments. Define interoperable formats for stateful AI containers Develop policy-aware state migration Establish benchmarks for stateful AI mobility.	Develop distributed operating-system concepts for AI services. Develop self-optimising AI mobility systems Establish European interoperability profiles for stateful AI mobility. Build open European reference implementations.

© 2024-2026 NexusForum.EU

		Link stateful AI migration to predictive handover.	Demonstrate stateful AI mobility in European use cases.	
	Federated and distributed AI	Adapting LLM designs to support federated computation Explore Split-Aware Model Training Develop generative semantic compression. Define stateful AI container formats that separate static weights (cache everywhere) and dynamic state (migrate) to enable liquid runtime migration.	Define industry standards for Split-Ready model architectures Ground Federated Green Learning in smart energy grid technologies For generative semantic compression, establish techniques to ensure the receiver's GenAI reconstructs the scene accurately based on shared context	A distributed Operating System concept where Process Migration of AI agents is a native primitive Semantic Edge Caching for Generative Assets at the edge
		← Timeline →		

6.1 Priority: Stateful continuous analytics and distributed state abstractions

As data is increasingly generated and processed at the edge, continuous analytics will need to operate in highly distributed environments. Applications and workflows that depend on prior data or span edge sites require efficient **state synchronization**, motivating new paradigms for state management and continuous analytics across the continuum. While central clouds offer state management, there is no technical solution for multi-provider distributed cloud-edge state management that ensures consistency and integrity beyond early research (e.g., conflict-free data types). The serverless approach abstracts infrastructure, and **stateful serverless** could also abstract distributed state management and synchronization.State management and operational tooling must work across multi-provider distributed cloud-edge environments where consistency/integrity and visibility are currently unsolved at scale.

Impact

Important AI applications, including agentic AI, depend on persistent state management. To support state management in a distributed computing continuum or cloud-edge federations, state handling must be turned into a portable platform abstraction, building on recent work that provides stronger guarantees for stateful serverless (e.g., causal consistency for shared state) and even transactional/serializable execution of workflows of stateful functions, while extending these ideas to heterogeneous, distributed continuum deployments.

If solved, the impact could be large: it would make stateful edge analytics and AI services *repeatable and less error-prone* across deployments and enable higher-value applications such as industrial digital twins and predictive maintenance, smart-grid monitoring and control,

© 2024-2026 NexusForum.EU

connected mobility/cooperative perception, and real-time video analytics and tracking across cameras/sites, with materially reduced network traffic and end-to-end latency by keeping state and compute close to the data.

Recommendations

Short-term

[Research & Innovation] Cloud-edge state synchronization mechanisms. Develop mechanisms for seamless state synchronization between cloud and edge, ensuring consistency and reliability.

[Research & Innovation] Programming models for stateful serverless at the edge. Simplify development of stateful serverless applications in edge/continuum environments by abstracting distributed state management and synchronization.

Medium-term

[Research & Innovation; Ecosystem] Developer platform tooling for stateful serverless in the continuum. Provide an integrated developer platform (SDKs, reusable components/templates, testbeds and simulation, CI/CD and deployment tooling) that (i) operationalises the stateful serverless programming models and (ii) exposes and automates run-time optimisation controls (e.g., policies/placement hints, observability hooks, and feedback loops), enabling developers to seamlessly exploit state placement and sync/migration trade-offs (latency/resources/energy). **Include AI integrations** (e.g., natural-language-to-configuration/policy generation, guided debugging and performance tuning, and automated generation of monitoring dashboards) to further lower the barrier to correct use and optimisation.

[Research & Innovation] Run-time optimisation considering account state sync/migration costs. Optimise state location, and resource and latency costs of state synchronization and migration (including trade-offs with energy efficiency).

Long-term

[Research & Innovation] Multi-provider state management ensuring consistency and integrity. Develop multi-provider state management solutions for distributed cloud–edge environments that provide robust consistency and integrity guarantees.

6.2 Priority: Portable AI applications and open AI technology stacks

AI applications are increasingly tied to specific hardware architectures, cloud services, model-serving platforms, data-management systems and operational toolchains. Many organisations adopt hyperscaler AI services because they provide integrated development environments, APIs, model hosting, storage, monitoring, security and deployment pipelines. This convenience can accelerate adoption, but it can also create vendor lock-in and make later migration costly.

© 2024-2026 NexusForum.EU

At the same time, European AI infrastructures such as EuroHPC systems, AI Factories, cloud-edge platforms, industrial edge environments and emerging European hardware stacks differ in their software environments, scheduling systems, data movement mechanisms and developer workflows. AI developers often face substantial friction when moving between cloud platforms, HPC systems, edge environments and specialised accelerator architectures.

This priority focuses on making AI applications more portable across heterogeneous computing environments. Portability does not mean that every AI workload should run everywhere, or that all workloads should move dynamically across providers. It means that AI models, applications and workflows should be easier to adapt, deploy and optimise across different environments when there is value in doing so.

The priority includes open model formats, portable ML libraries, retargetable compilers, hardware abstraction, common runtime interfaces, model-serving standards, reproducible deployment recipes, workflow portability and developer tools. It is especially important for reducing dependency on narrow hardware and software ecosystems, including CUDA-only stacks or proprietary cloud AI platforms.

Impact

Portable AI applications and open AI technology stacks would make European AI infrastructure easier to use and more attractive to developers, SMEs, public-sector organisations and industrial users. They would reduce the friction of moving AI workloads between cloud, HPC, edge and specialised hardware environments, and they would help avoid lock-in to specific cloud providers, model-serving platforms or proprietary accelerator ecosystems.

This priority is also essential for European hardware ambitions. RISC-V processors, AI accelerators, neuromorphic systems and other emerging architectures will only be adopted if they are supported by mature software stacks, compilers, libraries, runtimes, developer tools and model-serving frameworks. Without this software layer, even promising hardware may remain isolated from mainstream AI development.

Portability can also support sustainability and efficiency. If AI workloads can be deployed across different environments with less engineering effort, they can be placed where they are most efficient in terms of energy, cost, latency, privacy and performance.

Recommendations

Short-term

[Research & Innovation] Develop portable AI runtime and model-serving abstractions. Develop common abstractions for deploying AI models across cloud, HPC, edge and specialised accelerator environments. These should support model loading, inference serving, batching, quantisation, monitoring and lifecycle management.

[Research & Innovation] Improve support for heterogeneous AI execution. Improve support for efficient AI workload execution across GPUs, CPUs, RISC-V, ARM, NPUs, AI accelerators and emerging hardware. This should include hardware-aware optimisation without forcing developers into vendor-specific stacks.

© 2024-2026 NexusForum.EU

[Research & Innovation] Support retargetable ML compilers and portable ML libraries. Invest in compiler and runtime technologies based on open ecosystems such as LLVM/MLIR, WebAssembly and portable AI frameworks. These should reduce dependency on CUDA-only or x86-specific execution paths.

[Ecosystem] Create reference AI deployment recipes for European infrastructures. Develop reusable deployment templates for AI Factories, EuroHPC systems, European cloud providers, industrial edge systems and specialised hardware testbeds.

[Standardisation] Promote open model and workload formats. Support interoperable formats for models, metadata, quantisation, runtime requirements, evaluation results, provenance and deployment constraints.

Medium-term

[Research & Innovation] Develop cloud-HPC-edge workflow portability. Develop tools and services that make it easier to move AI workflows between cloud-native environments, HPC scheduling systems, AI Factories and edge deployments.

[Research & Innovation] Develop performance-portable AI software stacks. Create software stacks that can optimise AI workloads for different hardware targets while preserving a common developer experience.

[Ecosystem] Support open-source maintenance for strategic AI software components. Fund long-term maintenance of open libraries, compilers, runtimes, model-serving tools and benchmarking frameworks needed by European AI and hardware ecosystems.

[Testing and Benchmarking] Establish portability benchmarks. Benchmark how easily AI workloads can be deployed across different platforms, hardware targets and operational environments, including effort, performance, energy, cost and reproducibility.

[Ecosystem] Create application-porting programmes. Support SMEs, researchers and industrial users in porting AI workloads to European AI Factories, EuroHPC systems, European cloud platforms and emerging European hardware.

Long-term

[Research & Innovation] Develop technology-neutral AI execution environments. Develop mature execution environments where AI applications can target heterogeneous infrastructure without being rewritten for each platform.

[Standardisation and Governance] Establish European profiles for portable AI workloads. Define European interoperability profiles for AI workloads, including model formats, runtime requirements, metadata, observability, security and energy reporting.

[Ecosystem] Build a sustainable European open AI software stack. Create durable governance and funding mechanisms for the open-source components that underpin European AI deployment, including compilers, libraries, runtimes, model-serving tools and benchmarking infrastructure.

[Research & Innovation] Enable automatic workload adaptation across hardware targets. Develop AutoML and compiler-assisted methods that automatically quantise,

© 2024-2026 NexusForum.EU

compile, optimise and deploy AI models across heterogeneous fleets of devices, edge nodes, cloud systems and accelerators.

6.3 Priority: Stateful AI mobility and model migration Destinations

Modern AI applications are increasingly interactive, personalised and context-dependent. They may maintain user sessions, conversational history, key-value caches, retrieved context, local memory, sensor state, digital-twin state or agent workflow state. In mobile, industrial and edge environments, this state may need to move or be synchronised as users, devices, vehicles, robots or workloads move between devices, edge nodes, cloud platforms and HPC-backed services.

This priority focuses on **stateful AI mobility**: the ability to preserve continuity of AI services when execution moves between infrastructure locations. This includes migrating inference state, maintaining session continuity, synchronising model context, handling failures during migration, and deciding which parts of an AI workload should run locally, at the edge, in the cloud or on specialised infrastructure.

The need is especially clear in 5G/6G and edge environments. Applications such as connected vehicles, AR/VR, mobile AI assistants, cooperative robotics, emergency response, smart logistics and industrial field operations may require low-latency AI services while users or devices move across network and edge domains. Restarting an AI session after every handover would reduce quality of service and increase latency, bandwidth use and energy consumption. Instead, future AI systems will need mechanisms for transferring only the relevant dynamic state, such as key-value cache deltas, context summaries, tool state or agent memory, while keeping static model weights cached or replicated where appropriate.

Stateful AI mobility should not be understood as a universal requirement for all AI workloads. Many AI applications will remain stateless, batch-oriented or tied to one deployment environment. However, for mobile, interactive and edge-native AI systems, stateful migration and continuity will become an important capability.

Impact

Stateful AI mobility would improve the usability and reliability of AI services in mobile and distributed environments. It would allow AI applications to preserve context and continuity across handovers, failures, changes in connectivity and shifts between local, edge and cloud execution. This would benefit AR/VR, connected vehicles, robotics, industrial maintenance, smart-city services, mobile assistants, emergency response and other latency-sensitive applications.

The priority can also reduce bandwidth and energy use. If static model weights are cached close to users and only dynamic state is migrated, AI systems can avoid repeatedly transferring large models or restarting sessions. This can lower network traffic, reduce latency and improve quality of service.

© 2024-2026 NexusForum.EU

For Europe, stateful AI mobility is strategically relevant because it connects several European strengths: telco infrastructure, 5G/6G research, industrial edge, mobility, robotics, AI Factories, EuroHPC and emerging AI hardware. It also creates demand for open runtimes, portable model formats, state management abstractions and edge-cloud orchestration tools that are not locked into one proprietary cloud ecosystem.

Recommendations

Short-term

[Research & Innovation] Define stateful AI workload and state models. Develop common abstractions for the state of AI applications, including session state, key-value caches, retrieved context, user memory, tool state, agent workflow state and digital-twin state. Define which state can be replicated, summarised, discarded, encrypted, migrated or kept local.

[Research & Innovation] Develop stateful inference handover mechanisms. Research methods for transferring AI inference state between devices, edge nodes and cloud services without restarting the session. This should include differential key-value cache transfer, context summarisation, partial state migration, rollback and recovery.

[Research & Innovation] Separate static model assets from dynamic runtime state. Develop runtime and container patterns that distinguish static model weights, reusable embeddings and shared libraries from dynamic state such as session context, memory and key-value caches. Static assets should be cacheable or pre-positioned; dynamic state should be migrated only when needed.

[Research & Innovation] Study failure modes in stateful model migration. Develop methods for handling partial migration failure, edge node unavailability, inconsistent state, network interruption, rollback, session recovery and safe degradation.

[Research & Innovation] Link stateful AI migration to predictive handover. Coordinate with telco-edge work on predictive mobility and handover, so that application-layer state migration can be triggered before connectivity changes affect user experience.

Medium-term

[Ecosystem] Migrateable AI Container Standards. Developing a standard container format for "Stateful AI" that separates the static model weights (cached everywhere) from the dynamic user state (migrated).

[Research & Innovation] Develop migrateable AI runtimes for edge and mobile environments. Develop runtimes that support stateful migration of AI services across devices, edge nodes and cloud services. These runtimes should optimise for latency, bandwidth, privacy, energy, cost and service quality.

[Standardisation] Define interoperable formats for stateful AI containers. Support common formats that describe model assets, runtime dependencies, dynamic state, memory, permissions, provenance and recovery behaviour. This should avoid locking stateful AI mobility into one cloud, telco or hardware ecosystem.

© 2024-2026 NexusForum.EU

[Research & Innovation] Develop policy-aware state migration. Create mechanisms that decide whether state may move based on privacy, security, data governance, sovereignty, consent, sectoral regulation and organisational policy constraints.

[Testing and Benchmarking] Establish benchmarks for stateful AI mobility. Develop benchmarks for migration latency, bandwidth use, energy consumption, state consistency, recovery time, privacy preservation and user-perceived continuity.

[Ecosystem] Demonstrate stateful AI mobility in European use cases. Support pilots in connected vehicles, AR/VR, industrial field operations, robotics, emergency response and smart-city services.

Long-term

[Research & Innovation] Develop distributed operating-system concepts for AI services. Explore operating-system and runtime abstractions where process migration, model-state mobility, edge-cloud placement and AI service continuity are native capabilities rather than overlays.

[Research & Innovation] Develop self-optimising AI mobility systems. Create systems that automatically decide when to keep AI execution local, move it to an edge node, use cloud/HPC resources, or degrade gracefully, based on changing conditions.

[Standardisation and Governance] Establish European interoperability profiles for stateful AI mobility. Develop profiles for telco-edge, industrial edge, public-sector and mobility applications, including state formats, handover interfaces, privacy requirements, audit logs and fallback mechanisms.

[Ecosystem] Build open European reference implementations. Support open-source implementations of stateful AI migration runtimes, testbeds and developer tools to reduce dependency on proprietary cloud or mobile ecosystems.

6.4 Priority: Federated and distributed AI

Federated computation could play a significant role in the future of AI training by enabling more efficient and privacy-preserving methods to train foundational models. Developing efficient federated computation algorithms capable of handling LLM complexity is imperative for effective model aggregation. Additionally, optimizing resources ensures scalability and efficiency, while rigorous evaluation and testing across diverse data distributions and network conditions ensure performance. Training and deploying generative AI models can be resource-intensive, requiring significant computational power and storage. A computational strategy to address this challenge implies the usage of compression techniques to reduce computational cost, memory footprint and energy consumption.

Split Inference, where a massive model (like an LLM) is partitioned layer-by-layer between constrained edge devices and servers enable Foundation Models to run on device that would otherwise be impossible due to memory constraints. Split computing also enables early exits by stopping computation if confidence is high and distributes the inference workload, enabling complex AI on battery-powered devices.

© 2024-2026 NexusForum.EU

Immersive 6G applications like Holography and XR may reduce bandwidth requirements by transmitting *semantics* rather than pixels. Generative AI acts as a compression engine where the sender transmits a semantic description that is reconstructed with GenAI at the receiver. *Generative Semantic Compression* can potentially reduce bandwidth to enable high-quality experiences on limited networks. It is a nascent field where European research in semantic communications is strong and represents a paradigm shift from communication theory to *Communication/Computation Co-design* to enable immersive telepresence on mobile networks without requiring impossible bandwidths, fundamentally changing the economics of streaming. The implementation of such federated computing for LLMs requires distributed network infrastructure with reliable communication protocols.

For real-time AI, the current container migration (such as CRIU) is too sluggish. Europe should strive for liquid runtimes (such as WebAssembly-based state containers) that are optimized for AI state transfer in order to represent state migration among continuum nodes with minimal latency. In situations like vehicle-to-everything (V2X), this is essential for preserving QoS.

Impact

The radical shift of Federated computation for Foundation Models training could lead to more specialised and efficient AI architectures that can handle the complex demands of training large-scale neural networks while preserving privacy and leveraging distributed data sources. This shift is part of the broader trend towards more powerful and versatile AI systems that can be adapted for a multitude of business and consumer applications. Federated computation further enhances this by allowing for distributed training across various devices and data centres, potentially leading to more robust and widely applicable AI models.

Enables the deployment of GenAI on resource-constrained European IoT and mobile devices, reducing dependency on US hyperscaler APIs for every inference request. This also ensures seamless user experience for mobile AI applications (XR, Assistants) and prevents service interruption during handover. By enabling high-performance AI on legacy or low-power European hardware it could drastically reduce the carbon footprint of AI inference.

Federation computation also has the potential to comply with the stringent requirements of existing privacy regulations and promotes the possibility of keeping the data on local devices significantly reduces the risk of data breaches and unauthorized access. Further information is found e.g. in the reports by Woisetschläger et.al and Zhuang et.al.[45,46]

Recommendations

Short-term

[Research & Innovation] Adapting LLM designs to support federated computation may involve modifying them for decentralized data sources and partial updates. Research into transmitting only the *delta* of an LLM's KV-cache between edge nodes. Partial migration failure, e.g., if node goes offline mid-transfer requires sophisticated rollback and session

[45] *Woisetschläger, H., Isenko, A., Wang, S., Mayer, R., & Jacobsen, H. A. (2024). A survey on efficient federated learning methods for foundation model training. arXiv preprint arXiv:2401.04472.*

[46] *Zhuang, W., Chen, C., & Lyu, L. (2023). When foundation model meets federated learning: Motivations, challenges, and future directions. arXiv preprint arXiv:2306.15546.*

© 2024-2026 NexusForum.EU

recovery. A sub-task where the edge node caches completed LLM layers so that repeated queries do not re-compute cloud-side layers. Post-Training Quantization (PTQ) for Edge to quantize Foundation Models down to 2-4 bits with minimal accuracy loss, specifically tuned for heterogeneous edge hardware.

[Research & Innovation] Explore Split-Aware Model Training so that models are pre-trained with split points as first-class architecture parameters, not post-hoc heuristics. Explicitly link to Liquid AI for stateful handover of the partial computation. Develop algorithms that dynamically determine the optimal "split point" of a Neural Network graph based on real-time bandwidth and battery status.

[Research & Innovation: **Develop generative semantic compression.** *Generative Semantic Compression* requires *Semantic Common Ground* protocols for sender and receiver edge nodes to negotiate compatible generative models before transmission. It needs to support error bounds for reconstruction quality (analogous to rate-distortion theory). *Semantic Edge Caching with Invalidation* should be tested explored, where the edge node proactively pushes a cache-invalidation signal rather than waiting for the receiver to detect quality degradation, e.g., when the scene changes significantly. Standardize semantic symbols and prompt syntax to serve as the universal protocol for generative reconstruction across different vendor devices.

[Research & Innovation] Define stateful AI container formats that separate static weights (cache everywhere) and dynamic state (migrate) to enable *liquid runtime migration*. Introduce a compression-offloading co-optimisation framework to allow the system to select the best trade-off between running a compressed model locally vs. offloading a larger model to the edge server, as a function of task priority and network conditions.

Medium-term

Ecosystem: **Define industry standards for Split-Ready model architectures** (e.g., standardizing intermediate tensor formats) to allow a device from Vendor A to offload the second half of a model to an Edge Node from Vendor B.

Research & Innovation: **Ground Federated Green Learning in smart energy grid technologies**. Orchestration algorithms for Federated Learning that only select worker nodes powered by renewable energy or with excess battery capacity.

Ecosystem: **For *generative semantic compression*, establish techniques to ensure the receiver's GenAI reconstructs the scene accurately based on shared context** (e.g., a *Digital Twin* of the room), avoiding dangerous errors in professional applications.

Long-term

Support: A distributed Operating System concept where Process Migration of AI agents is a native primitive, not an overlay. Enabling a cluster of local devices (e.g., a room full of smartphones) to pool their NPU resources to run a single massive model collectively without touching the cloud.

Ecosystem: ***Semantic Edge Caching* for *Generative Assets* (LoRA adapters, 3D models) at the edge** so that only minimal semantic tokens need to be transmitted for reconstruction.

© 2024-2026 NexusForum.EU

NEXUS FORUM.EU

PART B

PILLAR II:

Cognitive Computing Continuum Convergence & Infrastructure

www.eucloudedgeiot.eu

7 Sustainable and Energy-efficient AI infrastructure

Primary destinations
1. Scalable and energy-efficient AI and data processing
Secondary destinations
2. An AI stack built on an open European hardware and computing ecosystem
4. Secure, sovereign, and interoperable European computing capabilities

Background and driving factors

Next-generation AI chips are generating thermal loads that challenge conventional air-cooling solutions. Modern high-density AI datacenters further require large amounts of power, which is becoming a bottleneck in identifying a location and acquiring permits and energy grid capacity.

New cooling and data-driven energy management solutions will be needed, as well as technical and regulatory innovations that can speed up permitting and deployment of scalable renewable energy solutions.

Furthermore, the greatest energy gains will not come through improved cooling and workload management, but from carefully co-designing and optimising the whole stack from hardware to software. Energy-efficiency is therefore a cross-cutting priority that should guide the design of AI architectures and algorithms, the design of hardware optimised to run them, and the development of energy-efficient software and code optimisation techniques.

