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

Revisiting Edge AI: Opportunities and Challenges

Tobias Meuser, *Technical University of Darmstadt, 64283, Darmstadt, Germany*

Lauri Lovén, *University of Oulu, 90014 Oulu, Finland*

Monowar Bhuyan, *Umeå University, 90187 Umeå, Sweden*

Shishir G. Patil, *UC Berkeley, California, Berkeley, CA, 94709, USA*

Schahram Dustdar, *Vienna University of Technology, Vienna 1040, Austria*

Atakan Aral, *Umeå University, 90187 Umeå, Sweden and University of Vienna, Vienna 1090, Austria*

Suzan Bayhan, *University of Twente, 7500 AE, Enschede, The Netherlands*

Christian Becker, *University of Stuttgart, 70569, Stuttgart, Germany*

Eyal de Lara, *University of Toronto, Toronto, ON, M5S 1A1, Canada*

Aaron Yi Ding, *TU Delft, 2600 AA, Delft, The Netherlands*

Janick Edinger, *University of Hamburg, 22527, Hamburg, Germany*

James Gross, *Royal Institute of Technology Stockholm (KTH), 100 44 Stockholm, Sweden*

Nitinder Mohan, *Technical University of Munich, 80333, Munich, Germany*

Andy D. Pimentel, *University of Amsterdam, 1098 GH, Amsterdam, The Netherlands*

Etienne Rivière, *UCLouvain, B-1348, Louvain-la-Neuve, Belgium*

Henning Schulzrinne, *Columbia University, New York, NY, 10027, USA*

Pieter Simoens, *Ghent University-imec, B-9052, Gent, Belgium*

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.

1089-7801 © 2024 IEEE
Digital Object Identifier 10.1109/MIC.2024.3383758
Date of current version 2 August 2024.

Edge computing is a significant paradigm shift that is reshaping the internet and applications landscapes by bringing data processing closer to the source of the data. This strategic evolution has the potential to enhance efficiency, responsiveness, and the respect of privacy.[2,3,4,21] Starting from mostly cloud-based solutions, more and more applications are now pushed along the computing continuum, closer and closer to edge devices. While there have been several definitions of the latter in the past, ranging from user end devices up to small, localized data centers, the general properties of edge devices are similar: their closeness to the user and the locality of processed data.[2] While the popularity of edge solutions increased in the recent years, the deployment of edge solutions is still relatively slow compared to the growth of the cloud market. This can be attributed to the high cost of building and managing a distributed infrastructure, but also to the relative complexity of building applications for the edge compared to building them only for the cloud.

The emergence of artificial intelligence (AI) and its significant demand for training data has made the use of edge devices for training and inference a clear subsequent development.[a] The requirements of machine learning (ML) applications for vast amounts of data make, indeed, the training and inference using that data at the edge an efficient and reasonable option in comparison to a cloud-centric approach. In addition, training and inference of ML models close to or at the edge come with significant advantages for end users, including better respect for data privacy and faster response times. However, the combination of artificial intelligence with edge computing also opens further challenges, especially due to the resource constraints and availability of those edge devices. These limitations are even more evident when comparing edge devices with the robust and omnipresent cloud infrastructure. Yet, applications like autonomous driving, which demand low-latency responses as well as processing of very high-dimensional data at very high rates, vividly illustrate the necessity of edge intelligence. In such safety-critical applications, even milliseconds matter, making it essential to have access to data sources and model decisions with minimal delay. Similarly, bringing learning and inference to the edge will enable new, innovative, and useful applications such as robotics, immersive multi-user applications (augmented reality), and smart health care, revolutionizing our way of living.

While exploring the synergy of AI and edge computing, it is crucial to address the unique challenges the integration of edge computing and intelligence presents. Despite its potential, edge intelligence can be influenced by resource constraints, notably in computing and storage resources, which are in significant contrast to the capabilities of traditional cloud infrastructures. Due to these limitations, protecting data and ensuring fast response times remain significant challenges, in which today's edge computing solutions are still outperformed by pure cloud-based computing on many occasions.[21] Edge infrastructure is usually deployed in physically accessible places and cannot benefit from the perimeter-based protection measures used in cloud computing. To make the edge a real augmentation for current cloud-only solutions, future research is necessary, focusing on the security, availability, and efficiency of edge intelligence. This articles not only reviews the decade-long journey to edge AI but also critically examines the viewpoints of various stakeholders and outlines the pressing challenges and exciting future research directions in this field.

THE DECADE-LONG JOURNEY TO EDGE AI

Edge intelligence emerged as an evolution of the edge computing paradigm, whose roots are traceable to the 2000s, primarily driven by the limitations of cloud computing in handling the burgeoning data generated by local devices, e.g., the Internet of Things (IoT). Edge computing decentralizes data processing, pushing it closer to data sources at the network's edge. This proximity reduces the distance data must travel, thereby decreasing latency and conserving bandwidth. Furthermore, edge computing alleviates the data load on central servers and enhances privacy by processing sensitive data locally.[4] Edge and cloud computing can complement each other and form the so-called continuum, with edge computing addressing immediate, localized processing needs while cloud computing remains essential for large-scale data storage and extensive computational tasks.

