AIoTwin @ Dubrovnik, Croatia. 16 Sept 2024

Sustainable & Trustworthy Edge AI for Future Computing

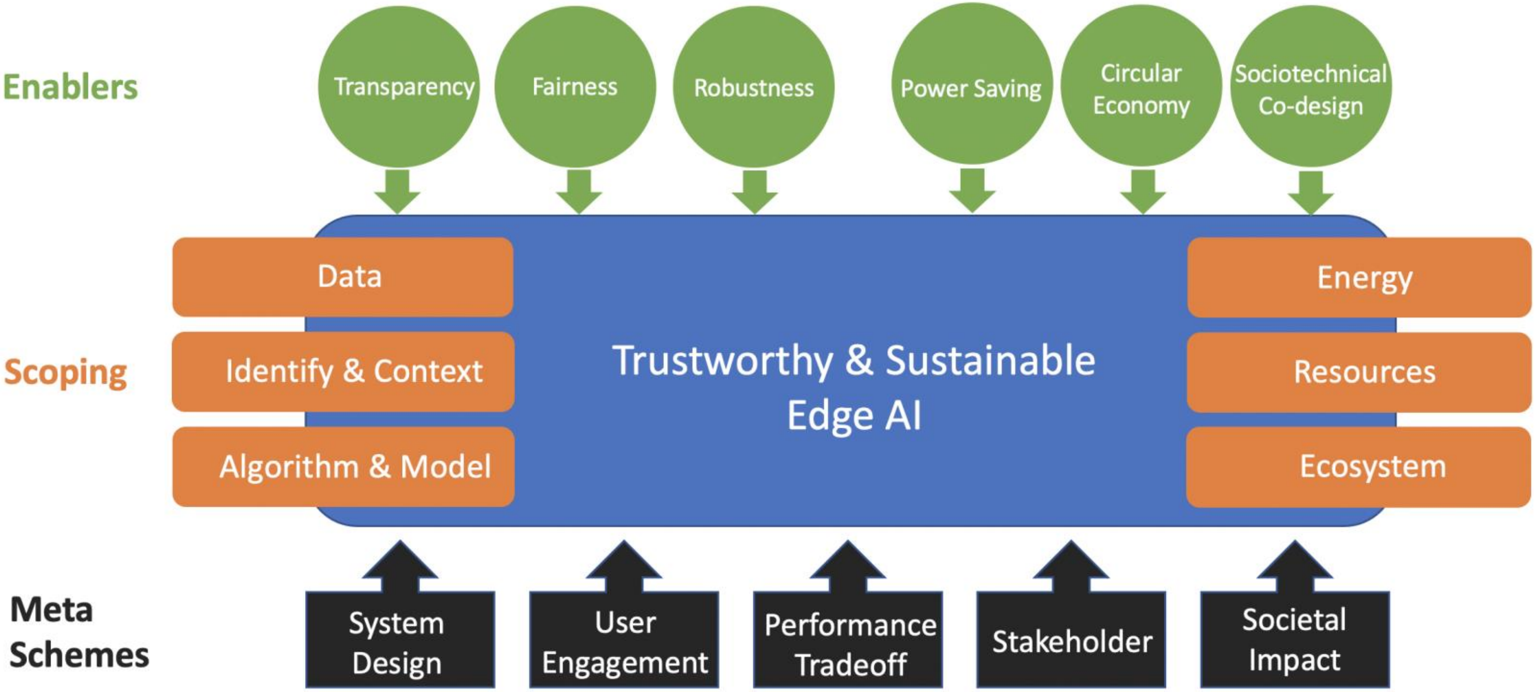
Aaron Ding

Director of CPI Lab, TU Delft

TUDelft

EU H2020 **SPATIAL** & Marie Curie **APROPOS**: Grant Agreement No.101021808 & No. 956090

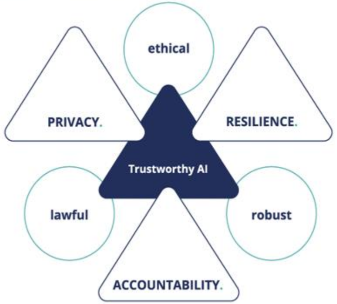
Complex Subject

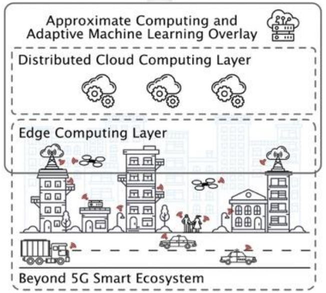


Aaron Ding, Marijn Janssen, Jon Crowcroft. “Trustworthy and Sustainable Edge AI: A Research Agenda”

CPI on Edge AI

- SPATIAL of €5M grant
- APROPOS of €4M grant




Approximate Computing and Adaptive Machine Learning Overlay
Distributed Cloud Computing Layer
Edge Computing Layer
Beyond 5G Smart Ecosystem

Trustworthy Edge AI

Score: 98/100 | Rate: 8%

Sustainable Edge AI

Score: 14.5/15 | Rate: 3%

Is Edge AI a Real Thing?

Edge AI is **Real**

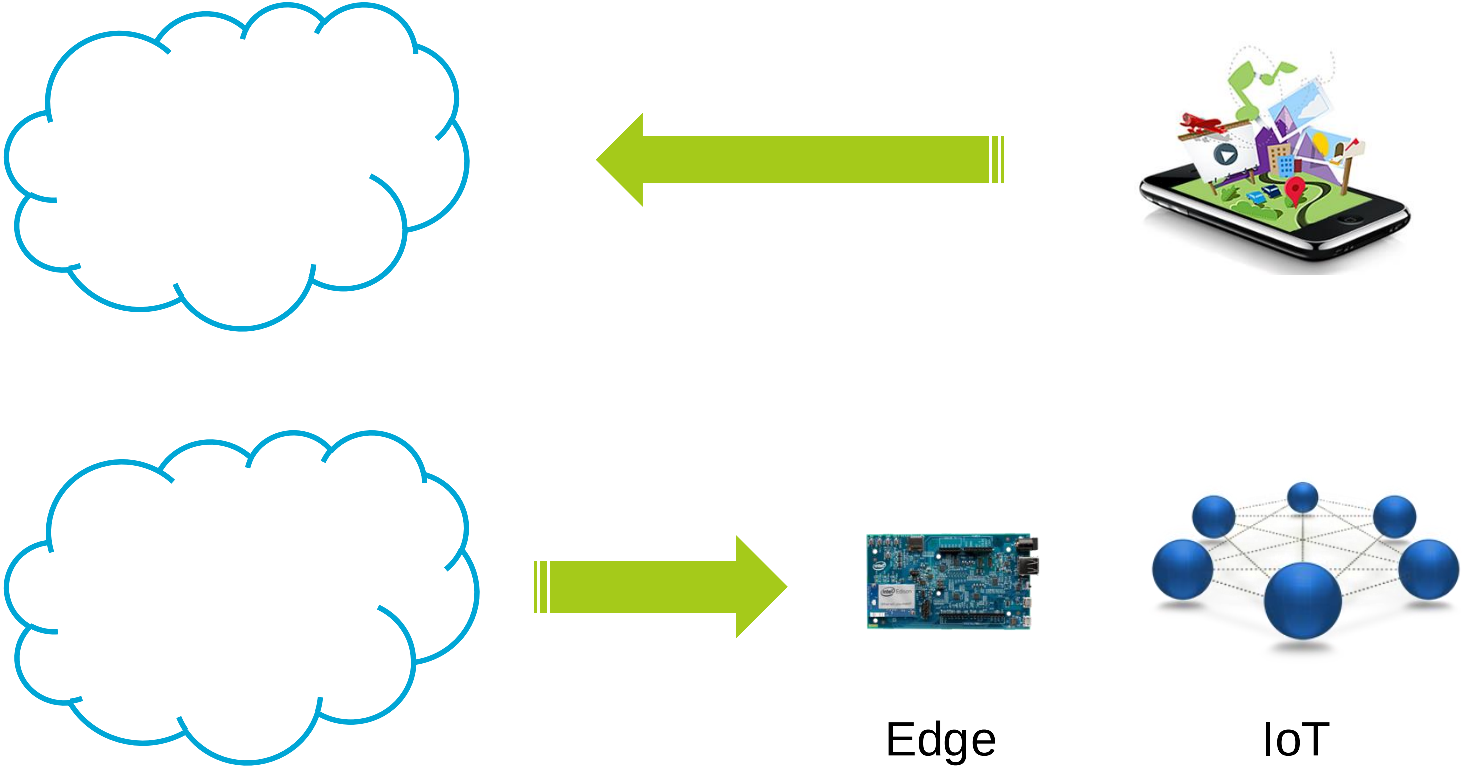
Provider	Dev Platform
Microsoft	Azure Data Box Edge
Intel	Movidius Neural Compute Stick
NVIDIA	Jetson Nano, TX, Xavier NX
Huawei	Atlas AI Computing Platform

Provider	Hardware
Google	Tensor Processing Units (TPU)
Intel	Movidius Vision Processing Units (VPUs) & Xeon D-2100
Qualcomm	Qualcomm Snapdragon 8 Series, Hexagon DSP
Huawei	Ascend Series & Kirin 600/900 NPU
Samsung	Exynos 9820 Neural Processing Unit (NPU)
NVIDIA	TURING GPU

Provider	Management Framework
Microsoft	Azure IoT Edge
Google	Google Cloud IoT
NVIDIA	NVIDIA EGX
Amazon	AWS IoT Greengrass
Alibaba	Link IoT Edge
Linux Found.	EdgeX & Akraino Edge Stack
Huawei	KubeEdge

What exactly is Edge AI ?