Europe in relation to the state of the art

Europe is a global leader in cooling solutions in general, and despite recent interest in implementing liquid cooling and investigate immersion cooling, adoption of liquid cooling is still limited in datacenters operated by European entities. One major challenge is cost, since liquid cooling requires high upfront investment (CAPEX), but has lower operational costs (OPEX), and requires a more complex HVAC system design than standard air cooling.

Europe is leading in waste heat recovery, for example in heat recovery technologies for both ventilation and liquid cooling, and integration with district heating. Despite the technology being there, the lack of financial incentives and clear business models between the involved actors, as well as the lack of established legal frameworks assigning obligations and ownerships, slow down the integration process and limit physical deployment of waste heat recovery.

Furthermore, European datacenters generally don't have the necessary sensor and monitoring solutions to perform data-driven optimisation of datacenter operation. In comparison,

© 2024-2026 NexusForum.EU

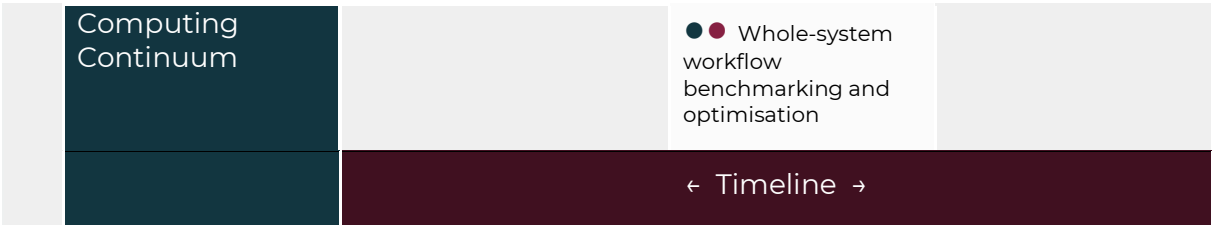
modern hyperscaler datacenters have advanced monitoring solutions, using machine learning algorithms to optimize operations.

Europe is also well-positioned globally in smart grid and renewable energy technologies, although European actors have fallen behind China in some areas, especially in terms of unit pricing for PV. However, Europe still faces regulatory and organisational challenges in implementing smart grid technologies and local energy sharing, especially regarding energy market permitting process, somewhat limiting new deployments of renewable energy integrated with existing energy infrastructures.

Overview of priorities and recommendations

Priority groups	Short-term	Medium-term	Long-term
Sustainable and Energy-efficient AI infrastructure			
Energy and cooling solutions for high-density datacenters	●● New cooling solutions for next-generation AI chips & datacenters. ●● Implement accurate energy & system performance measurements in datacenters. ●● Establish standardized energy performance & resilience metrics/benchmarks. ●● Support and incentivize investments in highly energy-efficient datacenter solutions..	●● Digital twins for holistic data-centre management. ●● Adopt a whole-system approach.	●● Extend holistic data-centre management to the computing continuum.
Solving renewable energy supply + datacenter workloads as a flexible energy asset in the smart grid	●● Simplify regulation for installing/sharing energy locally and incentivize waste-heat reuse. ●● On-site or local renewable generation at scale for high-power datacenters. ●● Define an energy flexibility concept for datacenters, and create a benchmarking system	●● Implement carbon- and energy-aware scheduling & orchestration.	
Energy-grid-aware scheduling in the	●● Develop an energy flexibility model for carbon- and grid-aware scheduling and orchestration	●● Data centres as active smart-grid participants	

© 2024-2026 NexusForum.EU



7.1 Priority: Energy and cooling solutions for high-density datacenters

Impact

Without advances and deployment of new cooling solutions, it will not be possible to use next-generation AI chips. Ensuring European datacenters are data-driven and use efficient datacenter (energy) management will lower their operational costs and make them more competitive. This could also strengthen Europe's attractiveness for large-scale AI and supercomputing investments.

Recommendations

Short-term

[Research & Innovation] New cooling solutions for next-generation AI chips and datacenters. For example liquid or even immersion cooling is needed to keep up with heat rejection of high-power AI chips. Immersion cooling in particular needs more research into hardware degradation over time.

[Guidelines & Regulation] Implementation of accurate energy and system performance measurements in datacenters. Develop strong guidelines, best practices, and requirements for performance measurements in new datacenters, and for retrofitting existing ones, that will prepare them for future data centre energy management solutions. This includes sensors and data collection systems.

[Benchmarking; Ecosystem] Establish a standardized set of energy performance and resilience metrics and benchmarks. While metrics like Power Usage Effectiveness (PUE) are valuable, they fall short in capturing the efficiency of processing per unit of energy, especially in the context of a highly heterogeneous continuum.

[Support] Support and incentivize large up-front investments in highly energy-efficient datacenter solutions, for example through special financial instruments.

Medium-term

[Research & Innovation] Digital twins for holistic data centre management. Develop solutions for jointly controlling cooling, management and workload optimisation of datacenters, using digital twin technologies and hybrid modeling approaches combining physical models and CFD, physics-based ML, and AI models.

© 2024-2026 NexusForum.EU

[Research & Innovation, Ecosystem] Adopt a whole-system approach. It's vital to recognize that power consumption isn't confined to data centres alone but extends to the network and specific applications, necessitating benchmarking considerations based on workflows. Innovation should also target energy-efficient software for IoT devices and energy-restricted devices to minimize power consumption.

Long-term

[Research & Innovation] Extend holistic data centre management to the Computing Continuum: Existing work has developed theoretical frameworks for managing computing continuum infrastructure, but most existing datacenter infrastructure (especially those owned by European operators) still lack the necessary data collection and monitoring systems to deploy them.

7.2 Priority: Solving renewable energy supply and datacenter workloads as a flexible energy asset in the smart grid

Impact

The high energy requirements of modern high-density AI datacenters have become a bottleneck in the planning and permissioning of datacenters. New technical solutions and regulatory frameworks can allow rapid deployment of local or on-site energy solutions at scale.

Furthermore, many datacenter workloads (such as batch training) allow some flexibility regarding when they need to run, which opens up the potential for datacenters to play an important role in energy-grid flexibility.

Recommendations

Short-term

[Guidelines & Regulation; Ecosystem] Simplify the regulatory framework for installing and sharing energy locally, and incentivise waste heat reuse. Current regulations slow down the installation of new energy supply at large scale, and makes it difficult to share energy surplus locally. Furthermore, waste heat reuse is complicated due to questions of guarantees of supply, financial incentives, who should pay for the infrastructure, contractual obligations, and so on. Streamlining regulations could make it easier for datacenters to buy (and sell) excess renewable energy from the local community. Establish common guidelines or frameworks to ensure waste heat reuse, especially in new datacenters.

[Research & Innovation] On-site or local renewable energy generation at scale for high-power datacenters. The high energy requirements of modern high-density AI datacenters have become a bottleneck in the planning and permissioning of datacenters. New technical solutions and regulatory frameworks can allow rapid deployment of local or on-site energy solutions at scale, achieved either through innovative large-scale power

© 2024-2026 NexusForum.EU

generation plants, or pooling of distributed energy resources (see, e.g., *energy communities* and *positive energy districts*).

[Research & Innovation; Benchmarking]: Develop an energy flexibility model for carbon- and grid-aware scheduling and orchestration. Previous work has investigated energy- and carbon-aware workload scheduling and orchestration algorithms. Future work should focus on defining a flexibility concept for datacenters, identify the types of workload that can contribute to energy flexibility (AI training is a strong candidate), and develop a benchmarking ecosystem of representative workloads. Efforts should focus on real and representative workloads.

Medium-term

[Research & Innovation; Support] Implementation of carbon- and energy-aware scheduling and orchestration. Adapt and implement scheduling and orchestration algorithms to shift workloads in time and space to react to local fluctuations in renewable energy supply, as well as take advantage of excess energy supply. Incentivize their adoption and implementation, while ensuring adherence with best practices and guidelines.

7.3 Priority: Energy-grid-aware scheduling in the Computing Continuum

Impact

Directly contributes to EU Green Deal objectives and climate neutrality by reducing the environmental footprint of digital infrastructure; prepares the continuum for smart grids and reduces contention for energy resources with other industries.

Recommendations

Short-term

Develop an energy flexibility model for carbon- and grid-aware scheduling and orchestration *(measure: research & innovation)*: Adapt scheduling/orchestration to shift workloads in time and space based on local renewable fluctuations and excess energy supply.

Medium-term

[Research & Innovation + Ecosystem] Data centres as active smart-grid participants *(measure:)*: Use open energy data, carbon-aware grids, and forecasting of local energy mixes to optimise carbon emissions (not only energy efficiency); further investigate and implement waste-heat recovery.

[Research & Innovation + Ecosystem] Whole-system workflow benchmarking and optimisation. Treat power consumption as spanning data centres, networks and

© 2024-2026 NexusForum.EU

applications; benchmark workflows and target energy-efficient software for IoT/energy-restricted devices.

© 2024-2026 NexusForum.EU

8 Telco Cloud–Edge: Telco as One of the Main Tenants and Infrastructure Providers

The convergence of networking and compute will be crucial in transforming Europe's telecommunications sector to support modern complex, data-driven applications. This strategic integration is essential for enhancing network performance by reducing latency, optimizing bandwidth utilization, and improving overall efficiency. This convergence is vital for ensuring that European infrastructure is robust enough to support an economy and society that are increasingly digital. By processing data closer to the source, these converged networks greatly reduce response times and increase the reliability of digital services, fostering technological resilience and operational efficiency across Europe.

The adoption of open network technologies, including aspects of Open Radio Access Networks (O-RAN), plays a part in this transformation by introducing more flexibility and vendor diversity into the network architecture. This approach allows for a more dynamic allocation of resources, adapting in real-time to varying traffic and service demands, which is crucial for the bandwidth-heavy and latency-sensitive applications prevalent today. The O-RAN Alliance[47] is defining a leading architecture that exemplifies this broader movement towards open and seamless integration of compute and networking capabilities in radio access networks, but it is important for European sovereignty to also support other approaches to this open networking landscape.

The goal of a fully converged network and compute infrastructure cannot be reached without first addressing some challenges. Technical issues like interoperability between different network components, complexity in managing distributed architectures, and heightened security risks are significant. However, Europe is well-positioned to approach these challenges through joint regulatory frameworks, strategic partnerships across the industry and society, and continuous innovation in cybersecurity and network management.

To remain competitive and secure, the EU needs to focus on strategic investments in digital infrastructure, which includes enhancing connectivity through advanced mobile networks and robust data services. This will not only support the immediate needs of European citizens and businesses but can also position Europe as a leading player in the global digital economy.

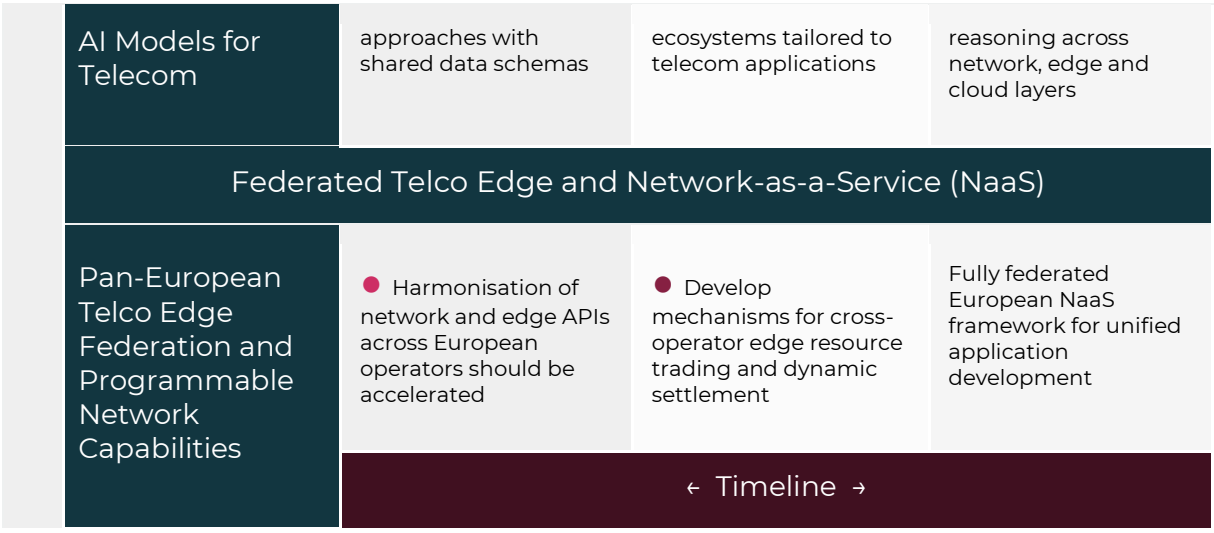
[47] *https://www.o-ran.org/*

© 2024-2026 NexusForum.EU

Overview of priorities and recommendations

Priority groups	Short-term	Medium-term	Long-term
Open Radio Access Networks (Open RAN)			
Secure and Governable Multi-Vendor Open RAN Architectures	● Define European interoperability and security frameworks for Open RAN deployments	● Develop runtime integrity verification and software supply-chain transparency mechanisms	● Establish harmonised European governance models for multi-vendor RAN ecosystems
AI-Native and Cross-Layer Optimisation in Open RAN	● Develop coordination mechanisms for RIC applications to prevent conflicting optimisation actions	● Develop energy-aware cross-layer optimisation across RAN, transport and edge	● Integrate RAN control fully into AI-native cloud-edge orchestration frameworks
Seamless data connectivity and predictive handover across different networks			
Predictive and Resilient Handover Across Heterogeneous Networks	● Sensing and quality-of-service monitoring across all available physical network channels	● Develop integrated hardware–software stacks that seamlessly combine terrestrial and satellite connectivity	● Seamless connectivity should evolve towards fully autonomous, self-optimising multi-network systems
Application-Aware Connectivity and Graceful Degradation Mechanisms	● APIs and development frameworks should be established to allow applications to detect connectivity conditions and adapt behaviour accordingly	● Tighter integration between connectivity management and AI-driven workload orchestration	● Connectivity resilience as an intrinsic property of the Cognitive Computing Continuum
AI-Native Telco Cloud-Edge			
AI-Driven Automation of Telco Operations	● Integrating high-quality, domain-specific data into AI pipelines	● Embed AI into network management and orchestration platforms in a modular and controlled manner	● Self-optimising networks with AI digital twins and auditable control
Domain-Specific and Trustworthy	● Develop telecom-specific model training	● Support open and federated AI model	● Develop AI to support cross-domain

© 2024-2026 NexusForum.EU



8.1 Open Radio Access Networks (Open RAN)

Primary destinations
4. Secure, sovereign, and interoperable European computing capabilities
Secondary destinations
2. An AI stack built on an open European hardware and computing ecosystem
3. A competitive European AI and machine learning ecosystem

Background and driving factors

The Radio Access Network remains one of the most vertically integrated and vendor-constrained parts of telecom infrastructure. Open RAN seeks to decouple hardware and software, introduce open interfaces and enable multi-vendor innovation. This shift is strategically important for Europe, as it determines whether future radio infrastructures remain dependent on a limited number of global suppliers or evolve into a more competitive ecosystem.

However, disaggregation increases system complexity. Multi-vendor integration, third-party RAN Intelligent Controller applications and AI-native optimisation introduce new challenges in security, coordination and accountability. Open RAN must therefore combine openness with governability.

Europe in relation to the state of the art

Europe has strong vendors and deep standardisation engagement. At the same time, global policy momentum around Open RAN is currently driven primarily by coalitions outside the European Union. Initiatives such as the Global Coalition on Telecommunications (GCOT) demonstrate a coordinated strategic positioning that does not yet include EU Member States. Without strategic alignment, there is a risk that “open” architectures become dominated by a few large integrators. Europe can differentiate through secure multi-vendor integration, AI-

© 2024-2026 NexusForum.EU

native optimisation aligned with regulation, and integration of RAN into the broader cloud-edge continuum.

Cross-cutting principles and dependencies

Open RAN must be understood as part of the broader European digital sovereignty agenda. Openness should not be limited to interface specifications but must extend to governance, interoperability and security. AI integration in Open RAN must be aligned with European regulatory principles, including the AI Act and critical infrastructure protection requirements. Furthermore, Open RAN development should be linked to European semiconductor initiatives and cloud-edge orchestration efforts to avoid siloed evolution.

8.1.1 Priority: Secure and Governable Multi-Vendor Open RAN Architectures

Impact

Ensuring security and accountability in multi-vendor RAN environments is essential to prevent fragmentation and vulnerability from undermining the benefits of openness.

Recommendations

Short-term
[Research & Innovation] Define European interoperability and security frameworks for Open RAN deployments, including certification mechanisms for third-party controller applications and secure onboarding procedures.
Medium-term
[Research & Innovation] Develop runtime integrity verification and software supply-chain transparency mechanisms adapted to cloud-native RAN stacks.
Long-term
[Ecosystem] Establish harmonised European governance models for multi-vendor RAN ecosystems, ensuring resilience and compliance in future 6G architectures.

8.1.2 Priority: AI-Native and Cross-Layer Optimisation in Open RAN

Impact

AI-driven optimisation can significantly improve spectral efficiency, energy consumption and adaptability, but requires coordination to prevent conflicts between distributed control applications.

© 2024-2026 NexusForum.EU

Recommendations

Short-term
[Research & Innovation] Develop coordination mechanisms for RIC applications to prevent conflicting optimisation actions. Supervisory coordination layers and arbitration policies are needed to ensure that local optimisations do not undermine global system performance. At the same time, shared European simulation and validation environments should be strengthened to test multi-vendor and multi-application interactions under realistic conditions. Such environments must go beyond isolated component testing and reflect full system behaviour.
Medium-term
[Research & Innovation] Develop energy-aware cross-layer optimisation across RAN, transport and edge. Enable cross-layer optimisation linking RAN, transport and edge compute, including energy-aware optimisation principles.
Long-term
[Research & Innovation] Integrate RAN control fully into AI-native cloud-edge orchestration frameworks within the European Cognitive Computing Continuum.

8.2 Seamless data connectivity and predictive handover across different networks

Primary destinations
4. Secure, sovereign, and interoperable European computing capabilities
Secondary destinations
1. Scalable and energy-efficient AI and data processing

Background and driving factors

An increasing number of industrial and public-sector applications rely on continuous connectivity across large and often remote areas. Robotics, autonomous systems, connected vehicles and drone-based services are becoming central to logistics, energy, agriculture and emergency response, frequently operating where communication infrastructure is sparse or heterogeneous. These systems often require uninterrupted links to human operators or backend control platforms, as edge and cloud infrastructures are essential for AI processing and supervisory control. Connectivity disruptions can therefore result not only in degraded performance, but also in safety risks and regulatory issues.

Seamless handover across heterogeneous networks, including 4G/5G/6G, WiFi and satellite systems, remains imperfect. Short communication blackouts may occur when transitioning

© 2024-2026 NexusForum.EU

between terrestrial and non-terrestrial networks or between public and private domains. While tolerable in consumer scenarios, such interruptions are unacceptable for mission-critical or autonomous applications. As terrestrial and satellite infrastructures become increasingly integrated, resilient and predictive multi-network connectivity becomes a foundational requirement of the Cognitive Computing Continuum and of Europe's digital ecosystem.

Europe in relation to the state of the art

Europe has strong capabilities in telecommunications, satellite communication and industrial IoT, with significant contributions to mobility management, software-defined networking and QoS optimisation. Emerging non-terrestrial network initiatives complement terrestrial 5G and future 6G systems. However, connectivity across heterogeneous networks remains fragmented. Interoperability between terrestrial and satellite systems is limited, and integration across public and private domains lacks common architectural principles. While global actors are advancing integrated connectivity solutions, including large-scale satellite constellations, Europe's coordination remains uneven. Predictive, AI-driven connectivity management is still largely experimental, and integration with edge-cloud orchestration is immature. Europe has the opportunity to define a coherent, sovereign multi-network architecture, but this requires stronger coordination across operators, satellite providers and cloud-edge stakeholders.

Cross-cutting principles and dependencies

Seamless data connectivity is structurally linked to the broader roadmap themes of cloud-edge orchestration and AI operationalisation. Predictive handover mechanisms intersect directly with work on AI-native orchestration, as connectivity decisions increasingly depend on real-time data analytics and cross-domain optimisation. There is also a natural connection to the cybersecurity transversal topic, since multi-network integration expands the attack surface and requires consistent identity and trust management across heterogeneous infrastructures

8.2.1 Priority: Predictive and Resilient Handover Across Heterogeneous Networks

Impact

If seamless and predictive handover mechanisms are not developed, future autonomous and mobile systems will remain vulnerable to connectivity disruptions. This will limit the scalability of robotics, drone operations, connected mobility and distributed industrial systems. It will also weaken Europe's ambition to deploy AI-driven services in remote and critical environments. Conversely, predictive handover mechanisms that anticipate network degradation and proactively switch communication channels can significantly enhance resilience. They can enable continuous service delivery, support safety-critical applications and strengthen Europe's strategic autonomy in connectivity infrastructures.

Recommendations

Short-term

© 2024-2026 NexusForum.EU

[Research & Innovation] Sensing and quality-of-service monitoring across all available physical network channels. This includes developing mechanisms for continuous performance assessment and early detection of degradation across terrestrial and satellite links. Predictive models, potentially AI-based, should be developed to anticipate short-term variations in channel quality and to support proactive handover decisions. At the same time, work is needed on standardised software protocols for fast communication channel switching. These protocols must ensure that handovers between heterogeneous networks occur with minimal packet loss and latency spikes. This requires close collaboration between telecom operators, device manufacturers and cloud-edge platform developers.

Medium-term

[Research & Innovation] Develop integrated hardware–software stacks that seamlessly combine terrestrial and satellite connectivity, enabling devices and edge platforms to maintain multiple channels in parallel and dynamically select optimal paths, including across multi-provider satellite services. Research should also address migration of computational state between edge nodes during connectivity transitions, ensuring continuity of latency-sensitive AI-driven functions. Connectivity management must be more tightly integrated with edge orchestration, so that handover decisions account not only for radio conditions, but also for workload placement, latency and energy considerations across the continuum.