Advent of Edge AI

Edge intelligence represents a further paradigm shift from edge computing, integrating AI to enhance the processing capabilities at the edge of the network. This integration further reduces latency and alleviates the bandwidth demand on central servers, while also providing additional benefits, such as enhanced privacy

[a] While there is debate about the differences between edge AI and edge intelligence, we use the terms edge AI and edge intelligence interchangeably in this article.

due to distributed approaches for ML like federated learning[9] and improved resilience due to local autonomy and decentralized control.[5] Edge intelligence has applications in various domains, including smart cities, health care, autonomous driving, and industrial automation, where low latency and local data processing are critical. This trend is further augmented by the increasing prevalence of 5G networks and the promises of future 6G networks, which offer the high-speed connectivity necessary for edge intelligence applications.[7] Figure 1 presents an illustration of the shift from a centralized, cloud-based use of AI for training and inference, to edge AI solutions in two representative use cases: autonomous driving and connected health solutions.

Edge AI Today

The state-of-the-art in edge intelligence can be divided into two main subfields: *AI on edge*, focusing on AI methods suitable for the decentralized, heterogeneous, and opportunistic edge environment, and *AI for edge,* focusing on the use of those methods for the benefit of the computing continuum.[6]

AI on edge has been propelled by advances in ML algorithms, particularly in deep learning, and their optimization for execution on constrained devices. The development of lightweight neural networks and techniques like model pruning and quantization are crucial in enabling complex AI models to run efficiently at the edge. In Figure 1, AI on edge allows model training and inference directly at the edge, either in a collaborative form through direct interaction between edge devices or using local edge servers close to these devices.

A notable trend is the emergence of distributed ML techniques for training and inference of AI models across multiple edge devices while preserving data privacy. For example, *federated learning* enables collaborative model training without the need to centralize data, aligning with the distributed nature of edge computing and addressing growing concerns around data security and privacy in AI.[9] To perform inference of large AI models at the edge without compressing them via pruning or quantization, these models can be split into several submodels. This allows for their distributed and collaborative execution on multiple, possibly heterogeneous, edge devices.[17,19] Alternatively, one may explore adaptive computation techniques where the inference cost is a function of the complexity of the data.[20] Finally, *hierarchical inference*[22] has been proposed where the interplay between larger and smaller neural network structures is leveraged toward accuracy, energy efficiency, and latency in edge-based inference scenarios.[23]

AI for edge, on the other hand, has seen significant advancements in integrating artificial intelligence with

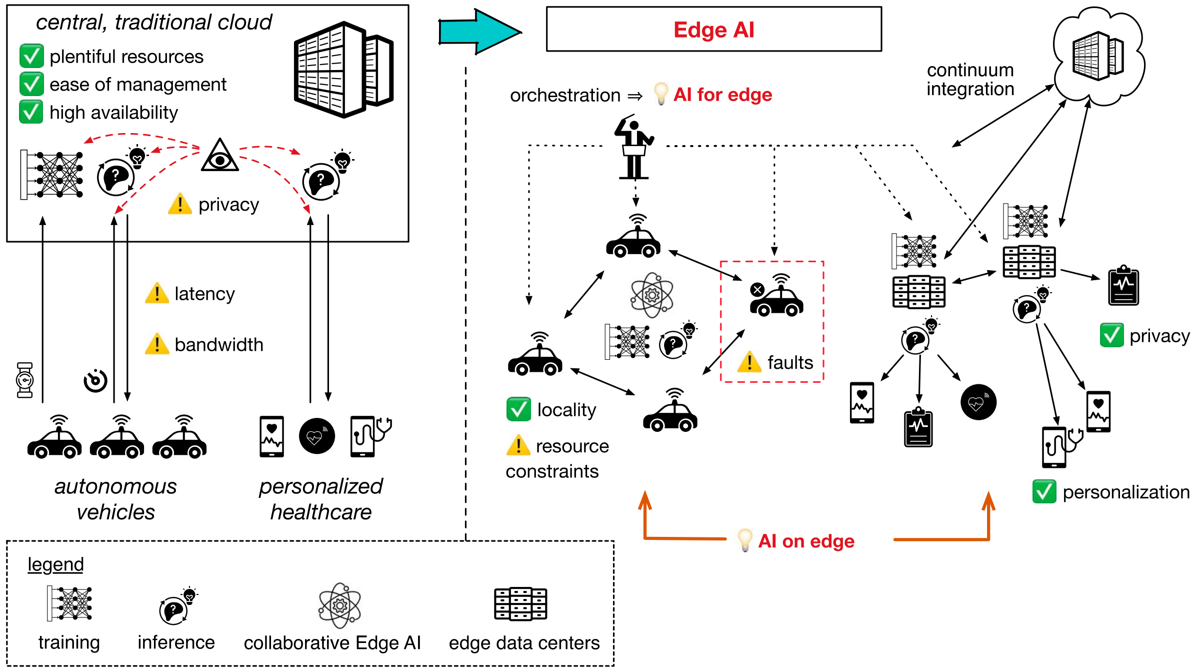

FIGURE 1. An illustration of the shift from centralized AI in the cloud (left) and Edge AI (right), and the associated challenges and opportunities, for two representative target applications: autonomous vehicles and personalized health care.

edge computing architectures, enhancing the capability of edge devices to perform sophisticated data processing and decision-making tasks, and paving the way for the intelligent orchestration of resources in the computing continuum,[10] as illustrated in Figure 1. Indeed, in addition to technological advancements, the current landscape of edge intelligence is shaped by an increasing focus on energy efficiency and sustainability.[8] Researchers and practitioners are actively exploring methods to reduce the energy footprint of edge AI systems. This is crucial for their widespread deployment, particularly in environments where power availability is a constraint. Necessary progress includes, for example, the development of energy-aware algorithms and hardware optimizations.

