Towards Native AI architectures in 6G

Professor Tasos Dagiuklas

SuITE Research Group

Head of Cognitive Systems Research Centre

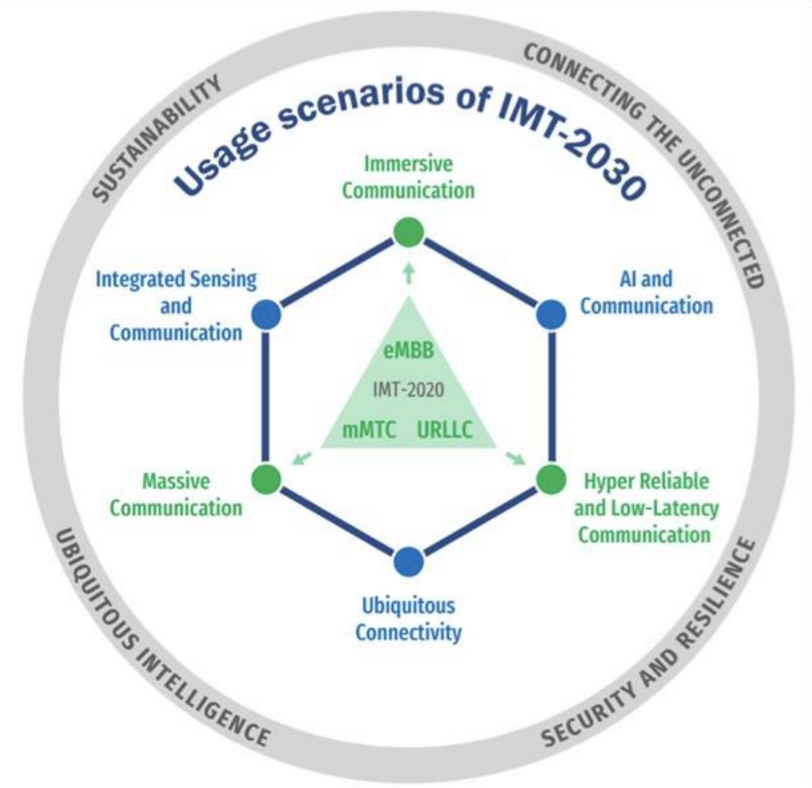
London South Bank University

UK

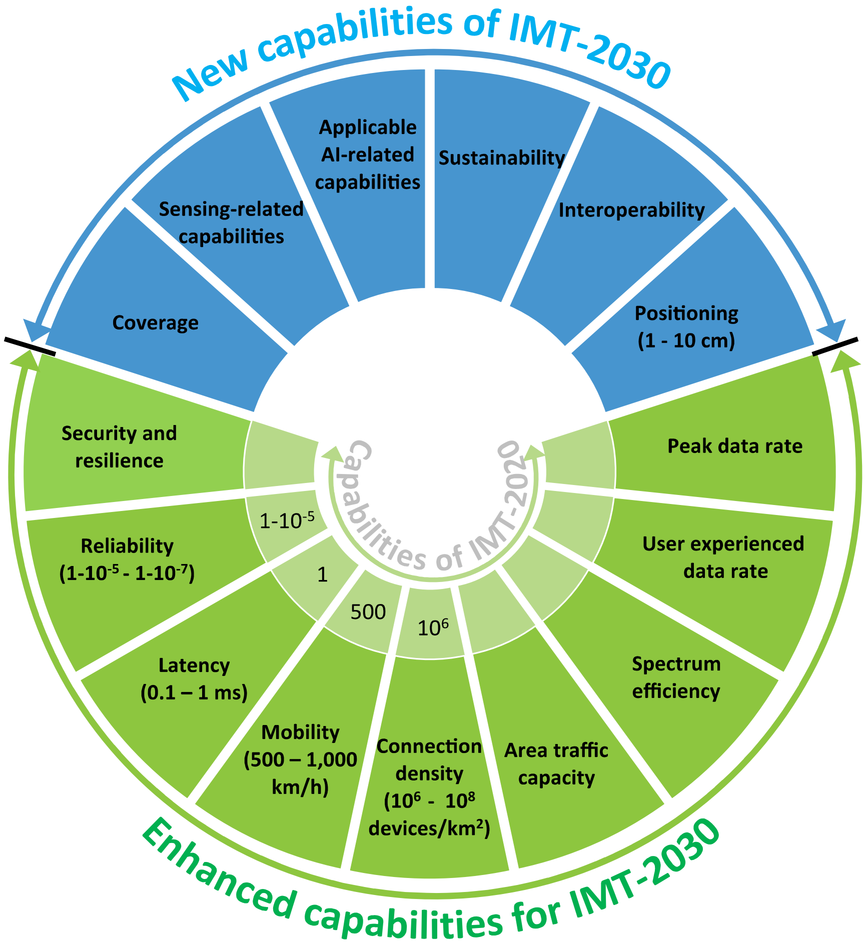
https://www.suitelab.org

London
South Bank
University
EST 1892

IMT-2030 Use Case Scenarios

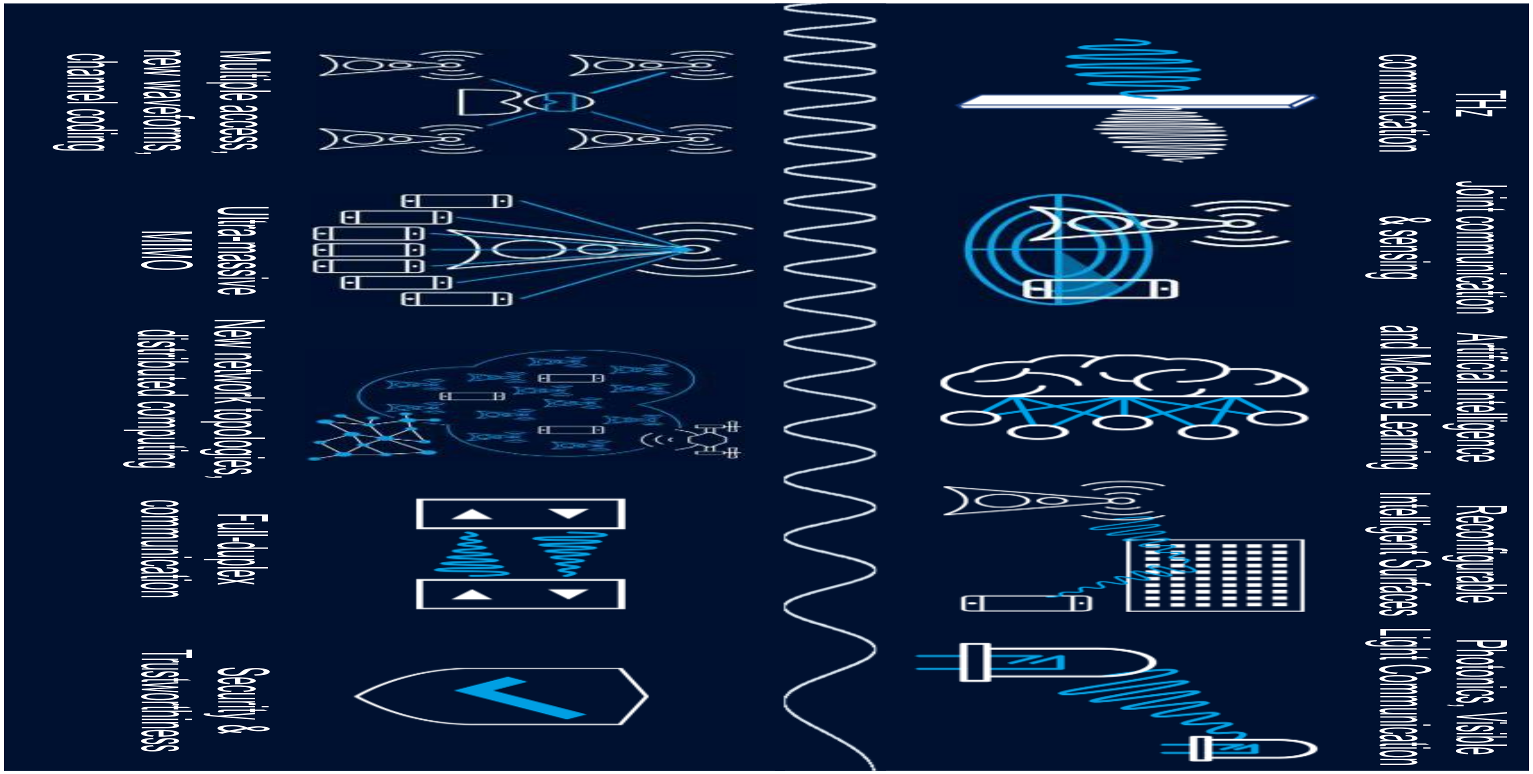


IMT-2030 capabilities


New capabilities of IMT-2030
Applicable AI-related capabilities
Sustainability
Sensing-related capabilities
Interoperability
Coverage
Positioning (1 - 10 cm)
Security and resilience
Peak data rate
Capabilities of IMT-2020
1-10-5
Reliability (1-10-5 - 1-10-7)
User experienced data rate
1
500
10^6
Latency (0.1 – 1 ms)
Spectrum efficiency
Mobility (500 – 1,000 km/h)
Connection density (10^6 - 10^8 devices/km^2)
Area traffic capacity
Enhanced capabilities for IMT-2030