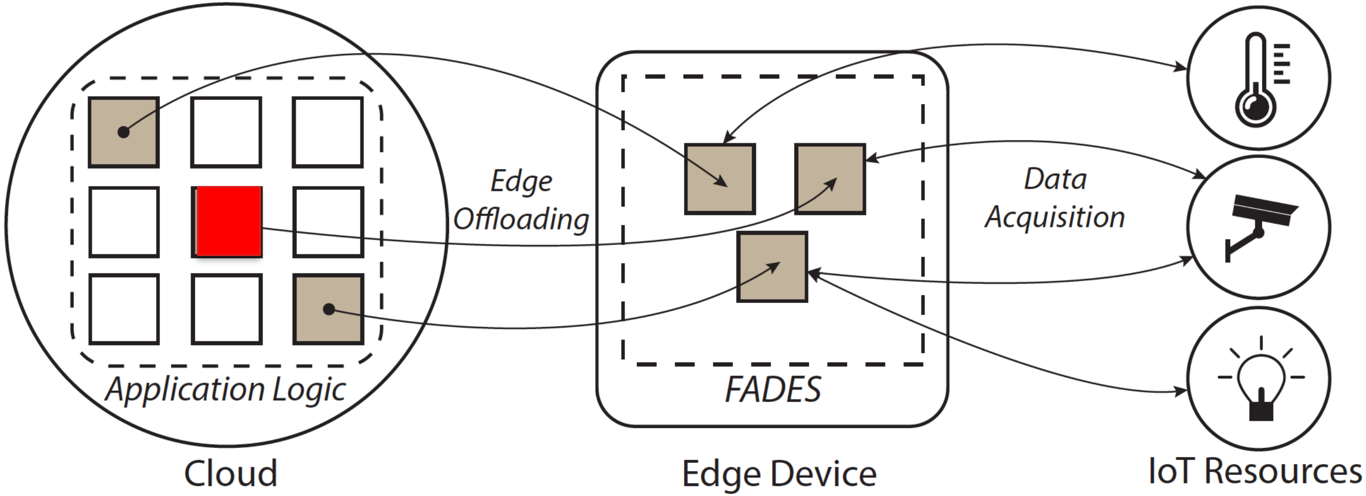
Edge Paradigm

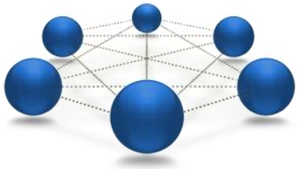


Edge

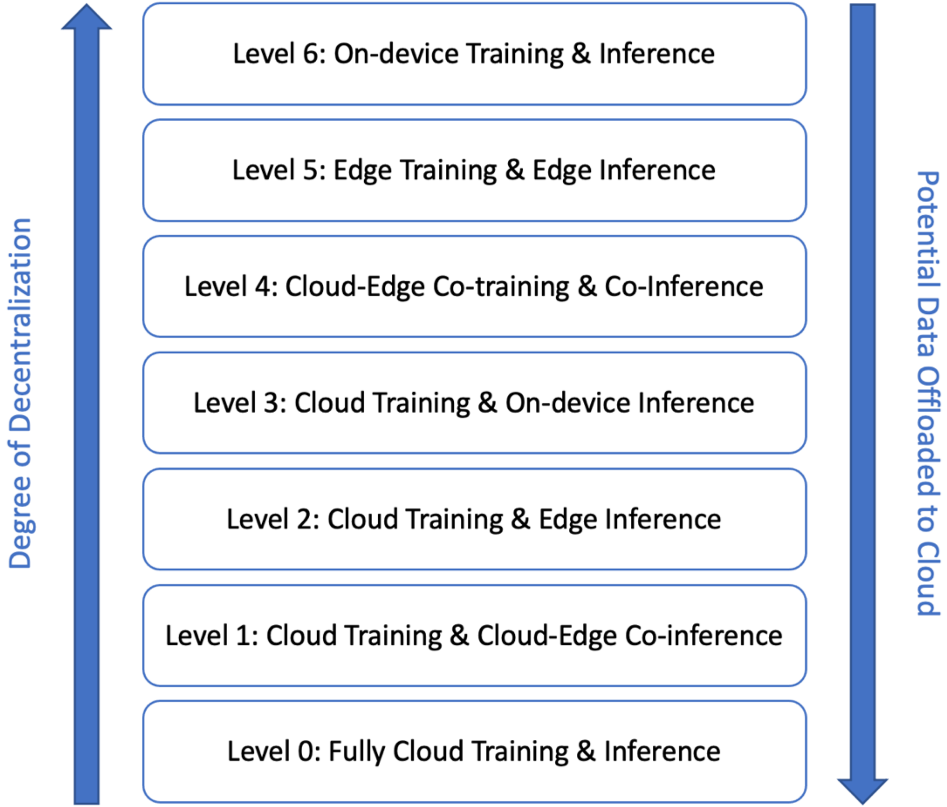
IoT

Example


Edge Offloading
Data Acquisition
Application Logic
FADES
Cloud
Edge Device
IoT Resources



Continuum Perspective



Degree of Decentralization

Level 6: On-device Training & Inference

Level 5: Edge Training & Edge Inference

Level 4: Cloud-Edge Co-training & Co-Inference

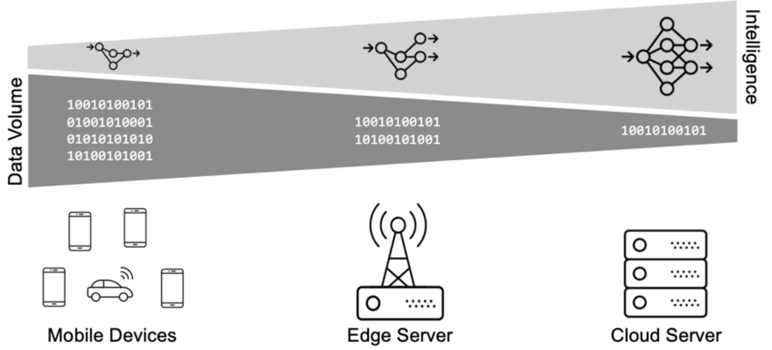
Level 3: Cloud Training & On-device Inference

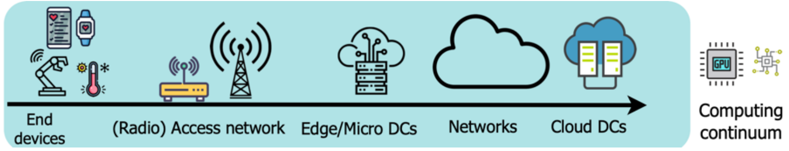
Level 2: Cloud Training & Edge Inference

Level 1: Cloud Training & Cloud-Edge Co-inference

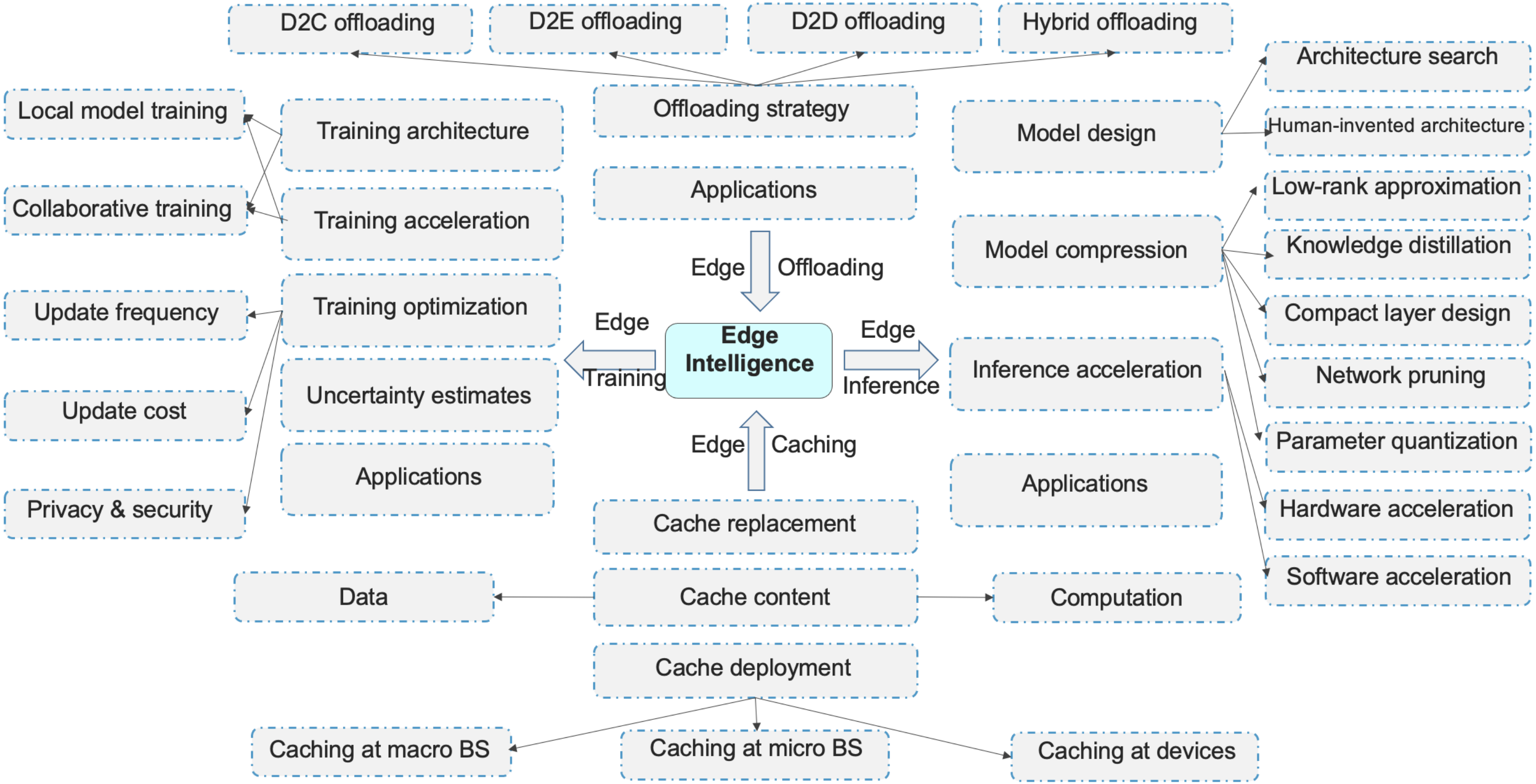
Level 0: Fully Cloud Training & Inference

Potential Data Offloaded to Cloud





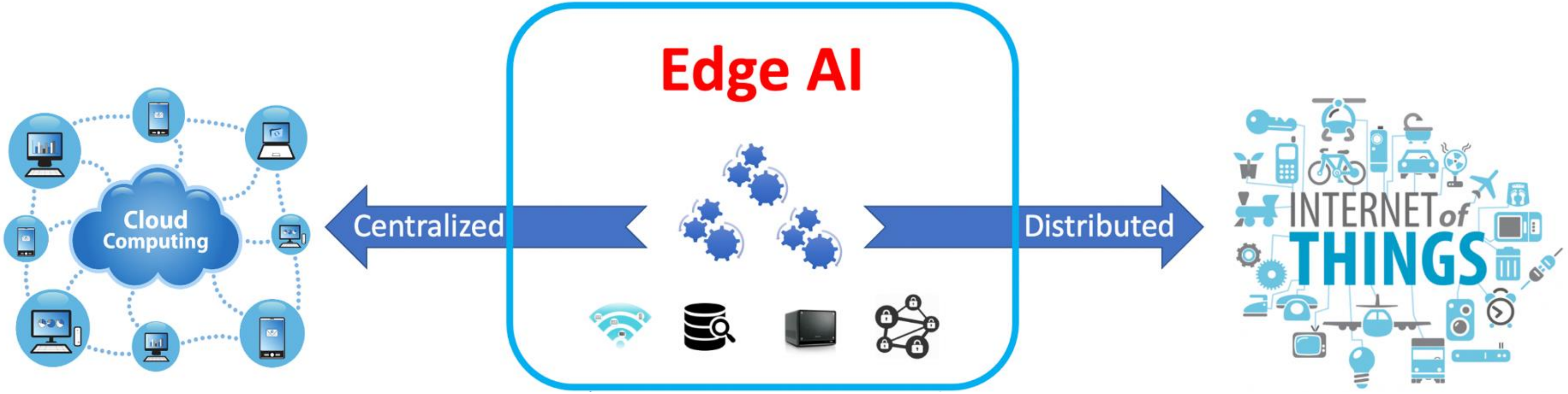
Enabling Techniques



“Edge Intelligence: Empowering Intelligence to the Edge of Network”, Proc of IEEE, Vol 109, No 11, 2021

Motivation

Bridge the Gap



Consolidate Cloud & IoT

Case: Crowd Intelligence on Edge

- **Societal impact** of past years
  - Responding and coping with emergency/pandemics
  - Urban activity/mobility sensing on the edge

Motivation

Gap: high fidelity entails high cost, infrastructure dependency, privacy intervention

- Low cost and scalable
  - User equipment
  - Deployment and coverage
- Passive (non-intrusive)
  - No need to force user interactions nor mandatory engagement
- Privacy-aware/friendly
  - Balance fidelity and data (local) regulations

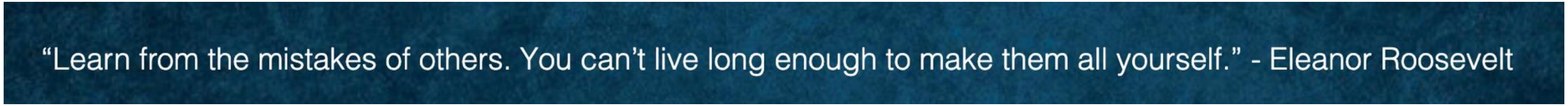
Unexpected

- Project ends…
  - Regulatory and legal considerations
  - Privacy in local context

Lessons

- Boundary + Awareness
  - Privacy on Edge? regulations and legal
  - Difference across countries

Where is my Tag? Unveiling Alternative Uses of Apple FindMy Service

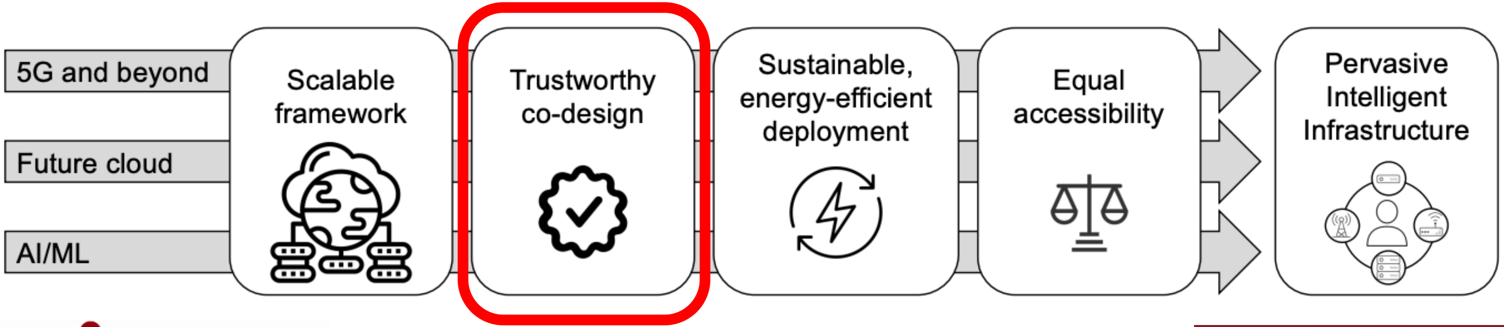


"Learn from the mistakes of others. You can't live long enough to make them all yourself." - Eleanor Roosevelt