Besides the characterization of AI on edge or AI for edge, we observe differences in provider models. Many applications of edge computing are extensions of multitier architectures that shift the processing along the continuum between sensors and actuators, coordination of the application domain, e.g., a production floor, and cloud services. Edge computing offers the opportunity to conquer communication load and latency requirements with the placement of processing along this continuum. As such, we see edge AI as a phenomenon in industrial applications.

WHO SHOULD CARE?

While the importance of edge computing and edge AI increased in the last decade with the introduction of increasingly challenging and data-driven applications like smart cities and industrial automation, different stakeholders have different perspectives on these paradigms and associated technologies. In the following, we introduce the perspectives of four stakeholders: the needs of *society* and *industry* are shaped into solutions by *developers*. These solutions are then subject to policies and regulations set by *governments*. Understanding the individual perspectives of these stakeholders is pivotal for shaping future research directions and enabling the sound development of Edge AI.

Figure 2 provides an overview of the interest of the societal, governmental, industrial, and developer perspectives for the different challenges of Edge AI. The plot should be interpreted as a general tendency.

Societal Perspective (Everyday Life of People)

Societally, the interest in Edge AI centers on its *practical applications* rather than on the underlying technological innovations. People are likely to appreciate the use of Edge AI in areas such as autonomous vehicles and

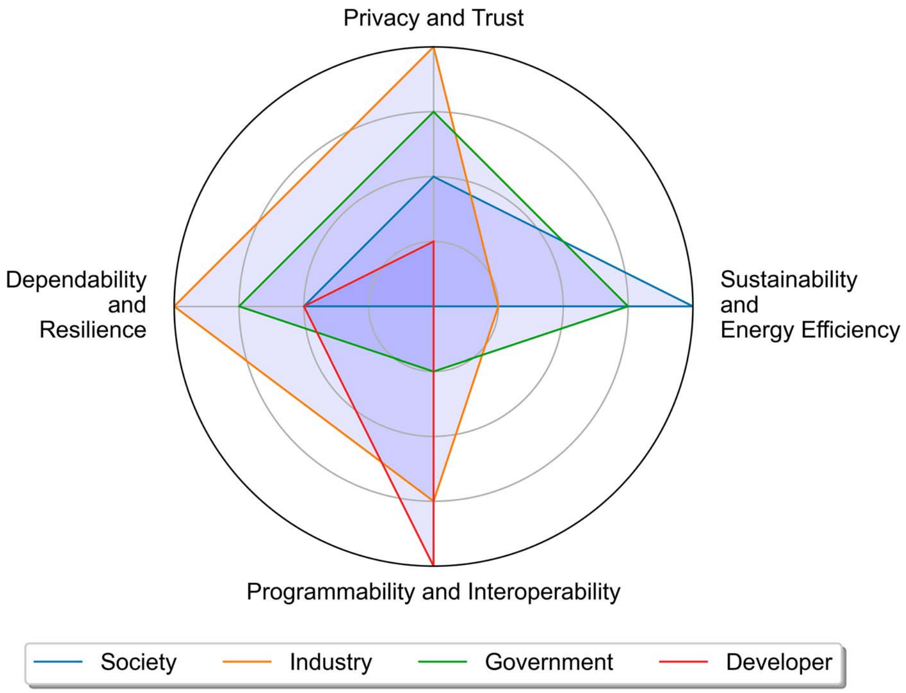

FIGURE 2. Demands of the societal, governmental, industrial, and developer perspectives. The darkness of the symbol illustrates the importance of each demand for the respective stakeholder's perspective (darker color means higher demand).

smart homes. Although the average users, particularly those without a technical background, may not notice the latency differences between cloud-based and edge-based execution, the accessibility of applications and their impact on daily life will be much more significant, as it will enable richer interactions and more complex applications. Especially in Europe, *privacy* is an important aspect, which is closely linked with edge computing and edge AI.

Industry Perspective

We can distinguish perspectives for two categories of industrial players: *consumers* of edge AI and *providers* of edge AI.

For consumers of edge AI, the question of *reliability and guarantees* is a major factor in deciding the future use of edge intelligence. The multitude of cloud and edge providers makes it hard to have confidence in the reliable operation of multiple systems and services. In addition, the ability to attribute system failures to specific components or providers diminishes, thereby limiting the potential for liability for such failures. Although this challenge can be mitigated by using combined cloud/edge providers, such as AWS Wavelength, using a single provider may substantially impact some of the benefits of edge computing, particularly in terms of system robustness and data protection. *Data protection* is a factor that, similarly to the societal perspective, is critical from an industrial viewpoint. This includes the protection of data while being processed at the edge, ensuring the *trustworthiness* of the edge network provider, and the protection of intellectual property, i.e., of the developed edge applications and trained AI models.[1]

For providers of edge AI, the question of *business cases* is pivotal for the success of edge AI. In the past, there already has been a transition from voice providers to data providers in telecommunication, who can now again transition, this time to computation providers. Especially in mobile networks, telecommunication operators are natural candidates to support the placement of computation close to the users and allow them to use AI services with small latency. However, if nongeneric models are required, this will result in the migration and placement of user models at the edge. As of now, it is unclear if this is a sound business case. The number of possible applications, e.g., assisted driving, support of the elderly, and on-the-fly translation services, are promising but require a business model to justify such an extension of telecommunication infrastructures. In addition, the multitude of cloud and edge providers, along with their interconnections, makes it challenging to ensure the system reliability that is rightfully expected by consumers. Similarly, ensuring the levels of data protection and trustworthiness that are requested by consumers can be challenging. Providers of edge AI solutions could also use edge AI solutions to improve their service, in which case they may again face challenges with the reliability of their solution. One important aspect is the consideration of consumer applications running on the edge, such that the operations used for improving their operation do not interfere with the regular operation of these applications.