Long-term

[Research & Innovation] Seamless connectivity should evolve towards fully autonomous, self-optimising multi-network systems. These systems would continuously evaluate available communication paths, anticipate failures and reconfigure routing across terrestrial and non-terrestrial networks in real time. Such capabilities will be essential for large-scale deployment of autonomous systems, remote industrial operations and resilient public infrastructures.

8.2.2 Priority: Application–Aware Connectivity and Graceful Degradation Mechanisms

Impact

Even with predictive handover, temporary disruptions or performance degradation may occur. Applications that assume continuous high-bandwidth connectivity risk failure when operating across heterogeneous networks. Without appropriate mechanisms, system reliability and safety may be compromised. By enabling applications to adapt dynamically to connectivity conditions, Europe can ensure robust operation of AI-enabled systems across diverse environments. This will support industrial digitalisation, smart mobility and public-sector services.

Recommendations

© 2024-2026 NexusForum.EU

Short-term

[Research & Innovation] APIs and development frameworks should be established to allow applications to detect connectivity conditions and adapt behaviour accordingly. This includes mechanisms for graceful performance degradation, buffering strategies on edge devices and rapid recovery once connectivity is restored. Research and innovation should focus on application-level resilience patterns that can be standardised across domains.

Medium-term

[Research & Innovation] Tighter integration between connectivity management and AI-driven workload orchestration. Edge nodes should be capable of temporarily assuming additional functionality when cloud connectivity is reduced, and synchronising state efficiently once full connectivity is re-established. This requires coordinated development of networking protocols, edge computing frameworks and AI model deployment strategies

Long-term

[Ecosystem] Connectivity resilience as an intrinsic property of the Cognitive Computing Continuum. Applications, edge platforms and network infrastructures should jointly adapt to varying connectivity conditions, ensuring continuity of critical services without manual intervention. This will enable robust deployment of autonomous systems across large geographical areas and heterogeneous infrastructures.

8.3 AI-Native Telco Cloud-Edge

Primary destinations

3. A competitive European AI and machine learning ecosystem

Secondary destinations

5. Advanced digitalisation and AI adoption in industry and public sectors

Background and driving factors

European telecommunications operators face stagnating revenues alongside high investment needs for 5G evolution, fibre rollout and future 6G infrastructure. At the same time, networks are becoming increasingly software-defined, cloud-integrated and data-intensive, creating both pressure and opportunity for transformation. The concept of an AI-native Telco Cloud reflects the convergence of cloud-native architectures and advanced AI techniques within telecom operations, where AI becomes an integral part of network control, orchestration and optimisation across RAN, edge and core domains. This includes emerging approaches such as AI-RAN, where AI-driven control loops are embedded directly into radio access network

© 2024-2026 NexusForum.EU

functions and coordinated with higher-layer orchestration. AI models are increasingly embedded into operational support systems, network management platforms and service control layers to enable higher levels of automation, adaptive optimisation and real-time decision-making. However, telecom systems require high reliability, regulatory compliance and strong security guarantees. Integrating AI into network operations therefore demands robust data governance, explainability and safeguards against unintended behaviour.

Europe in relation to the state of the art

Europe has strong telecom operators, established network vendors and a growing AI ecosystem. European research contributes significantly to network management, AI explainability and secure cloud architectures. In addition, regulatory frameworks such as the AI Act provide a structured governance environment. Nevertheless, large-scale AI-native telco systems are still emerging, and many operators rely on non-European cloud infrastructures and AI services for advanced analytics and automation. Without coordinated action, this may lead to increased dependency on such vendors in core operational systems. Europe's opportunity lies in combining its telecom engineering expertise with trustworthy, explainable and regulation-aligned AI solutions. By embedding generative AI directly into cloud-native telco stacks under European governance, Europe can shape an AI-enabled telecom ecosystem that remains competitive and sovereign.

Cross-cutting principles and dependencies

AI-native telco systems connect directly to the roadmap themes on AI. Embedding AI into telecom infrastructures requires alignment with broader efforts on operationalising AI, managing distributed and multi-agent systems, and ensuring explainability in critical infrastructures. There is also a strong linkage to cybersecurity and sovereignty, as AI integration reshapes data governance, control over operational intelligence and dependency structures across the continuum.

8.3.1 Priority: AI–Driven Automation of Telco Operations

Impact

AI-native automation can significantly improve operational efficiency, reduce costs and increase agility in service deployment. AI-driven methods can support configuration management, fault diagnosis, anomaly detection and dynamic optimisation of network behaviour across RAN, edge and core domains. If not developed strategically, however, telcos may become dependent on proprietary AI services and lose control over operational intelligence.

Recommendations

Short-term
[Research & Innovation] Integrating high-quality, domain-specific data into AI pipelines. Telecommunications data is heterogeneous, distributed and often noisy. Robust data architectures, preprocessing mechanisms and domain-aware model training are prerequisites for reliable generative AI deployment. Parallel efforts should address

© 2024-2026 NexusForum.EU

explainability and transparency in AI-assisted decision-making. Operational systems require traceable reasoning, especially when decisions affect service continuity or regulatory compliance.

Medium-term

[Research & Innovation] Embed AI into network management and orchestration platforms in a modular and controlled manner, supporting configuration, predictive maintenance and dynamic resource optimisation. Standardised interfaces between AI components and orchestration systems are essential to prevent vendor lock-in, while security and privacy safeguards must protect sensitive data during model training and inference. AI-driven orchestration should also incorporate energy-awareness as an optimisation objective, enabling distributed Telco Edge nodes to shift workloads in space or time to align with renewable energy availability and reduce operational costs and carbon intensity.

Long-term

[Research & Innovation] Self-optimising networks with AI digital twins and auditable control. AI-native telco systems should evolve towards self-optimising network environments where AI models continuously learn from operational data, simulate potential changes and recommend or execute improvements within defined safety boundaries. Such systems must remain auditable, controllable and aligned with European regulatory frameworks. AI-enabled network digital twins can further support predictive optimisation and resilience by enabling safe evaluation of control decisions.

8.3.2 Priority: Domain-Specific and Trustworthy AI Models for Telecom

Impact

Generic foundation models are not sufficient for mission-critical telecom operations. Domain-specific generative models trained on network data and operational contexts can improve performance, reliability and relevance. At the same time, inappropriate or opaque model behaviour may undermine trust and system stability. Developing trustworthy, telecom-specific generative models strengthens Europe's AI capacity and reduces dependency on external providers.

Recommendations

Short-term

[Research & Innovation] Develop telecom-specific model training approaches with shared data schemas that incorporate structured network knowledge, configuration data and performance metrics. Collaboration between operators, vendors and research institutions is essential to define shared data schemas and anonymisation strategies.

© 2024-2026 NexusForum.EU

Medium-term
[Ecosystem] Support open and federated AI model ecosystems tailored to telecom applications, enabling SMEs and research actors to develop specialised components without relying exclusively on non-European platforms. Alignment with open hardware and cloud initiatives can reinforce strategic autonomy. Federated and distributed learning approaches should be advanced as privacy-preserving alternatives to centralised training, with the Telco Edge serving as an aggregation layer that enables cross-sector collaboration without transferring sensitive raw data across borders.
Long-term
[Research & Innovation] Develop AI to support cross-domain reasoning across network, edge and cloud layers, enabling holistic optimisation of the Cognitive Computing Continuum. European leadership in trustworthy AI for critical infrastructure could become a distinguishing global capability.

8.4 Federated Telco Edge and Network-as-a-Service (NaaS)

Primary destinations
4. Secure, sovereign, and interoperable European computing capabilities
Secondary destinations
5. Advanced digitalisation and AI adoption in industry and public sectors

Background and driving factors

As cloud computing evolved, hyperscalers created unified global platforms that abstract infrastructure complexity behind programmable APIs. Developers can deploy services across continents without negotiating separately with infrastructure owners. In contrast, Europe's telecommunications landscape remains fragmented along national and operator boundaries. Even where edge computing capabilities exist within telco networks, they are typically exposed through heterogeneous interfaces and commercial models.

The concept of Network-as-a-Service (NaaS) extends beyond connectivity. It implies that network capabilities, such as guaranteed latency, slicing, edge compute location, and quality-of-service parameters, are exposed programmatically via standardized APIs. In a Cognitive Computing Continuum, the telco edge should not function merely as distributed infrastructure, but as a programmable platform.

However, without federation between operators, the telco edge cannot compete with hyperscale cloud offerings. A developer seeking to deploy a pan-European low-latency application must still integrate separately with multiple operators. This fragmentation undermines the Digital Single Market and limits Europe's ability to scale innovation.

© 2024-2026 NexusForum.EU

A federated Telco Edge, where operators interconnect their edge clouds and offer harmonized interfaces, would enable "compute roaming" analogous to traditional connectivity roaming. Such federation is not only a technical issue but also an economic and governance challenge, requiring orchestration, settlement, and trust mechanisms across domains.

Europe in relation to the state of the art

Europe is strong in telecom standardisation and has active initiatives around open APIs and edge platforms. Global API initiatives and operator alliances demonstrate awareness of the need for harmonization. At the same time, commercial hyperscalers continue to offer unified developer experiences that far exceed the current level of cross-operator coordination in Europe. While infrastructure investments under initiatives such as IPCEI-CIS are strengthening the hardware and cloud layer, the orchestration, brokerage and settlement layers required for dynamic, cross-border edge federation remain immature. Europe's challenge is therefore not only technical implementation, but also ecosystem alignment. If addressed strategically, federated Telco Edge could become a uniquely European model, leveraging existing telecom assets to create a sovereign, distributed alternative to centralized hyperscaler platforms.

Cross-cutting principles and dependencies

This topic bridges Telco Cloud-Edge with efforts to create an interoperable federated cloud-edge market in Europe. It directly connects to Open RAN (multi-vendor openness at the radio layer), to AI-Native Telco Cloud-Edge, and to AI-driven orchestration across federated domains and, as federation depends on strict adherence to common standards and open interfaces. It also intersects with cybersecurity considerations, since cross-operator federation requires zero-trust principles and coordinated incident response mechanisms.

8.4.1 Priority: Pan–European Telco Edge Federation and Programmable Network Capabilities

Impact

A federated Telco Edge would enable developers and enterprises to deploy latency-sensitive and AI-driven applications across Europe through unified APIs, without negotiating separately with each national operator. This would strengthen the Digital Single Market, reduce dependency on non-European cloud intermediaries, and create new business models for operators. Without federation, edge deployments risk remaining siloed, limiting scalability and reinforcing hyperscaler dominance in application-layer innovation.

Recommendations

Short-term
[Ecosystem] Harmonisation of network and edge APIs across European operators should be accelerated. This includes not only connectivity exposure but also standardized discovery of edge compute resources, quality-of-service guarantees and slice characteristics. Alignment with existing open API initiatives is essential to prevent further fragmentation. Parallel research and innovation efforts should explore intent-based

© 2024-2026 NexusForum.EU

orchestration layers that translate high-level application requirements into coordinated configurations across multiple operator domains.

Medium-term

[Research & Innovation] Develop mechanisms for cross-operator edge resource trading and dynamic settlement. This includes orchestration frameworks that support "East–West" interconnection between operator edge clouds and allow workloads to move transparently across national boundaries. Support for neutral interconnection points and edge exchange infrastructures may be required to reduce latency and operational complexity in cross-border scenarios.

Long-term

[Ecosystem] Fully federated European NaaS framework for unified application development. A fully federated European Network-as-a-Service framework should allow application developers to access edge and network capabilities across Member States as if operating within a single programmable environment. Regulatory alignment may be necessary to reduce legal friction for real-time cross-border edge processing, effectively enabling a coherent European computing continuum.

© 2024-2026 NexusForum.EU

9 Federations and Cloud–Edge AI Interconnect Framework

Primary destinations
4. Secure, sovereign, and interoperable European computing capabilities
Secondary destinations
1. Scalable and energy-efficient AI and data processing
2. An AI stack built on an open European hardware and computing ecosystem

Background and driving factors

Certain applications and AI workloads (e.g., an autonomous vehicle's inference task or a drone's video stream) must be able to move seamlessly from an edge node in one country to a node in a neighbouring country without renegotiating compliance or latency-inducing legal checks. There are ongoing initiatives to create a federated cloud-edge computing infrastructure in Europe, such as IPCEI-CIS, SIMPL, and GAIA-X.

Furthermore, managing LLMs in such a federated computing continuum requires an *AI Interconnect Framework* (not to be confused with interconnects in datacentres) that manages the lifecycle of AI models, including selection, placement, task coordination, and routing of prompts and inference results across a federated edge-cloud continuum. Unlike traditional MLOps, LLMOps must handle *foundation models* that are too large for single devices, requiring dynamic provisioning, fine-tuning, and retrieval-augmented generation (RAG) distributed across the network.

In the 6G era, *Learning* becomes a tradable resource in the 3C-L model.[48] In this paradigm, devices should act as intelligent agents that can buy and sell not just data, but gradient updates and inference tasks. This requires a *Market of Resources* where dynamic pricing mechanisms and smart contracts govern the trade of AI capabilities between micro-operators, verticals, and users.

Given these trends, efforts are needed to address cross-border and cross-provider exchange of workloads with seamless migration with a shared market and orchestration framework. Given this shared framework, there is an opportunity to address the unsustainable growth of AI energy consumption by adopting *Frugal AI*: algorithms designed to maximize accuracy per watt, rather than just raw accuracy.

Europe in relation to the state of the art

Currently, the European digital landscape is fragmented; while regulations like the GDPR and AI Act exist, their implementation and national interpretations often create *digital borders*. The

[48] *Peltonen, et al, « 6G white paper on edge intelligence » (2020), https://urn.fi/URN:ISBN:9789526226774*

lack of a genuine single market prevents European operators from achieving the scale necessary to compete with US hyperscalers, who operate on a global, borderless platform.

There are federation ecosystems emerging and maturing throughout Europe that need to be supported and built upon. There are also federation ecosystems outside of Europe, notably in Japan and South Korea, that are adopting European federation frameworks and solutions, including GAIA-X. The EU should seek to strengthen cooperation and integration between relevant EU-led initiatives and initiatives in other countries including Japan and South Korea.

Cross-cutting principle and dependencies

This directly ties with Telco Cloud-Edge federations and Network-as-a-Service (NaaS), as well as the broader topic of portable AI. There are also connections to sustainability and energy-efficiency, as we propose tighter integration between cloud-edge federation marketplaces and energy and sustainability metrics.

Overview of priorities and recommendations

	Short-term	Medium-term	Long-term
Priority groups	Federations and Cross-Border, Cross-Provider Market		
Create a cloud-edge AI Interconnect Framework	• Create a cross-border cloud-edge AI Interconnect Framework • Develop AI agents to ensure cross-border regulatory compliance	• Technical protocols for data packets and AI models to prove their origin and compliance • Establish a sovereign decentralized European repository for edge-optimized Foundation Models	• A binding regulatory framework that harmonizes data processing rules for federated edge workloads
Decentralized cross-provider marketplace	• Create a Decentralized Resource Auction Micro-Architecture • Explore SLA-Aware Pricing • Investigate Frugality-QoS Co-Scheduler to orchestrate edge components	• Standardization of Broker Agents that automatically negotiate price and SLA	• A mandatory energy efficiency rating system for cloud-edge AI services • Algorithmic trading rules for edge resources to prevent market manipulation
	← Timeline →		

© 2024-2026 NexusForum.EU

9.1 Priority: Create a cloud-edge AI Interconnect Framework

Impact

Standardizes the AI supply chain across edge providers in Europe, allowing operators and verticals to deploy GenAI services that are compliant, auditable, and independent of non-sovereign clouds. Taken together, this will help create a true Digital Single Market for AI, allowing European SMEs and Telcos to deploy services that scale across the continent immediately (e.g., "5G Corridors" for mobility). It reduces the compliance tax for cross-border innovation and creates a critical mass of computing resources capable of rivalling US/China.

Recommendations

Short-term

[Research & Innovation] Create a cross-border cloud-edge AI Interconnect Framework. Define how an orchestrator at one national edge node signals its counterpart to accept a live inference session. Introduce *Federated Compliance APIs* as a concrete software artefact alongside the regulatory proposal. Align *Automated Compliance Negotiators* explicitly with ETSI MEC and 3GPP edge application context transfer standards.

[Research & Innovation] Develop AI agents to ensure cross-border regulatory compliance, capable of instantly verifying if a workload complies with the local implementation of the AI Act in a neighbouring node before migrating, automating the *data border control*. Define semantic messaging protocols (Pub/Sub) for the AI Interconnect that carry not just data, but AI metadata (model version, prompt context, confidence scores) to enable interoperability between edge nodes.

Medium-term

[Research & Innovation] Technical protocols for data packets and AI models to prove their origin and compliance, using cryptographically signed *"AI visas"*, allowing them to pass through *digital borders* (firewalls/gateways) without deep packet inspection.

[Ecosystem] Establish a sovereign decentralized European repository for edge-optimized Foundation Models that are pre-validated for safety and energy efficiency, accessible via the AI Interconnect.

Long-term

[Regulation] A binding regulatory framework that harmonizes data processing rules for federated edge workloads, establishing that data processed in a participating EU state is legally equivalent to data processed domestically for industrial applications.

© 2024-2026 NexusForum.EU

9.2 Priority: Decentralized cross-provider marketplace

Impact

A decentralized marketplace for edge services allows smaller players (e.g., a city utility company) to monetize their idle edge compute, creating a decentralized "AI Cloud" that competes with hyperscalers. This is necessary to create a true Digital Single Market for AI, allowing European SMEs and Telcos to deploy services that scale across the continent immediately (e.g., "5G Corridors" for mobility).

Recommendations

Short-term

[Research & Innovation] Create a Decentralized Resource Auction Micro-Architecture for a 5 ms-latency edge resource auction implementation (distributed ledger vs. centralised auctioneer vs. gossip-based clearing). The implementation should address atomicity, e.g., what happens when an edge node commits to an inference task but fails mid-execution. Link to existing Data Space and Gaia-X economic models to avoid duplication.

[Research & Innovation] Explore SLA-Aware Pricing where price is a function of latency, model accuracy, and energy budget. Explore algorithms that determine the fair market value of a model update in Federated Learning based on its contribution to accuracy (Shapley values). Edge-Aware Prompt Routing to select which edge node runs a given prompt based on model availability, load, and user proximity.

[Research & Innovation] Investigate Frugality-QoS Co-Scheduler to orchestrate edge components to dynamically adjust model size/accuracy based on real-time energy budget and SLA requirements. Ground this in actual renewable energy prediction APIs (e.g., selecting nodes in regions with high solar/wind forecast), and connect eco-labelling to the AI Interconnect so that energy metadata is first-class in model routing decisions and allow developers to set an *energy budget* for an AI task, with the system automatically selecting the model size that fits the budget. Connect energy labelling to the AI Interconnect so that energy metadata travels with model artefacts across the continuum.

Medium-term

[Ecosystem] Standardization of *Broker Agents* that automatically negotiate price and SLA for offloading an inference task to a nearby edge node.

Long-term

[Regulation] A mandatory energy efficiency rating system for cloud-edge AI services, driving market demand for Frugal AI solutions. Develop a standard rating system (A to G) for AI models based on their Joules per Inference efficiency, driving market adoption of Green AI.

[Regulation] Algorithmic trading rules for edge resources to prevent market manipulation in automated high-frequency trading of compute/connectivity resources.

© 2024-2026 NexusForum.EU

NEXUS FORUM.EU

PART B

PILLAR III:

An AI-enabling Hardware-Software Stack

www.eucloudedgeiot.eu

10 European Semiconductor Design

Primary destinations
2. An AI stack built on an open European hardware and computing ecosystem
4. Secure, sovereign, and interoperable European computing capabilities
Secondary destinations
6. European leadership in disruptive and emerging computing paradigms

Background and driving factors

RISC-V, as an open specification, enables Europe to develop processors and domain-specific accelerators (including AI) benefitting from local and global ecosystems but without royalties or intellectual property dependencies. Together with trends such as chiplets, that increase yield and lower manufacturing costs, this lowers the barrier of entry to the microprocessor chip market for European companies and can drive demand for the manufacturing capabilities pursued in the European Chips Act.

Today's Compute Continuum is dominated by Intel's x86 in cloud/HPC and by ARM's 32- and 64-bit Instruction Set Architectures in edge and embedded domains, and most tools and techniques used in these domains are specialised for this architecture. ARM has recently made the push towards cloud and HPC which has resulted in significant efforts to port applications such as container managers, hypervisors, AI frameworks, as well as OS drivers for hardware. For broad adaption RISC-V also needs to address these software aspects as outlined in previous roadmaps and identified in the RISC-V community through collaborations such as RISE[49]. Recent results point to a performance gap to established non-European proprietary instruction set architectures[50] which must be closed to enable wide-spread adoption of RISC-V. Furthermore, the openness of RISC-V leads to faster innovation cycles which opens the question of how to manage and abstract increasing diversity of RISC-V computing hardware from end-users (and expose when necessary).[51]

Current AI hardware (GPUs) faces the "Memory Wall" and thermal limits. The "6G LLM" and "Edge Intelligence" white papers argue that for 6G, we need Processing-In-Memory (PIM) and Neuromorphic (brain-inspired) architectures to run event-driven AI at the extreme edge with micro-watt power consumption. Mapping standard Transformer models (LLMs) to these non-Von Neumann architectures requires a radical Hardware-Software Co-design approach to develop new compilers and sparsity-aware algorithms.