Roadmap

Roadmap for Edge AI: A Dagstuhl Perspective

Aaron Yi Ding[1]*, Ella Peltonen[2], Tobias Meuser[3], Atakan Aral[4], Christian Becker[5], Schahram Dustdar[6], Thomas Hiessl[6], Dieter Kranzlmüller[7], Madhusanka Liyanage[8], Setareh Magshudi[9], Nitinder Mohan[10], Jörg Ott[10], Jan S. Rellermeyer[11,1], Stefan Schulte[12], Henning Schulzrinne[13], Gürkan Solmaz[14], Sasu Tarkoma[15], Blesson Varghese[16], Lars Wolf[17]

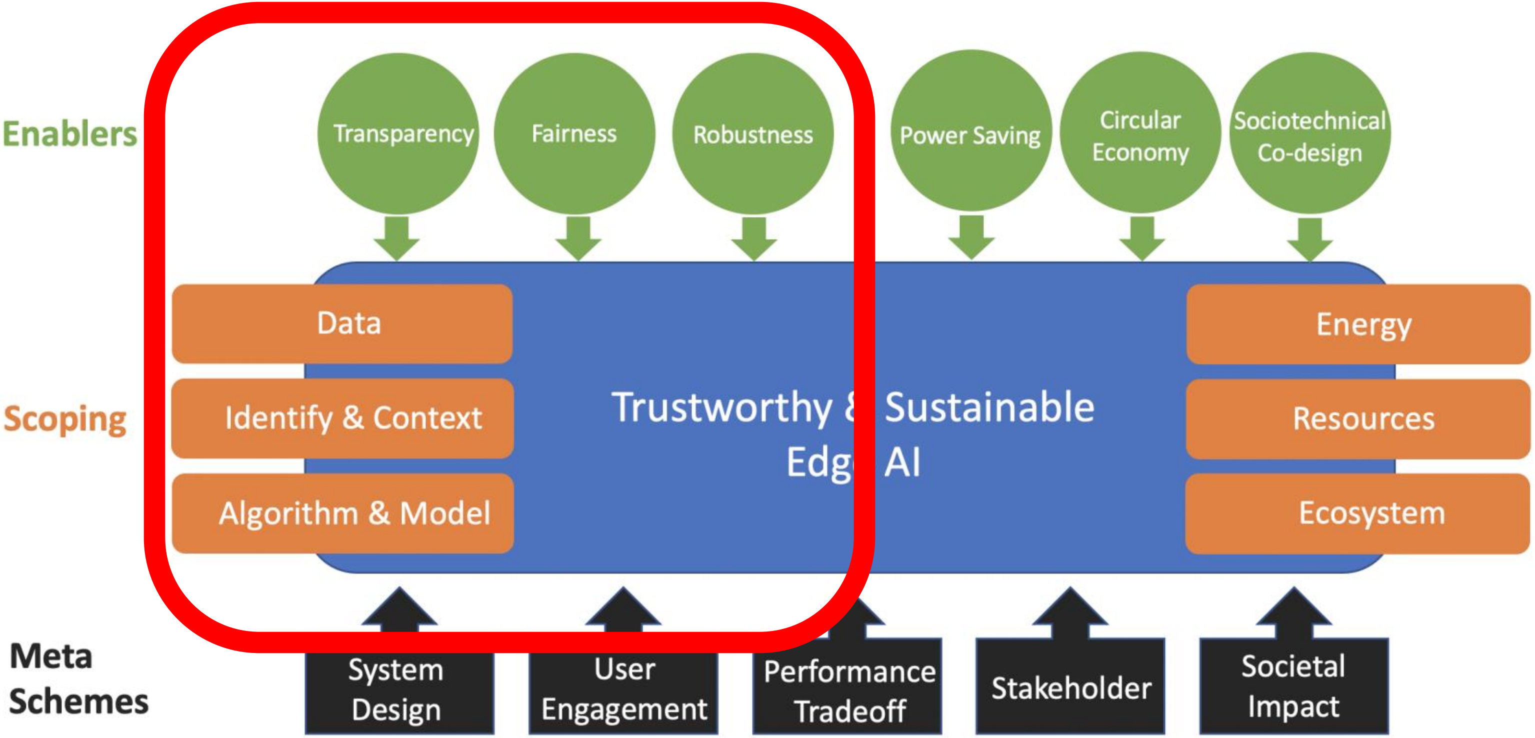


acm sigcomm

ACM SIGCOMM CCR, Vol. 52, No.1, 2022

COMPUTER COMMUNICATION REVIEW

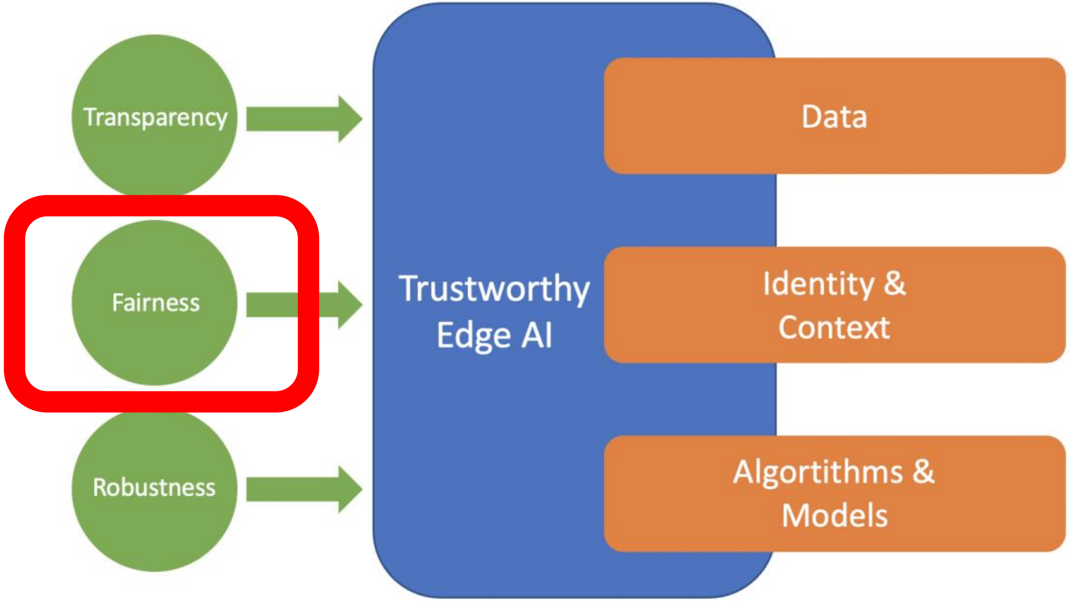
Research Agenda



Aaron Ding, Marijn Janssen, Jon Crowcroft. "Trustworthy and Sustainable Edge AI: A Research Agenda", IEEE TPS

Targets

- Enabler 🟢
  - Transparency
  - Fairness
  - Robustness
- Scope 🟠
  - Data
  - Identify & Context
  - Algorithm & Model



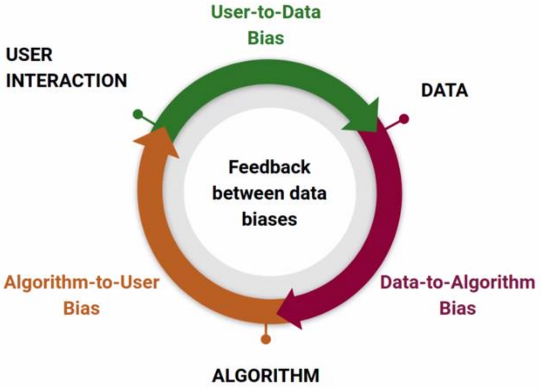
Aaron Ding, Marijn Janssen, Jon Crowcroft
“Trustworthy and Sustainable Edge AI: A Research Agenda”

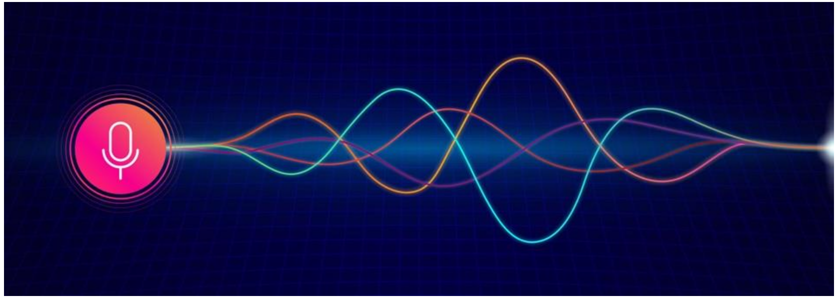
Trustworthy Edge AI

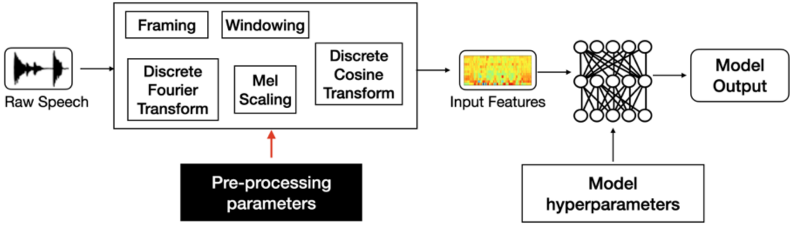
How ?

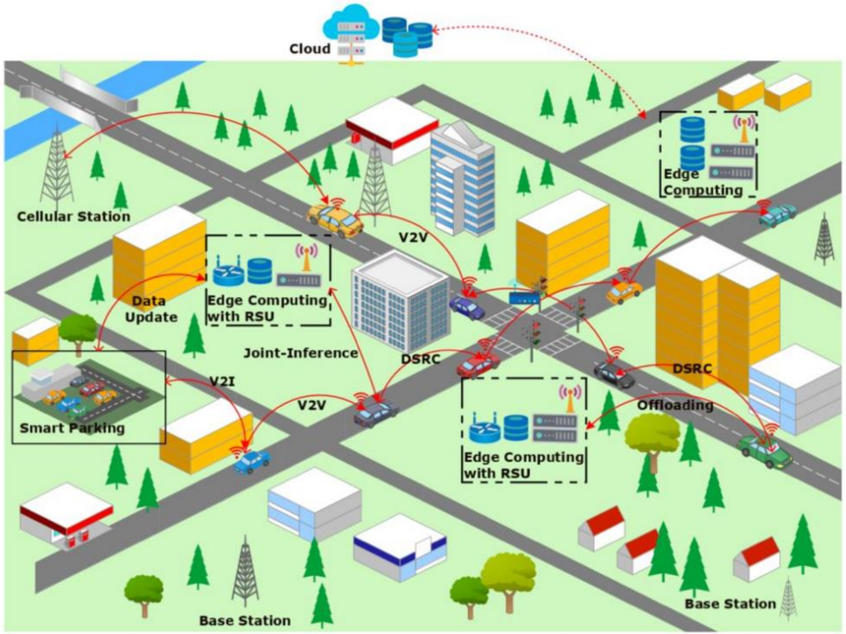
Academic Case Studies

- Voice-activated services
- Vehicular services









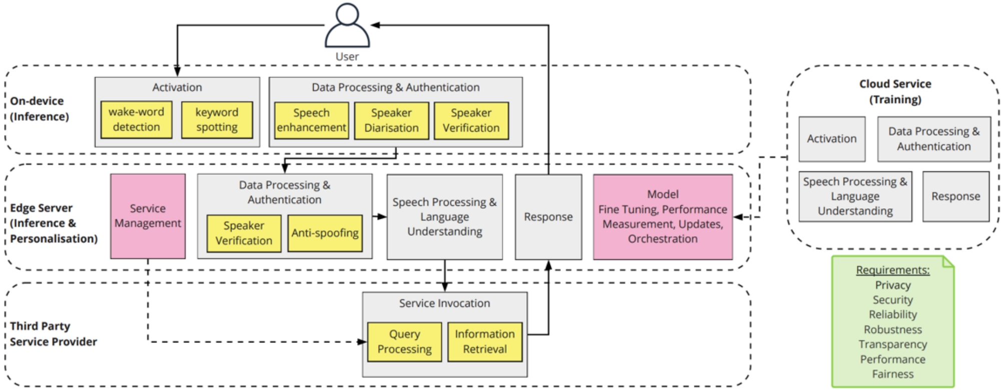
Case for Voice

- Poor situation…
- *“Bias exists at every development stage in the well-known VoxCeleb Speaker Recognition Challenge: model building, implementation, and data generation”*
- *“Most affected are female speakers & non-US nationalities, who experience significant performance degradation.”*