Governmental Perspective

The governmental view of edge AI is multifaceted, encompassing various aspects such as the enforcement of ethical and responsible use, the safeguarding of citizen privacy (echoing societal concerns), the *protection of the intellectual property* of companies, the setup of the necessary infrastructure, the promotion of *interoperability* through common standards, and the monitoring of data exchanges via lawful intercepts. The prioritization of these aspects varies for governments with different focuses. Notably, European nations, already pioneers in privacy regulations like General Data Protection Regulation (GDPR) and security regulations such as the EU Cyber Resilience Act, are likely to emphasize the ethical use of edge AI and the *protection of privacy and security*. We note, finally, that governmental actors can improve the development of edge intelligence through funding and regulations, enabling new services in the respective country.

Developer Perspective

From a developer's viewpoint, the *ease of programming* is crucial for adopting edge AI, particularly for creating distributed applications. Ideally, developers should expend minimal effort in addressing typical edge AI issues such as user and data mobility, distributed coordination, and synchronization. Therefore, to facilitate developer access to edge AI, a programming framework is necessary to simplify the development and configuration of edge AI applications. That includes managing computation and storage resources, automating the watermarking of deployed models, handling the distribution of sensor data, and providing program abstractions for new paradigms like quantum and neuromorphic computing.

Summary and Research Perspective

The research perspective combines all of the aforementioned views into a holistic one, in which future research aims to solve parts of today's problems of edge AI. Several works like Rausch et al.[24] and Nastic et al.[25] have proposed *programming models* for edge AI, such that training and inference can be executed in a decentralized manner; one example is the paradigm of federated learning. While the *locality of data* and the corresponding decentralization can be seen as a positive influence on *data privacy*, trust in edge devices can be limited at times. Thus, there are additional challenges related to the protection of data privacy, for which several ideas are currently being investigated. These include the use of homomorphic encryption, on-device filtering of task-relevant data in hardware-secured execution environments, and research on ensuring trust into edge AI solutions. Research results on these topics can provide valuable inputs to governments regulating edge processing systems.

CURRENT RESEARCH CHALLENGES AND OPPORTUNITIES

Edge AI offers a transformative approach to embedding intelligence into local devices. It is associated with challenges around resource constraints, security and privacy, sustainability, and dealing with the energy crisis. At the same time, it brings significant opportunities in real-time data processing, efficiency, and personalized experiences. The algorithms that make up AI are finding their way into a growing number of excellent services for users. The way this uptake happens and its technical potential was analyzed in several published studies.[2,11,13] This raises a number of issues in understanding the challenges and opportunities of edge AI, from which we highlight the most current and notable ones in the following.

Resource Limitations

Edge devices are characterized by limited computing and storage resources. While cloud-based applications can utilize a variety of computing devices, including CPUs, GPUs, and sometimes field-programmable gate arrays, edge devices commonly contain only a few hardware accelerators that are often tailored for a specific application or use case. In addition, the computing, memory, and storage of edge devices are significantly constrained, limiting the possibility of training and inference even further. This is especially a challenge when it comes to the application of edge AI solutions, as ML models commonly rely on dedicated hardware and require a large volume of memory and storage. In addition, the exchange of data is often critical and limited by the available network bandwidth. Thus, mechanisms need to be developed to limit the amount of exchanged information, not only with central infrastructure, but also between edge devices, e.g., by information-driven prioritization.[27] The training of ML models at the edge is particularly challenging due to these resource limitations, representing an ongoing challenge in the field.

Given that the location of inference is not always predetermined, spanning from powerful centralized devices to resource-constrained edge devices, the necessity for multiple ML models becomes apparent. Each deployment environment comes with its own set of constraints and requirements, whether it is real-time processing on edge devices or comprehensive analysis on robust computational platforms. As a result, developers often need to tailor and optimize models to suit diverse deployment scenarios, ensuring efficiency and effectiveness across the spectrum. Automated mechanisms are required to support this adaptation, such that edge AI solutions can seamlessly integrate into various contexts, catering to the specific needs and constraints of each deployment scenario while maintaining the best possible performance.

Privacy and Trust

Ensuring reliability, security, privacy, and ethical integrity is key to establishing trustworthiness in both edge AI applications and connected systems. This is crucial as edge devices handle sensitive data, and the consequences of breaches can be severe.

Essential to establishing trust is secure processing and storage combined with robust encryption and stringent access controls. AI models must be reliable and accurate, despite the limited resources of edge devices, and robust against adversarial attacks. The use of hardware-supported, trusted execution environments is sometimes considered but comes with its own set of challenges regarding performance and integration. Additionally, transparency and explainability in AI decision-making are increasingly important, especially in critical applications. Compliance with regulations like GDPR, mandating data privacy and security, is also a key aspect of edge AI to be addressed.

Sustainability and Energy Efficiency

The growing need for AI applications emphasizes the importance of creating energy-efficient and sustainable edge AI algorithms. Advanced AI, particularly deep learning, consumes substantial energy,[26] presenting a sustainability challenge. Balancing performance with energy efficiency is crucial for edge AI. While achieving higher levels of accuracy may seem like the ultimate goal, it is imperative to recognize that each incremental improvement in accuracy often demands a substantial increase in energy consumption. This tradeoff becomes particularly apparent in scenarios where ultrahigh accuracy might not be crucial. In such cases, allocating excessive energy resources for marginal gains in accuracy could be inefficient and environmentally unsustainable. Thus, developers and researchers must conscientiously evaluate the necessity of heightened accuracy against the energy footprint it entails.

