Edge AI: Towards a European Roadmap

EPoSS/INSIDE Edge AI Working Group

Dr. Inessa Seifert

05.12.2024

Motivation and the structure of the presentation

What is Edge AI? Why a new Edge AI roadmap is needed?

What is a value chain behind Edge AI?

Where is the potential for Edge AI ?

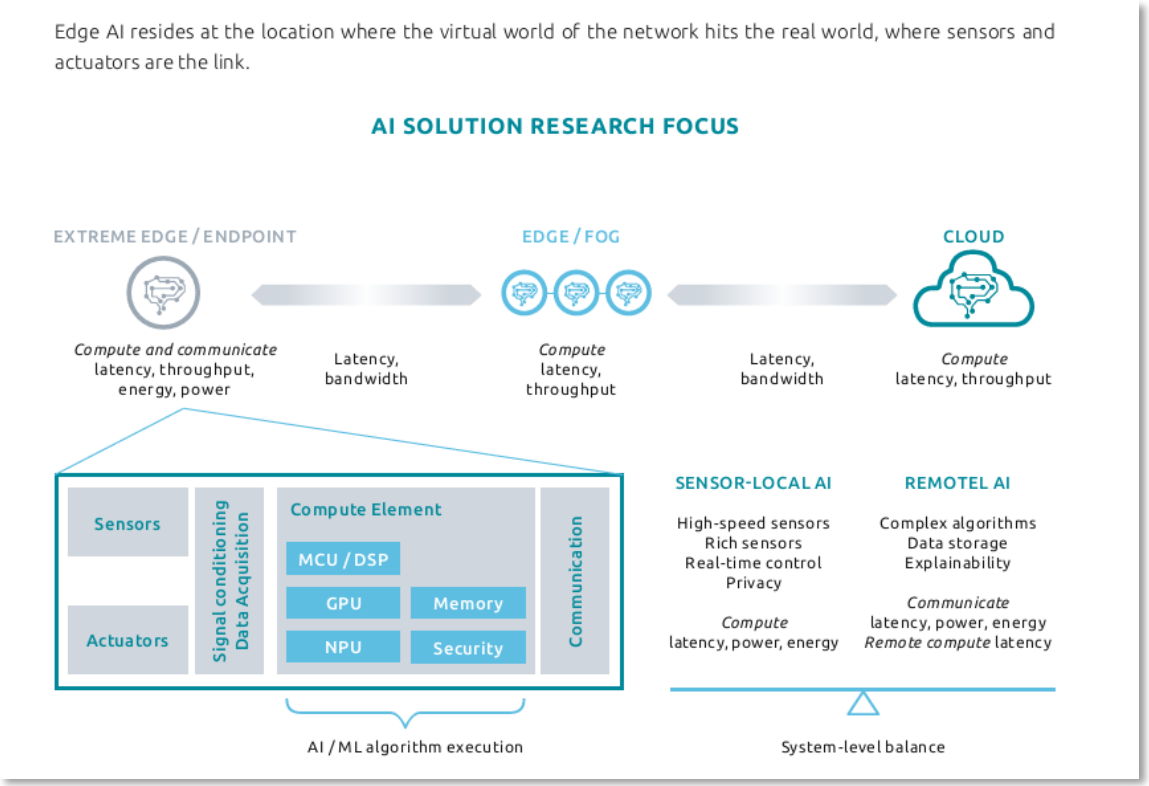
What are funding activities of Chips JU in the area of Edge AI?

What are the next steps?

What is Edge AI? Why a new Edge AI roadmap is needed?

In intelligence and artificial intelligence, an **intelligent agent (IA)** is an agent that **perceives its environment**, **takes actions autonomously** in order to **achieve goals**, and **may improve its performance with learning or acquiring knowledge** (Russel & Norvig, Artificial Intelligence: A Modern Approach, 4th US ed. modified 2022).

Low latency requirements combined with privacy concerns have shaped sub-fields application areas such as **Edge AI**, enabling processing and reasoning at the edge of the digital continuum that covers **cloud edge and IoT (Internet of Things) connected devices**.



[AI at the Edge, White paper (2021)]

Evolution of AI systems towards General Artificial Intelligence

Sam Altman, CEO of OpenAI, predicts we will **reach level five within ten years**, while some in the space believe it could take up to fifty years. The actual timeline remains uncertain, but the rapid pace of AI development is undeniable … (Forbes, 2024)

- **conversational AI:** computers can interact in conversational language with people
- **reasoning AI:** can perform basic problem-solving tasks comparable to a human
- **autonomous AI:** "agents" can operate autonomously on a user's behalf
- **innovative AI:** AI helps with inventions
- **organizational AI:** can develop innovations independently - not just running processes, but improving them

Large Language Models

Digital twins

Metaverse/ Omniverse

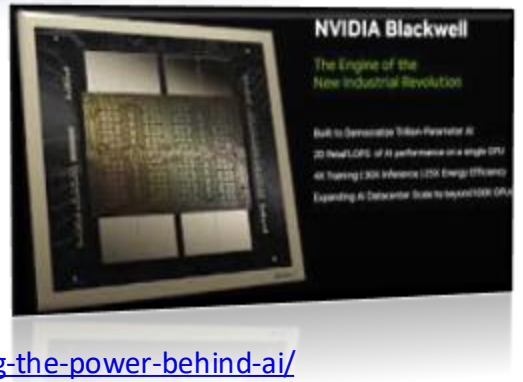
Virtual worlds

Status quo: Large Language Models

Definition: A large language model (LLM) is a type of computer model developed for natural language processing tasks, such as language generation. As language models, LLMs acquire these capabilities by learning statistical relationships from large amounts of text during a self-supervised or semi-supervised training process.

- '**Hallucinations**' pose a problem for natural language generation systems that use LLMs (such as ChatGPT1, Gemini2): Users cannot trust the correctness of certain output (see [1]).
- LLMs require a **lot of energy** due to the enormous size of the models [2].
- The **hardware** and **energy investments are enormous** [2].

Planned construction of nuclear power plants: Amazon, Google & Microsoft [4]



Overheating of Blackwell-Prozessor [3]

[1] https://www.nature.com/articles/s41586-024-07421-0
[2] https://adasci.org/how-much-energy-do-llms-consume-unveiling-the-power-behind-ai/
[3] https://www.hostzealot.de/blog/news/nvidias-uberhitzungsproblem-bei-blackwell-prozessoren-wurde-aufgedeckt
[4] https://www.nytimes.com/2024/10/16/business/energy-environment/amazon-google-microsoft-nuclear-energy.html

AI

Big tech's dominant approach is to prioritize closed flagship models while also releasing lighter-weight open models

Public company	Market cap	Country	Dominant approach	Notable closed activity	Notable open activity
NVIDIA	$3.60T	US	Open	Invested in multiple closed model developers	Introduced NVLM 1.0 multimodal frontier-level LLM family (September 2024)
	$3.45T	US	Closed	Announced proprietary on-device and server foundation models (June 2024)	Released OpenELM model family (April 2024)
Microsoft	$3.17T	US	Closed	Multi-billion-dollar investment in OpenAI; rumored to be working on 500B parameter MAI-1 model (May 2024)	Released Phi-3 small language models (April 2024)
amazon	$2.22T	US	Closed	Amazon Titan foundation model family available on Amazon Bedrock	Supports open-source AI models on AWS infrastructure
Google	$2.01T	US	Closed	Announced flagship Gemini 1.5 model (February 2024)	Introd (Febru
Meta	$1.46T	US	Open		Introd family
Tencent 腾讯	$532.9B	China	Closed	Announced Hunyuan Turbo foundation model (September 2024)	Text-t releas
Alibaba	$214.3B	China	Open	Flagship Qwen language models available via API	Launc model

Source: CB Insights company data; company releases.
Note: Market cap data as of 11/15/2024. Companies selected based on market cap & regional relevance. Developers open-sourcing AI models do so on a spectrum, sharing some combination of model weights, underlying source code, and original training data.

European AI start ups

ALEPH ALPHA

MISTRAL AI_

AI

High compute costs, limited moats, and big tech competition have created a market ripe for a shake-up

Recent pivots and quasi-exits among foundation model players validate the trend

Pivots > to lighter-weight models, while layering paid services on top

ALEPH ALPHA AI21labs

Both moved away in 2024 from competing on general-purpose LLMs to building smaller and/or optimized models and related AI tools.

Quasi-exits > collapsing into big tech

ADEPT Inflection character.ai

All essentially "acqui-hired" by big tech companies, with the founders and large portions of teams going to the acquirer.

The deals reflect the high costs of model development, with licensing payments going to investors.

Paywall frontier models

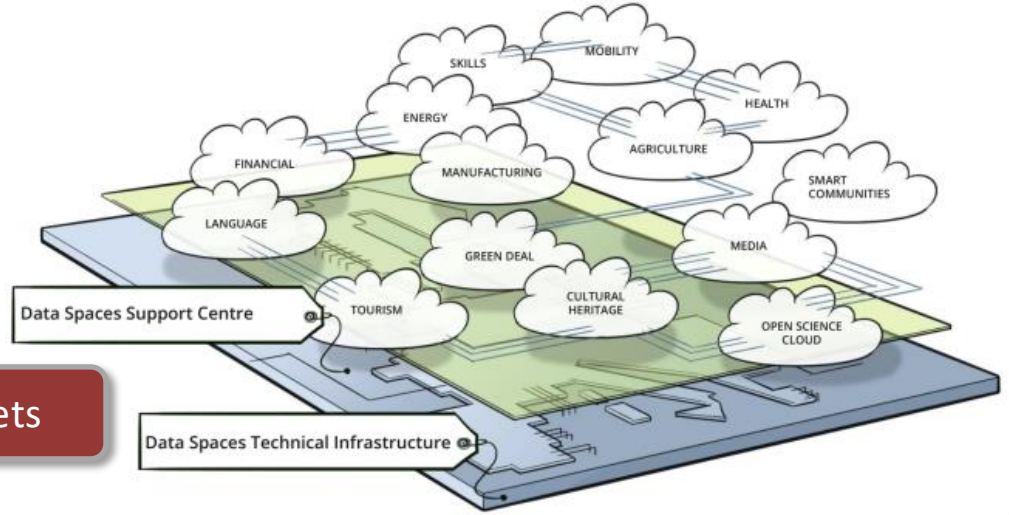
MISTRAL AI_ xAI

For open-source AI developers without a clear path to revenue, selling access to their best models while open-sourcing their lower tiers is one approach companies are taking — much like big tech.

Source: CB Insights

61 / CBINSIGHTS

Cloud Edge IoT Continuum ecosystem perspective



[1] Image credits: Data space for cultural heritage, Jeroen Meijer, Atelier Compass, 2022-04-28, Europeana Foundation, The Netherlands. CC-BY-SA

GenAI

foundation models

high quality data sets

Cloud

intelligent agent

(LLM-based) multi-agent systems

frugal AI

orchestration and brokering as enabler for **secure, distributed, energy efficient, trustworthy and safe AI at the Edge**

Digital twins

Metaverse/omniverse

Virtual worlds

IoT and resource constrained devices

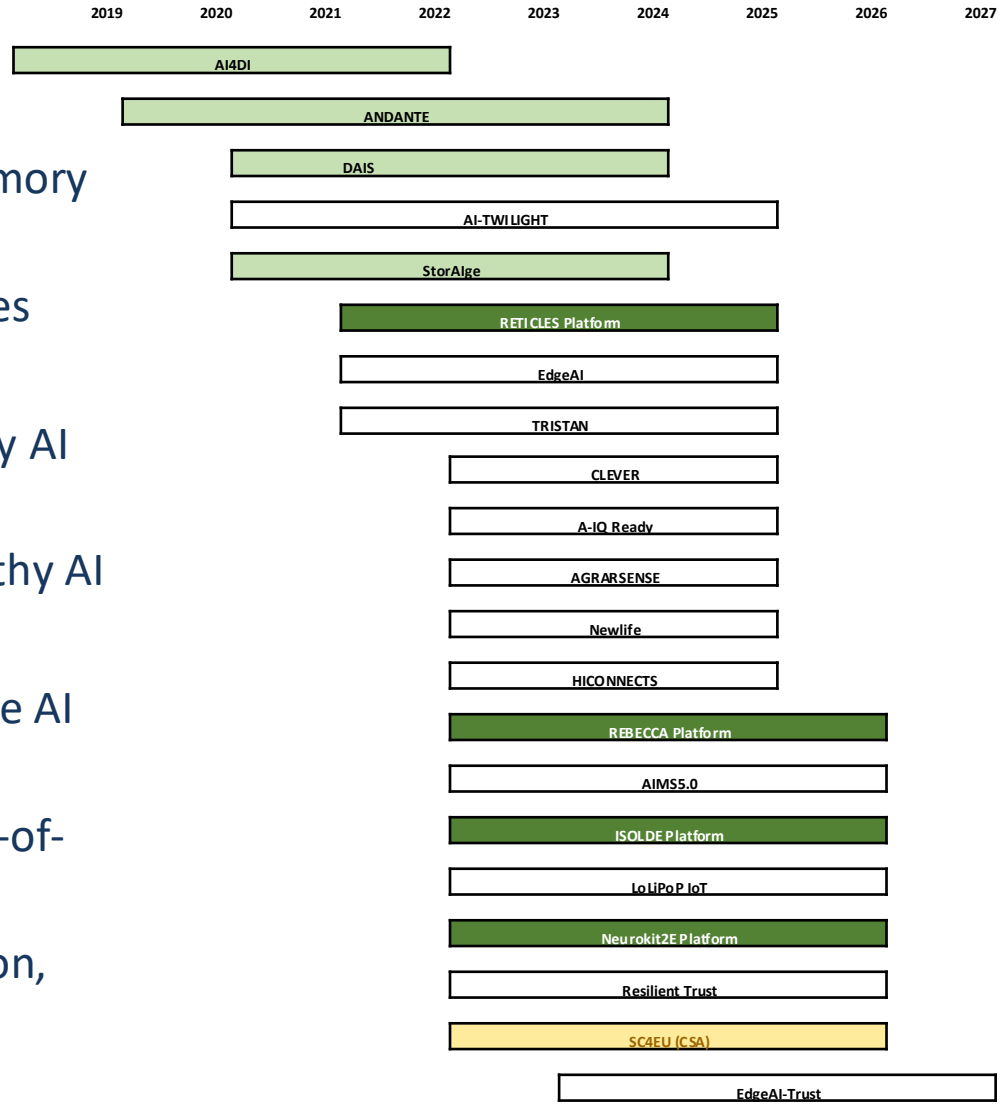
- Neuromorphic Accelerators
- In-Memory Computing (Memristive Technologies)
- ASICs (Application-Specific Integrated Circuits) and SoCs (System-on-Chip)
- FPGAs (Field Programmable Gate Arrays)

Systems Integrators' Interaction with Stakeholder Groups in the Cloud-Edge-IoT Continuum

Stakeholders	Tasks and tools for Systems Integrators
Cloud-Edge-IoT Infrastructure providers	Design, Testing and Deployment Tools
Telko Edge, Connectivity providers	Energy efficient safe and secure data transmission, resource managment and orchestration
Chip designers, Hardware vendors	(AI assisted) Hardware and Software Co-Design for sector-specific applications
AI researchers and Solution providers	Safe and secure training, deployment and re-use of explainable and intreperable AI models
IT security providers	Safety and security testing, certification and validation tools
End users in the vertical domains	Deployment, Maintenance and Support Tools
Data space providers, Data intermediaries, Data rights owners and Data providers	Safe and secure access to (sensor) data, secure orchestration and brokering
Digital Twin providers/associations/vendors	Interoperability, standardisation and certification
Metaverse/Omniverse – Providers and users of simulation environments	Safe and secure access to real-time data, testing and deployment of AI models at the Edge

Key Digital Technologies and Chips JU R&I projects towards an Edge AI Roadmap

- Moving processing on the edge (e.g. adv. memory management, in-memory computing accelerators)
- Distributed Edge AI: foundational models, data and learning technologies
- AI verification and certification
- AI chips supporting multiple computing paradigms and multi-technology AI (e.g. classical, neuromorphic, deep learning)
- AI explainability, interpretability, verification and certification: trustworthy AI
- Interoperability, scalability, modularity, self-x functionalities
- Engineering tools for designing, training, updating and maintaining edge AI
- Support for entire lifecycle from requirement specification to design, development, deployment, operation, maintenance, evolution and end-of-life
- Intent driven optimization, multi-agents, machine-to-machine interaction, interaction with digital twins and simulation environments



Next steps

CEI-Sphere part of EUCloudEdgeIoT.eu a Coordination and Supporting Action funded by the EU Commission

1. **Consultation with the Chips JU community on the Edge AI Roadmap**
2. **CEI-Sphere: a series of online webinars and onsite workshops:**

- Emerging data-driven business models for service providers opened up by Data Act, Data Governance Act, AI Act, ...
- Guidelines for Privacy Enhancing Technologies in Cloud-Edge-IoT Infrastructures
- Recommendations for collaboration and support in open multi-stakeholder CEI value chains and networks

Would you like to join the Edge AI Working Group?

inessa.seifert@vdivde-it.de

5th December 2024