“Bias Propagation in On-device ML”
ACM TOSEM 2023

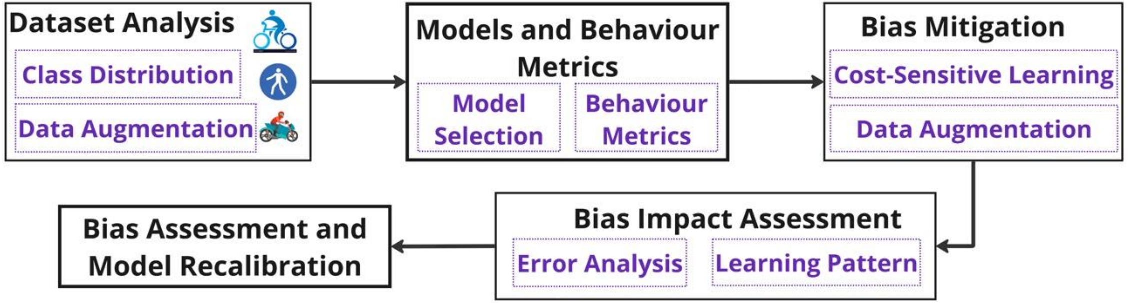
“Bias in Automated Speaker Recognition”
ACM FAccT 2022

“Characterising the Role of Pre-Processing Parameters in Audio-based Embedded Machine Learning”
ACM SenSys 2021



Case for Cars

- Poor situation too …
- *“Biased-car dataset leads to algorithmic bias, e.g., towards pedestrians and cyclists”*
- *“Poor data diversity… Vulnerable classes (e.g., pedestrians and cyclists) generally have less representation within the dataset”*



Dataset Analysis

Class Distribution

Data Augmentation

Models and Behaviour Metrics

Model Selection

Behaviour Metrics

Bias Mitigation

Cost-Sensitive Learning

Data Augmentation

Bias Impact Assessment

Error Analysis

Learning Pattern

Bias Assessment and Model Recalibration

“Bias Detection and Generalization in AI Algorithms on Edge for Autonomous Driving”
ACM/IEEE SEC 2022

"Approximate Edge AI for Energy Efficient Autonomous Driving Services"
IEEE COMST 2023

"Adaptive Approximate Computing in Edge AI and IoT Applications"
Elsevier JSA 2024

Not Enough…

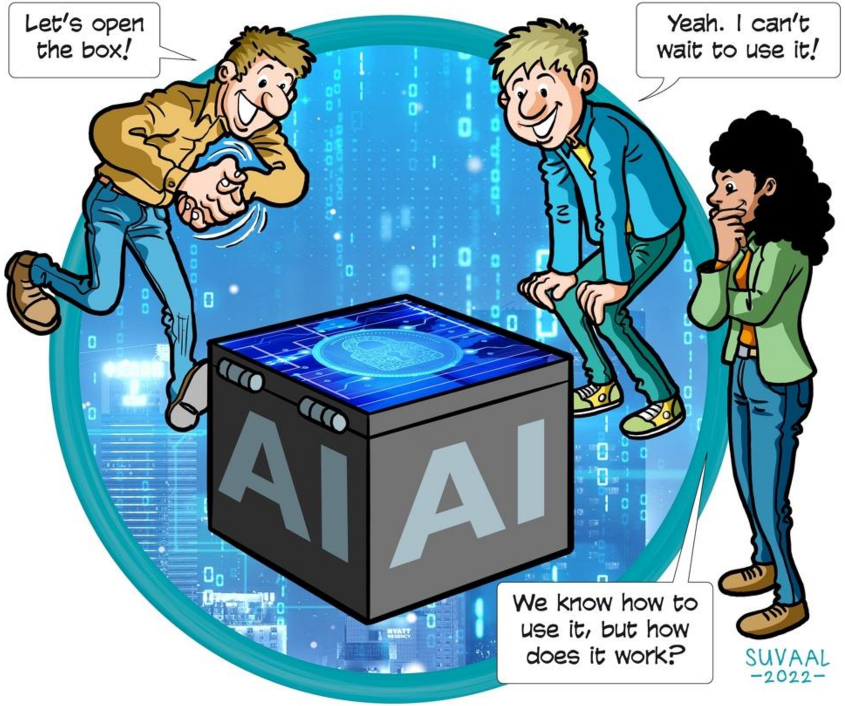
Real case please

Real Cases

Industry + Academia

EU Horizon Project

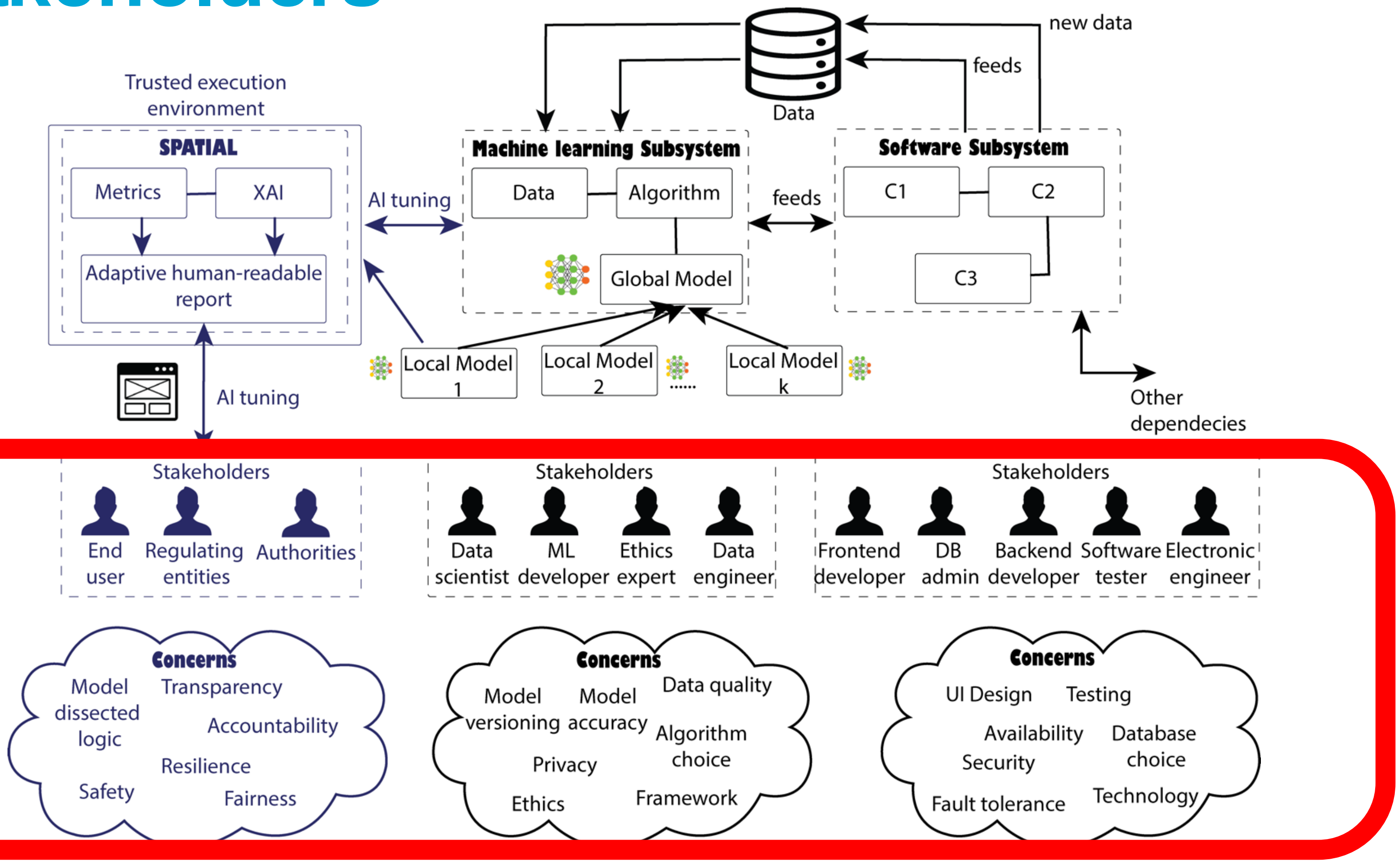
Score: 98/100 | Rate: 8%


Let's open the box!
Yeah. I can't wait to use it!
AI
AI
We know how to use it, but how does it work?
SUVAAL -2022-

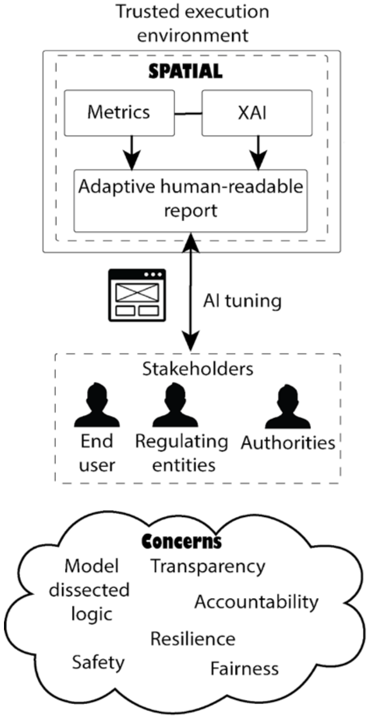
Funded by the European Union

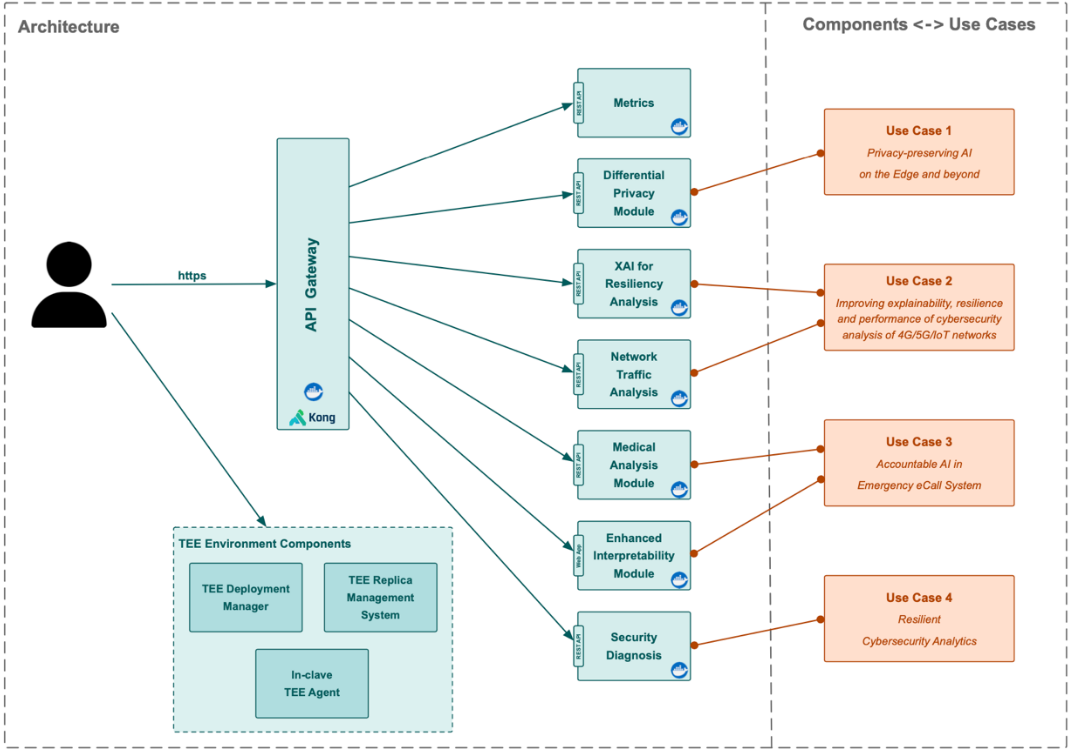
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

Stakeholders



Trust as a Service





Architecture
Components <-> Use Cases
https
API Gateway
Kong
Metrics
Differential Privacy Module
XAI for Resiliency Analysis
Network Traffic Analysis
Medical Analysis Module
Enhanced Interpretability Module
Security Diagnosis
REST API
Web App
TEE Environment Components
TEE Deployment Manager
TEE Replica Management System
In-clave TEE Agent
Use Case 1
Privacy-preserving AI on the Edge and beyond
Use Case 2
Improving explainability, resilience and performance of cybersecurity analysis of 4G/5G/IoT networks
Use Case 3
Accountable AI in Emergency eCall System
Use Case 4
Resilient Cybersecurity Analytics

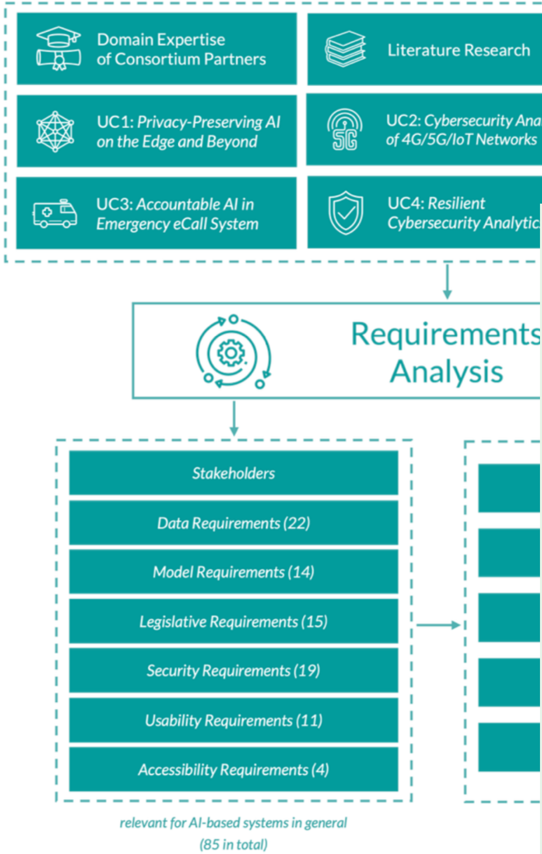
Telefónica

montimage

Fraunhofer FOKUS

WITH secure

Know-How



Domain Expertise of Consortium Partners

Literature Research

Workshops

UC1: Privacy-Preserving AI on the Edge and Beyond

UC2: Cybersecurity Analysis of 4G/5G/IoT Networks

Meetings

UC3: Accountable AI in Emergency eCall System

UC4: Resilient Cybersecurity Analytic

Requirements Analysis

Stakeholders

Data Requirements (22)

Model Requirements (14)

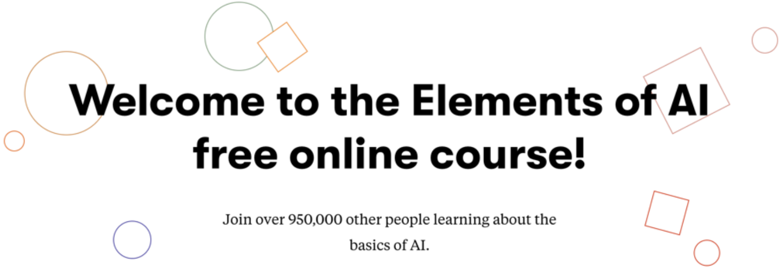
Legislative Requirements (15)

Security Requirements (19)

Usability Requirements (11)

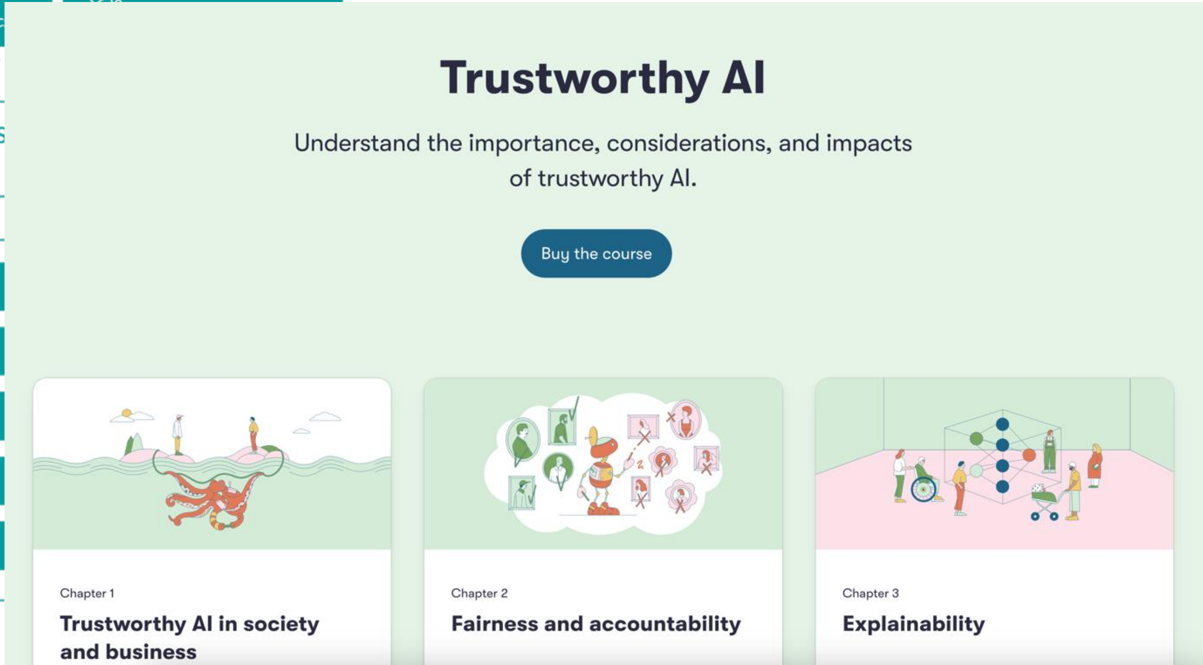
Accessibility Requirements (4)

relevant for AI-based systems in general (85 in total)



Welcome to the Elements of AI free online course!

Join over 950,000 other people learning about the basics of AI.



Trustworthy AI

Understand the importance, considerations, and impacts of trustworthy AI.

Chapter 1

Trustworthy AI in society and business

Chapter 2

Fairness and accountability

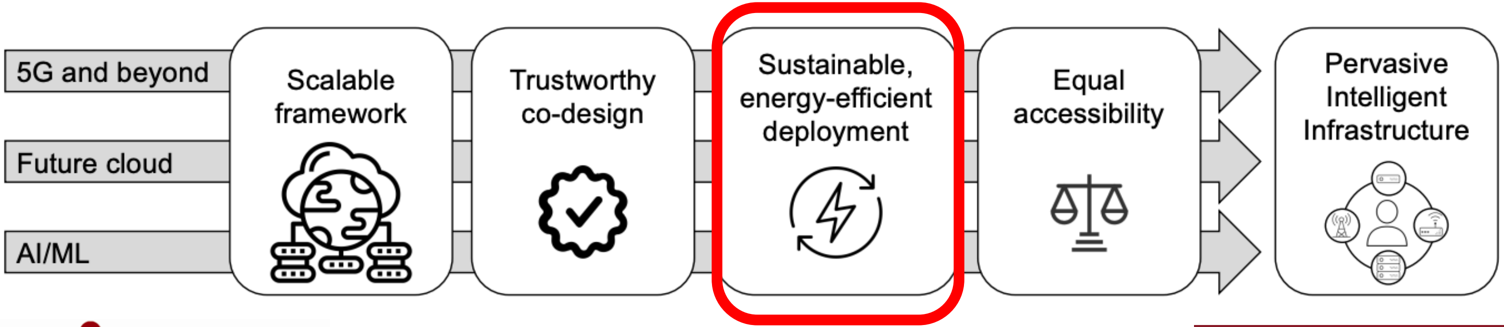
Chapter 3

Explainability

Roadmap for Future

Roadmap for Edge AI: A Dagstuhl Perspective

Aaron Yi Ding[1*], Ella Peltonen[2], Tobias Meuser[3], Atakan Aral[4], Christian Becker[5], Schahram Dustdar[6], Thomas Hiessl[6], Dieter Kranzlmüller[7], Madhusanka Liyanage[8], Setareh Magshudi[9], Nitinder Mohan[10], Jörg Ott[10], Jan S. Rellermeyer[11,1], Stefan Schulte[12], Henning Schulzrinne[13], Gürkan Solmaz[14], Sasu Tarkoma[15], Blesson Varghese[16], Lars Wolf[17]



5G and beyond

Future cloud

AI/ML

Scalable framework

Trustworthy co-design

Sustainable, energy-efficient deployment

Equal accessibility

Pervasive Intelligent Infrastructure

acm sigcomm

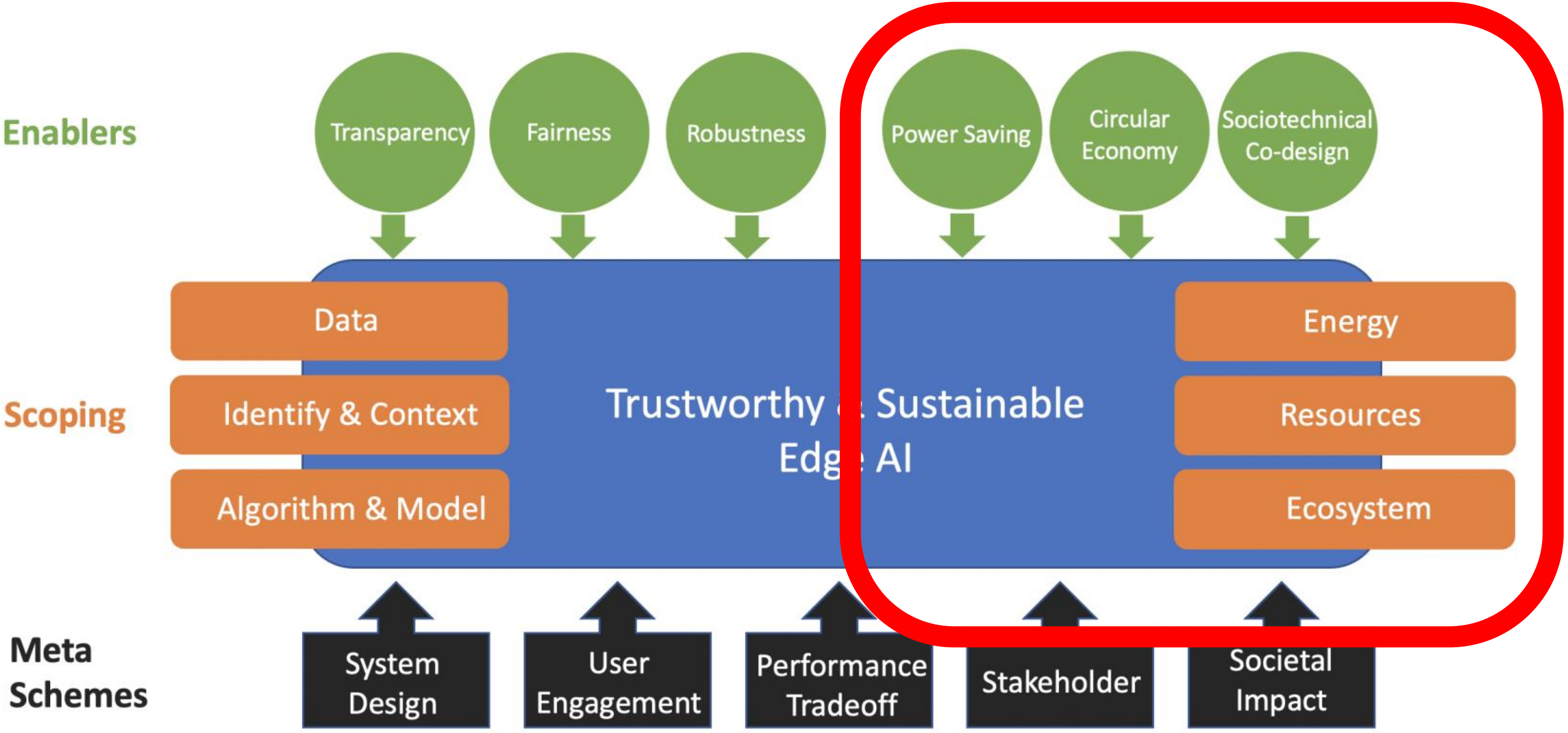
ACM SIGCOMM CCR, Vol. 52, No.1, 2022

COMPUTER COMMUNICATION REVIEW

Sustainable Edge AI

How ?

Research Agenda



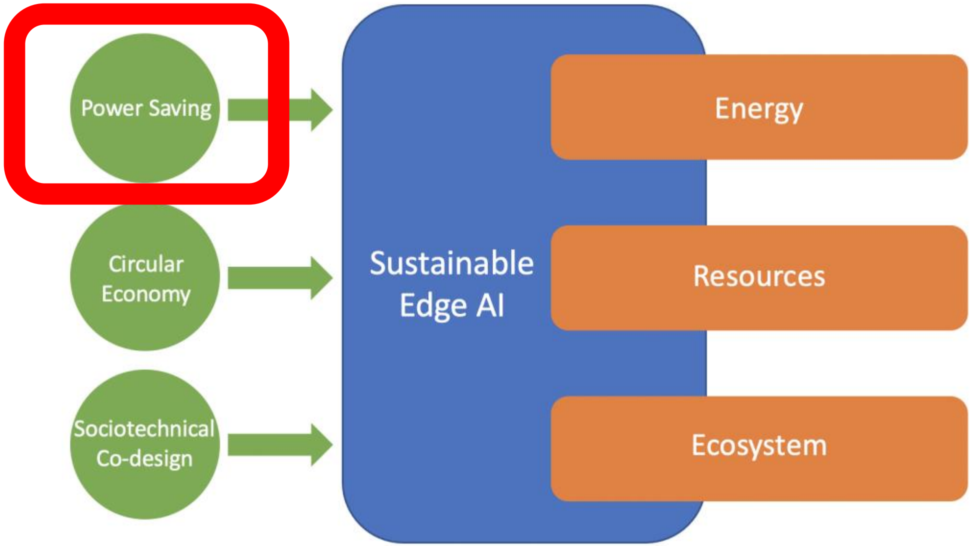
Aaron Ding, Marijn Janssen, Jon Crowcroft. "Trustworthy and Sustainable Edge AI: A Research Agenda", IEEE TPS

Sustainable is not a slogan

- Energy optimization for Edge AI
  - Full pipeline: data acquisition, transfer, computation, storage

Enabler 🟢

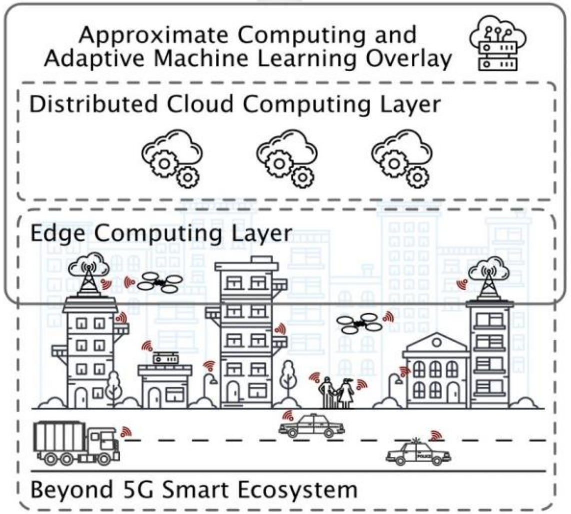
Scope 🟠



Aaron Ding, Marijn Janssen, Jon Crowcroft, “Trustworthy and Sustainable Edge AI: A Research Agenda”

Sustainable Edge AI

- EU Marie Curie ITN: grant of €4M
- 15 Marie Curie Researchers
- 20+ industrial and academic partners


Approximate Computing and Adaptive Machine Learning Overlay

Distributed Cloud Computing Layer

Edge Computing Layer

Beyond 5G Smart Ecosystem

APROPOS Project

Sustainable AI

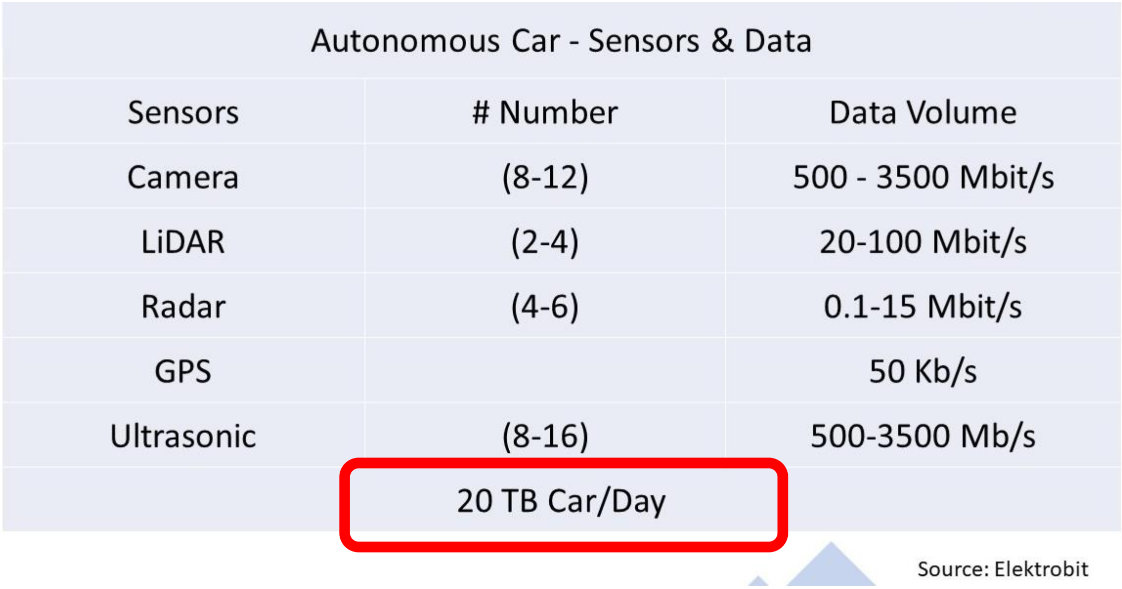
Score: 14.5/15 Rate: 3%

Case of Future Cars

Vehicular Data

- Data increases
- Electric cars: battery life matters!

750MB per second, as Google's driverless car prototype reported



Autonomous Car - Sensors & Data		
Sensors	# Number	Data Volume
Camera	(8-12)	500 - 3500 Mbit/s
LiDAR	(2-4)	20-100 Mbit/s
Radar	(4-6)	0.1-15 Mbit/s
GPS		50 Kb/s
Ultrasonic	(8-16)	500-3500 Mb/s
	20 TB Car/Day	

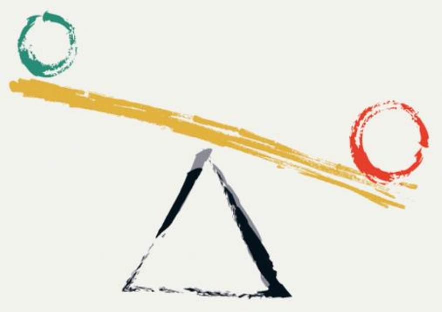
Source: Elektrobit

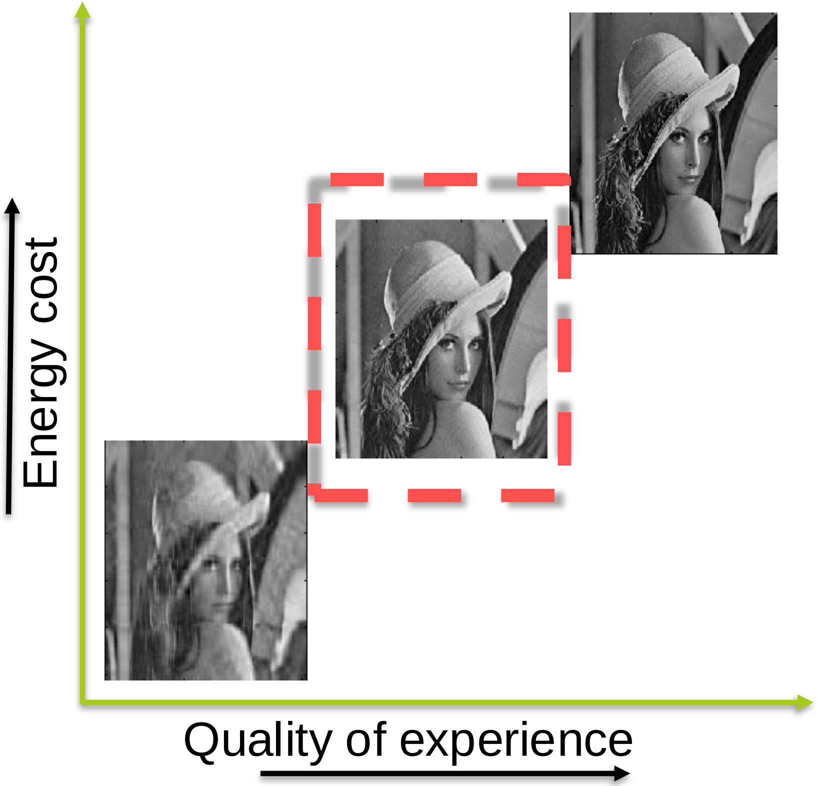
Energy Awareness

Sweet spot:

- Tolerance on accuracy & latency
- Less safety critical
- Relax the accuracy but still with acceptable experience

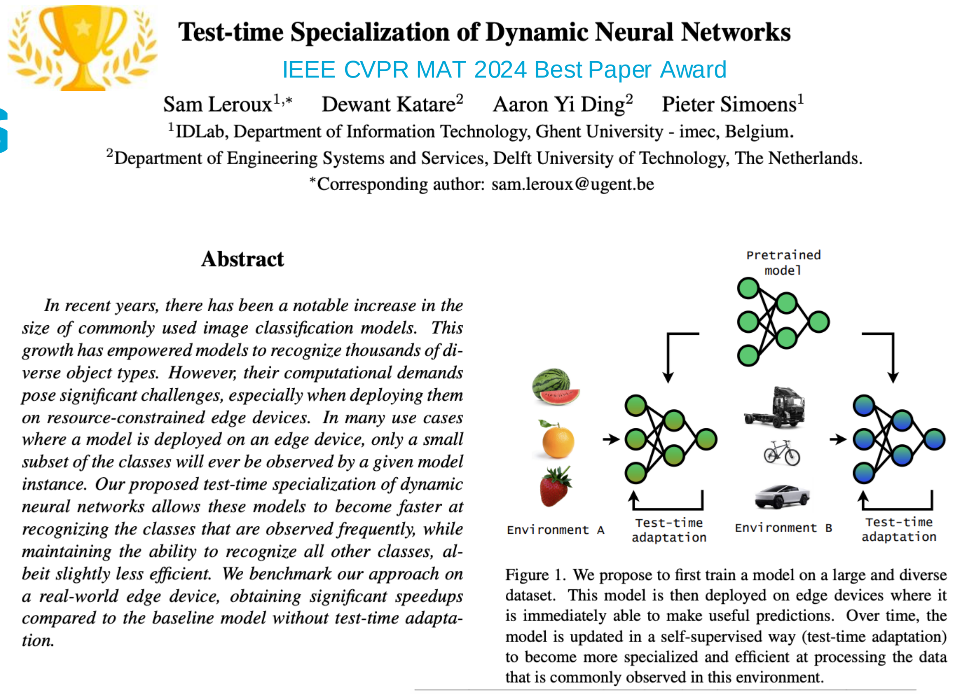
How good is good enough?





Energy Awareness

- Adaptive Approximation for



Test-time Specialization of Dynamic Neural Networks

IEEE CVPR MAT 2024 Best Paper Award

Sam Leroux[1,*] Dewant Katare[2] Aaron Yi Ding[2] Pieter Simoens[1]

[1] IDLab, Department of Information Technology, Ghent University - imec, Belgium.

[2] Department of Engineering Systems and Services, Delft University of Technology, The Netherlands.

*Corresponding author: sam.leroux@ugent.be

Abstract

In recent years, there has been a notable increase in the size of commonly used image classification models. This growth has empowered models to recognize thousands of diverse object types. However, their computational demands pose significant challenges, especially when deploying them on resource-constrained edge devices. In many use cases where a model is deployed on an edge device, only a small subset of the classes will ever be observed by a given model instance. Our proposed test-time specialization of dynamic neural networks allows these models to become faster at recognizing the classes that are observed frequently, while maintaining the ability to recognize all other classes, albeit slightly less efficient. We benchmark our approach on a real-world edge device, obtaining significant speedups compared to the baseline model without test-time adaptation.

Figure 1. We propose to first train a model on a large and diverse dataset. This model is then deployed on edge devices where it is immediately able to make useful predictions. Over time, the model is updated in a self-supervised way (test-time adaptation) to become more specialized and efficient at processing the data that is commonly observed in this environment.

“Approximate Edge AI for Energy Efficient Autonomous Driving Services”
IEEE COMST 2023 Impact Factor 35,6

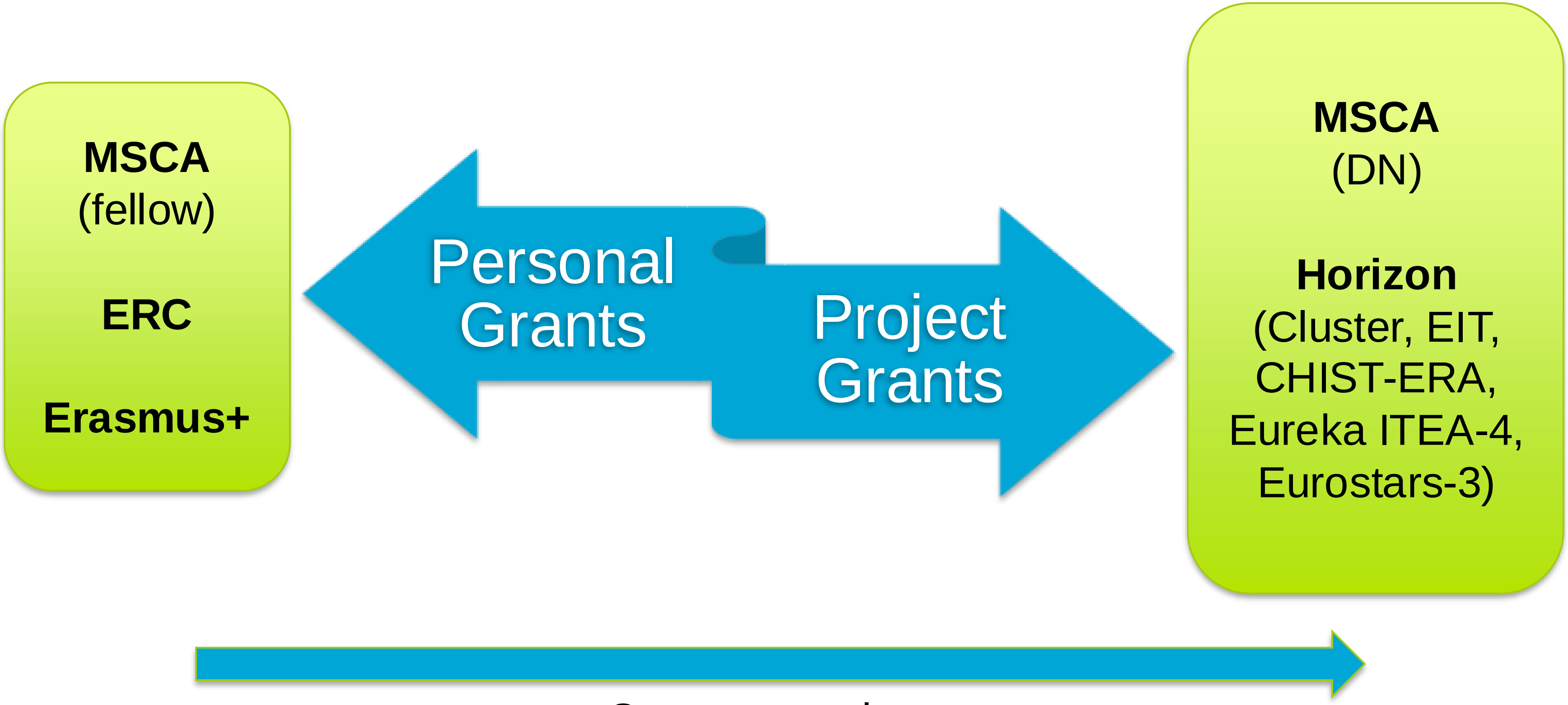
“Nimbus: Towards Latency-Energy Efficient Task Offloading for AR Services”
IEEE Transactions on Cloud Computing 2022

Independence… Grant

Have you ever applied ?

How many times / much ☺

Binary View



MSCA
(fellow)

ERC

Erasmus+

Personal Grants

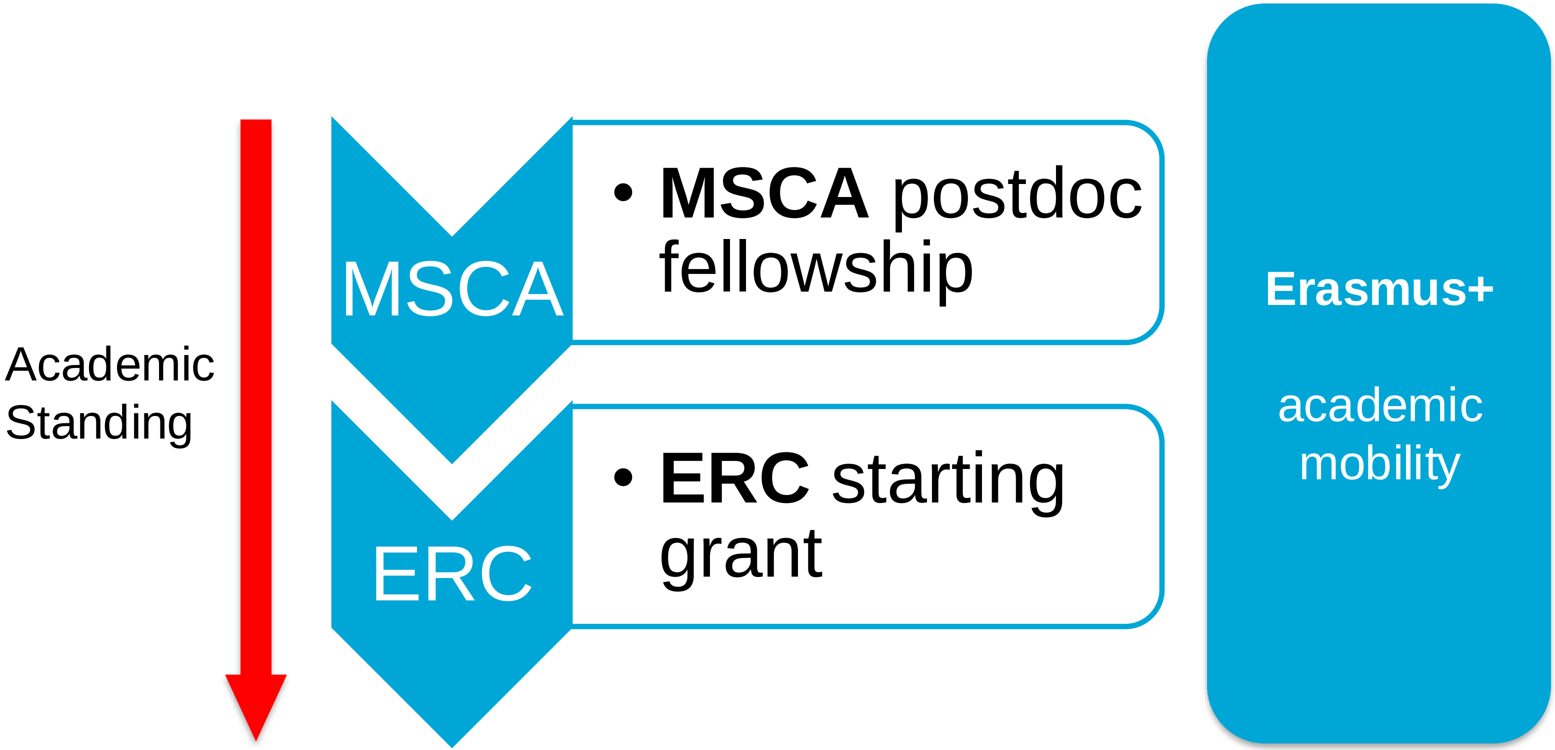
Project Grants

MSCA
(DN)

Horizon
(Cluster, EIT, CHIST-ERA, Eureka ITEA-4, Eurostars-3)

Career maturity

Personal Grants



Academic Standing

MSCA

- **MSCA** postdoc fellowship

ERC

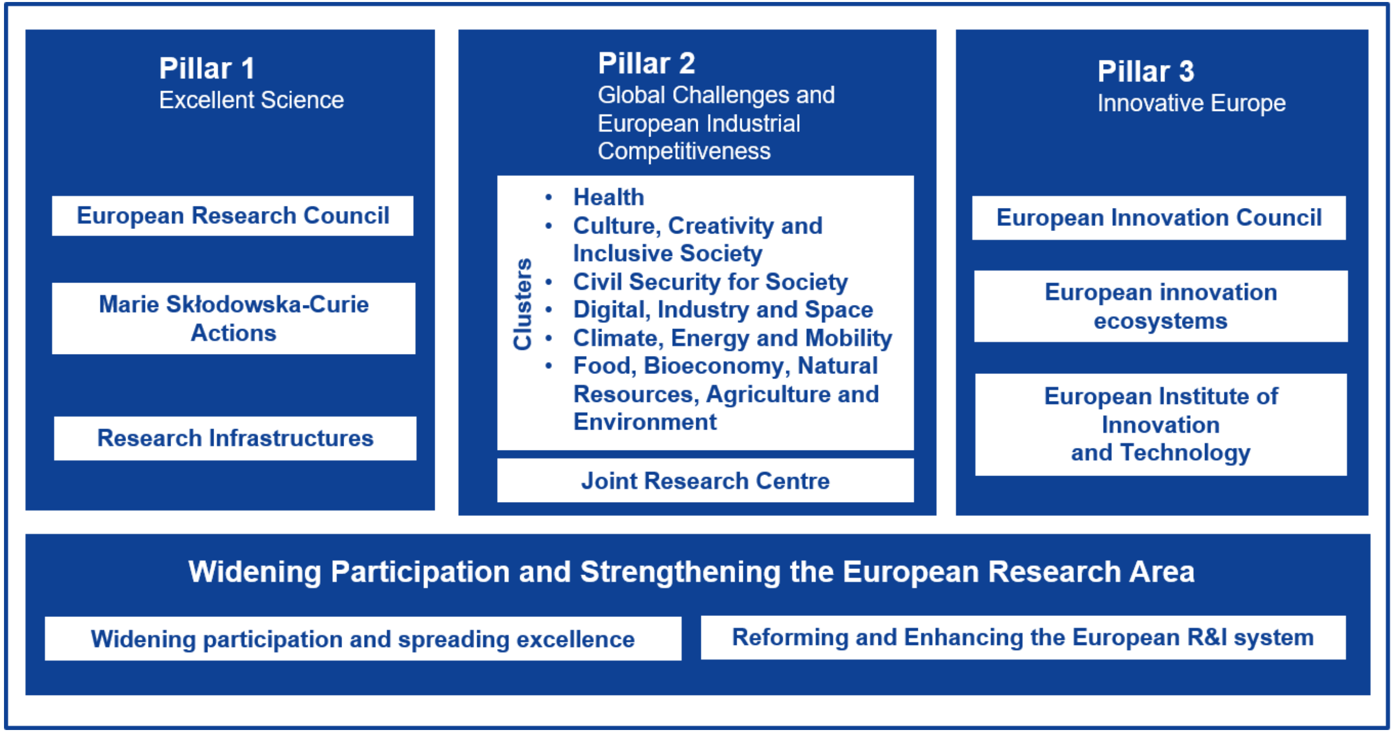
- **ERC** starting grant

Erasmus+

academic mobility

Horizon Europe

- Time frame 2021-2027



Pillar 1
Excellent Science

- European Research Council
- Marie Skłodowska-Curie Actions
- Research Infrastructures

Pillar 2
Global Challenges and European Industrial Competitiveness

Clusters
- Health
- Culture, Creativity and Inclusive Society
- Civil Security for Society
- Digital, Industry and Space
- Climate, Energy and Mobility
- Food, Bioeconomy, Natural Resources, Agriculture and Environment

Joint Research Centre

Pillar 3
Innovative Europe

- European Innovation Council
- European innovation ecosystems
- European Institute of Innovation and Technology

Widening Participation and Strengthening the European Research Area

- Widening participation and spreading excellence
- Reforming and Enhancing the European R&I system

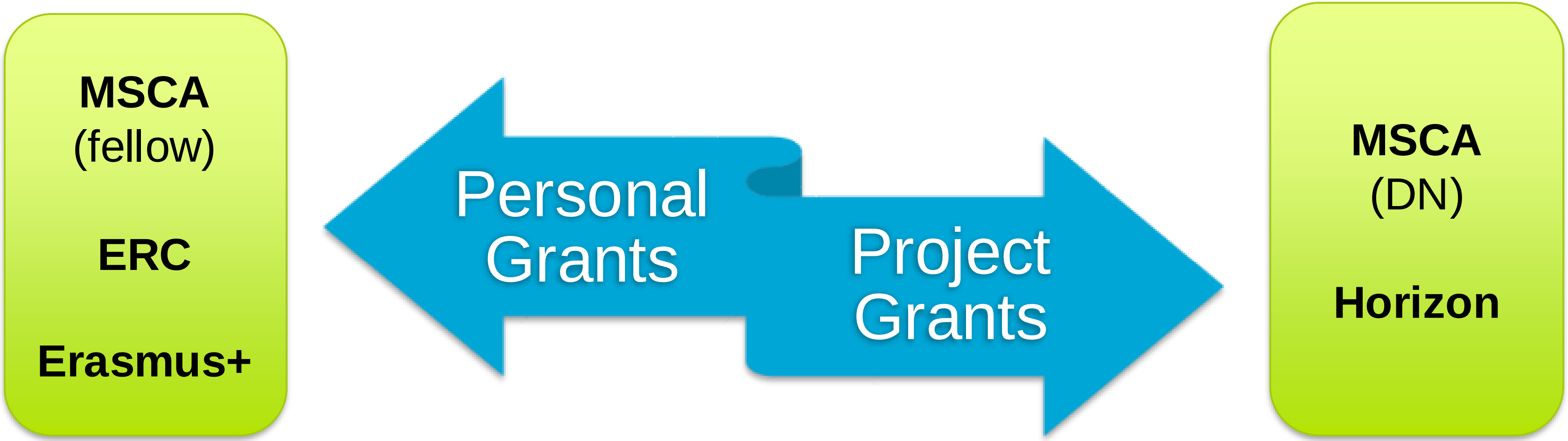
Takeaway

Academic Freedom

Independence

Leadership

Industrial Connections



MSCA
(fellow)

ERC

Erasmus+

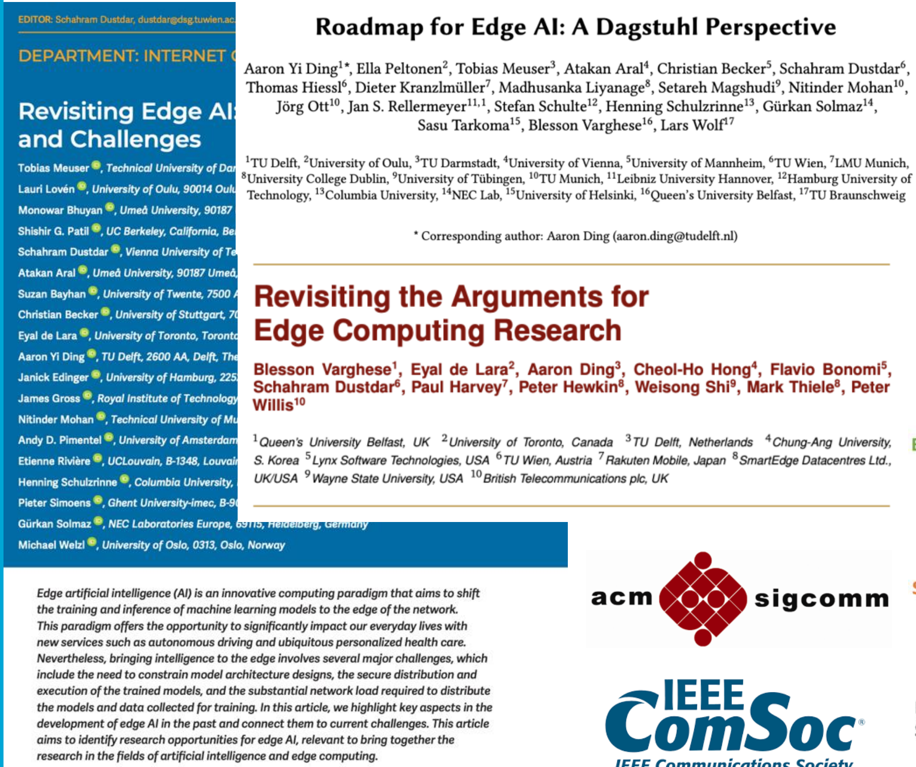
Personal Grants

Project Grants

MSCA
(DN)

Horizon

Takeaway



EDITOR: Schahram Dustdar, dustdar@dsg.tuwien.ac

DEPARTMENT: INTERNET C

Revisiting Edge AI:
and Challenges

Tobias Meuser, Technical University of Dar
Lauri Lovén, University of Oulu, 90014 Oulu
Monowar Bhuyan, Umeå University, 90187
Shishir G. Patil, UC Berkeley, California, Be
Schahram Dustdar, Vienna University of Te
Atakan Aral, Umeå University, 90187 Umeå,
Suzan Bayhan, University of Twente, 7500 A
Christian Becker, University of Stuttgart, 70
Eyal de Lara, University of Toronto, Toronto
Aaron Yi Ding, TU Delft, 2600 AA, Delft, The
Janick Edinger, University of Hamburg, 225
James Gross, Royal Institute of Technology
Nitinder Mohan, Technical University of Mu
Andy D. Pimentel, University of Amsterdam
Etienne Rivière, UCLouvain, B-1348, Louvain
Henning Schulzrinne, Columbia University,
Pieter Simoens, Ghent University-imec, B-90
Gürkan Solmaz, NEC Laboratories Europe, 69115, Heidelberg, Germany
Michael Welzl, University of Oslo, 0313, Oslo, Norway

Edge artificial intelligence (AI) is an innovative computing paradigm that aims to shift the training and inference of machine learning models to the edge of the network. This paradigm offers the opportunity to significantly impact our everyday lives with new services such as autonomous driving and ubiquitous personalized health care. Nevertheless, bringing intelligence to the edge involves several major challenges, which include the need to constrain model architecture designs, the secure distribution and execution of the trained models, and the substantial network load required to distribute the models and data collected for training. In this article, we highlight key aspects in the development of edge AI in the past and connect them to current challenges. This article aims to identify research opportunities for edge AI, relevant to bring together the research in the fields of artificial intelligence and edge computing.

Roadmap for Edge AI: A Dagstuhl Perspective

Aaron Yi Ding[1]*, Ella Peltonen[2], Tobias Meuser[3], Atakan Aral[4], Christian Becker[5], Schahram Dustdar[6], Thomas Hiessl[6], Dieter Kranzlmüller[7], Madhusanka Liyanage[8], Setareh Magshudi[9], Nitinder Mohan[10], Jörg Ott[10], Jan S. Rellermeyer[11,1], Stefan Schulte[12], Henning Schulzrinne[13], Gürkan Solmaz[14], Sasu Tarkoma[15], Blesson Varghese[16], Lars Wolf[17]

[1]TU Delft, [2]University of Oulu, [3]TU Darmstadt, [4]University of Vienna, [5]University of Mannheim, [6]TU Wien, [7]LMU Munich, [8]University College Dublin, [9]University of Tübingen, [10]TU Munich, [11]Leibniz University Hannover, [12]Hamburg University of Technology, [13]Columbia University, [14]NEC Lab, [15]University of Helsinki, [16]Queen's University Belfast, [17]TU Braunschweig

* Corresponding author: Aaron Ding (aaron.ding@tudelft.nl)

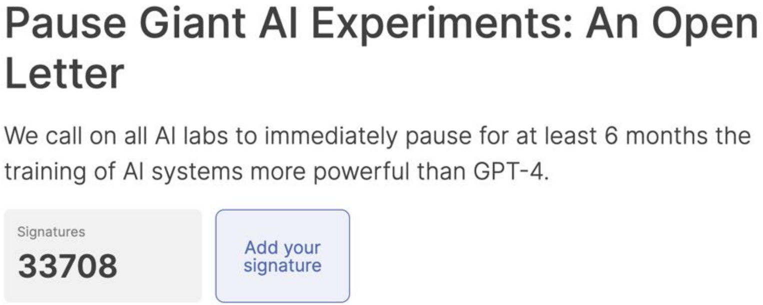
Revisiting the Arguments for Edge Computing Research

Blesson Varghese[1], Eyal de Lara[2], Aaron Ding[3], Cheol-Ho Hong[4], Flavio Bonomi[5], Schahram Dustdar[6], Paul Harvey[7], Peter Hewkin[8], Weisong Shi[9], Mark Thiele[8], Peter Willis[10]

[1]*Queen's University Belfast, UK* [2]*University of Toronto, Canada* [3]*TU Delft, Netherlands* [4]*Chung-Ang University, S. Korea* [5]*Lynx Software Technologies, USA* [6]*TU Wien, Austria* [7]*Rakuten Mobile, Japan* [8]*SmartEdge Datacentres Ltd., UK/USA* [9]*Wayne State University, USA* [10]*British Telecommunications plc, UK*

acm sigcomm

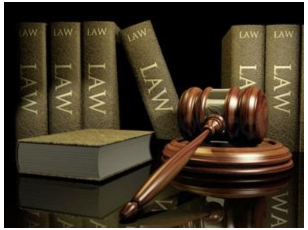
IEEE ComSoc
IEEE Communications Society

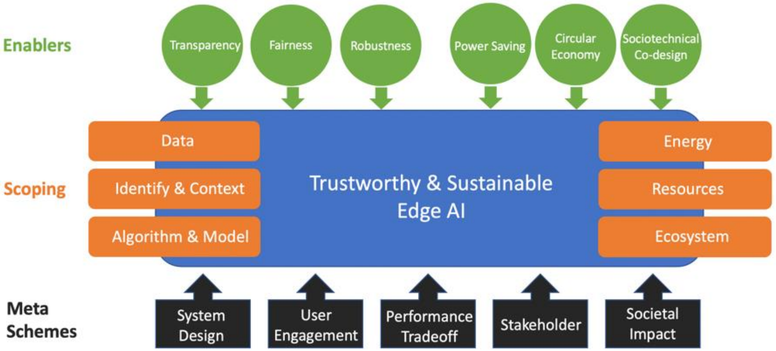


Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures
33708





"Trustworthy and Sustainable Edge AI: A Research Agenda"

Outlook

EdgeSys 2025

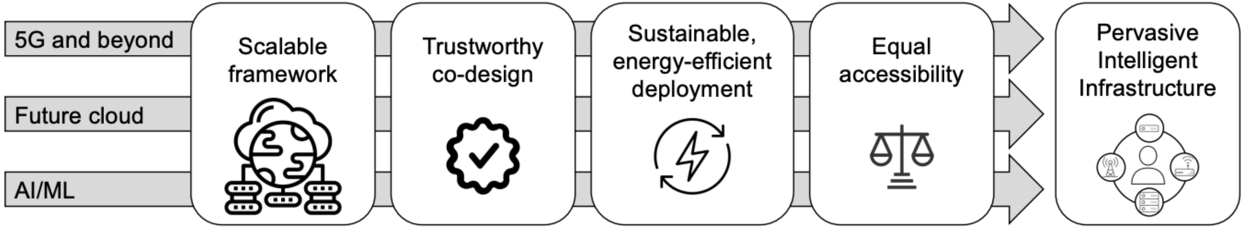
@ Rotterdam, NL

acm Association for Computing Machinery

IEEE Advancing Technology for Humanity

SEC in Rome!

ACM/IEEE Symposium on Edge Computing



acm sigcomm

ACM SIGCOMM CCR, Vol. 52, No.1

COMPUTER COMMUNICATION REVIEW