Another important aspect is the growing importance of renewable energy sources to the energy grid. Since most renewable sources are dependent on environmental conditions (like sunshine for solar cells), there are times, e.g., on a hot summer day with a lot of wind, when power is abundant. While conservative energy usage remains a significant challenge in edge AI, another key challenge is the possibility of performing non-time-critical calculations like model training when there is an energy surplus. Executing these calculations at times of excess power can help balance out spikes in energy generation and compensate for the fluctuating nature of most renewable energy sources. Moreover, distributing the energy demand geographically can help alleviate supply problems faced by large-scale data centers accumulated in certain regions such as Northern Virginia, USA and Amsterdam, The Netherlands. As our energy storage capacities are limited and often inefficient, this can greatly improve the efficiency of the power grid and edge devices.

While energy consumption during operation is an important challenge, so is the production and lifecycle of the deployed edge devices. Designing more durable, upgradeable, and recyclable devices is of critical importance to improve the environmental footprint of edge AI solutions. Additionally, implementing policies to encourage energy-efficient AI and regulating the environmental impact of device manufacturing and disposal is essential.

Programmability and Interoperability

Edge AI involves diverse devices like smartphones, IoT devices, and industrial machinery, each with unique constraints. Creating programmability frameworks for edge AI is challenging due to the need to orchestrate services across this varied hardware efficiently.[17] Developers face the complexity of differing device capabilities in terms of CPU power, GPU availability, memory, and energy consumption.[12] This complexity makes the deployment of services at a large scale such as smart cities a major and already continuing challenge.[14] The lack of standardized tools further complicates development, often requiring the use of incompatible tools and platforms, leading to longer development times and integration issues.

The programmability challenge of edge AI is even further intensified by the need for interoperability, i.e., combining operations on a variety of devices and systems, like sensors, smartphones, and industrial machinery. These devices should work together seamlessly despite different operating systems, software, and hardware. A key issue is the lack of standardized protocols and data formats, making it crucial to develop universal standards for effective communication. Integrating edge AI with existing systems poses challenges because of unsupported software and hardware components. As the number of interconnected devices grows, scalability and easy integration of new devices become important and difficult. Minimizing delays caused by interoperability is crucial in real-time processing scenarios like environmental monitoring, autonomous driving, and industry 4.0. However, managing resources efficiently in this interconnected environment, together with resources available in the continuum, is also a key challenge to be addressed.

To summarize, unified programmability frameworks are essential for deploying edge AI algorithms effectively, ensuring efficient service orchestration, resource management, and device interoperability across the continuum.

Dependability and Resilience

Dependability focuses on the reliability, security, and robustness of AI systems operating on edge computing devices used for AI decision-making. It encompasses ensuring these systems perform consistently and accurately, even in challenging or unpredictable environments.[18] Such systems, crucial in cyber-critical sectors like health care and industrial automation, must always be operational with robust design and effective failover strategies.[16] Developed systems must protect data and AI model integrity against various threats, be capable of handling more data, and accommodate more devices or geographical areas. Systems must autonomously detect and resolve faults and adapt to changing conditions and emerging threats for dependable operation.

In addition to dependability, the resilience of edge AI is pivotal to guarantee its operability at all times. Resilience involves ensuring reliable functioning against offensive security and disruptions under various conditions. Edge devices must be robust against physical challenges like extreme temperatures and mechanical impacts and maintain data integrity and security. Even in poor network conditions, edge systems should either be able to provide reliable connectivity (via alternative communication technologies or robust protocols) or operate offline until connectivity is available again. In addition, these systems need to be fault-tolerant, possibly with backup solutions. AI models should adapt to changing data patterns without needing extensive retraining. As edge AI networks grow, scalability and manageability become key, alongside efficient resource management to handle varying workloads across the continuum.

Measurability

Defining generalized metrics for evaluating performance across the cloud-edge continuum is challenging due to the unique characteristics (e.g., distribution in training and inference, shared resources) and constraints (e.g., resource limitations, real-time requirements) of different edge AI applications and connected systems. This is particularly true for the challenges mentioned earlier, which are currently difficult to measure and quantify. To allow for research in those areas, identifying metrics that accurately measure the development is pivotal.

Additionally, a significant challenge in edge AI involves balancing tradeoffs among accuracy, latency, resource usage, and privacy across this continuum. While simulations or emulations can predict the performance of approaches in certain scenarios, it is essential to verify the validity of developed approaches in real-world testbeds. Developing benchmark evaluation frameworks in such real-world environments and considering real-world use cases remains an open challenge. These benchmarks must rely on generalized metrics to precisely measure and evaluate the unique characteristics of Edge AI.

FUTURE RESEARCH DIRECTIONS

In this section, we identify the most promising research directions: the integration of large language models (LLMs) into edge AI applications, low-latency inference for autonomous vehicles, shifting focus toward energy and privacy in our society, enhancing edge interoperability, and finally advancing trust and security in edge AI systems. We detail each of these directions next.