Source: ITU-R M.2160

6G Research Areas

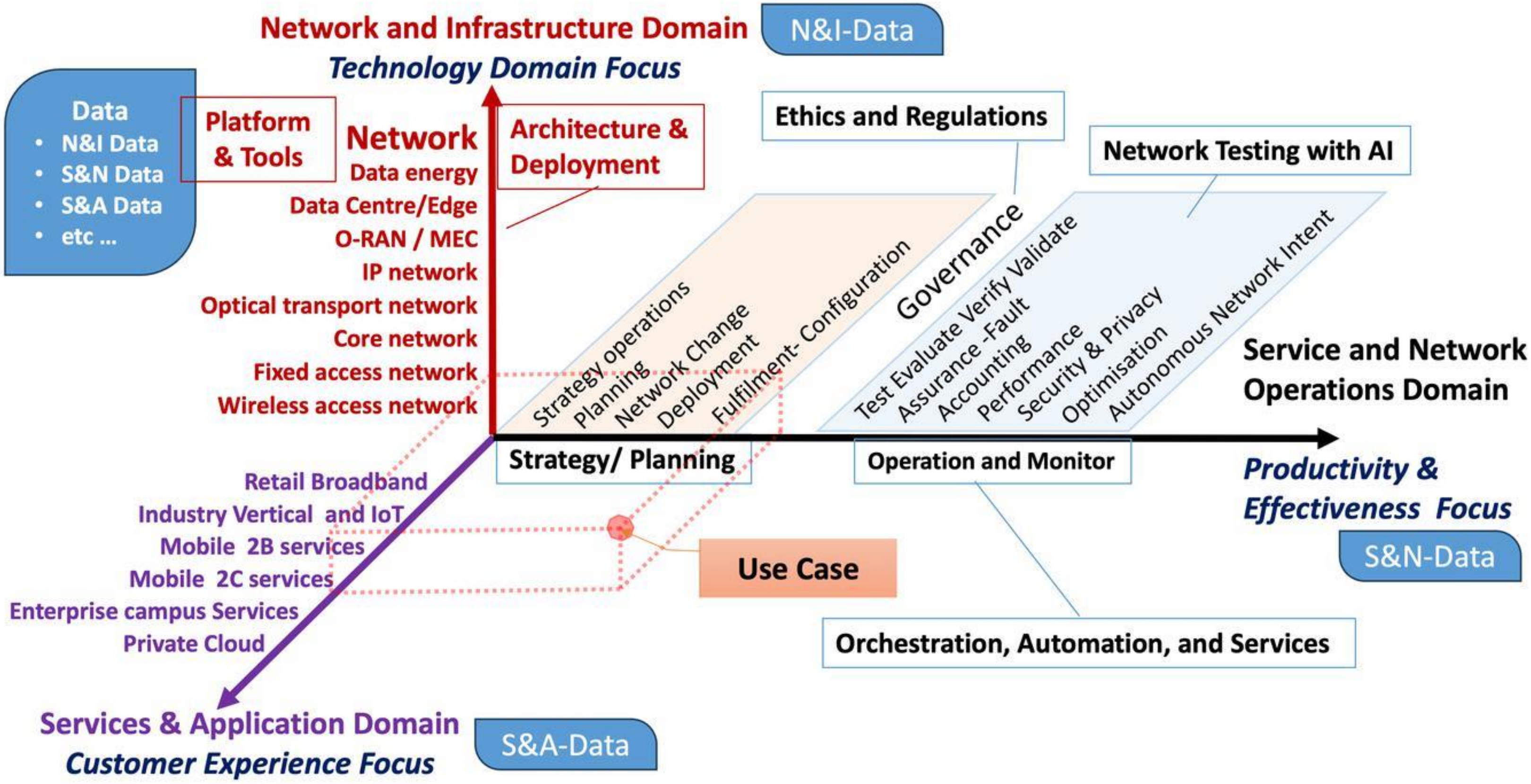

THz communication
Joint communication & sensing
Artificial intelligence and Machine Learning
Reconfigurable Intelligent Surfaces
Photonics, Visible Light Communication
Multiple access, new waveforms, channel coding
Ultra-massive MIMO
New network topologies, distributed computing
Full-duplex communication
Security & Trustworthiness