Europe in relation to the state of the art

[49] *RISE: RISC-V Software Ecosystem, Linux Foundation, https://riseproject.dev/*

[50] *Francesco Lumpp, et al, "On the Containerization and Orchestration of RISC-V architectures for Edge-Cloud computing," ESAAM'23, 2023. Robert Balas, et al, "CV32RT: Enabling Fast Interrupt and Context Switching for RISC-V Microcontrollers," IEEE Transactions on VLSI Systems, 2024. Andrea Bartolini, et al, "Monte Cimone: Paving the Road for the First Generation of RISC-V High-Performance Computers," SOCC'22, 2022.*

[51] *https://digital-strategy.ec.europa.eu/en/library/recommendations-and-roadmap-european-sovereignty-open-source-hardware-software-and-risc-v*

The European efforts, spanning the Compute Continuum, expose a strong European commitment to a joint European ecosystem based on RISC-V. Over the past decade and a half there has been significant investment in implementing European processors targeting both HPC and embedded and IoT applications (e.g., KDT/Chips JU, EuroHPC JU). The EU goal is the production of cutting-edge and sustainable semiconductors in Europe, with at least 20% of world production in value by 2030, including manufacturing capacities below 5nm nodes, aiming at 2nm, with an aim to improve energy efficiency by a factor 10. To achieve this the RISC-V ISA plays a central role in EUs strategy.

The efforts rely on retained momentum in ongoing RISC-V hardware efforts, including for AI, supported by EuroHPC JU and Chips JU to drive demand and generate necessary spillover effects in European RISC-V hardware. Industry sectors across the Compute Continuum may have other needs than e.g., HPC, automotive, and space.

European processor designs are available and even leading in some markets, e.g., telecommunications and automotive, but faces hard competition from other developed economies, like US and China, especially with emerging open standard processor architectures like RISC-V. Ongoing European efforts in selected industries serve as key locomotives in this strategy, but there is additional potential for the European RISC-V ecosystem in making it adoptable by a larger share of the economy.

Neuromorphic computing research in Europe is of the highest caliber (e.g., SpiNNaker, Human Brain Project). However, they work in a fundamentally different way from “standard” deep learning models that are broadly used. Standard AI developers used to PyTorch/TensorFlow can’t use these odd processors because there is a gap in the toolchain and expertise in how to work with these new AI architectures.

Europe has a strong position in the production of machinery necessary for semiconductor and silicon manufacturing (i.e., ASML in the Netherlands) but relies on manufacturing and packaging in other parts of the world to produce broader market silicon products. South Korea and Japan are also world-leading in some areas of semiconductor chip manufacturing, especially memory and certain materials and chemicals necessary in the production.

© 2024-2026 NexusForum.EU

Overview of priorities and recommendations

		Short-term	Medium-term	Long-term
	European Semiconductor Design			
Priority groups	Develop competitive RISC-V processors and systems for European and global markets	Support development of RISC-V processors and AI accelerators within the EU Adopters and potential adopters of RISC-V processors to engage in the RISC-V ecosystem	The availability of trained engineers and researchers with RISC-V expertise and the capability of integrating new RISC-V products with existing infrastructure and devices is needed	European RISC-V hardware and software ecosystem can be extended to encompass the entire cognitive Compute Continuum Widespread implementation of Trusted Execution Environments (TEEs) specifically optimized for distributed LLM inference
	Ensure availability of necessary Software Stacks	Develop technology-neutral and efficient AI execution on different platforms Expertise in AI compilers, e.g. from Python to native, to facilitate cross-platform by design Efficient resource allocation for AI workloads in the HPC integrated continuum	Open-source sparse computation libraries (like a European alternative to CUDA) Standardized machine learning and AI pipelines for heterogeneous compute infrastructure	Ability to maintain commodity software that underlies the European computing ecosystem
		← Timeline →		

10.1 Priority: Develop competitive RISC-V processors and systems for European and global markets

Develop competitive RISC-V processors in Europe. This includes high-performance out-of-order processors for high-performance systems, all the way to niche market custom processors at the far edge. This does not only include the promotion of the design and manufacturing of advanced processors in the EU but also ensuring market demand for European processors during the critical scale-up phase.

Impact

The RISC-V ecosystem is currently experiencing tremendous growth, driven by the capacity to use the adaptable instruction set architecture to fill new market niches, lessen vendor lock-in, and establish new industries in new regions of the world, including the EU, to bolster digital sovereignty. Promoting the development of processors in Europe for Europe both ensures availability of strategically important components, but also the opening of completely new markets in which Europe today is underrepresented globally. Europe consumes approximately

© 2024-2026 NexusForum.EU

1/5 of all semiconductors globally but produces only 1/10,[52] indicating that there is a significant opportunity to increase European market share.

Recommendations

Short-term
[Research & Innovation] Support development of RISC-V processors and AI accelerators within the EU, including organizations developing the processors themselves (e.g., financial or coordination support) but may in the short term include stimulating demand for EU processors through e.g., public procurement or other available mechanisms, to reduce the risk of moving from established processor vendors. **[Ecosystem]** Adopters and potential adopters of RISC-V processors to engage in the RISC-V ecosystem, including participation in the Special Interest Groups of RISC-V International to drive the standardization efforts of the ISA to meet the needs of European existing and emerging industry.
Medium-term
[Support] The availability of trained engineers and researchers with RISC-V expertise and the capability of integrating new RISC-V products with existing infrastructure and devices is needed. This includes training the next generation of engineers, but also the ability for engineers already part of the work force to be trained.
Long-term
[Ecosystem] European RISC-V hardware and software ecosystem can be extended to encompass the entire cognitive Compute Continuum, extending the European ecosystem and leveraging technology spillover effects from R&D in HPC (EuroHPC JU), automotive, and space (Chips JU), and strengthening the ability to adopt RISC-V across more European industrial sectors. Key use cases from cognitive computing continuum industry sectors, and their supply chains, are identified and used to drive additional software and hardware research and development needs for the European ecosystem. **[Support]** Widespread implementation of Trusted Execution Environments (TEEs) specifically optimized for distributed LLM inference, ensuring model weights and user inputs are encrypted during processing.

10.2 Priority: Ensure availability of necessary Software Stacks

Ensure the necessary software stacks, such as operating systems, compilers, and runtime libraries are available to meet customer demands to integrate them into their products. Software tools and technologies to handle challenges and opportunities with fast-innovation

[52] https://digital-strategy.ec.europa.eu/en/factpages/chips-act

© 2024-2026 NexusForum.EU

open instruction set architectures should be made available to abstract and manage increase in hardware diversity. Additional development is required to port software tools and technologies relevant to the Compute Continuum to RISC-V and close the performance gap to legacy architectures.

If there is to be a shift towards EU-made RISC-V processors, accelerators, interconnects, etc., it is unlikely to occur all at once, and there will be a period during which this transition will need to be supported. This provides a challenge as expertise is needed in both RISC-V and already established domains. In the longer term it is highly likely that the compute continuum will consist of high levels of heterogeneity both within individual systems, and especially across the continuum, to exploit specialized hardware to perform computation with highest efficiency (e.g., time to solution, energy usage, etc.).

This roadmap does not address Electronic Design Automation (EDA) tools, necessary for Hardware Design, and represented in the architectural figure as "Hardware Design Tooling", as it is out of scope of this Roadmap. For the interested reader we refer to existing efforts, e.g., the Recommendations and Roadmap for Open-Source EDA in Europe produced by the FOSSi foundation and the GoIT project at the request of the Chips Joint Undertaking[53].

Impact

The software stack is the main enabler for application builders to utilize any hardware developed or marketed in the previous sections of this chapter.

Recommendations

Short-term

[Research & Innovation] Develop technology-neutral and efficient AI execution on different platforms. Software-hardware co-design and solutions for optimising the AI stack as new hardware architectures become available.

Research & Innovation Expertise in AI compilers, e.g. from Python to native, to facilitate cross-platform by design. This includes increased contributions to existing open-source compilers by EU stakeholders but may also include the development of novel AI compiler components.

[Research & Innovation] Efficient resource allocation for AI workloads in the HPC integrated continuum: Develop strategies for dynamic resource allocation based on workload requirements to optimize resource usage in the Compute Continuum for AI.

Medium-term

[Ecosystem] Open-source sparse computation libraries (like a European alternative to CUDA) that natively support unstructured sparsity on edge CPUs/NPUs.

[Ecosystem] Standardized machine learning and AI pipelines for heterogeneous compute infrastructure, that automatically compile, quantize, and deploy a new model version to a heterogeneous fleet of devices if performance drops.

[53] *FOSSi Foundation, Roadmap and Recommendations for Open-Source EDA in Europe, https://fossi-foundation.org/resources/eu-roadmap*

© 2024-2026 NexusForum.EU

Long-term

[Ecosystem] Ability to maintain commodity software that underlies the European computing ecosystem. The development of open-source tools for broad industry adoption. After reaching initial maturity and adoption successful open-source projects require institutions for governance, conflict resolution, and maintenance.

© 2024-2026 NexusForum.EU

11 AI inference hardware: AI Accelerators, Memory, and Interconnects

Primary destinations
1. Scalable and energy-efficient AI and data processing
3. A competitive European AI and machine learning ecosystem
Secondary destinations
2. An AI stack built on an open European hardware and computing ecosystem
6. European leadership in disruptive and emerging computing paradigms

Background and driving factors

The inference stage of large language models and agentic AI is mainly memory-bound, rather than compute-bound, which opens up a new market opportunity for AI inference hardware that is not served by state-of-the-art GPUs. The reason is that a major portion of inference for large language models (the decode stage) can't be parallelized; and for agentic AI the reason is that inference involves orchestrating storage, CPU, memory, and AI accelerators. Thus, LLMs and agentic AI require new architectures, covering both hardware and software, to solve this challenge.

Current large language models work by generating tokens (such as words) one-by-one, starting from a given context (the prompt). That is, it runs the prompt through an LLM, which generates a new word which is appended to the prompt, and the process is repeated with the new prompt (previous + one new word) over and over again until the process stops. This means that LLM inference can be considered as two stages: the pre-fill stage, and the decode stage.

© 2024-2026 NexusForum.EU

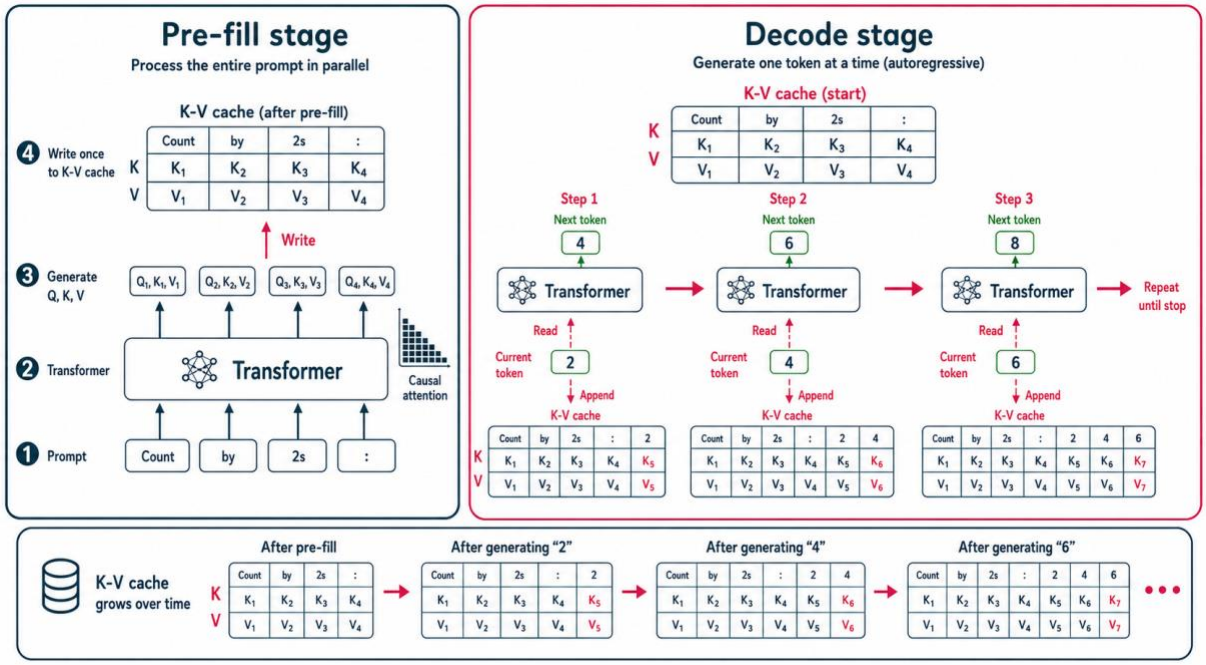


Figure 4: Illustration of the prefill and decode stages in LLM inference

The prefill stage is when the original prompt is processed. At this stage, all the words in the prompt are known at the start, and can be processed in parallel. GPUs and accelerators with high compute-throughput excel in parallel computations. That is, the prefill-stage is a compute-bound operation.

The decode stage is the iterative process of generating "one more token" (or word). At this stage, the previously known words don't need to be fully re-processed again: the K-V pairs for the previously known tokens, that are expensive to compute, have already been computed and stored in a so-called K-V cache. Therefore, only the K-V pair of the new token needs to be recomputed, and stored in the K-V cache. That is, computations during the decode stage are sequential rather than parallel. At each iteration, the previously computed K-V pairs must be fetched from memory, causing idle time for the processor. Therefore, GPUs and traditional AI accelerators are not ideal for the decode stage. Instead, the decode stage is memory-bound, and requires advances in memory-compute architectures.

Europe in relation to the state of the art

The global state of the art in AI inference hardware is currently dominated by non-European GPU and accelerator ecosystems, especially mature CUDA-based hardware and software stacks. This dominance is reinforced not only by chip performance, but also by optimised kernels, model-serving frameworks, developer tools, cloud availability and large developer communities. For Europe, this creates a dependency risk across both hardware and software.

At the same time, the memory-bound nature of LLM decoding creates an opportunity. While the prefill stage benefits from high-throughput accelerators, the decode stage is limited by memory bandwidth, K-V cache movement, interconnect efficiency and energy consumption. This opens space for new architectures beyond conventional GPU scaling, including processing-near-memory, processing-in-memory, chiplets, high-bandwidth interconnects, memory-aware accelerators, quantised inference and hardware-software co-design.

© 2024-2026 NexusForum.EU

Europe has relevant strengths in HPC, embedded and industrial systems, telecommunications, automotive, low-power electronics, RISC-V, open hardware, compiler technologies and safety-critical engineering. These strengths can support differentiated AI inference systems for data centres, edge, mobile, industrial, automotive and robotics use cases.

The key challenge is to turn these strengths into usable platforms. European AI inference hardware will need mature software stacks, compiler and runtime support, model-serving frameworks, benchmarks, developer tools, application-porting programmes and early demand through AI Factories, EuroHPC, industrial pilots and public procurement. Europe's opportunity is therefore not simply to replicate existing GPU ecosystems, but to build open, energy-efficient, memory-aware and software-accessible AI inference systems aligned with European needs for sustainability, interoperability and technological autonomy.

Overview of priorities and recommendations

Priority groups	Short-term	Medium-term	Long-term
AI inference hardware: AI Accelerators, Memory, and Interconnects			
Processing and memory architectures for AI inference chips	Develop processing and memory architectures optimized for LLM and agentic AI inference Develop high-bandwidth, energy-efficient, and reconfigurable interconnects for AI systems Develop chiplet-based AI inference architectures Link European AI accelerator and processing-in-memory development to existing chip initiatives, open software stacks and ecosystems	Advance processing-near-memory and processing-in-memory for AI inference Create European AI hardware benchmarking suites Develop open compiler stacks for processing-in-memory and AI accelerator architectures Support application-porting and developer access programmes	Scale up post-von-Neumann AI inference architectures Develop integrated European AI inference platforms Develop adaptive hardware-software systems for heterogeneous AI inference Establish open European interfaces for AI accelerator integration
Develop optical/photonics AI inference chips	Make optical computing a core pillar of the European chip strategy	Develop a European strategy for supporting photonics/optical processing technologies	
	← Timeline →		

© 2024-2026 NexusForum.EU

11.1.1 Priority: Processing and memory architectures for AI inference chips

With the move towards more AI-enabled systems the importance of energy-efficient and performant systems to host these workloads more focus moves from the main system processors to specialized accelerators, on-chip and off-chip data interfaces. At the same time, there is high global demand for memory solutions to store weights and other model parameters for AI workloads, leading to significant increases in prices and global bottlenecks in supply chains. For Europe to be successful in its adoption of a European RISC-V AI ecosystem these bottlenecks should be addressed, for both the inference and training phases.

Impact

With the development of AI accelerators, the efficiency of AI inference can be significantly increased. Instead of running complex control-flow oriented pipelines in general purpose processors, specialized accelerators can be tailored to meet the data-flow oriented pipelines of modern AI. Today's main data-parallel accelerator for AI are GPUs, but there are several other promising software-programmable acceleration directions being marketed by e.g., Sambanova, Semidynamics, Axelera AI, and many others. In addition to this, models are being deployed directly in hardware, without the need for software implementations, examples of this include the Taalas HC1 which places a Llama model directly in silicon, promising orders of magnitude faster AI inference than what can be achieved by software implementations.

At the same time, increasing the amount of processing that can be done per unit of time moves bottleneck of AI systems to data interconnects, on- and off-chip networks, and memory.[54] AI inference is already from the get-go a significantly data-bound operation and is already today hitting a memory wall. The most performant systems move to High Bandwidth Memory, but it still faces significant cost barriers and shortages, while traditional DRAM cannot keep pace with the increased demands. Addressing these shortcomings would allow increased performance scaling in AI accelerators, without starving them from data and thus lowering their utilization.

Lastly, with increased performance in both processing (acceleration) and memory, the final bottleneck is the interconnects (on and off chip) that are tasked with moving data from storage to the point where it is processed. This includes both high-speed on-chip interconnects and networks on chip (NoC), as well as specialized high-bandwidth off-chip interconnects and networking. Alleviating this bottleneck is one of the next enablers for more advanced AI models, and the capability to deploy them.

Recommendations

Short-term
[Research & Innovation] Develop processing and memory architectures optimized for LLM and agentic AI inference. This includes chip-level, node-level and cluster-level architectures. Focus on strong hardware-software co-design and ecosystems. Cluster- and node-level architectures can initially focus on integrating existing chip technologies for high-

[54] https://arxiv.org/pdf/2601.05047 *Xiaoyu Ma and David Patterson (Google), « Challenges and Research Directions for Large Language Model Inference Hardware », 2026*

© 2024-2026 NexusForum.EU

performance data centres, while chip-level design should be considered a longer-term investment. Current needs include efficient support for attention mechanisms, key-value cache management, sparse inference, quantised models, retrieval-augmented generation, tool-use loops and multi-agent execution patterns.

[Research & Innovation] Develop high-bandwidth, energy-efficient, and reconfigurable interconnects for AI systems. Support research on on-chip networks, chiplet interconnects, accelerator-to-memory interfaces, accelerator-to-accelerator communication, and node-level interconnects for AI inference clusters. Initially focus on higher-level interconnects between existing chip technologies, and treat chip-level as longer term. Support emerging developments on reconfigurable photonic interconnects.

[Research & Innovation] Develop chiplet-based AI inference architectures. Explore modular chiplet architectures combining compute, memory, interconnect and specialised acceleration. This can help European actors innovate in parts of the AI hardware stack without needing to own every element of advanced semiconductor manufacturing.

[Support, Ecosystem]: Link European AI accelerator and processing-in-memory development to existing chip initiatives, open software stacks and ecosystems. Ensure that these investments address not only peak performance, but also energy efficiency, programmability, software maturity and deployability. Ensure that European accelerator and memory architectures are supported by compiler, runtime and framework integration from the start. This should include, where appropriate, LLVM/MLIR-based flows, vLLM and SGLang support, RISC-V extensions where appropriate, open model formats, quantisation toolchains and integration with widely used AI frameworks.

Medium-term

[Research & Innovation] Advance processing-near-memory and processing-in-memory for AI inference. Support scaling up of R&D in architectures that reduce data movement by bringing computation closer to memory. Prioritise use cases where memory bandwidth and energy consumption dominate inference cost, which currently include LLM inference, embedding search, recommendation, graph-based retrieval and edge AI.

[Ecosystem] Create European AI hardware benchmarking suites. Develop benchmarks that measure AI inference performance, memory efficiency, energy consumption, latency, cost, programmability and software maturity across different hardware targets. Benchmarks should include LLM inference, multimodal inference, edge AI, agentic workflows, and industrial AI workloads.

[Research & Innovation; Support]: Develop open compiler stacks for processing-in-memory and AI accelerator architectures. Develop open-standard compiler and runtime stacks that allow European hardware vendors to support mainstream AI frameworks without proprietary lock-in. These stacks should support model partitioning, memory placement, quantisation, sparsity, scheduling and hardware-specific optimisation while exposing portable abstractions to developers.

[Ecosystem]: Support application-porting and developer access programmes. Fund programmes that help researchers, SMEs and industrial users port AI workloads to European accelerators, RISC-V systems, processing-in-memory prototypes and edge AI

© 2024-2026 NexusForum.EU

platforms. This should include developer kits, reference systems, cloud-accessible testbeds and documentation.

Long-term

[Research & Innovation; Support]: Scale up post-von-Neumann AI inference architectures. Support scaling up of alternatives to GPU-dominated AI inference, such as processing-near-memory computing, neuromorphic accelerators, dataflow architectures, photonic interconnects, hybrid-quantum approaches, and processing-in-memory.

[Research & Innovation; Ecosystem]: Develop integrated European AI inference platforms. Develop complete AI inference platforms combining European processors, accelerators, memory technologies, interconnects, compiler stacks, runtimes and developer tools. These platforms should be usable across data centre, edge, industrial and embedded environments.

[Research & Innovation] Develop adaptive hardware-software systems for heterogeneous AI inference. Create systems that can automatically select, partition and optimise AI workloads across heterogeneous compute and memory resources according to latency, energy, cost, privacy, reliability and accuracy requirements.

[Standardisation; Ecosystem] Establish open European interfaces for AI accelerator integration. Develop and promote open interfaces for accelerator discovery, programming, runtime integration, memory management, telemetry and performance portability. This should allow European hardware components to be integrated into wider AI systems without dependence on closed proprietary ecosystems.