Integration of LLMs in Edge AI

The integration of LLMs into applications on the edge presents an exciting avenue for future research. LLMs have traditionally been considered too computationally expensive for inference on the edge, relegating them to cloud-based inference. Running them on edge devices introduces a paradigm shift. Increasingly, edge devices are powered with energy-efficient accelerators. For instance, Apple neural engines (ANEs) are available in iPhones and edge-tensor processing units from Google are available as submodules for embedded devices. Running LLMs on these edge accelerators offers the advantage of "free" inference to these companies since the "cost" (primarily: energy consumption) now occurs on end-user devices. This approach could benefit applications with relaxed latency requirements, such as social media platforms, where immediate response is not critical. However, the challenge lies in adapting these computationally intensive models to the constraints of edge devices, including limited processing power and energy efficiency. Classical learning techniques such as distillation, neural architecture search, and systems techniques such as quantization and sparsification are all potential candidates yet "unproven" in their effectiveness. The challenge in evaluating LLMs makes this no easier. Future research should focus not on optimizing LLMs for edge environments but could also lead to innovations in customized hardware that meets the power profiles of the edge.

Edge Computing for Autonomous Agents

Autonomous agents are becoming increasingly important for our society. This includes, e.g., autonomous robots in a smart factory and autonomous vehicles. Even while both the first self-driving vehicles and the first autonomous robots are deployed, these agents currently only function with constant network connectivity, in certain areas, or under certain conditions. But even today, a multitude of sensors, both external and on-board sensors, generate vast amounts of data that could overwhelm traditional internet infrastructures. This data, if shared with low delay, can improve the driving behavior of other agents, allowing for even higher levels of automation. Edge computing allows local processing of data, reducing the need to transfer massive volumes over the network. This not only enhances response times and operational efficiency but also supports real-time decision-making at "internet blind spots" crucial for the reliable operation of the agents. This is not merely a "nice to have," but essential. The multisensor inputs, perception at different times of the day and in different environments, across diverse weather conditions, and social elements not only advocate for an edge-focused solution but open up new avenues for research in both training and inference aspects on the edge.

Focus on Energy Efficiency and Privacy in Society

The increasing societal awareness of energy consumption and privacy concerns presents a unique opportunity for edge AI. Local processing on edge devices ensures data privacy by keeping sensitive information within the device, thereby guaranteeing data privacy. Moreover, in power-linear systems used on the edge, such as deeply embedded deployments, communication costs (Wi-Fi, Bluetooth, and so on) can be higher than computational expenses. Thus, there is a pressing need for research focused on developing energy-efficient edge AI solutions that balance communication overhead with computational efficiency. This involves exploring energy-aware algorithms, sustainable hardware designs, and optimizing network protocols for energy conservation. Low-power wide-area networks are a promising direction, trading off throughput with power. While this presents one design point in the wide Pareto curve, how to develop solutions that are general-purpose enough to lower production costs, while being tailored enough to support the unique needs of applications is an open research question.

Enhancing Edge AI Interoperability

As edge AI systems become more prevalent, ensuring their scalability and interoperability is a new and unexplored frontier. An open question is how the different AI-enabled edge devices should talk to each other. Beyond the perennial debate of decentralized versus centralized, hub-and-spoke versus circular, one exciting thrust could be on developing standardized protocols and frameworks that enable seamless integration of diverse edge devices and systems. This includes creating universal data formats and communication standards to facilitate efficient interaction and more critically discovery between different types of edge devices, such as sensors, wearables, smartphones, industrial equipment, autonomous vehicles, and so on.

The evolution of edge AI also brings along the need to address interoperability with emerging non-von Neumann architectures (e.g., quantum and neuromorphic computing).[15] It is essential to develop protocols and standards that enable effective communication and collaboration with traditional computing systems. This involves not only the translation of data formats and communication protocols but also the understanding and alignment of the fundamentally different ways in which different non-von Neumann architectures process and interpret data. Neuromorphic computing systems, for instance, mimic the neural structure of the human brain to achieve extreme parallelism and energy efficiency; however, they are event-driven and operate with analog data (spikes). Bridging this gap is crucial in creating a truly interconnected edge AI ecosystem, where devices can leverage the unique strengths of both current and future computing paradigms.

Advancing Trust and Security in Edge AI Systems

Ensuring the trustworthiness and security of edge AI systems are paramount, especially as they become integral to critical infrastructures and personal devices. Future research should focus on developing robust security protocols and encryption methods to protect sensitive data processed at the edge. This includes enhancing the resilience of edge AI systems against cyber threats and ensuring that AI decision-making processes are transparent, explainable, and compliant with regulatory standards like GDPR. This becomes challenging given edge devices sometimes lack trusted environments, which are pivotal for protecting privacy-sensitive data. Addressing these aspects will not only improve the security and reliability of edge AI systems but also foster public trust in their deployment and usage.

CONCLUSION

In this article, we revisited the history and current state of edge AI solutions, ranging from its origin as a combination of edge computing with AI to its current state with decentralized interference and training of AI on resource-constraint edge devices. We highlighted the different challenges and research opportunities of edge AI today, including the perspectives of the relevant stakeholders in the edge AI area. Finally, we envision future research directions for researchers in the field.

REFERENCES

1. Y. Li, H. Wang, and M. Barni, “A survey of deep neural network watermarking techniques,” *Neurocomputing*, vol. 461, pp. 171–193, Oct. 2021, doi: 10.1016/j.neucom.2021.07.051.
2. A. Y. Ding et al., “Roadmap for edge AI: A Dagstuhl perspective,” *ACM SIGCOMM Comput. Commun. Rev.*, vol. 52, no. 1, pp. 28–33, 2022, doi: 10.1145/3523230.3523235.
3. B. Varghese et al., “Revisiting the arguments for edge computing research,” *IEEE Internet Comput.*, vol. 25, no. 5, pp. 36–42, Sep./Oct. 2021, doi: 10.1109/MIC.2021.3093924.
4. W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, “Edge computing: Vision and challenges,” *IEEE Internet*

Things J., vol. 3, no. 5, pp. 637–646, Oct. 2016, doi: 10.1109/JIOT.2016.2579198.