Source: Rhode & Scharz, 2023

Why AI in Telecommunications?-(1)

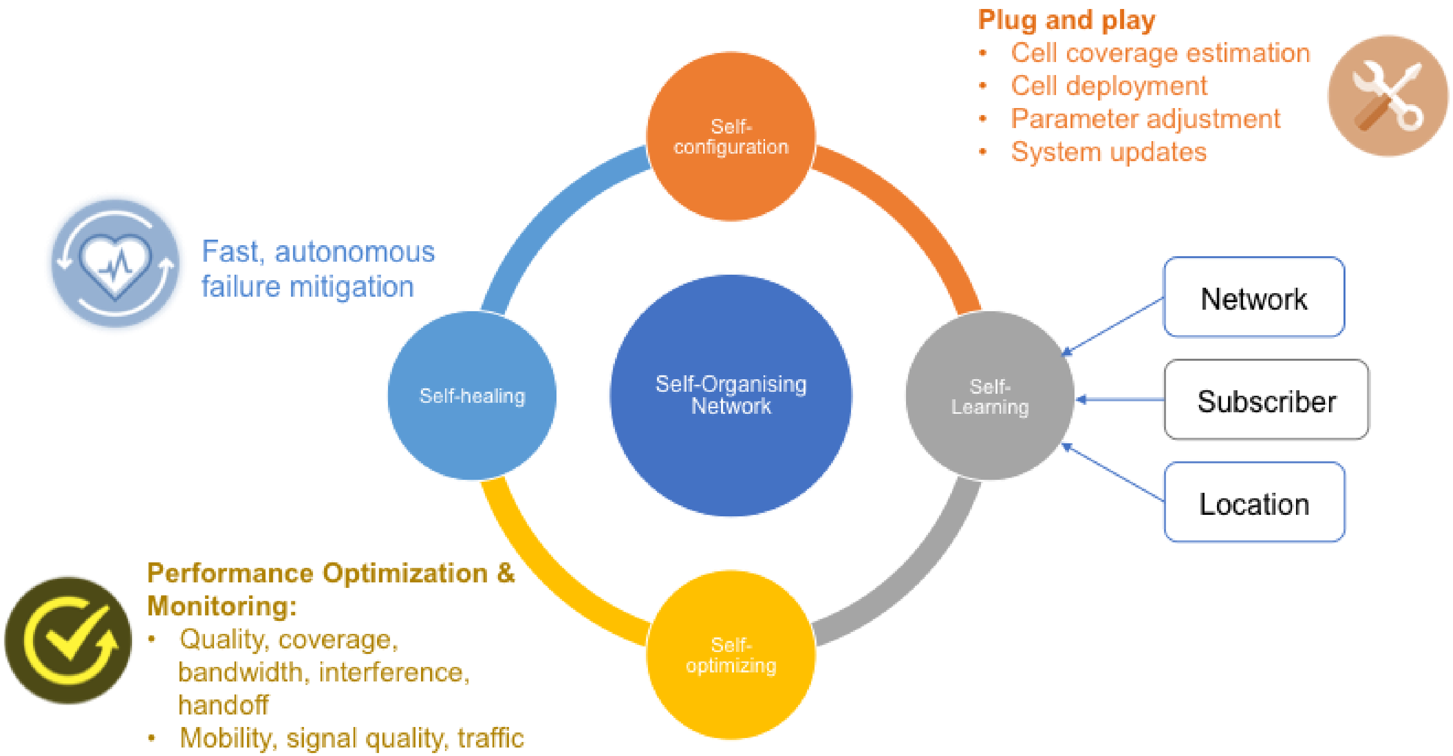
- Network and infrastructure optimisation that is needed to optimise services and provide operational efficiency.
- Lifecycle and operational process for managing services, applications, network, and infrastructure.
- Services and application offered to customer and enterprises.

Why AI in Telecommunications?-(2)



Source: UKTIN, 2024

Knowledge Defined Networking



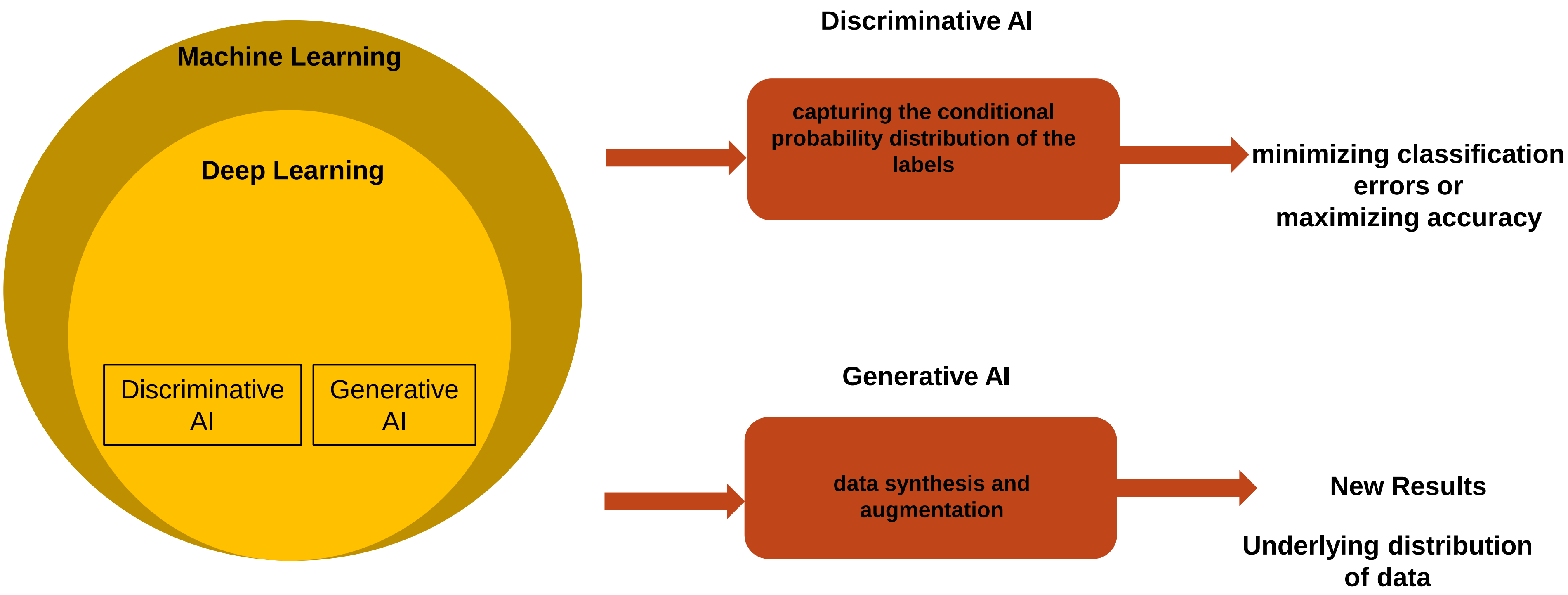
Plug and play
- Cell coverage estimation
- Cell deployment
- Parameter adjustment
- System updates

Fast, autonomous failure mitigation

Performance Optimization & Monitoring:
- Quality, coverage, bandwidth, interference, handoff
- Mobility, signal quality, traffic

SON-KDN cycle

AI technological options



Machine Learning

Deep Learning

Discriminative AI

Generative AI

Discriminative AI

capturing the conditional probability distribution of the labels

minimizing classification errors or maximizing accuracy

Generative AI

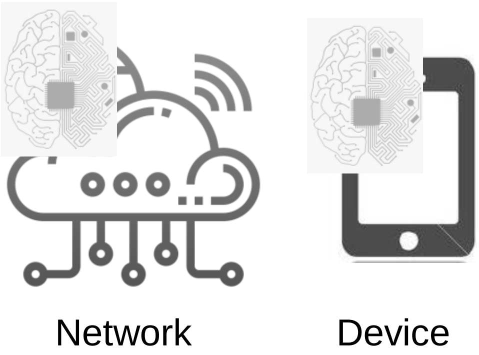
data synthesis and augmentation

New Results

Underlying distribution of data

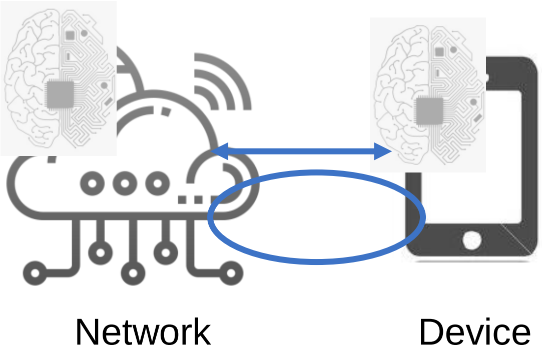
AI/ML Deployment in Networking

Independent AI/ML



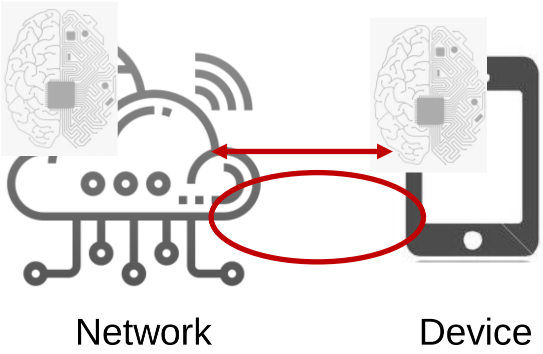
- ML can be deployed independently either at the network or at the device
- Proprietary ML deployment
- Proprietary data collection

Co-ordinated AI/ML



- Co-ordination between network & device
- Proprietary & standardized ML procedures
- Data collection for both training and monitoring

Native AI/ML

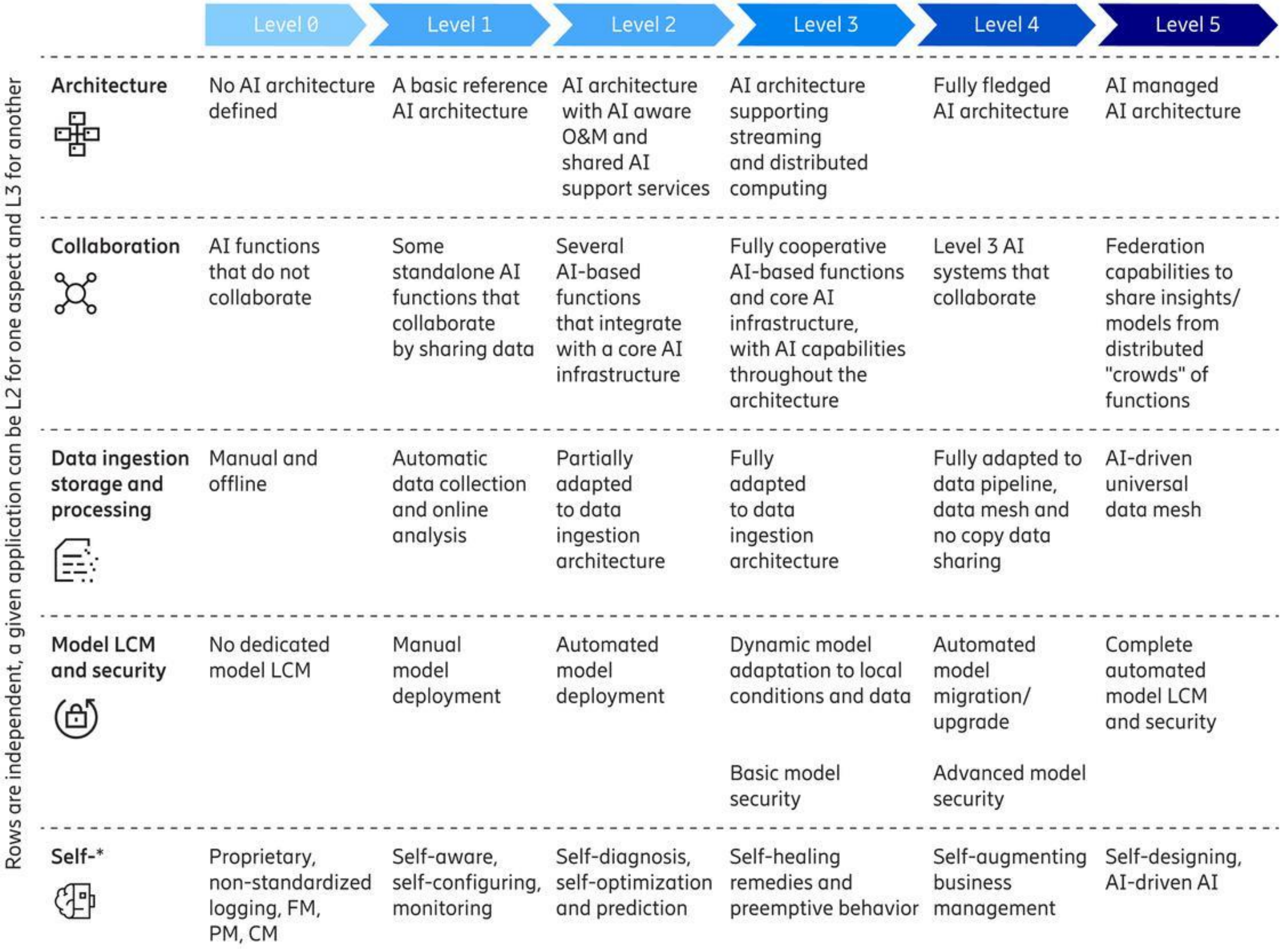


- Autonomous ML deployment between network and devices across all layers
- ML procedures to train performance and adapt to different environments
- From DevOps to MLOps

Towards Native AI

- AI capabilities, available or exposed for network or services.
- Splitting the entire AI system into multiple subsystems based on the specific objectives of the service.
  - Each component is then integrated into the service function of the service, to provide a cohesive system.
  - The split AI approach can utilize a distributed architecture where different parts of the system handle
    - Data pre-(processing)
    - model training,
    - model inference

Native AI Capabilities

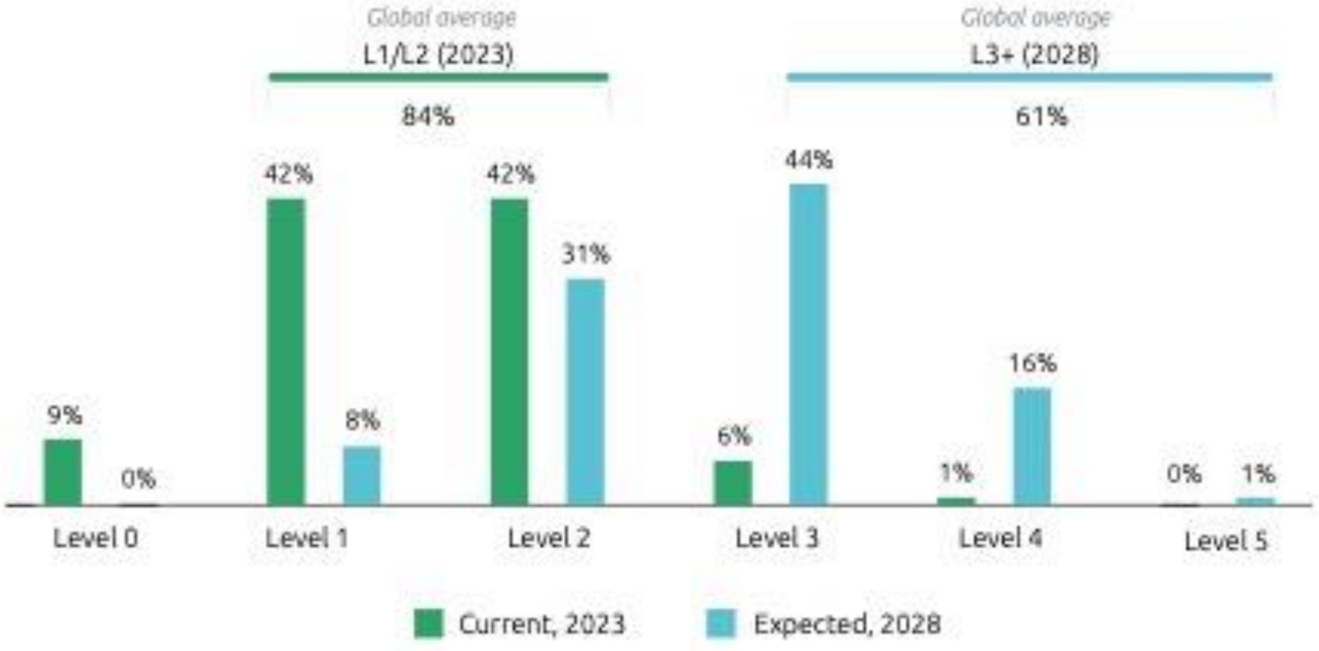


Rows are independent, a given application can be L2 for one aspect and L3 for another

	Level 0	Level 1	Level 2	Level 3	Level 4	Level 5
Architecture	No AI architecture defined	A basic reference AI architecture	AI architecture with AI aware O&M and shared AI support services	AI architecture supporting streaming and distributed computing	Fully fledged AI architecture	AI managed AI architecture
Collaboration	AI functions that do not collaborate	Some standalone AI functions that collaborate by sharing data	Several AI-based functions that integrate with a core AI infrastructure	Fully cooperative AI-based functions and core AI infrastructure, with AI capabilities throughout the architecture	Level 3 AI systems that collaborate	Federation capabilities to share insights/ models from distributed "crowds" of functions
Data ingestion storage and processing	Manual and offline	Automatic data collection and online analysis	Partially adapted to data ingestion architecture	Fully adapted to data ingestion architecture	Fully adapted to data pipeline, data mesh and no copy data sharing	AI-driven universal data mesh
Model LCM and security	No dedicated model LCM	Manual model deployment	Automated model deployment	Dynamic model adaptation to local conditions and data Basic model security	Automated model migration/ upgrade Advanced model security	Complete automated model LCM and security
Self-*	Proprietary, non-standardized logging, FM, PM, CM	Self-aware, self-configuring, monitoring	Self-diagnosis, self-optimization and prediction	Self-healing remedies and preemptive behavior	Self-augmenting business management	Self-designing, AI-driven AI

Source: Ericsson White paper, https://www.ericsson.com/en/reports-and-papers/white-papers/ai-native, 2023

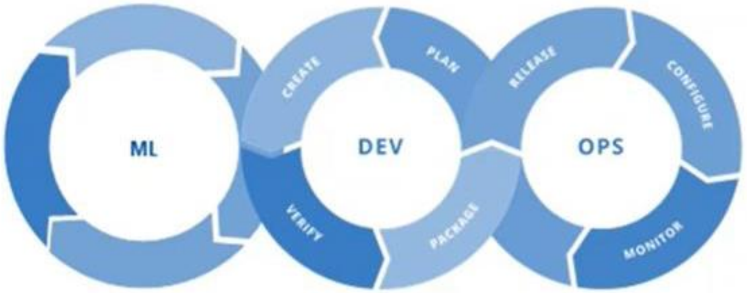
Overall Network Autonomy: Current vs Expected



Source: Capgemini Research Institute, Autonomous Telco Networks Survey, December 2023–January 2024, N = 113 large CSP

From DevOps to MLOps

MLOps = ML + DEV + OPS



Experiment
Data Acquisition
Business Understanding
Initial Modeling

Develop
Modeling + Testing
Continuous Integration
Continuous Deployment

Operate
Continuous Delivery
Data Feedback Loop
System + Model Monitoring

Native AI Requirements

- Novel computing architectures and infrastructures for handling extensive data volumes and intricate algorithms is required
- Use of heterogeneous (GPUs, FPGAs, NPUs, DPUs) H/W infrastructure.
- Workload Management
  - Use of AI to manage outages and initiate migration
  - Orchestration Frameworks need to be extended to handle Dynamic and Multi-Tenant Resources in a secure manner
- Intent-Based Automation using LLM

AI Challenges

- Managing Data Errors
  - Imprecise Measurements, with added Noise
  - Missing Values or Entire Records
  - Data Anomalies
  - Records which are communicated with a significant delay (e.g. online measurements).
- Growing Demand for AI Area Networking: massive data transfers and instantaneous processing, without bottleneck
  - DPU and AI Accelerators
  - Infiniband vs. Ultra Ethernet

AI in Telecommunications: SDOs-(1)

- ITU: Setting the International goals for IMT 2030
- ETSI:
    - Securing AI (SAI)
    - Experiential Networked Intelligence (ENI)
    - Zero touch network & Service Management (ZSM)
    - Network Functions Virtualisation (NFV)
    - Open CAPIF
    - Open Slice

AI in Telecommunications: SDOs-(2)

- 3GPP: Towards 6G (IMT-2030) recommendations
  - 3GPP: Systems Architecture, SA-WG1, WG2, WG5
    - AI/ML operation splitting
    - AI/ML model/data distribution & sharing
    - Distributed and Federated training
    - Management services for managing AI/ML capabilities / Intent
  - 3GPP: Radio Access Network, RAN-WG1, WG3
    - CSI feedback / Beamforming / Positioning
    - Energy Saving / Load balancing / Mobility optimisation

Questions

Email: tdagiuklas@lsbu.ac.uk
URL: www.suitelab.org