11.1.2 Priority: Develop optical/photonics AI inference chips

Photonics/optical computing is an emerging processing paradigm that works by manipulating light to perform computations. Combining photonics-based processing with photonics-based communication could reduce for example latency or energy use, compared to traditional processors.

Impact

Optical processing is a highly promising highly energy-efficient technology, and there are opportunities to integrate optical processing with photonics-based networking technologies to reduce energy use and latency. For example, the Japanese telecom giant NTT is investing in their vision for the “All-Photonics Network”.[55]

Both Europe and Japan have strong research and industrial foundations in photonics, and there are companies developing optical computing technologies in both regions. As a leading region in the foundational technologies, the EU should make it a top priority to include and support optical/photonics processing pathways in the European chip strategy.

[55] *https://iowngf.org/the-all-photonics-network-enables-the-next-generation-digital-economy/*

© 2024-2026 NexusForum.EU

Recommendations

Short-term

[Ecosystem & support] Make optical computing a core pillar of the European chip strategy. Photonics/optical processing (or computing) is an area of emerging importance, that could eliminate some of the bottlenecks of current processing technologies, including data transfer and high energy use/heat rejection. Prototypes have recently demonstrated architectures for AI inference at lab scale. There is a significant opportunity for Europe to build on its strong foundation in the core technologies and establish leadership in this emerging technology.

Medium-term

[Ecosystem & Support] Develop a European strategy for supporting photonics/optical processing technologies. It is critical to create a concrete strategy for how these technologies can be developed, manufactured, and adopted, both from a market demand and supply chain perspective. In particular, it is necessary to better understand potential use cases of these technologies and trade-offs with other technologies as they are scaled up.

© 2024-2026 NexusForum.EU

12 Emerging Processor Architectures

12.1 Neuromorphic systems

Primary destinations
1. European leadership in disruptive and emerging computing paradigms
Secondary destinations
2. Scalable and energy-efficient AI and data processing
4. Secure, sovereign, and interoperable European computing capabilities

Background and driving factors

Traditional digital computing is increasingly challenged by the performance, energy, and latency requirements of modern AI and edge workloads, as well as by the optimisation and simulation of complex systems (e.g., chemistry and pharmaceuticals, process optimisation in production lines, and traffic and freight transport).

Neuromorphic computing seeks to emulate aspects of the brain's self-organising and self-learning behaviour and introduces an event-driven computing paradigm that can improve power–performance efficiency for such workloads. Despite its potential, neuromorphic hardware has not yet been widely adopted in commercial AI data centres, in part due to limited opportunities for application-oriented testing and rapid transition into prototypes and small series, and due to the lack of standardised neuron models and common training techniques.

Because neuromorphic systems differ from the commonly used von Neumann architecture, existing software tools and solutions are generally not compatible, and the event-based nature of these systems creates new requirements for the computing continuum in terms of communication and computing patterns.

If integrated successfully, neuromorphic hardware at the edge and in data centres could improve the sustainability and energy efficiency of AI processing and enable new classes of always-on sensing applications. In parallel, advancing neuromorphic software is important to realise application-level benefits (not only hardware efficiency) and may also create synergies with neuroscience as brain-like principles are recreated in electronics. Increased hardware–software co-design and novel interfaces between neuromorphic and digital computing domains are therefore needed to enable seamless integration across edge and datacentre environments, including considerations such as edge constraints and GDPR compliance.

Europe in relation to the state of the art

Europe has strong assets across the neuromorphic value chain, spanning EU-funded system development, unique research infrastructure, emerging industrial offerings, and HPC integration pathways. For example, the roadmap highlights HYBRAIN as an EU-funded effort toward a brain-inspired computing system, and notes growing momentum for neuromorphic

© 2024-2026 NexusForum.EU

capabilities across the continuum. A distinctive European differentiator is BrainScaleS-2, a spiking, accelerated mixed-signal (analog/digital) neuromorphic platform where neuron and synapse dynamics are realised as a physical emulation in hardware (largely analog circuits) with digital periphery and event routing, rather than being purely computed as software on standard processors. It is operated at the European Institute for Neuromorphic Computing and is accessible online for users through the EBRAINS infrastructure.

On the industrial side, European companies are targeting sensor-edge / consumer-device integration with neuromorphic processors and development kits aimed at ultra-low-power inference close to sensors (e.g., SynSense; Innatera), including deployed audio and vision inference use cases.

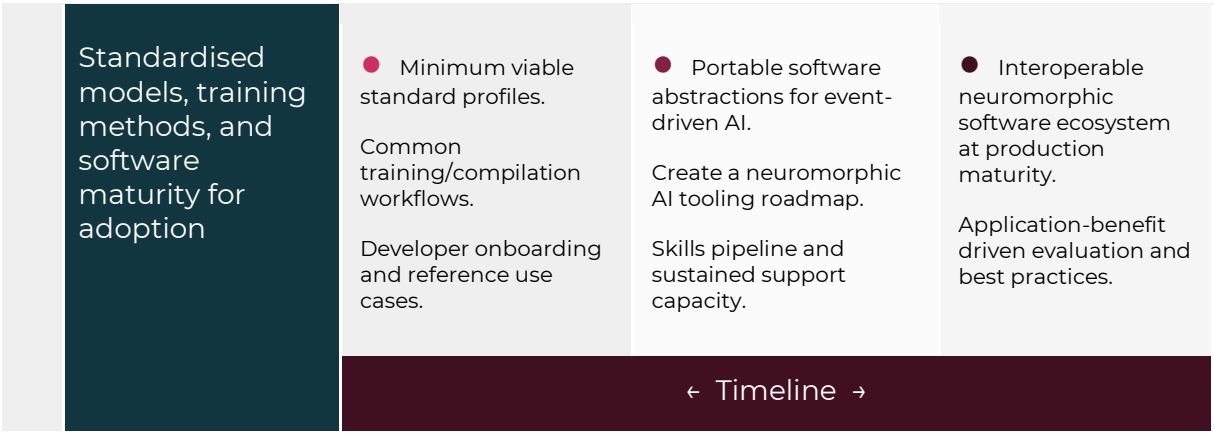
In the HPC domain, the first EU exascale supercomputer JUPITER will include a neuromorphic module, indicating a European pathway toward modular supercomputing integration that explicitly includes neuromorphic capabilities. Other communities, such as ESA, are also exploring neuromorphic-oriented calls, suggesting potential synergies beyond the core cloud-edge-HPC ecosystem.

Taken together, Europe's opportunity is to convert these assets into scalable uptake by focusing on hardware–software co-design, application-oriented test and evaluation pathways, and interoperability with mainstream software stacks and continuum orchestration, which the roadmap identifies as key blockers to broader adoption.

Overview of priorities and recommendations

Priority groups	Short-term	Medium-term	Long-term
Neuromorphic systems			
Hardware–software co-design and interoperability for continuum integration	● Neuromorphic–digital interface specifications. Application-oriented integration pilots (edge + datacentre) Continuum requirements for event-driven computing Spiking Neural Network (SNN) Converters.	● Address the neuromorphic-to-cloud interface with hybrid orchestration patterns and reference architectures. ● Interoperability test suites and benchmarks. Standard integration stack components.	● Production-grade neuromorphic services in the continuum. Neuromorphic Edge Pilot Lines.

© 2024-2026 NexusForum.EU



12.1.1 Priority: Hardware–software co–design and interoperability for continuum integration

Develop and integrate neuromorphic accelerators as an important resource in the cognitive computing continuum. Enable neuromorphic capabilities to be used as part of real edge–cloud–HPC systems by creating workable interfaces between neuromorphic and digital domains. This accelerates integration into data centres and edge deployments, improves energy efficiency/sustainability, and reduces fragmentation caused by incompatible architectures and tooling.

Impact

Neuromorphic computing makes ultra-efficient, event-driven processing of sensor data at the edge possible, enabling always-on inference and new applications (e.g., wearables, smart home, IoT). Neuromorphic hardware is a pathway to more sustainable and energy-efficient AI processing at the edge and datacentres. There are additional spillover benefits to neuroscience through hardware emulation of neural principles.

Recommendations

Short-term
[Research & Innovation + Ecosystem] Neuromorphic–digital interface specifications. Define practical interface contracts between neuromorphic modules and conventional compute (data formats for events, synchronisation, control APIs, memory/data movement patterns). This reduces “one-off” integrations and enables repeatable hybrid pipelines.
[Support + Ecosystem] Application-oriented integration pilots (edge + datacentre). Fund pilots that integrate neuromorphic hardware into (i) sensor-adjacent edge devices and (ii) data-centre/HPC environments, with explicit benchmarks for power, latency, and accuracy. Include a pathway from prototype to small series.
[Research & Innovation] Continuum requirements for event-driven computing. Document how event-driven communication/computation changes continuum needs (networking patterns, orchestration triggers, monitoring/observability).

© 2024-2026 NexusForum.EU

[Research & Innovation] Spiking Neural Network (SNN) Converters. Develop automated tools that convert pre-trained Deep Neural Networks (DNNs) into Spiking Neural Networks (SNNs) compatible with European neuromorphic hardware.

Medium-term

[Research & Innovation + Ecosystem] Address the neuromorphic-to-cloud interface with hybrid orchestration patterns and reference architectures. Develop repeatable "blueprints" for splitting workloads across neuromorphic + conventional compute (e.g., event-driven edge pre-processing feeding conventional analytics/training), including scheduling and resource management patterns. Define how spike-based outputs from neuromorphic edge sensors are converted for further processing by conventional AI, for example token/vector representations for consumption by transformer-based cloud models.

[Ecosystem] Interoperability test suites and benchmarks. Establish shared benchmark suites for always-on sensing and edge pipelines that allow comparison across platforms and integration approaches (including workload-level energy/latency metrics).

[Ecosystem + Research & Innovation] Standard integration stack components. Develop reusable middleware/adapters (drivers, runtimes, monitoring hooks) so neuromorphic devices can plug into mainstream deployment toolchains.

Long-term

[Ecosystem + Regulations + Support] Production-grade neuromorphic services in the continuum. Mature operational readiness: lifecycle tooling (deployment, updates, observability), security hardening, certification pathways where relevant, and multi-vendor portability to support sustained deployments at scale.

[Support] Neuromorphic Edge Pilot Lines. Funding for pilot manufacturing lines of commercial-grade neuromorphic edge accelerators to reduce reliance on imported GPUs.

12.1.2 Priority: Standardised models, training methods, and software maturity for adoption

Position Europe as a frontrunner in the adoption of neuromorphic technologies, and develop adoption ecosystems for European neuromorphic SMEs and scale-ups enabling application-oriented testing, prototyping, and transition to small-series deployments.

Impact

Reduces barriers that prevent neuromorphic hardware from moving beyond niche use: lack of standard neuron models, lack of common training techniques, and incompatibility with existing tools. Shifts the value case from hardware novelty to reliable application benefit.

Recommendations

Short-term

© 2024-2026 NexusForum.EU

[Ecosystem] Minimum viable standard profiles. Define agreed "baseline" neuron/synapse model profiles and operating conventions that multiple platforms can support, enabling repeatable experimentation and reducing fragmentation.

[Research & Innovation + Support] Common training/compilation workflows. Establish a starter set of training and deployment workflows that developers can reuse across use cases (especially always-on sensing).

[Support] Developer onboarding and reference use cases. Provide reference implementations (audio/vision always-on sensing) and documentation that demonstrate end-to-end development and deployment.

Medium-term

[Research & Innovation + Ecosystem] Portable software abstractions for event-driven AI. Create portable representations/abstractions that reduce rewriting effort when moving between neuromorphic platforms and allow smoother integration with existing software stacks.

[Ecosystem] Create a neuromorphic AI tooling roadmap. Close gaps vs mainstream AI tooling (profiling, debugging, monitoring, CI/CD integration) so neuromorphic development becomes operationally tractable.

[Support + Ecosystem] Skills pipeline and sustained support capacity. Build durable European capacity (training curricula, support organisations, community governance) so neuromorphic development and deployment can scale beyond specialist teams.

Long-term

[Ecosystem + Research & Innovation] Interoperable neuromorphic software ecosystem at production maturity. Establish widely adopted standards and de-facto portable tooling so applications can be developed, trained/compiled, validated, and deployed across multiple neuromorphic platforms with minimal re-engineering. This includes stable runtimes, versioning, verification/validation practices, and operational tooling comparable to mainstream AI stacks.

[Ecosystem + Support] Application-benefit driven evaluation and best practices. Institutionalise practices that demonstrate and quantify task-level benefits (accuracy/latency/energy/reliability) across representative use cases, shifting the narrative from hardware novelty to repeatable application value.

12.2 Hybrid quantum and classical computing fusion

Primary destinations

1. European leadership in disruptive and emerging computing paradigms

© 2024-2026 NexusForum.EU

Secondary destinations
2. Scalable and energy-efficient AI and data processing
4. Secure, sovereign, and interoperable European computing capabilities

Background and driving factors

The fusion of quantum and classical computing is driven by the complementary strengths of the two paradigms. Classical computing relies on binary bits and struggles with certain problem classes involving exponential variables, while quantum computing uses qubits that can represent a richer set of possibilities and can be more efficient for specific problem types. In practice, the most promising near-term algorithms are hybrid, combining quantum and classical computation to leverage strengths of both.

A major constraint is that quantum computers require specialised operating environments and cryogenic temperatures, making them impractical for personal devices. As a result, widespread access is expected to come via cloud delivery: quantum calculations are performed in data centres and consumed through cloud services. This enables users to benefit from quantum capability without local quantum hardware. Over time, convergence in cloud environments is expected to increase effective computational capability and allow software to choose dynamically between classical, AI, and quantum resources within a single workflow.

Europe in relation to the state of the art

Europe's opportunity is to translate quantum infrastructure investments into usable hybrid services by ensuring that quantum computing is accessible through cloud and HPC delivery models and integrated into practical workflows. The differentiator will be operational fusion: making hybrid algorithms deployable with reliable orchestration, predictable service models, and a user experience that does not require deep quantum expertise. As the hybrid paradigm becomes the main route to value, Europe can position itself by building robust cloud/HPC integration and tooling that enables "quantum where it helps" inside broader computational pipelines.

© 2024-2026 NexusForum.EU

Overview of priorities and recommendations

	Short-term	Medium-term	Long-term
Priority groups	**Hybrid quantum and classical computing fusion**		
Hybrid algorithm workflows and orchestration for quantum–classical fusion	● Develop hybrid-quantum architectures and workflow reference patterns, including for AI workloads. Develop tooling for hybrid execution and orchestration.	● Automated orchestration and optimization of hybrid workflows. Performance and efficiency evaluation for hybrid workflows.	● Production-grade orchestration for hybrid services.
Cloud delivery model for quantum-enabled computing services	● Create a QC-as-a-service integration into cloud platforms. User abstraction and developer experience.	● Service model maturity for hybrid quantum computing. Cloud-side integration with classical/AI resources.	● Fully integrated quantum–classical–AI cloud services. Industrial-grade service operation at scale.
	← Timeline →		

12.2.1 Priority: Hybrid algorithm workflows and orchestration for quantum–classical fusion

Impact

Makes quantum computing practically useful by enabling repeatable hybrid workflows where quantum and classical steps are combined efficiently, increasing performance for suitable problem classes while keeping complexity manageable.

Recommendations

Short-term
[Research & Innovation + Ecosystem] Develop hybrid-quantum architectures and workflow reference patterns, including for AI workloads. Define and share canonical hybrid architectures and workflow patterns that specify how quantum and classical components interact (iteration structure, data exchange, convergence criteria).

© 2024-2026 NexusForum.EU

[Research & Innovation + Support] Develop tooling for hybrid execution and orchestration. Provide basic tooling support to run hybrid algorithms across quantum backends and classical compute, focusing on usability and reproducibility.

Medium-term

[Research & Innovation] Automated orchestration and optimization of hybrid workflows. Develop orchestration logic that can decide which paradigm (classical or quantum) to use for different sub-tasks based on efficiency/performance considerations and manage execution across resources.

[Ecosystem] Performance and efficiency evaluation for hybrid workflows. Establish evaluation practices that quantify the end-to-end benefit of hybrid approaches (not just component-level metrics).

Long-term

[Ecosystem + Support] Production-grade orchestration for hybrid services. Mature orchestration, monitoring, reliability, and cost/efficiency control so hybrid quantum services can be operated at scale in cloud/HPC environments.

12.2.2 Priority: Cloud delivery model for quantum-enabled computing services

Impact

Enables broad access to quantum computing by delivering it through cloud services, making quantum capability available without local specialised hardware and lowering barriers for users and organisations.

Recommendations

Short-term

[Ecosystem + Support] QC-as-a-service integration into cloud platforms. Ensure that quantum execution is consumable through cloud service interfaces with clear onboarding and usage patterns.

[Support] User abstraction and developer experience. Provide abstractions that let users benefit from hybrid computing without managing quantum-specific operational details.

Medium-term

[Ecosystem + Support] Service model maturity for hybrid quantum computing. Develop predictable service models (access, reliability expectations, documentation, support) that make hybrid quantum computing usable beyond expert users.

© 2024-2026 NexusForum.EU

[Research & Innovation + Ecosystem] Cloud-side integration with classical/AI resources. Strengthen integration so hybrid workflows can compose cloud AI/classical compute with quantum backends seamlessly.

Long-term

[Ecosystem + Research & Innovation] Fully integrated quantum–classical–AI cloud services. Offer mature cloud services where hybrid workflows can seamlessly compose classical, AI, and quantum components, and where the platform can select and orchestrate the most efficient paradigm for each sub-task without exposing complexity to end users.

[Ecosystem + Support + Regulations] Industrial-grade service operation at scale. Establish service reliability, monitoring/observability, security controls, and predictable cost/usage management so quantum-enabled cloud services can be used routinely in production contexts.

12.3 Integration of quantum computing infrastructure

Primary destinations

1. European leadership in disruptive and emerging computing paradigms

Secondary destinations

2. Scalable and energy-efficient AI and data processing

4. Secure, sovereign, and interoperable European computing capabilities

Background and driving factors

Quantum computing is emerging as a transformative technology with the potential to affect sectors such as cryptography, materials science, and complex system simulation, and it is attracting substantial global investment in scalable hardware and efficient quantum algorithms. A central direction globally is to integrate quantum computing with established HPC and cloud infrastructures so that quantum resources can be used as part of larger computing workflows rather than as isolated systems.

In Europe, the strategic objective is to build a robust quantum computing ecosystem that leverages Europe's strengths in HPC and supports innovation across Member States. This includes development across multiple quantum hardware modalities (including superconducting qubits, trapped ions, and photonics) and ecosystem-building actions such as Quantum Excellence Centres to grow quantum programming facilities, application libraries, and a skilled workforce.

A key driver for infrastructure integration is that current QC capacity and integration maturity are still limited: the lack of robust infrastructure, spanning both hardware and software,

© 2024-2026 NexusForum.EU

makes it difficult to fully leverage QC capabilities or to combine them efficiently with HPC and cloud services, including for demanding workloads. The core challenges include scaling quantum hardware, developing efficient algorithms, and building the operational and engineering foundations required for practical QC–HPC–cloud usage (testbeds, workflow integration, benchmarking, and user support).

Looking ahead, the transition from today's prototype systems toward fault-tolerant quantum computing reinforces the need to treat QC as a full-stack, system-level scaling problem. Progress will depend not only on improving component performance, but also on system integration aspects such as measurement and feedback, decoding and classical processing layers, control and cryogenic/electronics stacks, and overall reliability at scale. This makes coordinated infrastructure planning, shared test and evaluation capabilities, and tight hardware–software co-design central drivers for Europe's quantum infrastructure roadmap.

Europe in relation to the state of the art

Europe is pursuing a diversified approach across major quantum hardware modalities (including superconducting qubits, trapped ions, and photonics) and has launched major ecosystem and infrastructure initiatives to translate research capability into usable systems. These include the Quantum Flagship (including QUCATS – Quantum Flagship Coordination and Support Action), OpenSuperQ – An Open Superconducting Quantum Computer (A Quantum Computer for Europe), and hybrid/integration efforts such as HPCQS – High Performance Computer and Quantum Simulator hybrid, alongside EuroHPC Joint Undertaking (EuroHPC JU) deployments hosting European quantum computers. Europe is also strengthening the software and skills layer through structures such as European Quantum Excellence Centres (QECs), and coordination/industrial platforms such as the European Quantum Industry Consortium (QuIC), as well as wider strategic initiatives including the Quantum Internet Alliance (QIA) and the European Quantum Communication Infrastructure (EuroQCI) Initiative.

At the same time, Europe's limiting factor remains infrastructure maturity and integration depth: capacity is still limited and fragmented, and it remains difficult to operate quantum computing as a reliable service integrated with HPC and cloud workflows at scale. The transition from today's prototype (Noisy Intermediate-Scale Quantum, NISQ) systems toward fault-tolerant quantum computing (FTQC) raises the bar further and requires explicit planning for error-correction readiness, control and cryogenic/electronics stacks, and operational reliability in QC–HPC–cloud service environments. A relevant policy development is that the EU Quantum Act is under preparation and scheduled for adoption in 2026.

Cross-cutting principles and dependencies

Treat quantum computing (especially fault-tolerant quantum computing) as a system-level, full-stack endeavour: progress must be measured with shared metrics and comparable benchmarks, not isolated component claims. Maintain tight hardware–software co-design anchored in platform roadmaps, while building federated access and operational maturity so quantum computing becomes a real continuum capability. Use demand-side pull (pilots → procurement pathways) to enable industrial scaling and reduce dependency risks, while keeping SME/scale-up participation practical through access, support, and predictable stage gates.