5. L. Lovén et al., “EdgeAI: A vision for distributed, edge-native artificial intelligence in future 6G networks,” in *Proc. 6G Wireless Summit*, Levi, Finland, 2019, pp. 1–2.
6. Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, “Edge intelligence: Paving the last mile of artificial intelligence with edge computing,” *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, Aug. 2019, doi: 10.1109/JPROC.2019.2918951.
7. E. Peltonen et al., “The many faces of edge intelligence,” *IEEE Access*, vol. 10, pp. 104,769–104,782, 2022, doi: 10.1109/ACCESS.2022.3210584.
8. O. L. A. López et al., “Energy-sustainable IoT connectivity: Vision, technological enablers, challenges, and future directions,” *IEEE Open J. Commun. Soc.*, vol. 4, pp. 2609–2666, 2023, doi: 10.1109/OJCOMS.2023.3323832.
9. J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik, “Federated optimization: Distributed machine learning for on-device intelligence,” 2016, *arXiv:1610.02527*.
10. H. Kokkonen et al., “Autonomy and intelligence in the computing continuum: Challenges, enablers, and future directions for orchestration,” 2022, *arXiv:2205.01423*.
11. A. Y. Ding, M. Janssen, and J. Crowcroft, “Trustworthy and sustainable edge AI: A research agenda,” in *Proc. 3rd IEEE Int. Conf. Trust, Privacy Secur. Intell. Syst. Appl. (TPS-ISA)*, Atlanta, GA, USA, 2021, pp. 164–172, doi: 10.1109/TPSISA52974.2021.00019.
12. S. G. Patil, P. Jain, P. Dutta, I. Stoica, and J. Gonzalez, “POET: Training neural networks on tiny devices with integrated rematerialization and paging,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, PMLR, 2022, pp. 17,573–17,583.
13. R. Singh and S. S. Gill, “Edge AI: A survey,” *Internet Things Cyber-Physical Syst.*, vol. 3, pp. 71–92, Jan. 2023, doi: 10.1016/j.iotcps.2023.02.004.
14. B. Cheng, G. Solmaz, F. Cirillo, E. Kovacs, K. Terasawa, and A. Kitazawa, “FogFlow: Easy programming of IoT services over cloud and edges for smart cities,” *IEEE Internet Things J.*, vol. 5, no. 2, pp. 696–707, Apr. 2018, doi: 10.1109/JIOT.2017.2747214.
15. D. Kimovski et al., “Beyond von Neumann in the computing continuum: Architectures, applications, and future directions,” *IEEE Internet Comput.*, early access, 2023, doi: 10.1109/MIC.2023.3301010.
16. A. Khalil et al., “Dependability: Enablers in 5G campus networks for industry 4.0,” in *Proc. 19th Int. Conf. Des. Reliable Commun. Netw. (DRCN)*, Vilanova i la Geltru, Spain, 2023, pp. 1–8, doi: 10.1109/DRCN57075.2023.10108299.
17. X. Guo, A. D. Pimentel, and T. Stefanov, “Automated exploration and implementation of distributed CNN inference at the edge,” *IEEE Internet Things J.*, vol. 10, no. 7, pp. 5843–5858, Apr. 2023, doi: 10.1109/JIOT.2023.3237572.
18. X. Guo, A. D. Pimentel, and T. Stefanov, “RobustDiCE: Robust and distributed CNN inference at the edge,” in *Proc. 29th Asia South Pacific Des. Automat. Conf. (ASP-DAC)*, Incheon Songdo Convensia, South Korea, Jan. 2024, pp. 26–31, doi: 10.1109/ASP-DAC58780.2024.10473970.
19. E. De Coninck et al., “DIANNE: A modular framework for designing, training and deploying deep neural networks on heterogeneous distributed infrastructure,” *J. Syst. Softw.*, vol. 141, no. 7, pp. 52–65, Jul. 2018, doi: 10.1016/j.jss.2018.03.032.
20. S. Leroux, T. Verbelen, P. Simoens, and B. Dhoedt, “Iterative neural networks for adaptive inference on resource-constrained devices,” *Neural Comput. Appl.*, vol. 34, no. 13, pp. 10,321–10,336, 2022, doi: 10.1007/s00521-022-06910-5.
21. N. Mohan, L. Corneo, A. Zavodovski, S. Bayhan, W. Wong, and J. Kangasharju, “Pruning edge research with latency shears,” in *Proc. 19th ACM Workshop Hot Topics Netw. (HotNets)*, New York, NY, USA: ACM, 2020, pp. 182–189, doi: 10.1145/3422604.3425943.
22. G. Al-Atat, A. Fresa, A. Behera, V. Moothedath, J. Gross, and J. Champati, “The case for hierarchical deep learning inference at the network edge,” in *Proc. 1st Int. Workshop Netw. AI Syst. (NetAISys)*, 2023, pp. 1–6, doi: 10.1145/3597062.3597278.
23. V. Moothedath, J. Champati, and J. Gross, “Getting the best out of both worlds: Algorithms for hierarchical inference at the edge,” *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 2, pp. 280–297, 2024, doi: 10.1109/TMLCN.2024.3366501.
24. T. Rausch, W. Hummer, V. Muthusamy, A. Rashed, and S. Dustdar, “Towards a serverless platform for edge AI,” in *Proc. 2nd USENIX Workshop Hot Topics Edge Comput. (HotEdge)*, 2019, pp. 1–7.
25. S. Nastic, P. Raith, A. Furutanpey, T. Pusztai, and S. Dustdar, “A serverless computing fabric for edge and cloud,” in *Proc. IEEE 4th Int. Conf. Cogn. Mach. Intell. (CogMI)*, Piscataway, NJ, USA: IEEE Press, 2022, pp. 1–12, doi: 10.1109/CogMI56440.2022.00011.
26. D. Katare, D. Perino, J. Nurmi, M. Warnier, M. Janssen, and A. Y. Ding, “A survey on approximate edge AI for energy efficient autonomous driving services,” *IEEE Commun. Surveys Tuts.*, vol. 25, no. 4, pp. 2714–2754, 4th Quart. 2023, doi: 10.1109/COMST.2023.3302474.
27. D. Bischoff, F. A. Schiegg, D. Schuller, J. Lemke, B. Becker, and T. Meuser, “Prioritizing relevant

information: Decentralized V2X resource allocation for cooperative driving," *IEEE Access*, vol. 9, pp. 135,630–135,656, 2021, doi: 10.1109/ACCESS.2021.3116317.

TOBIAS MEUSER is a sensor researcher and heads the research group of adaptive communication networks in the Chair of Communication Networks at the Technical University of Darmstadt, 64283, Darmstadt, Germany. Contact him at tobias.meuser@kom.tu-darmstadt.de.

LAURI LOVÉN is a postdoctoral researcher of electrical engineering and the coordinator of the Distributed Intelligence strategic research area in the 6G Flagship research program, at the Center for Ubiquitous Computing (UBICOMP), University of Oulu, 90014 Oulu, Finland. Contact him at lauri.loven@oulu.fi.

MONOWAR BHUYAN is an assistant professor of computer science and heads the Cyber Analytics and Learning Group within the Autonomous Distributed Systems Lab at Umeå University, 90187 Umeå, Sweden. Contact him at monowar@cs.umu.se.

SHISHIR G. PATIL is a Ph.D. student in computer science at UC Berkeley, California, Berkeley, CA, 94709, USA. He is affiliated with the Sky Computing Lab (previously RISE), Lab11, and Berkeley AI Research (BAIR). Contact him at shishirpatil@berkeley.edu.

SCHAHRAM DUSTDAR is a full professor of computer science and heads the Research Division of Distributed Systems at Vienna University of Technology, Vienna 1040, Austria. Contact him at dustdar@dsg.tuwien.ac.at.

ATAKAN ARAL is an assistant professor in the Department of Computing Science and a member of the Autonomous Distributed Systems Lab and Green Distributed Computing Group at the Autonomous Distributed Systems Lab at Umeå University, 90187 Umeå, Sweden. Contact him at atakan,aral@umu.se.

SUZAN BAYHAN is an associate professor at Design and Analysis of Communication Systems (DACS) and affiliated with EDGE research center at the University of Twente, 7500 AE, Enschede, The Netherlands. Contact her at s.surname@utwente.nl.

CHRISTIAN BECKER is a full professor for computer science and heads the Institute for Parallel and Distributed Systems at the University of Stuttgart, 70569, Stuttgart, Germany. Contact him at christian.becker@ipvs.uni-stuttgart.de.

EYAL DE LARA is a professor of computer science at the University of Toronto, Toronto ON, M5S 1A1, Canada. Contact him at delara@cs.toronto.edu.

AARON YI DING is an associate professor at TU Delft and University of Helsinki (permanent) and leads the Cyber Physical Intelligence (CPI) Lab, 2600 AA, Delft, The Netherlands. Contact him at aaron.ding@tudelft.nl.

JANICK EDINGER is a professor of computer science and head of the research group for Distributed Operating Systems at the University of Hamburg. 22527, Hamburg, Germany. Contact him at janick.edinger@uni-hamburg.de.

JAMES GROSS is a professor with the Electrical Engineering and Computer Science School of KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden. Contact him at jamesgr@kth.se.

NITINDER MOHAN is a senior researcher in Chair of Connected Mobility at the Technical University of Munich, 80333, Munich, Germany. Contact him at mohan@in.tum.de.

ANDY D. PIMENTEL is a full professor at University of Amsterdam, 1098 GH, Amsterdam, The Netherlands, and chairs the Parallel Computing Systems (PCS) group within the Systems and Networking Lab at the Informatics Institute, 1000 GG, Amsterdam, The Netherlands. Contact him at a.d.pimentel@uva.nl.

ETIENNE RIVIÈRE is a professor of Computer Science and heads the Cloud and Large Scale computing group at UCLouvain, B-1348, Louvain-la-Neuve, Belgium. Contact him at etienne.riviere@uclouvain.be.

HENNING SCHULZRINNE is a professor in the Dept. of Computer Science at Columbia University, New York, NY, 10027, USA. Contact him at schulzrinne@cs.columbia.edu.

PIETER SIMOENS is an assistant professor at the Internet Technology and Data Science Lab at Ghent University-imec, B-9052, Gent, Belgium. Contact him at pieter.simoens@ugent.be.

GÜRKAN SOLMAZ is a senior researcher at NEC Laboratories Europe, 69115, Heidelberg, Germany. Contact him at gurkan.solmaz@neclab.eu. Contact him at.

MICHAEL WELZL is a full professor in the Networks and Distributed Systems group of the Department of Informatics at University of Oslo, 0313, Oslo, Norway. Contact him at michawe@ifi.uio.no.