© 2024-2026 NexusForum.EU

Overview of priorities and recommendations

	Short-term	Medium-term	Long-term
	Integration of quantum computing infrastructure		
Priority groups: Federated QC–HPC–cloud infrastructure and service readiness	● Federated access and user support across European QC sites. Reference QC–HPC–cloud integration stack. Baseline hybrid benchmarking and evaluation. SME/scale-up access pathway.	● Scaled testbeds with operational tooling. Application enablement via libraries and competence centres. Workflow interoperability conventions.	● Production-grade QC as a service. Pan-European federation and resource brokering.
Transition to fault-tolerant quantum computing through full-stack scale-up and procurement pull	● Define an FTQC "Grand Challenge" with staged gates. Independent Test & Evaluation capability. Platform-specific full-stack roadmaps.	● Fund engineering bottlenecks to scale. Utility pilots with a route to procurement. Capital crowd-in at phase gates.	● Procurement and deployment of EU-built FTQC systems. European system-integration capacity.
	← Timeline →		

12.3.1 Priority: Federated QC–HPC–cloud infrastructure and service readiness

Impact

Turns quantum computing from an isolated capability into a usable infrastructure layer integrated with HPC and cloud: repeatable hybrid workflows, reliable access, comparable evaluation, and faster translation of pilots into operational services.

Recommendations

Short-term
[Ecosystem + Support] Federated access and user support across European QC sites. Harmonise onboarding, access models, documentation, and helpdesk/user support so QC can be consumed similarly to HPC services.

© 2024-2026 NexusForum.EU

[Research & Innovation + Ecosystem] Reference QC–HPC–cloud integration stack. Deliver baseline integration patterns for job submission/scheduling, data movement, identity and access, and security boundaries for hybrid workflows.

[Ecosystem] Baseline hybrid benchmarking and evaluation. Establish a starter set of benchmarks and reporting practices for QC–HPC–cloud workflows (time-to-solution, quality, cost; workload-relevant metrics).
[Support + Ecosystem] SME/scale-up access pathway. Provide low-friction access (vouchers, sandbox time, integration coaching) so quantum SMEs can validate use cases on real infrastructure and mature prototypes.

Medium-term

[Ecosystem + Support] Scaled testbeds with operational tooling. Expand into shared testbeds with monitoring/observability, reproducibility practices, accounting/cost controls, and well-defined service interfaces.

[Support + Ecosystem] Application enablement via libraries and competence centres. Build and maintain reusable application libraries, reference workflows, and training programmes that directly support uptake.

[Ecosystem] Workflow interoperability conventions. Define conventions for workflow metadata, provenance, and portability so hybrid pipelines can move across sites with less re-engineering.

Long-term

[Ecosystem + Support + Regulations] Production-grade QC as a service. Mature to dependable services (service levels, security posture, long-term support, lifecycle management), suitable for sustained industrial use.

[Research & Innovation + Ecosystem] Pan-European federation and resource brokering. Enable brokering/scheduling across multiple QC resources and sites as an integrated European capability.

12.3.2 Priority: Transition to fault-tolerant quantum computing through full-stack scale-up and procurement pull

Impact

Accelerates Europe's progression from prototypes to fault-tolerant quantum computing (FTQC) by treating it as a full-stack, system-integration and industrialisation challenge, and by creating credible demand-side pull and scale-building mechanisms.

Recommendations

Short-term

© 2024-2026 NexusForum.EU

[Ecosystem + Regulations] Define an FTQC "Grand Challenge" with staged gates. Set system-level milestones and down-selection logic (logical-level progress, integration readiness, manufacturability signals, operational readiness), with clear governance and decision accountability.

[Ecosystem + Support] Independent Test & Evaluation capability. Establish a credible T&E function that defines common metrics, validates claims, and runs comparable benchmarks on shared test infrastructure (EuroHPC and national testbeds).

[Research & Innovation] Platform-specific full-stack roadmaps. Require hardware-aware plans spanning QPU, control/electronics (incl. cryo where relevant), error correction/decoding, compilers/middleware, and HPC/cloud integration.

Medium-term

[Research & Innovation] Fund engineering bottlenecks to scale. Target scalable I/O and control stacks, high-bandwidth readout/data movement, real-time calibration, decoding/classical processing layers, packaging, and manufacturability.

[Support + Regulations + Procurement] Utility pilots with a route to procurement. Run milestone-driven pilots on EuroHPC and EU cloud pathways to demonstrate credible utility and integration readiness, and to create first-buyer visibility.

[Ecosystem + Support] Capital crowd-in at phase gates. Use staged funding and procurement visibility to attract private investment and support scale-up (especially for SMEs/scale-ups), while encouraging consolidation over time rather than fragmentation.

Long-term

[Procurement + Regulations + Ecosystem] Procurement and deployment of EU-built FTQC systems. Execute deployment once maturity thresholds are met, with sustained operational capability and protected critical supply chain choke points.

[Ecosystem + Support] European system-integration capacity. Build durable capability to integrate full-stack systems (hardware + control + software + operations) so Europe is not only a component supplier but also a system provider.

© 2024-2026 NexusForum.EU

NEXUS FORUM.EU

PART B

PILLAR IV:

AI-Tooling for Intelligent Infrastructure Management and Developer Productivity

www.eucloudedgeiot.eu

13 Federation and system-level optimisation for the Computing Continuum

Primary destinations
4. Secure, sovereign, and interoperable European computing capabilities
Secondary destinations
1. Scalable and energy-efficient AI and data processing

Background and driving factors

The adoption of cloud-edge computing and AI services is increasing rapidly, driving up the energy demand and carbon footprint of Europe's digital infrastructure. Achieving climate goals requires the computing continuum to become not only more energy-efficient, but also **carbon-aware**. That is, able to shift and optimise workloads based on the availability of renewable energy and, over time, interact more directly with smart grids and broader energy systems. Since data centers are not running at full capacity at all times, there is potential to define an energy flexibility model for data centers with smart-grid-aware workload schedulers.

Improving sustainability outcomes also requires **holistic infrastructure management** that spans both hardware and software. That is, moving beyond static operating practices and simple facility-level efficiency metrics toward optimisation that reflects heterogeneous workloads and the full end-to-end system. This includes better sensing and actuation, more advanced cooling approaches, and the use of models (including digital twins) to optimise cooling and workload placement jointly. Today, many data centers in Europe are not equipped with the proper sensing and actuation installations to achieve this.

At the same time, performance and cost optimisation in the continuum becomes a **system-level coordination problem**. In multi-provider settings, optimisation is distributed: each entity has partial visibility and limited control, and optimising individual layers in isolation can lead to poor global outcomes. Layering and standardised interfaces support interoperability, but **cross-layer optimisation** can improve efficiency and performance, while creating new design trade-offs and integration challenges across network, data, compute, and facility layers.

Finally, there is a potential evolution toward a more **federated and even hyper-decentralised continuum**, with many independent nodes and new providers contributing resources. This aligns with Europe's ambition to reduce dependency on large centralised providers and enable a more competitive market, but it challenges today's approaches to service discovery, trust, lifecycle management, data transfer, and scalability, especially if the market becomes highly fragmented or includes very large numbers of small and privately operated edge nodes.

Europe in relation to the state of the art

In multi-provider and federated settings, Europe needs holistic optimisation approaches that work across organisational boundaries, not only within a single provider domain.

© 2024-2026 NexusForum.EU

Europe has strong research capability and many projects on orchestration/scheduling, but a persistent gap is operationalisation: deploying, validating, and fine-tuning methods using real operational data and reference use cases at scale. Many previous EU projects have developed AI-enabled orchestration tools and AI-optimization solutions for the computing continuum. While this has led to many theoretical advances, the lack of operational continuum testbeds and true computing continuum applications are limiting further advances. Further advances require new operational datasets and testbeds to evaluate and further develop these AI-optimization and orchestration solutions.

Cross-cutting principles and dependencies

Computing continuum testbeds and true computing continuum applications are required to further advance AI-native management, orchestration, and optimization solutions, beyond theoretical advances.

Overview of priorities and recommendations

Priority groups	Short-term	Medium-term	Long-term
Federation and system-level optimisation for the Computing Continuum			
Continuum and cross-provider optimisation	● Cross-provider coordination mechanisms. Set up and operate large-scale testbeds simulating real-life continuum systems operation and continuum-native applications.	● Standard services and APIs.	● Digital shadows and digital twins of the continuum.
Federated and decentralised continuum orchestration, trust, and market mechanisms	● Decentralised service discovery. Fault tolerance and resilience for decentralised scenarios.	● Peer-to-peer orchestration and lifecycle management. Data sovereignty, security, and trust mechanisms.	● Business models for small operators.
	← Timeline →		

13.1 Priority: Continuum and cross-provider optimisation

Impact

Improves resource utilisation and reduces energy use across the continuum by coordinating optimisation across facility, infrastructure, orchestration/workflow, and application/service layers. This will help reduce the gap with hyperscaler full-stack advantages and enable a more competitive EU alternative. At the same time, it addresses the expanded attack surface created by diverse APIs and integration points.

Recommendations

Short-term

- **[Research & Innovation] Cross-provider coordination mechanisms.** Develop protocols/mechanisms to coordinate optimisation across (1) cooling + server utilisation, (2) meta-orchestration/workflow execution, and (3) application/service levels, coordinating network, data, and compute.
- **[Support] Set up and operate large-scale testbeds simulating real-life continuum systems operation and continuum-native applications.** Offer projects to test and evaluate their orchestration and optimization solutions using the testbed and continuum-native applications.

Medium-term

- **[Ecosystem; Support] Standard services and APIs**. Establish "standard" services/APIs to facilitate optimisation across the continuum and support service discovery and interoperability.

Long-term

- **[Research & Innovation] Digital shadows and digital twins of the continuum.** Develop continuum-level digital shadows/twins focusing on data flows and integrations needed to optimise the management of network, data, and compute resources.

13.2 Priority: Federated and decentralised continuum orchestration, trust, and market mechanisms

Impact

Enables a more competitive and open cloud-edge market with greater resilience and autonomy, including new players and smaller operators contributing resources (potentially leveraging local renewables), while addressing the challenges of discovery, trust, lifecycle management, and scalability in highly decentralised settings.

© 2024-2026 NexusForum.EU

Recommendations

Short-term

[Research & Innovation] Decentralised service discovery. Develop service discovery mechanisms and explore suitable payment methods for highly dynamic multi-provider scenarios.

[Research & Innovation] Fault tolerance and resilience for decentralised scenarios. Advanced fault detection, self-healing, and redundancy strategies tailored to decentralised environments to ensure high availability and reliability.

Medium-term

[Research & Innovation] Peer-to-peer orchestration and lifecycle management. Develop robust and scalable decentralised orchestration techniques for managing workloads, data flows, and resource allocation and application lifecycles.

[Research & Innovation] Data sovereignty, security, and trust mechanisms. Advance privacy-preserving computing and secure multiparty computation; explore trustworthy certification of program execution and proof of execution in decentralised environments.

Long-term

[Ecosystem; Support] Business models for small operators. Develop mechanisms for smaller operators to offer available compute resources when available (e.g., pay-per-use).

© 2024-2026 NexusForum.EU

14 AI-native management and application development for a heterogeneous computing continuum

Primary destinations
4. Secure, sovereign, and interoperable European computing capabilities
Secondary destinations
3. A competitive European AI and machine learning ecosystem
5. Advanced digitalisation and AI adoption in industry and public sectors

Background and driving factors

Systems in the computing continuum are becoming increasingly complex and distributed, raising the barrier to developing, deploying, and operating applications and increasing the risk of failures and bugs. For deployments that extend to **multi-provider** configurations, heterogeneity and coordination costs are further amplified. To make the continuum easy to use, we need **developer, deployment, and operations abstractions** that reduce cognitive load and provide consistent lifecycle support across cloud-edge environments.

From a **developer experience (DevX)** perspective, the continuum remains hard to program due to heterogeneous runtimes, targets, and constraints. **Developer platforms** are therefore essential to provide user-friendly toolchains, reusable building blocks, and integrated workflows that make development and deployment repeatable. **Agentic AI** can further improve DevX by translating high-level intent (requirements and constraints) into concrete artefacts (i.e., scripts, configuration, and deployment workflows), and refining them using runtime feedback and incident learnings.

An obstacle for operations teams is the gap between human operator insight and system telemetry: it is difficult to extract actionable information from logs and metrics. Tooling is needed to collect lifecycle telemetry across the continuum and to synthesize and interpret it for debugging, root-cause analysis, and user-approved remediation. AI can lower the operational burden by generating operational scripts and temporary monitoring dashboards from natural language (and releasing resources when no longer needed), enabling “zero-touch” deployments under parameterised constraints, and composing specialised agents into end-to-end operations applications. A related need is improved security and operations center (SOC) visibility and AI-assisted log analysis to extend threat monitoring capabilities and reduce analyst workload.

Finally, **safety-critical applications** (autonomous transport, healthcare monitoring, industrial automation, smart grids, emergency response) must remain reliable and safe as resources become more distributed. Distributed sensors, strict safety requirements, and energy/cooling constraints combined with high-performance AI components increase the need for orchestration and operational tooling that supports real-time considerations, fault tolerance, and secure execution, while addressing privacy.

© 2024-2026 NexusForum.EU

Europe in relation to the state of the art

European cloud alternatives have still not reached the level of ease-of-use of US hyperscalers. On the other hand, there are no software engineering, application development, and management tools for a highly heterogeneous computing continuum. This is an opportunity to develop European offerings and market for AI-native tooling for the European computing ecosystem.

At the same time, key enabling capabilities remain immature at continuum scale: (i) AI-assisted O&M that can synthesize distributed telemetry and propose remediations; (ii) SOC visibility across cloud-edge aligned to real-world threat techniques; and (iii) state management and synchronization across multi-provider distributed cloud-edge environments ensuring consistency and integrity (not available as a technical solution today).

Cross-cutting principles and dependencies

AI-assisted code generation and DevOps tooling are excellent use cases for agentic AI. Such software engineering tools are also required to lower the barriers for developers to use the computing continuum or cloud-edge federations.

Computing continuum testbeds and true computing continuum applications are required to further advance AI-native management, orchestration, and optimization solutions, beyond theoretical advances.

Overview of priorities and recommendations

	Short-term	Medium-term	Long-term
AI-native management and application development for a heterogeneous computing continuum			
Priority groups: “Zero-touch” deployment and lifecycle automation	Natural-language to operational scripts	Constraint-based “zero-touch” deployment	End-to-end lifecycle automation with composable AI agents
Priority groups: AI-native operations maintenance and incident response for continuum systems	AI for debugging and root-cause analysis Distributed telemetry synthesis with AI-integrated operator query interface Security Operations Centre (SOC) visibility mapping to real-world threat scenarios	Human-in-the-loop remediation and temporary observability tooling Operationalisation of AI-assisted SOC log analysis and triage	Composable specialised agents for end-to-end operations workflows
	← Timeline →		

© 2024-2026 NexusForum.EU

		Short-term	Medium-term	Long-term
Priority groups	"Zero-touch" deployment and lifecycle automation	Natural-language to operational scripts	Constraint-based "zero-touch" deployment	End-to-end lifecycle automation with composable AI agents
	AI-native operations maintenance and incident response for continuum systems	AI for debugging and root-cause analysis Distributed telemetry synthesis with AI-integrated operator query interface Security Operations Centre (SOC) visibility mapping to real-world threat scenarios	Human-in-the-loop remediation and temporary observability tooling Operationalisation of AI-assisted SOC log analysis and triage	Composable specialised agents for end-to-end operations workflows
		← Timeline →		

14.1 Priority: "Zero-touch" deployment and lifecycle automation

Impact

Makes the Computing Continuum easier to use by reducing expert-only operational steps and enabling repeatable deployment and lifecycle management under parameterised constraints (e.g., latency, resource budgets, and non-functional requirements). Such "zero touch" interfaces are arguable one of the main competitive advantages of hyperscaler offerings. For Europe, this is a strategic lever for **digital sovereignty**: interoperable "zero-touch" automation that works across heterogeneous cloud–edge environments and providers can **reduce dependency on single-vendor control planes**, improve portability, and limit lock-in compared to hyperscaler offerings that are typically optimised for and coupled to one provider ecosystem. This also creates a key integration point for **agentic AI** as a European capability layer: agents can translate intent into deployable artefacts (scripts/configuration/policies) and orchestrate lifecycle workflows in a controlled and auditable manner across providers and domains.

Recommendations

Short-term

© 2024-2026 NexusForum.EU

- **[Research & Innovation] Natural-language to operational scripts**. Use AI to translate human-language intent into operational scripts and configuration artefacts to support deployment and operational tasks, with human-in-the-loop approval.

Medium-term

- **[Research & Innovation] Constraint-based "zero-touch" deployment.** AI agents interpret natural-language requirements and execute deployments that satisfy parameterised constraints, aiming for an optimal solution given the constraints.

Long-term

- **[Research & Innovation] End-to-end lifecycle automation with composable AI agents.** Develop agents that learn from deployments and operational outcomes and can be **composed into end-to-end lifecycle workflows** spanning deployment, monitoring, and remediation.

14.2 Priority: AI-native operations maintenance and incident response for continuum systems

Impact

Reduces downtime and operational cost and improves reliability and security of continuum services by closing the gap between operator insight and system telemetry. Enables faster debugging and root-cause analysis and supports user-approved remediation through AI synthesis and interpretation of distributed logs/metrics across cloud–edge environments.

Recommendations

Short-term

- **[Research & Innovation] AI for debugging and root-cause analysis.** Apply LLMs and multimodal models to extract actionable insights from semi-structured operational data (e.g., logs and traces).
- **[Research & Innovation] Distributed telemetry synthesis with AI-integrated operator query interface**. Develop distributed pipelines that collect, process, and synthesize large volumes of logs/metrics/traces across the continuum and expose an AI-assisted interface for interactive investigation and insight extraction.
- **[Research & Innovation] Security Operations Centre (SOC) visibility mapping to real-world threat scenarios**. Assess SOC visibility for different real-world TTPs by identifying relevant cloud/edge data sources, using AI to classify what can be extracted and its completeness, and mapping gaps to established frameworks (e.g., MITRE ATT&CK).

Medium-term

© 2024-2026 NexusForum.EU

- **[Research & Innovation] Human-in-the-loop remediation and temporary observability tooling.** Automatically interpret failure contexts, propose remediation actions for approval, and generate temporary/ephemeral monitoring dashboards from natural-language descriptions (with resources released when no longer needed).
- **[Research & Innovation; Ecosystem] Operationalisation of AI-assisted SOC log analysis and triage.** Support SOC teams with AI-assisted log analysis to reduce workload and improve the speed and quality of decisions.

Long-term

- **[Research & Innovation; Ecosystem] Composable specialised agents for end-to-end operations workflows.** Develop specialised agents that learn from incidents and can be composed into end-to-end workflows spanning investigation, remediation execution (with guardrails/approval), and integration with operational tooling.

© 2024-2026 NexusForum.EU

NEXUS FORUM.EU

PART B

PILLAR V:

Sectoral Adoption, Support Structures, Testbeds & Benchmarks

www.eucloudedgeiot.eu

15 Convergence of Operational Technologies and Information Technologies

Primary destinations
5. Advanced digitalisation and AI adoption in industry and public sectors
Secondary destinations
3. A competitive European AI and machine learning ecosystem
4. Secure, sovereign, and interoperable European computing capabilities

Background and driving factors

Operational Technology (OT), such as PLCs and SCADA systems, are widely used for automation and monitoring applications in many industry sectors. These have traditionally used industry-specific hardware components, programming languages and toolchains, and communication protocols, such as Modbus or CAN bus. Each vendor often has its own implementations, leading to vendor lock-in effects, and difficulties upgrading legacy systems as modern technologies become available.

The trend among many vendors in recent years is to offer cloud platforms, connecting their monitoring and automation systems to offer data-driven cloud services and applications. Many have even developed platforms to allow third party vendors to create services and applications that can be integrated with their cloud platforms.

IoT technology and contemporary communication protocols have become more widely used as a result of this development. In addition to the traditional vendors, many SMEs have entered the market to compete with them, providing gateways to integrate legacy systems, offering IoT solutions that include hardware and analytics platforms in the cloud, and other cloud services that can be deployed on top of existing data infrastructure and platforms.

There may be a compelling case for the creation of an on-premises OT edge paradigm that divides functions between edge and cloud based on operational constraints, even though some OT vendors have shifted their trend toward the cloud. For example, many critical system functions must be maintained even if connectivity to the cloud is lost. Moreover, there are security concerns in use cases with highly distributed automated systems, which is the case for example for building owners that own many buildings spread out geographically.

In the future, the expectation is to see further convergence between OT and IT, as adoption of commodity hardware in OT increases, and monitoring and automation systems are further integrated with cloud platforms. Hence, traditional automation systems incorporating modern IoT and cloud technologies could be seen as an emergence of a new edge paradigm.

© 2024-2026 NexusForum.EU

The convergence of OT and IT technologies should therefore not take the route of moving everything to a central cloud provider but rather seek to combine the strengths of local processing with the peak power of cloud to perform aggregated analytics and advanced AI tasks.

Europe in relation to the state of the art

Europe has a strong industrial base in automation, manufacturing, energy systems, mobility, buildings, logistics and critical infrastructure. This gives Europe an important starting point for OT and IT convergence: many of the environments where edge AI, digital twins, industrial IoT and autonomous operations will be deployed are already European industrial strengths. Europe also has established expertise in safety-critical systems, industrial control, embedded systems, telecommunications, cybersecurity, robotics and standards-based engineering.

At the same time, the state of the art remains fragmented. Industrial OT systems often rely on long-lived proprietary platforms, vendor-specific toolchains, closed communication stacks and legacy protocols. Modern IT systems, by contrast, are increasingly based on cloud-native software, APIs, containerisation, DevOps, observability and AI-enabled analytics. Bridging these worlds is difficult because OT environments have different requirements from cloud IT environments: deterministic behaviour, high availability, safety certification, offline operation, long equipment lifecycles and strict change-control processes.

European industry is also exposed to a strategic risk. If OT and IT convergence is driven mainly through proprietary cloud platforms offered by incumbent automation vendors or global hyperscalers, industrial users may become locked into closed ecosystems for their operational data, analytics, AI services and digital twins. This could limit interoperability, raise switching costs, and make it harder for European SMEs and technology providers to offer specialised services on top of industrial data.

Europe’s opportunity is therefore to develop an open, secure and interoperable OT and IT edge paradigm. This should combine local control and resilience with cloud, HPC or AI Factory capabilities where they add value. Local OT edge systems should support time-critical control, safety, privacy-sensitive processing and offline continuity. Cloud and HPC resources should support aggregated analytics, model training, digital-twin simulation, fleet optimisation and large-scale AI services. The key challenge is to define architectures, interfaces, data models, security mechanisms and lifecycle tools that allow these layers to work together without forcing all industrial functions into a centralised cloud model.

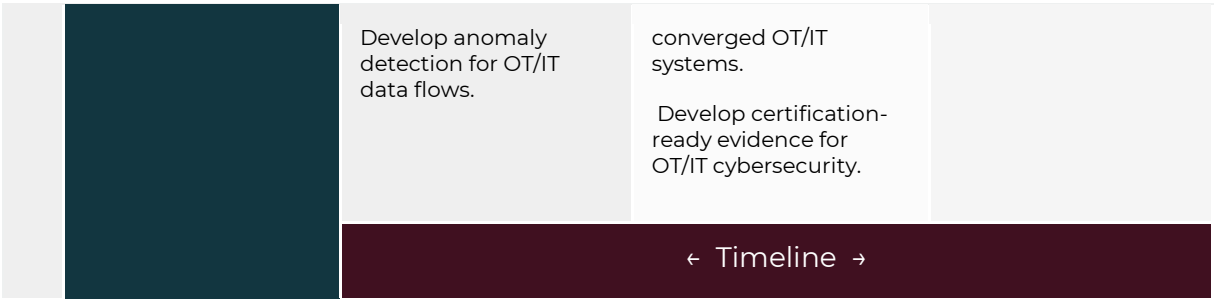
This presents a strategic opportunity for Europe to differentiate itself. Rather than adopting a generic 'cloud-first' model, Europe can pioneer trusted industrial edge architectures tailored to its rigorous standards for cybersecurity, digital sovereignty, and energy efficiency. By leveraging its deep domain expertise, Europe can transform key sectors such as manufacturing, mobility, and logistics into global benchmarks for resilient infrastructure.

© 2024-2026 NexusForum.EU

Overview of priorities and recommendations

	Short-term	Medium-term	Long-term
Priority groups	**Convergence of Operational Technologies and Information Technologies**		
On-premises OT edge and resilient industrial AI	Develop reference architectures for OT-safe industrial edge systems. Develop open industrial edge gateways for legacy OT systems. Develop local AI inference for industrial monitoring and anomaly detection. Establish testbeds for offline-first OT edge operation.	Develop certifiable AI components for OT environments. Develop lifecycle management for long-lived OT/IT systems.	Develop self-optimising industrial edge systems with safety constraints. Establish European reference platforms for OT edge AI.
Secure OT/IT data integration and industrial data spaces	Develop semantic mappings for OT systems. Develop secure data extraction from legacy OT environments. Develop edge-based data quality and provenance tools.	Integrate industrial edge systems with data spaces. Develop common semantic models for industrial and building data. Develop privacy- and confidentiality-preserving analytics for industrial data.	Enable cross-site industrial intelligence. Adopt trusted industrial data-sharing frameworks.
Industrial digital twins and AI-enabled operational optimisation	Develop edge-connected digital twin architectures. Combine AI models with engineering and physics-based models. Develop digital twin validation methods.	Develop real-time and near-real-time digital twins. Develop AI-assisted optimisation for industrial systems. Establish digital twin testbeds for European sectors.	Develop self-updating digital twins. Establish interoperable digital twin ecosystems.
Cybersecurity and lifecycle management for converged OT/IT systems	Develop security patterns for OT/IT edge architectures. Develop vulnerability and asset management for long-lived OT systems.	Develop secure update and rollback mechanisms for industrial edge systems. Establish cyber-resilience testbeds for	Develop resilient-by-design OT/IT architectures. Establish European OT/IT security competence networks.

© 2024-2026 NexusForum.EU



15.1 Priority: On-premises OT edge and resilient industrial AI

Many industrial and critical-infrastructure environments require local processing and control even when connectivity to central cloud systems is unavailable. OT and IT convergence should therefore support on-premise edge architectures that can host analytics, AI inference, digital-twin components and automation support functions close to physical systems, while preserving safety and operational continuity.

Impact

On-premise OT edge systems would allow European industry to deploy AI and data-driven services without compromising resilience, safety or control over operational data. They would enable predictive maintenance, anomaly detection, energy optimisation, process monitoring, robotics support and local decision support in factories, buildings, energy systems and infrastructure. They would also reduce dependency on central cloud connectivity and support sectors where latency, data sensitivity or safety requirements make cloud-only approaches unsuitable.

Recommendations

Short-term

[Research & Innovation] Develop reference architectures for OT-safe industrial edge systems. Define architectures that separate safety-critical control loops from AI-assisted monitoring, analytics and advisory functions. These architectures should clarify which functions must remain local, which may be delegated to cloud continuum systems, and how fallback modes operate.

[Research & Innovation; Ecosystem]: Develop open industrial edge gateways for legacy OT systems. Support the development of open gateways that connect industrial and building automation systems, including PLCs, SCADA systems, building management systems, and legacy fieldbus protocols, to modern data, analytics and AI platforms while preserving security, reliability and operational constraints.

[Research & Innovation] Develop local AI inference for industrial monitoring and anomaly detection. Create lightweight AI models and runtimes for predictive maintenance, fault detection, quality control, energy optimisation and safety monitoring at the industrial edge.

© 2024-2026 NexusForum.EU

[Testing and Benchmarking] Establish testbeds for offline-first OT edge operation. Develop testbeds that evaluate how industrial AI and analytics services behave under degraded connectivity, partial failure, delayed updates and local-only operation.

Medium-term

[Research & Innovation] Develop certifiable AI components for OT environments. Create methods for testing, validating and documenting AI components used in industrial environments, especially where AI supports safety-relevant decisions or operational optimisation.

[Research & Innovation] Develop lifecycle management for long-lived OT/IT systems. Support secure update, rollback, versioning and compatibility management for edge software deployed alongside industrial systems with long equipment lifecycles.

[Research & Innovation] Develop human-supervised autonomy for industrial edge. Develop AI systems that can recommend actions, support operators and automate low-risk tasks while preserving human oversight and clear escalation paths.

Long-term

[Research & Innovation] Develop self-optimising industrial edge systems with safety constraints. Create industrial edge systems that can optimise energy, production quality, maintenance schedules or resource use while respecting safety, process and regulatory constraints.

[Ecosystem] Establish European reference platforms for OT edge AI. Develop open or semi-open reference platforms for industrial edge AI that can be adopted by SMEs, system integrators, automation vendors and industrial users.

15.2 Priority: Secure OT/IT data integration and industrial data spaces

OT and IT convergence depends on the ability to extract, structure, govern and share operational data. Many industrial environments produce valuable data, but it is often locked in vendor systems, poorly documented, difficult to access securely, or not semantically harmonised. Europe needs secure and interoperable data integration mechanisms that support both local use and controlled sharing through industrial data spaces.

Impact

Secure OT/IT data integration would allow industrial users to unlock operational data for analytics, AI, digital twins and optimisation while retaining control over sensitive information. It would enable SMEs and technology providers to build services on top of industrial data, reduce vendor lock-in, support data-space participation and improve cross-site learning across factories, buildings, energy systems and infrastructure.

© 2024-2026 NexusForum.EU

Recommendations

Short-term

[Research & Innovation] Develop semantic mappings for OT systems. Create methods for mapping data from PLCs, SCADA, building management systems, sensors, robots and industrial equipment into common semantic models.

[Research & Innovation] Develop secure data extraction from legacy OT environments. Support approaches that allow operational data to be accessed without weakening safety, availability or cybersecurity. This includes read-only integration patterns, secure gateways, protocol translation and anomaly detection for data flows.

[Research & Innovation] Develop edge-based data quality and provenance tools. Develop tools that assess data quality, timestamping, lineage, calibration status and provenance close to the source before data is used in AI systems or shared externally.

Medium-term

[Research & Innovation] Integrate industrial edge systems with data spaces. Develop connectors and governance mechanisms that allow industrial edge platforms to participate in European data spaces while enforcing access control, usage policies, provenance and contractual restrictions.

[Standardisation] Develop common semantic models for industrial and building data. Support standardised ontologies and data models for priority sectors such as manufacturing, energy, mobility, buildings, logistics and water systems.

[Research & Innovation] Develop privacy- and confidentiality-preserving analytics for industrial data. Support methods such as federated analytics, secure aggregation, confidential computing and synthetic data generation where raw industrial data cannot be shared.

Long-term

[Ecosystem] Enable cross-site industrial intelligence. Develop mechanisms that allow organisations to learn across many sites without exposing sensitive operational data, supporting benchmarking, predictive maintenance, energy optimisation and fleet-wide digital twins.

[Standardisation and Governance] Adopt trusted industrial data-sharing frameworks. Establish and adopt reusable governance, contractual and technical models for controlled sharing of OT data between asset owners, technology providers, SMEs, researchers and public authorities.

© 2024-2026 NexusForum.EU

15.3 Priority: Industrial digital twins and AI-enabled operational optimisation

Digital twins are one of the most important use cases for OT and IT convergence. They connect operational data, engineering models, simulations, AI analytics and decision-support tools. However, many digital twins remain siloed, manually maintained or disconnected from real-time OT systems. Europe should support digital twins that are operationally useful, trustworthy and integrated with industrial edge and cloud/HPC capabilities.

Impact

Better industrial digital twins would improve design, monitoring, optimisation and maintenance across manufacturing, buildings, mobility, energy and infrastructure. They would support safer AI deployment by combining learned models with engineering constraints and physical models. They would also help European industry reduce energy use, improve productivity, manage assets more effectively and test interventions before applying them to real systems.

Recommendations

Short-term

[Research & Innovation] Develop edge-connected digital twin architectures. Create architectures that connect local OT data streams to digital twins while keeping time-critical functions local and secure.

[Research & Innovation] Combine AI models with engineering and physics-based models. Develop hybrid modelling approaches that use machine learning and AI for prediction, anomaly detection and optimisation while integrating and preserving physical constraints and engineering knowledge.

[Research & Innovation] Develop digital twin validation methods. Create methods for assessing whether a digital twin remains accurate as equipment, processes, sensors and operating conditions change.

Medium-term

[Research & Innovation] Develop real-time and near-real-time digital twins. Support digital twins that can ingest live OT data, simulate alternative operations and support operational decision-making under latency and reliability constraints.

[Research & Innovation] Develop AI-assisted optimisation for industrial systems. Develop tools that recommend process changes, maintenance actions, energy optimisation strategies or resource allocations, with human oversight and explainability.

[Testing and Benchmarking] Establish digital twin testbeds for European sectors. Create sector-specific testbeds for manufacturing, buildings, mobility, energy and infrastructure, including benchmark datasets, simulation models and evaluation protocols.

Long-term

© 2024-2026 NexusForum.EU

[Research & Innovation] Develop self-updating digital twins. Create digital twins that can detect model drift, request validation, update parameters and incorporate new data while preserving auditability and human control.

[Ecosystem] Establish interoperable digital twin ecosystems. Support interoperability between digital twins, industrial data spaces, edge platforms, AI services and simulation/HPC environments.

15.4 Priority: Cybersecurity and lifecycle management for converged OT/IT systems

The convergence of OT and IT significantly expands the attack surface of critical industrial systems. As once-isolated environments integrate with cloud services, AI platforms, and third-party ecosystems, they face unprecedented exposure. Given that OT lifecycles far outlast IT update cycles, patching remains a structural challenge. Consequently, integrating cybersecurity, lifecycle management, and operational resilience must be a cornerstone of the R&I agenda.

Impact

Improved cybersecurity and lifecycle management would reduce the risks associated with connecting OT systems to modern IT and AI environments. It would support safer industrial digitalisation, protect critical infrastructure, reduce downtime, and make it easier for organisations to adopt AI and cloud-edge technologies without compromising security or operational continuity.

Recommendations

Short-term

[Research & Innovation] Develop security patterns for OT/IT edge architectures. Define secure architectures for connecting legacy OT systems to edge and cloud services, including network segmentation, zero-trust access, identity management, secure gateways and monitoring.

[Research & Innovation] Develop vulnerability and asset management for long-lived OT systems. Create tools for identifying, classifying and monitoring OT assets, firmware, software dependencies and known vulnerabilities across heterogeneous industrial environments.

[Research & Innovation] Develop anomaly detection for OT/IT data flows. Support AI-assisted detection of abnormal communication patterns, data exfiltration, unsafe commands and compromised edge gateways.

Medium-term

© 2024-2026 NexusForum.EU

[Research & Innovation] Develop secure update and rollback mechanisms for industrial edge systems. Create methods for patching, updating and rolling back edge software without interrupting critical operations or violating safety requirements.

[Testing and Benchmarking] Establish cyber-resilience testbeds for converged OT/IT systems. Create environments where industrial users can test cyber incidents, degraded operation, recovery procedures and AI-assisted incident response.

[Standardisation] Develop certification-ready evidence for OT/IT cybersecurity. Support tools that generate audit evidence for compliance with cybersecurity and resilience requirements.

Long-term

[Research & Innovation] Develop resilient-by-design OT/IT architectures. Create architectures that maintain safe operation under cyberattack, connectivity loss, partial system failure or compromised AI components.

[Ecosystem] Establish European OT/IT security competence networks. Support knowledge sharing, training, reference architectures and incident-learning mechanisms for SMEs, industrial users, public infrastructure operators and technology providers.

© 2024-2026 NexusForum.EU

16 AI Operationalization

Primary destinations
2. A competitive European AI and machine learning ecosystem
5. Adoption of advanced digitalisation and AI in industry and public sectors
Secondary destinations
4. Secure, sovereign, and interoperable European computing capabilities

Background and driving factors

Deploying modern AI technologies in the continuum is challenging as Large Language Models (LLMs) and Foundation Models (FMs) are very large and costly in power usage, execution time, and storage needs. Most LLMs are highly demanding and mainly use HPC-grade equipment where the training of the foundational models (FMs) can take advantage of high-performance GPUs and other power-hungry accelerators. The term GenAIOps refers to successfully implementing, overseeing, and refining AI applications in a heterogeneous and constrained networked computing environment.[56] The term encompasses best practices from other operational best practices such as DataOps (Data Operations), LLMOps (Large Language Models Operations), and DevOps[57], essential to managing AI implementation.[58]

For both federated all-continuum device usage and classical cloud-centred LLM training and inference, scale is critical to achieve operational efficiency. Infrastructure needs for training and inference, i.e., pre-trained on vast amounts of unstructured data to learn complex concepts, need to be addressed.[59] [60]

Deploying large models on the lower compute end of the continuum, e.g., far-edge or IoT devices, is much more difficult. The minimization of the various elements that comprise the computing continuum can facilitate the operationalization of Generative AI applications. Implementing federated computation for LLMs requires careful consideration of technical, ethical, and legal aspects. It's a multi-disciplinary effort that involves advancements in machine learning, data privacy, and distributed systems. The high complexity of foundational models and the resource heterogeneity found in federated computation makes communication efficiency, scalability, data privacy and security paramount.

A shift from passive AI "tools" to active AI Agents and Agent Swarms that can plan, reason, and execute tasks autonomously requires new Agentic orchestration layers beyond Kubernetes. In the Cloud-Edge continuum, these agents will negotiate resources, migrate workloads, and collaborate to solve complex problems (e.g., autonomous driving fleets)[61]. As AI moves to the

[56] Continuum AI: Integrating Foundational AI Agents with the Cognitive Computing Continuum

[57] https://www.karini.ai/blogs/navigating-genaiops-in-enterprises

[58] The Operationalization of AI

[59] https://dl.acm.org/doi/full/10.1145/3625289

[60] Autonomy and Intelligence in the Computing Continuum: Challenges, Enablers, and Future Directions for Orchestration

[61] Lovén, et al, « Large Language Models in the 6G-Enabled Computing Continuum: a White Paper » (2025), https://urn.fi/URN:NBN:fi:oulu-202501211268

© 2024-2026 NexusForum.EU

edge to control critical systems (autonomous driving, health), *black box* models are unacceptable and instead require Explainable AI (XAI) and Trustworthiness.[62] Users and regulators need to know why an AI agent made a specific decision (e.g., "Why did the drone land here?"). This explanation must be generated locally at the edge, in real-time, often using Neuro-Symbolic techniques to map neural outputs to logical concepts.

Europe in relation to the state of the art

Current LLM deployments rely on centralized US hyperscaler APIs. Europe lacks a unified, open-standard "AI Interconnect" that allows European SMEs to chain together models across a federated infrastructure. Instead, Europe leads in regulatory frameworks (AI Act, Data Act, NIS2) but lags in the operationalization of a unified market. Current cross-border data flows often rely on complex bilateral agreements or are routed through non-EU hyperscaler clouds that abstract these borders effectively but pose sovereignty risks. European initiatives like Gaia-X and Simpl attempt to solve this via federation, but a legal free movement zone for AI execution, where a "European" label supersedes "National" labels for edge computing, is missing. Europe's focus on Data Spaces and Fair Data Economy aligns with this. However, the economic primitives (pricing algorithms, clearing mechanisms) for trading ephemeral AI tasks at the edge do not exist.

Advances in AI, HPC, simulation, and sensor ecosystems now allow robots to learn, adapt, and perform complex tasks with remarkable accuracy. Europe has strong industrial robotics sectors (Industry 4.0). However, current approaches to edge computing and robotics remain fragmented, with limited interoperability between sensing, AI processing, and control systems. Many deployments rely on centralized orchestration and static configurations, limiting scalability, adaptability, and efficiency. State-of-the-art solutions in edge AI often struggle with hardware heterogeneity and resource constraints, while robotic platforms typically operate in isolation, without seamless integration into distributed computing infrastructures, current solutions are often siloed.

While US research often focuses on training larger models (scaling laws), Europe has a strategic interest in Small Models and efficiency to run AI on its industrial edge infrastructure in an energy-efficient way, aligned with the European Green Deal. Europe has strong industrial Digital Twin initiatives (Industry 4.0). However, using Generative AI to populate these twins with synthetic training data is a frontier where Europe can lead, leveraging its strict privacy stance (using synthetic data avoids GDPR issues). The EU AI Act mandates transparency for high-risk AI. Europe is the global leader in *Trustworthy AI* regulation. The challenge is implementing technical XAI solutions that are lightweight enough for edge devices. There is a need for reproducibility and certification in LLMOps to comply with the EU AI Act, ensuring audit trails for autonomous agents.

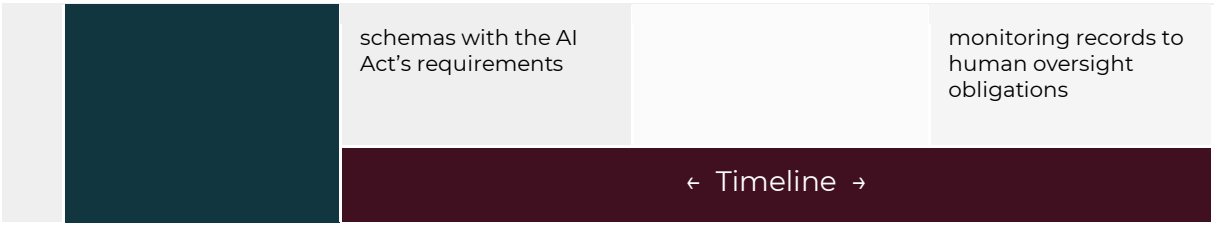
[62] *Ylianttila, et al, « 6G white paper : research challenges for trust, security and privacy » (2020), https://urn.fi/URN:ISBN:9789526226804*

© 2024-2026 NexusForum.EU

Overview of priorities and recommendations

	Short-term	Medium-term	Long-term
Priority groups	**AI Operationalization**		
Large-Scale Testbeds	Develop large-scale testbeds to accurately simulate real-world continuum conditions. Explore cloud-generated, edge-validated architectures for Generative Digital Twins. Explore Continuum Explanation Chaining when inference is split across device, edge, and cloud	Provide Digital Twin Data Spaces with shared Synthetic Data Lakes for vertical industries Standardize Trust Score for edge AI models that dynamically updates based on their performance and explainability	Certification of Synthetic Training Mandating immutable ledgers (blockchain) that record the decision logic of autonomous edge agents
Autonomous AI Agents, Multi-Agent Orchestration, and Robotics Support	Develop approaches for coordinating autonomous agents in the continuum Develop bandwidth-efficient Gossip Learning algorithms Define a Swarm-to-cloud escalation policy to determine under what conditions a swarm delegates to the cloud Employ Privacy-Preserving Drift Detection to detect concept drift	Create mechanisms to verify that an autonomous AI agent is authorized Standardize APIs for "Ad-hoc Edge Formation"	Establishing legal and technical frameworks for "Agent Responsibility" Self-Evolving AI Swarms
Evaluation and Monitoring of Human–AI Collaboration in Deployed Systems	Develop and validate a common set of domain-agnostic metrics for human–AI collaboration quality Define lightweight, embeddable logging schemas and instrumentation APIs that can be integrated into AI application pipelines Engage European standardization bodies to align HAIC metric definitions and logging	Invest in open, reusable benchmarking infrastructure implementing collaboration evaluation pipelines Establish simulation and synthetic pre-deployment evaluation methods Develop cross-domain benchmark datasets and sector-specific best practices for collaboration evaluation	Integrate collaboration-quality feedback into the AI development lifecycle Establish regulatory guidance and conformity assessment procedures that require HAIC evaluation evidence for high-risk AI systems Define accountability and audit frameworks that link HAIC

© 2024-2026 NexusForum.EU



16.1 Priority: Large-Scale Testbeds

Prepare technology and service providers for the single European digital market and its emerging technologies. Large-scale testbeds are needed to accelerate EU-scale deployment of AI services, and to provide technology and service providers with a testing ground for new technologies (such as 6G) before they are deployed at scale. Additionally, there is demand for an international testbed for cross-border data flows, for example with Japan and South Korea, considering the Cross-border Data flow deal signed between EU and Japan. Large scale testbeds address the needs of several emerging application areas.

Edge devices often lack sufficient local data to train robust models, and privacy laws prevent uploading raw data to the cloud. Simultaneously, Generative AI can produce synthetic data to train models for rare events (e.g., network failures, cyberattacks). This drives the need for Generative Digital Twins, virtual replicas that use GenAI to simulate diverse scenarios, to create synthetic datasets to train edge AI agents before they are deployed in the physical world. This data generation is inherently privacy-preserving and can be used to train models for use in scenarios where real data is not only sparse but highly regulated.

Large Scale Testbeds provide the necessary means to establish explainable AI (XAI) in large settings. The testbed needs to encourage user participation through transparency and trust-building measures, and implement continuous monitoring for issues like data drift. Furthermore, ecosystem collaboration between academia, industry, and regulatory bodies is essential for standardizing federated computation methodologies and advancing LLMs.

Impact

Large scale testbeds support accelerated deployment of AI services on EU-scale and prepare technology and service providers for the single European digital market. Access to large-scale testbeds that can simulate real-world conditions for testing AI services, pipelines, and workflows in the Computing Continuum is also critical to scaling up and advancing results from European research and innovation projects. Without such test environments, it's challenging to validate and optimize solutions and services for use in operational conditions.

Privacy-preserving data generation could solve the cold start problem for Edge AI by allowing European SMEs to train industrial AI models without needing years of historical data, accelerating the adoption of AI in sectors like manufacturing and logistics.

Advancing a human-centric approach, for example human-AI collaboration and AI transparency, is essential for social acceptance of emerging technologies such as AI and 6G. It allows humans to trust autonomous systems and enables compliance with EU regulations, giving European AI products a "Quality of Trust" competitive advantage.

© 2024-2026 NexusForum.EU

Recommendations

Short-term

- **[Ecosystem] Develop large-scale testbeds to accurately simulate real-world continuum conditions.** Set up coordination between multiple international partners to facilitate large-scale international testbeds to ensure the testbed provides environments where information can be shared in real-world conditions. Testbeds must integrate 6G technologies. Provide means to turn working testbeds into market products, reusing standardized and established technologies, such that working scaled up test bed applications can be realized. A potential candidate is public-private partnerships that enable successful pilots to acquire assets. Establish of large-scale international testbeds (e.g., between France, Germany, Italy, Japan, South Korea) specifically to stress-test the legal and technical handover of live AI services across national fiber/5G/6G boundaries.
- **[Research & Innovation] Explore cloud-generated, edge-validated architectures for *Generative Digital Twins*** that run in the cloud but validation of synthetic data quality runs at the edge using lightweight discriminators. Integrate *Generative Digital Twins* into the continuum operational fabric as live co-simulators alongside physical edge nodes. Develop metrics to mathematically verify that synthetic data generated by GenAI is statistically representative of real-world physics and constraints.
- **[Research & Innovation] Explore *Continuum Explanation Chaining* when inference is split across device, edge, and cloud** to facilitate trustable and explainable AI. The explanation must be assembled from partial explanations at each tier and presented coherently. Define Trust Score Propagation, i.e., how a trust score computed at the edge tier is validated and updated as a model migrates across continuum nodes. Development of Local Interpretability algorithms for simplified explanations for specific inferences on constrained devices without running the full model analysis. Investigate the possibility to add a dedicated lightweight XAI hardware target (e.g., an explanation accelerator as a co-processor alongside the NPU) as an accelerator (e.g., connecting to AI hardware development in Europe in Section 11). Develop tools to verify the robustness of models against adversarial examples before they are deployed, and how to contain it.

Medium-term

- **[Ecosystem] Provide *Digital Twin Data Spaces* with shared *Synthetic Data Lakes* for vertical industries** (e.g., automotive crash scenarios generated by AI) to train safety-critical edge models.
- **[Ecosystem] Standardize *Trust Score* for edge AI models that dynamically updates based on their performance and explainability** in the local context to support *Continuum Explanation Chaining*.

Long-term

- **[Regulation] Certification of Synthetic Training.** Regulatory frameworks that accept models trained on synthetic data for safety-critical certification (e.g., in autonomous driving), provided the Digital Twin fidelity is proven.

© 2024-2026 NexusForum.EU

- **[Regulation] Mandating immutable ledgers (blockchain) that record the decision logic of autonomous edge agents** for post-incident forensic analysis.

16.2 Priority: Autonomous AI Agents, Multi-Agent Orchestration, and Robotics Support

The rapid growth of Artificial Intelligence (AI) and Generative AI (GenAI) is transforming autonomous systems such as robots and vehicles (drones, cars etc). These systems must process massive volumes of data in real time, where decisions (e.g., navigation, manipulation, or safety actions) must be executed with minimal latency. While AI training relies on cloud and High-Performance Computing (HPC), the actionable phase, AI inference, must increasingly occur at the edge or directly on-device, closer to where data is generated. This shift is driven by latency and bandwidth limitations, privacy requirements, and context-aware intelligence.

Integrating on-device (on-robot), edge, cloud, and HPC resources into a coherent continuum remains challenging due to heterogeneity, fragmented technology stacks, and differences in device capabilities. Robotic systems, operating as extreme-edge resources, introduce additional complexity due to mobility, energy constraints, and intermittent connectivity, requiring new approaches to orchestration, optimization, and AI lifecycle management.

Managing millions of AI models at the edge is impossible manually and requires Automated AI Lifecycle Management where the system autonomously triggers retraining, model swapping, and pruning based on data drift and resource availability. This is *AI managing AI* to ensure the continuum remains operational and efficient. Aligned with ETSI ZSM standards, Europe needs to extend ZSM concepts from *network functions* to *AI functions* (AI-as-a-Service).

We are seeing a shift from centralized orchestration to "Swarm Networking" and "Collective Intelligence"[63]. In scenarios like autonomous driving platoons or industrial robotics, devices must form ad-hoc local clouds to solve problems jointly without relying on backhaul connectivity. This requires Agentic AI capable of decentralized planning and execution. The network must support *horizontal* collaboration between devices (Device-to-Device), not just vertical offloading. A Swarm Intelligence layer that allows robots from different vendors to form a temporary edge cluster is crucial for resilience in disconnected/islanded modes.

Impact

Enables Zero-Touch operations where the network optimizes itself through agent negotiation, and allows complex vertical applications (e.g., rescue swarms) to operate autonomously. This in turn increases resilience of critical infrastructure (e.g., rescue drones) by enabling them to operate as an autonomous swarm when the central network fails.

Automated AI Lifecycle Management drastically reduces the operational cost (OpEx) of deploying edge AI and ensures that edge models remain accurate over time without manual intervention.

[63] *Peltonen, et al, « 6G white paper on edge intelligence » (2020), https://urn.fi/URN:ISBN:9789526226774*

© 2024-2026 NexusForum.EU

Recommendations

Short-term

[Research & Innovation] Develop approaches for coordinating autonomous agents in the continuum, combining Device Agents (lightweight reflexes), Edge Agents (coordination, resource brokering), and Cloud/HPC Agents (long-horizon planning). Add a research item on “Cross-Tier Agent Message Passing” with bounded latency guarantees.

[Research & Innovation] Develop bandwidth-efficient *Gossip Learning* algorithms where edge nodes exchange model updates in a peer-to-peer fashion without a central aggregator. Use hierarchical swarm topologies to allow edge nodes to form a meta-swarm with cloud as a silent coordinator. Quantify gossip protocol bandwidth overhead under realistic 5G sidelink constraints.

[Research & Innovation] Define a *Swarm-to-cloud escalation policy* to determine under what conditions a swarm delegates to the cloud (e.g., sustained resource shortage, detected OOD data). Develop lightweight protocols for AI agents to communicate intent and status (e.g., "I need a GPU for 5ms") without the overhead of human-language prompts. Define a *Dual-Authority Model* to distinguish decisions that Automated AI Lifecycle Management can make autonomously at the edge (e.g., model quantisation, local fine-tuning) from those requiring cloud-side governance approval (e.g., deploying a new base model).

[Research & Innovation] Employ Privacy-Preserving Drift Detection to detect concept drift of the AI-managing-AI without exposing raw edge data to the cloud. Ensure that autonomous retraining in a TEE produces verifiable model provenance. Employ Lightweight statistical methods to detect "Concept Drift" (data changing) on edge devices, triggering an automated request for retraining or model updates.

Medium-term

[Ecosystem] Create mechanisms to verify that an autonomous AI agent is authorized to purchase compute resources or access data on an edge node.

[Ecosystem] Standardize APIs for "Ad-hoc Edge Formation," allowing a device to discover nearby compute resources and initiate a swarm computation task dynamically.

Long-term

[Regulation] Establishing legal and technical frameworks for "Agent Responsibility", determining who is liable when an autonomous agent makes a mistake in the continuum. Establish legal frameworks to determine accountability when a decentralized swarm of AI agents makes a collective decision that causes physical harm.

[Research & Innovation] Self-Evolving AI Swarms: Edge clusters that autonomously share data and re-train themselves (collective learning) to adapt to new environments without any cloud instruction.

© 2024-2026 NexusForum.EU

16.3 Priority: Evaluation and Monitoring of Human–AI Collaboration in Deployed Systems

The deployment of AI systems in operational settings increasingly takes the form of collaborative workflows, where human operators and AI components interact iteratively to accomplish tasks — in domains such as healthcare decision support, smart city services, industrial operations, and autonomous assistance. Current practice in AI operationalization provides limited tooling for evaluating the quality of these collaboration processes: evaluation remains predominantly model-centric, focusing on accuracy and latency of the AI component in isolation, while the interaction dynamics that determine real-world effectiveness are rarely measured in a structured or reproducible way.

A first fundamental gap is the **absence of standardized evaluation metrics** for the collaboration process itself, capturing properties such as operator trust calibration, cognitive effort, reliance dynamics, and collaborative efficiency, as distinct from model accuracy and evaluation during training. Existing metrics are designed to assess AI model outputs, not the quality of the human–AI interaction that occurs when those outputs are acted upon by a human decision-maker.

A second challenge is the **lack of lightweight, embeddable instrumentation** that can be integrated into existing AI application pipelines with minimal overhead. Without standardized logging schemas and instrumentation APIs, collecting the interaction data needed to compute collaboration-quality indicators requires bespoke per-application development. This barrier effectively limits evaluation to well-resourced research pilots rather than routine operational practice.

Even where evaluation is performed**, there is no established mechanism for continuous post-deployment monitoring** of collaboration quality across model versions, operator cohorts, and evolving task distributions. This makes it difficult to detect degradation in human–AI interaction dynamics after initial deployment, a critical gap given that model updates, changes in user populations, and shifts in task context can all alter collaboration quality without any corresponding change in model accuracy.

Finally, **current evaluation frameworks show limited adaptability across domains.** Metrics and logging schemas are typically designed for specific experimental setups and cannot be reused across healthcare, manufacturing, public administration, or other sectors without significant re-engineering. This fragmentation prevents the accumulation of cross-domain evidence and raises the cost of adoption for organizations that lack dedicated evaluation expertise.

The ability to include standardized, easily integrated, and adaptable metrics for the evaluation of human-AI collaboration directly within application development enables the out-of-the-box documentation and logging of the necessary interaction data—reducing the barrier for developers to instrument their systems and making collaboration-quality evidence a natural by-product of the development and deployment process rather than a separate evaluation effort.

© 2024-2026 NexusForum.EU

Table 1. Overview of priorities, and their relationship to research challenges in the field and AI policy relevance.

R&I Priority	Research Challenge	Policy Relevance
Standardised collaboration metrics	Domain-agnostic metric definitions and structured logging contracts computable from interaction logs	AI Act transparency & human oversight documentation
Open benchmarking infrastructure	Modular, API-accessible evaluation pipelines integrated with MLOps and CI/CD	Reproducibility and auditability of AI deployments
Lightweight embedding & adaptation	Low-overhead instrumentation patterns for heterogeneous AI applications across cloud-edge environments	Broad sectoral adoption; interoperability across continuum deployments
Simulation & surrogate-agent evaluation	Pre-deployment HAIC quality assessment via calibrated synthetic agents	Risk assessment for high-risk AI systems (AI Act Annex III)
Lifecycle integration & continuous monitoring	Longitudinal collaboration monitoring with degradation detection and feedback into model/workflow improvement	AI Act Art. 9 risk management; ongoing human oversight obligations

Impact

Enables operational monitoring for arbitrary HAIC metrics directly within deployed AI workflows, allowing organizations to track collaboration quality in production without specialized infrastructure or offline post-processing.

Enables model evolution monitoring from a human–AI collaboration perspective, providing evidence of whether iterative model updates improve or degrade the quality of the collaboration process, beyond standard accuracy benchmarks.

Recommendations

Short-term

[Research & Innovation] Develop and validate a common set of domain-agnostic metrics for human–AI collaboration quality, operationalized as computable functions over structured interaction logs and covering efficiency, trust, cognitive load, adaptability, and fairness. Ensure metrics are defined independently of any specific AI model or application architecture to allow reuse across sectors.

[Research & Innovation] Define lightweight, embeddable logging schemas and instrumentation APIs that can be integrated into AI application pipelines with minimal

© 2024-2026 NexusForum.EU

overhead, enabling collaboration-quality evidence to be collected as a natural by-product of development and deployment without bespoke per-application tooling.

[Ecosystem] Engage European standardization bodies (CEN/CENELEC, ETSI) to align HAIC metric definitions and logging schemas with the AI Act's requirements for human oversight documentation and transparency, establishing a common interoperability baseline for deployed AI systems.

Medium-term

[Research & Innovation] Invest in open, reusable benchmarking infrastructure implementing collaboration evaluation pipelines as modular, API-accessible services that integrate with existing MLOps and CI/CD workflows, enabling HAIC evaluation to function as a standing operational capability rather than a one-off measurement exercise.

[Research & Innovation] Establish simulation and synthetic pre-deployment evaluation methods using agents calibrated to realistic human behavior profiles, enabling assessment of collaboration quality before live rollout and supporting risk evaluation for high-stakes AI applications.

[Ecosystem] Develop cross-domain benchmark datasets and sector-specific best practices for collaboration evaluation in healthcare, public administration, industrial operations, and education, providing reference implementations that lower adoption barriers for organizations without dedicated evaluation expertise.

Long-term

[Research & Innovation] Integrate collaboration-quality feedback into the AI development lifecycle, closing the loop between operational monitoring and model/workflow improvement, including automated mechanisms for detecting when collaboration degradation should trigger human review, retraining, or interface redesign.

[Regulation] Establish regulatory guidance and conformity assessment procedures that require HAIC evaluation evidence for high-risk AI systems under the EU AI Act (Annex III), including longitudinal collaboration monitoring records as part of mandatory risk management and post-market surveillance obligations.

[Regulation] Define accountability and audit frameworks that link HAIC monitoring records to human oversight obligations, enabling organizations and regulators to verify that human control over AI-assisted decisions is substantive and not merely nominal across the operational lifetime of deployed systems.

© 2024-2026 NexusForum.EU

17 Conclusions

Europe's digital competitiveness will increasingly depend on its ability to build, operate and control advanced capabilities across AI, data, computing infrastructure, hardware, software, security, energy and sectoral adoption. In this roadmap, the European Cognitive Computing Continuum describes this connected capability landscape, recognising that recent advances in AI will require development of new European capabilities across a range of complementary platforms, infrastructures and technology approaches.

A central message of the roadmap is that Europe should not organise all research and innovation priorities around a single architecture. Federation is essential where data, services and infrastructure must operate across providers, organisations or borders. Other priorities require specialised computing resources, integrated hardware–software development, open technology stacks or locally deployed industrial and edge capabilities. These approaches should be treated as complementary and connected through appropriate standards, interfaces and shared ecosystems.

Taken together, the six Strategic Destinations define the outcomes Europe should pursue, while the five Pillars identify the technological, operational and adoption-related capabilities needed to reach them. Progress will require coordinated advances in foundational AI, continuum infrastructure, hardware–software stacks, intelligent infrastructure management and sectoral deployment, supported by cybersecurity, sustainability, interoperability, open source and digital sovereignty as transversal priorities.

The roadmap also highlights that **Europe's AI strategy must go beyond model development alone**. While Europe needs the capacity to train, fine-tune, evaluate and operate advanced AI models, the next phase of AI competition will increasingly concern the systems built around models. Agentic AI, workflow execution graphs, harness engineering, memory and knowledge integration, verified tool use, observability, small models and on-device AI will determine how AI becomes useful in real industrial, scientific and public-sector environments. Neuro-symbolic AI is especially important in this context, because it can combine the flexibility of learned models with structured knowledge, reasoning, domain constraints and traceability.

Emerging computing paradigms will also shape Europe's long-term position. Neuromorphic computing offers a pathway towards event-driven, low-power and brain-inspired AI systems, particularly relevant for edge intelligence, robotics, sensing and autonomous systems. Quantum computing, and especially hybrid quantum-classical computing, may open new possibilities for optimisation, simulation, cryptography, scientific discovery and AI-enabled workflows. These paradigms are not immediate replacements for existing systems, but they are strategic options that Europe must mature through sustained research, testbeds, software stacks and early application pilots.

A further conclusion is that **the hardware-software stack is strategic**. European investments in RISC-V, AI accelerators, memory architectures, neuromorphic systems, quantum technologies and post-exascale computing must be matched by mature software stacks, compilers, runtimes, benchmarks, developer tools and adoption pathways. Without software maturity and demand creation, European hardware and emerging-computing initiatives risk remaining isolated from real deployment.

© 2024-2026 NexusForum.EU

Sustainability must also be treated as a design requirement from the start. AI infrastructure will place growing pressure on energy systems, data centres, cooling, hardware supply chains and grid capacity. Europe should therefore prioritise energy-efficient AI, memory-aware inference, carbon- and grid-aware scheduling, efficient data centres, waste-heat reuse and whole-system optimisation.

As digitalisation and AI become embedded across a growing range of sectors, coordination will also need to extend beyond individual project clusters. Coordination and Support Actions play an important role in aligning projects within a common thematic area, but many of the roadmap's priorities cut across established programme, technology and sector boundaries. Europe should therefore support stronger coordination between project clusters, including mechanisms that connect and align the activities of different CSAs. This would help identify shared dependencies, reduce duplication, coordinate standards and testbeds, exchange reusable results and ensure that advances in AI, data, computing infrastructure, cybersecurity, energy and sectoral applications are developed as part of a coherent European capability agenda.

Finally, the roadmap shows that **research and innovation must be connected more directly to adoption.** AI Factories, EuroHPC, testbeds, industrial pilots, public procurement, open-source communities, standardisation and international cooperation are essential for turning European R&I into usable capabilities.

The overall message is clear: Europe should not bet on one architecture, one platform or one technology path. It should build the capabilities, expertise, and ecosystems across research and business needed to drive development in both current and emerging AI and computing technologies.

© 2024-2026 NexusForum.EU

18 List of Contributors

Editor-in-chief

Thomas Ohlson Timoudas, RISE Research Institutes of Sweden

Editorial team

Anastasiia Andriievska, RISE Research Institutes of Sweden

Björn Forsberg, RISE Research Institutes of Sweden (Cloud-Edge for AI WG co-lead)

Anders Forsgren, RISE Research Institutes of Sweden (Telco Cloud-Edge WG co-lead)

Ian Marsh, RISE Research Institutes of Sweden (AI for Cloud-Edge WG co-lead)

Sachiko Muto, RISE Research Institutes of Sweden (Sovereignty & Open Source WG co-lead)

Jon Summers, RISE Research Institutes of Sweden (Sustainability and Energy-efficiency WG co-lead)

Iraklis Symeonidis, RISE Research Institutes of Sweden (Cybersecurity WG co-lead)

Named contributors

Michael Abdou, OpenNebula

Andrew Adams, Meiji university

Juncal Alonso, Tecnalia

Antonio Alvarez Romero, OpenNebula

Enrique Areizaga, Tecnalia

Philippe Bonnet, University of Copenhagen

Michele Capobianco, Capobianco

Paris Carbone, KTH Royal Institute of Technology, RISE Research Institutes of Sweden

Marta Chinnici, ENEA C.R. CASACCIA

Josu Diaz de Arcaya, Tecnalia

Geir Horn, University of Oslo

Olatz Ibañez, Tecnalia

JeongGil Ko, Yonsei university

Johan Kristiansson, Luleå University of Technology

Alberto Pablo Martí Martínez, OpenNebula

Fernanda Milla de León, OpenNebula

Michele Nicosia, Oracle

© 2024-2026 NexusForum.EU

Nicolae Paladi, Ph.D, Canary Bit AB

Francesco Panella, Martel Innovate

María Paula Pecho, OpenNebula

Asier Garcia Perez, Tecnalia

Giovanni Rimassa, Martel Innovate

Qian Wang, KTH Royal Institute of Technology

Lander Bonilla Viana, Tecnalia

Thor Wikfeldt, RISE Research Institutes of Sweden

Cagatay Yilmaz, RISE Research Institutes of Sweden

Tarik Zakaria Benmerar, Acigna Inc.

Chiara Zincone, OpenNebula

Unnamed contributions from the NexusForum.EU Working Groups

AI for Cloud-Edge

Cloud-Edge for AI

Telco Cloud-Edge

Cloud-Edge Use Cases

Sovereignty & Open Source

Sustainability

Interoperability

Cybersecurity

© 2024-2026 NexusForum.EU

NEXUS FORUM EU

Consolidating Research and Policy along the Cognitive Computing Continuum

eucloudedgeiot.eu

Funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI