Alliance for AI, IoT and Edge Continuum Innovation

AI in Manufacturing

Release 1.0

AIOTI WG Manufacturing

16 July 2025

Executive Summary

The integration of Artificial Intelligence (AI) in manufacturing is fundamentally transforming industrial paradigms, driving the convergence of productivity objectives with regulatory compliance, sustainability imperatives, and human-centric innovation. This transformation is being accelerated by the forthcoming EU Artificial Intelligence Act, which establishes a comprehensive risk-based classification framework and mandates stringent compliance requirements for high-risk AI applications, including quality control systems, predictive maintenance solutions, and human safety monitoring platforms. While this regulatory landscape presents significant governance and transparency challenges, particularly for small and medium enterprises (SMEs), it simultaneously creates unprecedented opportunities for the development of ethical, human-aligned AI solutions.

The transition toward Industry 5.0 represents a paradigmatic shift from pure digitalisation to human-centric manufacturing, where AI transcends its traditional role as an automation tool to become a collaborative enabler between humans and machines. Advanced technologies such as Large Language Models (LLMs), sophisticated speech-to-text systems, and egocentric video understanding are creating adaptive interfaces that conform to human cognitive patterns rather than forcing workers to adapt to machine constraints. This paradigm fundamentally transforms operator interactions, enhances knowledge-intensive processes, and promotes unprecedented inclusivity across manufacturing environments.

The Cloud-Edge-IoT (CEI) continuum architecture serves as the technological foundation for this evolution, enabling optimal production efficiency through the strategic balance of cloud-based model training capabilities with edge-based real-time inference processing. This distributed architecture supports critical functions including predictive maintenance, automated quality inspection, and adaptive process control while simultaneously reducing operational latency and enhancing system resilience. Nevertheless, significant challenges remain in areas of data interoperability, cybersecurity frameworks, and AI model retraining protocols across heterogeneous distributed infrastructures.

From a business perspective, AI functions as a critical enabler of innovative servitisation models, including "Manufacturing-as-a-Service" (MaaS) and comprehensive predictive maintenance services. These business models decouple asset ownership from operational usage, enabling flexible pricing mechanisms and advancing sustainability objectives. AI technologies unlock additional revenue streams through advanced analytics platforms, digital twin implementations, generative design capabilities, and AI-assisted documentation systems, providing particular competitive advantages for SMEs seeking market differentiation.

The energy implications of AI deployment—spanning both training and inference phases—present substantial sustainability considerations. Achieving sustainable AI operations requires strategic optimisation across computational, storage, and network resource demands. Emerging methodologies, including federated learning architectures, edge-aware model placement strategies, and comprehensive energy profiling frameworks, are essential for balancing performance requirements with environmental responsibility.

The emerging agentic AI paradigm promises the development of fully autonomous systems capable of sophisticated goal-oriented behaviour, complex decision-making processes, and independent task execution. These AI agents—whether virtual or physically embodied—will revolutionise operations across production, supply chain management, maintenance protocols, and beyond. However, the successful implementation of such systems fundamentally depends on robust explainability frameworks (XAI), ethical design principles, and the systematic integration of human-in-the-loop oversight, formalised through established control architectures such as MAPE-K (Monitor, Analyse, Plan, Execute, Knowledge).

Table of Contents

Introduction

The global manufacturing sector stands at an unprecedented inflection point. As Artificial Intelligence (AI) technologies achieve unprecedented maturity and penetrate industrial ecosystems with increasing sophistication, their impact transcends traditional process automation to fundamentally redefine operational paradigms, human-machine interactions, and value creation mechanisms. This transformation represents a convergence of technological advancement, organisational restructuring, and cultural evolution, propelling the industry beyond the established frameworks of Industry 4.0 toward the emerging **Industry 5.0** paradigm, where human-centric innovation, environmental sustainability, and operational resilience emerge as defining strategic imperatives.

The manufacturing sector's digital transformation journey over the preceding decade established a comprehensive foundation of sensors, Industrial Internet of Things (IIoT) infrastructure, and automation technologies. This technological substrate now enables AI's revolutionary capabilities: the evolution from systems that merely sense to those that **comprehend**, **anticipate**, and **execute** autonomous actions. Contemporary AI implementations empower manufacturing systems to perform sophisticated defect detection, optimise complex production parameters, deliver personalised maintenance interventions, and support strategic decision-making processes with unprecedented levels of autonomy and precision.

However, AI's role in manufacturing extends far beyond performance enhancement—it fundamentally transforms **the distribution and application of intelligence** throughout factory ecosystems. The Cloud-Edge-IoT (CEI) continuum architecture enables manufacturers to strategically combine cloud-based computational resources with edge device responsiveness. AI models developed and trained within centralised cloud infrastructures can be seamlessly deployed across distributed local nodes, facilitating real-time quality control processes, dynamic energy optimisation, and agile production adaptations. This hybrid architectural approach supports not only operational efficiency improvements but also enhanced **resilience and adaptability**—capabilities that have become increasingly critical within today's volatile supply chain environments and unpredictable market dynamics.

The strategic deployment of AI in manufacturing occurs within an increasingly complex **regulatory** and **social** environment that demands careful navigation. The European Union's Artificial Intelligence Act, scheduled for comprehensive implementation in 2025, establishes a comprehensive risk-based regulatory framework that categorizes critical manufacturing applications—including computer vision-based quality inspection systems and human-robot collaborative platforms—as high-risk deployments. Compliance with these evolving regulations necessitates the implementation of robust governance frameworks, comprehensive traceability mechanisms, and explainable AI (XAI) methodologies to ensure **transparency and trust** throughout the complete AI lifecycle.

Beyond regulatory compliance, the manufacturing sector faces significant cultural transformation challenges. As factories increasingly adopt AI systems characterized by growing autonomy and algorithmic complexity, ensuring that **humans remain central to industrial innovation** becomes paramount. The Industry 5.0 paradigm advocates for AI systems designed to adapt to existing human competencies and established workflows, rather than requiring workforce adaptation to machine constraints. This human-centric approach necessitates the development of intuitive interfaces powered by Large Language Models, advanced natural language processing capabilities, and egocentric perception systems, enabling operators to interact with manufacturing systems through natural and inclusive interaction modalities.

AI's transformative impact extends beyond the manufacturing floor to catalyse fundamental business model innovation. The growing servitisation trend—where traditional products evolve into comprehensive service delivery platforms—is gaining substantial momentum across the manufacturing landscape. AI technologies enable sophisticated predictive maintenance capabilities, dynamic pricing algorithms, advanced demand forecasting, and comprehensive digital twin implementations, supporting flexible, usage-based monetisation strategies. For small and medium enterprises (SMEs), these AI-enabled business models represent transformative opportunities, unlocking sustainable recurring revenue streams while reducing traditional barriers to advanced automation adoption. However, successful implementation requires strategic investments in **data-driven infrastructure architectures**, comprehensive workforce reskilling programs, and extensive customer education initiatives.

As AI systems achieve increasing sophistication, energy consumption patterns and environmental impact considerations emerge as critical strategic concerns. Sustainable AI implementation practices—encompassing federated learning architectures, edge computing optimisations, and comprehensive model efficiency improvements—are essential for aligning technological innovation with environmental stewardship responsibilities. Simultaneously, the emerging frontier of agentic AI introduces autonomous agents capable of sophisticated goal decomposition, complex workflow execution, and seamless collaboration with both human operators and machine systems. These intelligent agents will extend beyond production optimisation to manage comprehensive service ecosystems, encompassing financial processes, logistics coordination, and regulatory compliance management.

This paper comprehensively examines how AI is systematically reshaping manufacturing across five critical dimensions: regulatory compliance and human-centric integration strategies, enabling infrastructure architectures, innovative servitisation business models, sustainability challenges and solutions, and emerging paradigms including agentic AI implementations. Through rigorous analysis grounded in current technological trends, European policy initiatives, and documented industrial use cases, we demonstrate that AI's impact transcends traditional manufacturing enhancement—it represents a fundamental **redefinition of industrial paradigms** for the digital age.

1. Interaction between AI and Humans

The European Union's Artificial Intelligence Act (AI Act), expected to become law in 2025, is set to significantly impact the manufacturing sector by introducing a comprehensive regulatory framework for AI systems. The Act clusters AI applications based on risk levels – unacceptable, high, limited, and minimal – imposing strict obligations on high-risk systems commonly used in manufacturing, such as those for quality control, predictive maintenance, and worker safety monitoring. Manufacturers deploying these systems will be required to conduct conformity assessments, maintain detailed technical documentation, and ensure transparency and human oversight throughout the AI lifecycle. Non-compliance could result in substantial fines, up to 7% of global turnover or €35 million, emphasising the importance of adherence to these new regulations.

While the Act aims to enhance trust and safety in AI applications, it also presents challenges, in particular when it comes to Small and Medium Enterprises (SMEs) in the manufacturing sector, for which the compliance may imply significant investments in AI governance frameworks, regular audits, and stakeholder education. In front of this, the Act also offers opportunities for innovation by encouraging the development of transparent, accountable, and human-centric AI systems. By aligning with the EU AI Act, manufacturers can mitigate legal risks, as well as enhance their competitiveness in a market increasingly valuing ethical and responsible AI usage.

Regarding the competitiveness of companies, the AI deployment can be seen as a "usual" technology upgrade aimed at improving performance (e.g., exploiting computer vision in quality control, or implementing proactive routines to decrease the machine downtime for maintenance stops). One of the main questions which is interesting the communities of industries, researchers, and policy makers is however how AI will cope with the new industrial trends, which in Europe is currently being shaped after the concept of human-centric manufacturing (also known as "Industry 5.0").

This new paradigm disrupts indeed the current trends: while with the increasing digitalisation which characterised the last decade, manufacturing companies embodied an increasing amount of digital technologies in their shopfloors, asking operators to upskill or reskill in order to master and exploit these tools, human-centric paradigm changes the perspective, forcing the technology content of industrial assets to adopt human-friendly interfaces and reducing their impact on the required skills for being employed.

Under this aspect, recent trends of AI, in particular for what concerns the advancements of Large Language Models, can be seen as a catalyser for the realisation of the human-centric paradigm: those models can indeed easily elaborate instructions provided in natural language by the operators and cluster them into machine-ready instructions, easing the interaction between humans and machines, while mimicking a person-to-person correspondence. Correspondence which can become an actual dialogue when these tools are coupled with algorithms for speech-to-text conversion [Ivanko2023].

Apart from easing the interaction between human operators and machines, AI can also play an important role when dealing with a challenge introduced by the so-called "Industry 4.0", where one of the main targets was constituted by the production flexibility, exploiting the capabilities of IoT to reach mass customisation until one-piece-flow levels. This increase of flexibility multiplied the product portfolios of manufacturing companies, but implied a significant drawback in the body of knowledge operators were supposed to master in order to realize hundreds of different recipes and configurations. Also in this point, recent advancements of AI in the context of egocentric perception is supposed to be able to ease the operators' ongoing tasks and to be able to supervise them by warning in case of errors.

This technology is already available in the behavioural research field, with the so-called "egocentric video understanding" which typically leverages on 2D, and 3D convolutions to

retrieve information from video streams [Wang2016, Lin2019]. However, with the popularity of VLM, there have been a huge number of works that focus on the use of VLM to recognize and understand human behaviour [Carreira2017], these methods are inspired by common VLM in video understanding such as Video-LLaMA [Zhang2023], or BLIP-2 [Li2023].

The usage of video streams as input data in environments where the lighting and scene are not controlled (like in first-person point of view), however, poses several challenges for AI tools, in particular for what concerns the issue of domain shift, where, currently, two main approaches are exploited: domain generalisation [Munro2020] and unsupervised domain adaptation [Ganin2015]. But the problem of domain shift is not only characteristic of vision systems, and can also regard time series and other signals. It is the case, for example, of process industry, where shifts in the composition of raw materials can make AI-based controllers to shift and to produce undesired effects in the production processes.

To cope with this issue, a novel approach is under investigation and involves the embodiment of autonomic computing in the production control and monitoring loops. This discipline implies the addition of functionalities such as self-Configuring, self-Healing, self-Optimizing, and self-Protecting (self-CHOP or self-X) [Kephart2003] in the control loops and are supposed to efficiently trigger retraining of AI models in order to increase the resilience of control loops.

Given, in any case, the complexity of these tools, part of the aforementioned self-X functionalities can be taken into account by operators and data analysts overviewing and supervising the process. This approach, which can be formalised into a MAPE-K control modelling, can highlight the importance of human agents in the production control loops, formalising their role in the decision making at a production level [Cuzzola2025].

However, whenever the human actors are involved in close cooperation with AI tools, it becomes important – in particular considering the AI Act – the ability to enhance trust and transparency. As AI systems take on increasingly complex and autonomous roles, from quality inspection to human-robot collaboration, users need to understand and trust the decisions these systems make. This calls for the development and integration of explainable AI (XAI) techniques, which allow stakeholders to trace the reasoning behind a model's outputs. In practice, this can range from intuitive visual dashboards that highlight decision pathways, to algorithms specifically designed to offer human-interpretable justifications for classification or control actions [Onari2025].

Beyond technical explainability, trust also includes ethical design principles, considering fairness, accountability, and data privacy. In industrial settings, this means ensuring that AI tools do not introduce bias in operator evaluation or job assignment, that they are deployed transparently with clear responsibilities for oversight, and that the data they rely on are managed in accordance with privacy and security regulations. In this sense, ethical AI is not only a compliance goal, but a strategic asset: by embedding transparency and fairness into AI systems from the outset, manufacturers can enhance workforce acceptance, reduce resistance to innovation, and ensure alignment with the broader human-centric goals of Industry 5.0 [Jourabchi2025].

References:

[Outeda2024] Outeda, C. C. (2024). The EU's AI act: a framework for collaborative governance. Internet of Things, 101291.

[Ivanko2023] Ivanko, D., Ryumin, D., & Karpov, A. (2023). A review of recent advances on deep learning methods for audio-visual speech recognition. *Mathematics*, *11*(12), 2665.

[Wang2016] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016, September). Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision* (pp. 20-36). Cham: Springer International Publishing.

[Lin2019] Lin, J., Gan, C., & Han, S. (2019). Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7083-7093).

[Carreira2017] Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6299-6308).

[Zhang2023] Zhang, H., Li, X., & Bing, L. (2023). Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

[Li2023] Li, J., Li, D., Savarese, S., & Hoi, S. (2023, July). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (pp. 19730-19742). PMLR.

[Munro2020] Munro, J., & Damen, D. (2020). Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 122-132).

[Ganin2015] Ganin, Y., & Lempitsky, V. (2015, June). Unsupervised domain adaptation by backpropagation. In *International conference on machine learning* (pp. 1180-1189). PMLR.

[Kephart2003] Kephart, J. O., & Chess, D. M. (2003). The vision of autonomic computing. *Computer*, *36*(1), 41-50.

[Cuzzola2025] Cuzzola, F. A., Quadrini, W., Teocchi, S., Pinzone, M., & Fumagalli, L., (2025). Evaluating humachine dynamics in process industry AI applications, SUBMITTED TO Journal of Manufacturing Technology Management.

[Onari2025] Onari, M. A., Baer, G., Zhang, C., Grau, I., Nobile, M. S., & Zhang, Y. (2025, March). The Dynamics of Trust in XAI: Assessing Perceived and Demonstrated Trust Across Interaction Modes and Risk Treatments. In *The 3rd World Conference on eXplainable Artificial Intelligence*.

[Jourabchi2025] Jourabchi Amirkhizi, P., Pedrammehr, S., Pakzad, S., & Shahhoseini, A. (2025). Generative Artificial Intelligence in Adaptive Social Manufacturing: A Pathway to Achieving Industry 5.0 Sustainability Goals. *Processes*, *13*(4), 1174.

1.1 Interaction of Edge/Cloud with AI infrastructures

The evolution of Artificial Intelligence (AI) has profoundly transformed the manufacturing landscape, fostering synergies between Cloud Computing, Edge Computing, and the Internet of Things (IoT) - collectively referred to as the Cloud-Edge-IoT (CEI) continuum. This continuum delivers diverse range of functional capabilities that support end-to-end data lifecycle management. These technologies have redefined the architecture of industrial information systems, enabling a more efficient and adaptive approach to production processes. Historically, the limited computational power of edge devices necessitated the execution of AI tasks on centralized cloud servers. However, advancements in specialized processors and optimized frameworks now allow AI inference to be performed directly on edge devices. Furthermore, the concept of the edge has evolved beyond its technical definition to be perceived abstractly as an organisational edge, encompassing computing resources and data sources located close to the point of operation. In this context, a factory unit or manufacturing plant can function as an edge node, leveraging on-premises infrastructure to process data and execute AI workloads locally. This shift has given rise to a more efficient and responsive CEI paradigm: the cloud focuses on training AI models and large-scale analytics, while the edge handles real-time processing, reducing latency and enhancing operational responsiveness. This strategic CEI integration enables seamless information flow across cloud and edge systems through IoT-enabled communication networks, driving smarter and more autonomous manufacturing environments.

The CEI Continuum offers substantial benefits for manufacturing by sensing physical environments/objects for monitoring of events, and enabling real-time decision-making at the edge while leveraging centralized cloud resources for broader analytics. This balance optimizes resource allocation, minimizes data transfer costs, and reduces latency. For example, IoT devices generate lots of data and edge computing facilitate real-time data processing for monitoring and optimisation of manufacturing processes, significantly improving efficiency and productivity. Predictive maintenance becomes feasible as CEI technologies can detect equipment failures before they occur, reducing downtime and maintenance costs. Additionally, these technologies enhance product quality by providing real-time data analytics to identify and correct defects early in production. The CEI approach also supports sustainable manufacturing practices by fostering human-machine collaboration, improving supply chain agility, and building resilience against disruptions.

The adoption of CEI technologies is not without challenges. Interoperability between heterogeneous platforms remains a critical issue, as does ensuring data security across distributed systems. Efficiently updating AI models without disrupting operations is another significant hurdle. Addressing these challenges requires the development of open standards and frameworks that facilitate integration across diverse systems. In Europe, initiatives such as Horizon Europe and the EUCloudEdgeIoT initiative are actively driving research into innovative CEI solutions. These programs aim to enhance industrial competitiveness while promoting sustainability and resilience through advanced digital technologies.

The European manufacturing sector has already begun embracing CEI technologies at scale. For instance, manufacturers spent €8.7 billion on edge computing, €30.8 billion on cloud services, and €66 billion on IoT solutions in 2024 alone. Projects like IoTwins exemplify how CEI principles can be applied to create digital twins that integrate data from sensors, APIs, historical records, and open data sources. These digital twins enable manufacturers to optimize production processes by simulating real-world conditions and diagnosing potential issues before they arise.

Looking ahead, the future of manufacturing will see deeper integration of AI within production environments through CEI Continuum. Adaptive models capable of autonomously responding to changing operational conditions will improve quality control, reduce costs, and enable more flexible and sustainable production methods. Emerging technologies such as 5G connectivity, augmented reality (AR), robotics, and advanced analytics along with distributed edge architectures will further enhance the capabilities of smart factories under the CEI framework. The convergence of IT (Information Technology) and OT (Operational Technology) will play a pivotal role in ensuring seamless integration across all layers of manufacturing operations.

In conclusion, the CEI Continuum represents a transformative vision for the manufacturing industry by aligning technological innovation with economic growth and environmental stewardship. By adopting CEI principles, manufacturers can achieve greater flexibility, operational efficiency, and sustainability while maintaining Europe's leadership in global industrial innovation.

References:

AIOTI WG Manufacturing. Cloud-Edge-IoT Continuum: Opportunities and Challenges for Manufacturing Industry. AIOTI Report Release 1.0 (2024). Available at: AIOTI Website.

IDC Report on European Spending on CEI Technologies (2024). Referenced in AIOTI WG Manufacturing Report.

IoTwins Project Overview: Digital Twins for Manufacturing Optimisation (2023). Available at: IoTwins Project.

EUCloudEdgeIoT Initiative: Strategic Research Agenda (2023). Available at: EUCloudEdgeIoT Website.

Aeros Project Webinar Report: Cloud-Edge-IoT Innovations in Manufacturing (2023). Available at: Aeros Project.

HIPEAC News: AIOTI Paper on Edge IoT Industrial Immersive Technologies (2024). Available at: HIPEAC Website.

2. Business Models

The integration of Artificial Intelligence (AI) in the manufacturing sector is transforming business models, driving significant evolution in production and decision-making processes. AI can be applied both as a direct technology, such as for quality control, anomaly detection, condition monitoring, and the creation of digital twins, and as an indirect tool within the production cycle, optimising process parameters, verifying key performance indicators, and generating trend-based insights for process planning. This technology enables the adoption of new business models, including servitisation, which can take various forms and offer substantial benefits for both providers and users.

One of the most relevant developments is predictive manufacturing, where AI analyses historical and real-time data to predict failures and optimise maintenance, thereby reducing downtime and improving operational efficiency. Another emerging model is **Manufacturing-as-a-Service (Maas)**, which enables the real-time customisation and adaptation of production processes, offering manufacturing as a service and optimising resource utilisation. Cognitive automation, enabled by AI integration with industrial robots and autonomous systems, allows for more flexible and adaptive production, reducing the need for human supervision while increasing productivity. AI is also a key enabler of quality control, leveraging computer vision and machine learning algorithms to detect defects with greater accuracy than traditional methods, thereby minimising waste and improving product quality.

The use of AI-driven digital twins facilitates the creation of virtual models of production lines, enabling continuous simulation and optimisation of manufacturing processes, ultimately reducing costs and enhancing overall performance. Additionally, AI can optimise supply chain management, enabling dynamic control of procurement chains, anticipating demand fluctuations, and improving logistics and inventory management.

AI surrogate models are increasingly being adopted to replace resource-intensive simulation methods. Furthermore, generative AI models enable the automation of creative tasks such as conceptual design generation, marketing content creation (images and videos), and coding tasks such as code generation and bug fixing. Additionally, LLM chatbots have proven to be supportive tools for complex document reasoning, making technical documentation, norms, and standards more accessible in production environments.

Returning to the topic of servitisation, this business model enables the transformation of physical products into services with flexible payment structures. Companies can adopt usage-based pricing models, including pay-per-use, fixed daily, weekly, monthly, or annual fees, or a structure based on the number of users or work groups. Another option is function-based pricing, allowing customers to pay only for the features they actively use.

This approach applies not only to AI-driven software but also to the physical assets that utilise it. For instance, a quality control system or a press for screw production could be offered as a service, with payments tied to actual machine or system usage. This model provides advantages not only in terms of economic efficiency but also in sustainability.

Servitisation incentivises machine manufacturers to design increasingly reliable products since maintenance costs shift from the user to the provider. As a result, this leads to reduced waste and material disposal, as compensation is often linked to the long-term performance of the provided solution. Consequently, a virtuous cycle is established in which reliability and efficiency become key drivers of economic and environmental sustainability. Furthermore, servitisation fosters the development of more durable products, mitigating premature obsolescence and promoting a circular economy. In a landscape where environmental regulations and market expectations push for greater ecological responsibility, servitisation emerges as a strategy that effectively integrates technological innovation, production efficiency, and sustainability.

This transformation is particularly relevant for small and medium-sized enterprises (SMEs), which face unique challenges when implementing servitisation models.

2.1 Understanding Servitisation

Servitisation refers to the process of integrating products and services to create comprehensive solutions for customers. For example, a company that manufactures machinery might not only sell the machine but also offer a rental service accompanied by predictive maintenance, ensuring continuous operation and efficiency. This model allows customers to pay for actual usage rather than ownership, reducing waste and enhancing productivity. An example is tool wear monitoring in line production, where predictive analytics support and optimise tool changes and ensure efficient tool utilisation.

2.2 Benefits of Servitisation

The advantages of servitisation for manufacturing companies are numerous:

1. **Increased Competitiveness**: By offering high-value personalised services, companies can differentiate themselves from competitors and capture new market shares.

2. **Customer Loyalty**: A closer relationship with customers through ongoing support fosters satisfaction and loyalty.

3. **New Revenue Streams**: The sale of additional services generates recurring revenue, enhancing overall business income.

4. **Operational Efficiency**: Services provided can optimize production processes and reduce costs.

5. **Sustainability**: Servitisation contributes to more sustainable business models through predictive maintenance and optimized product use.

2.3 Challenges for SMEs

While the benefits of servitisation are clear, SMEs often encounter significant hurdles when attempting to implement this model.

1. Limited Resources

Many SMEs lack the financial and human resources necessary to invest in the technologies required for effective servitisation. Implementing AI-driven solutions, IoT systems, and data analytics tools requires substantial upfront investment, which can be daunting for smaller firms.

2. Customer Education

Transitioning customers from traditional purchasing models to servitisation can be challenging. Many customers may be unfamiliar with the concept of paying for usage rather than ownership, leading to resistance. Educating customers about the value proposition of servitized offerings is crucial but often requires extensive marketing efforts.

3. Cultural Shift

Adopting a servitisation model necessitates a cultural shift within the organisation. Employees must be trained not only in new technologies but also in customer service skills that emphasize relationship building over transactional sales. This shift can be difficult to achieve in organisations accustomed to traditional manufacturing paradigms.

4. Complexity in Implementation

Implementing a servitisation strategy involves redesigning business processes, which can be complex and time-consuming. SMEs may struggle with integrating new services into existing workflows without disrupting operations.

5. Competition from Larger Firms

Larger manufacturers may have more resources to invest in advanced technologies and marketing strategies for their servitised offerings, putting SMEs at a disadvantage. Competing effectively requires SMEs to find niche markets or unique value propositions that differentiate their services.

6. Lack of Data Augmented Processes

Many small and medium-sized enterprises (SMEs) still rely on outdated mechanical infrastructure that does not support modern technological capabilities. As a result, implementing AI-driven processes becomes cumbersome due to the lack of sufficient data. To address this challenge, data augmentation across various business and manufacturing processes is essential. As a substitute for fine-tuned or end-to-end trained models on custom data, foundation models (FMs) may provide an alternative solution, particularly in scenarios where resources or data are scarce.

2.4 Strategies for Successful Servitisation

To successfully implement servitisation, SMEs must focus on several key strategies:

Understanding Customer Needs

A deep understanding of customer needs is essential for developing effective servitised offerings. Engaging with customers through surveys or feedback mechanisms can provide insights into their preferences and pain points.

Investing in Technology

Investing in enabling technologies such as AI, IoT, and data analytics is critical for delivering personalized services. These technologies allow SMEs to monitor equipment health, predict failures, and optimise service delivery based on real-time data.

Training Employees

Training staff to adapt to new service-oriented roles is vital. This includes not only technical training but also customer service skills that foster long-term relationships with clients.

Developing Strategic Partnerships

Forming partnerships with technology providers or other businesses can help SMEs access resources they may lack internally. Collaborations can also enhance service offerings by combining expertise from different sectors.

Focusing on Sustainability

Emphasizing sustainability within the servitisation model can attract environmentally conscious customers. Offering solutions that reduce waste and promote efficient resource use aligns with growing market demands for sustainability.

2.5 AI as an Enabler of Servitisation

Artificial Intelligence (AI) plays a crucial role in supporting and enhancing business models like servitisation, contributing to a significant transformation in how manufacturing companies operate and interact with customers. This new paragraph will explore how AI can facilitate the transition to servitised models, improving operational efficiency, service personalisation, and sustainability.

1. Process Optimisation and Automation

AI enables companies to optimise their production and operational processes through automation. By utilising machine learning algorithms, businesses can analyse large volumes of data to identify inefficiencies and improve service quality. For example, AI tools can monitor machine conditions in real-time, predict failures, and schedule preventive maintenance interventions, reducing downtime and enhancing the reliability of the services offered.

2. Service Personalisation

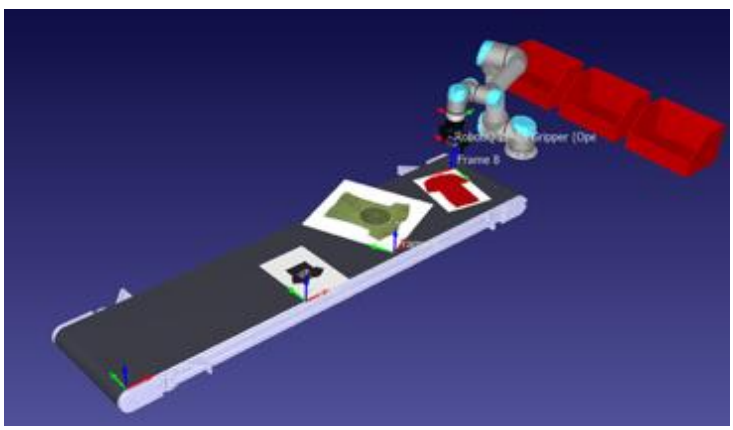
Servitisation involves a shift from product to service, and AI is fundamental in providing a personalised experience for customers. Through the analysis of user behaviour data and preferences, companies can develop tailored offerings that precisely meet individual customer needs. This approach not only increases customer satisfaction but also fosters loyalty, as customers perceive greater value in personalised services.

3. Sustainability and Innovation

The integration of AI into servitisation models also contributes to sustainability. AI technologies can help companies monitor resource consumption and optimise asset usage, reducing waste and costs. Additionally, AI facilitates the development of circular models where products are designed to be reused or recycled, promoting a more environmentally friendly approach to manufacturing.

4. AI innovation as Enabler for Sustainability and Digital Circular Economy

AI innovation as Enabler for Sustainability and Digital Circular Economy : Sustainability goes hand in hand with circular economy. AI innovation in various forms such vibration monitoring for predictive maintenance, machine vision for quality inspection, deep vision based robotic autonomous assembly and vison-based scrap, or second-hand material sorting are true digital enables for sustainability and circularity at manufacturing industries. The figure below demonstrated a digital twin case of computer vision enabled robotic application for sorting second hand textiles and t-shirts according to colour and material type for enabling a circular economy.



5. Data Augmentation and Data-Driven Decisions

Data augmentation in physical processes by digital means refers to the use of digital technologies, such as IoT, to enhance, simulate, or supplement real-world (physical) processes by generating additional data points or modifying existing data to improve outcomes, efficiency, or decision-making.

For SMEs, identifying opportunities for data augmentation, mapping relevant physical processes, and conducting thorough data analysis are essential steps. These efforts enable businesses to optimise operations, foster innovation, and support sustainable practices, including the adoption of emerging technologies and business models such as servitisation. Augmented data also empowers AI systems to effectively apply predictive analytics, regression, classification, computer vision, and other data-driven capabilities.

AI allows companies to make more informed decisions by providing access to insights derived from data analysis. This advanced analytical capability enables businesses to quickly adapt to market dynamics and customer needs, improving their competitiveness. Companies that use AI to gather and analyse data can identify emerging trends and anticipate future customer requirements, making their business model more agile and responsive.

6. Creation of New Revenue Opportunities

Finally, implementing AI in servitised models opens new revenue streams through innovative services such as Predictive Maintenance as a Service (PMaaS) or AI-based solutions for quality improvement. These new services not only generate recurring revenue but also allow companies to differentiate themselves in the market by offering added value to customers.

In summary, artificial intelligence is a key element in the transition toward servitisation models in the manufacturing sector. Through process optimisation, service personalisation, sustainability promotion, and support for data-driven decision-making, AI not only improves operational efficiency but also creates new opportunities for businesses. Companies that embrace these technologies can better position themselves to face future challenges and fully leverage the benefits of servitisation.

In conclusion, while servitisation presents significant opportunities for the manufacturing sector—especially for SMEs—it also poses unique challenges that must be navigated carefully. By understanding customer needs, investing in technology, training employees, forming strategic partnerships, and focusing on sustainability, SMEs can successfully transition to this innovative business model. As the industry continues to evolve, those who embrace servitisation will likely find themselves better positioned to compete in an increasingly complex market landscape.

References:

AIOTI WG Manufacturing. AIOTI WG Manufacturing. Cloud-Edge-IoT Continuum: Opportunities and Challenges for Manufacturing Industry. AIOTI Report Release 1.0 (2022). Available at: AIOTI Website.

Techmec. (2025). Servitizzazione: Il futuro dell'industria manifatturiera.

Econopoly - Il Sole 24 Ore. (2023). Servitizzazione: La mossa vincente per l'industria manifatturiera del futuro.

Agenda Digitale. (2024). Servitizzazione: Prospettive e tendenze per il 2024.

Technology Review Italia. (2024). Servitizzazione: La trasformazione del settore manifatturiero.

Exorint. (2025). Che cos'è la servitizzazione?

BeanTech. (2025). La servitizzazione: come utilizzarla al meglio per il post-vendita.

Universe IT Blog. (2025). Servitisation: Cos'è e come trasforma il settore manifatturiero.

Innovation Post. (2025). Servitizzazione e digitale: quali modelli per l'industria manifatturiera.

AI4Business. (2023). "AI: Technology for deriving business value from servitisation."

ESG360. (2023). "Artificial Intelligence and Servitisation: How to Improve Efficiency."

Innovation Post. (2024). "Artificial Intelligence and Machine Learning Accelerate Service Transformation."

Industria Italiana. (2024). "Business Model Patterns, Servitisation, and AI: The Pillars of Modern Industries."

Innovation Post. (2024). "Servitisation and Digital Technologies as Accelerators for Sustainability."

3. Power Consumption of AI

Artificial-intelligence workloads already draw an estimated 415 TWh (Terawatt-hour) of electricity worldwide (1.5 % of total electricity consumption in 2024) - and the IEA projects a doubling by 2030 in its conservative base case [IEA2025]. Frontier model training compete with heavy energy consuming industries like metal and chemical manufacturing [IEA2025]. GPT-4 for example was trained on 25000 GPUs and consumed an estimated 22 MW, including 12 MW consumed by supporting IT and cooling infrastructure [IEA2025], additionally according to Frymire and Rahman, "*the power required to train frontier AI models is doubling annually*" [Frymire2024]. In addition to energy costs during AI training, inference also consumes significant amounts of energy, especially as large language models have been rapidly adopted by both consumers and the industry.

AI services are not only CPU/GPU intensive but more so data Input/Output (I/O) intensive, and reliant on consuming large data assets (datasets). This implies novel dependencies between functional and data assets, which impacts the sustainability of running AI in terms of: energy consumption, green ratio of consumed energy a.o. CO2 footprint, economic cost, performance, privacy and cybersecurity.

From an infrastructure perspective, the energy profile of an AI workflow can be broken down to the combination of sub-profiles of:

- functional components of a workflow (e.g. data I/O, data preprocessing, model training, model evaluation, inference, and implementation);
- storage components of datasets feeding to a workflow; and,
- data-function mapping in a workflow, which implies data transmission that can be instigated over the network.

The above translates to energy consumption in relation to CPU/GPU (functional), memory (storage) and transmission (network). Existing work so far has proven that minimising all three – for maximum environmental and economic sustainability – is not feasible. Instead, trade-offs need to be considered to achieve efficient operation that supports AI sustainability. In this context, energy measurement is an important evaluation parameter for both model training and deployment/inference.

Moving towards cloudified AI services over an edge-to-cloud continuum, renders these trade-offs nontrivial. Because the energy cost, supply and CO2 footprint of edge-cloud Points of Presence (PoPs) differ and so is the different edge/core networks that connect them. The edge is known to have higher unitary cost of energy (e.g. per CPU rack, per sqm) [Bertoldi2017, Shehenaz2022]. The geographic density and distribution of PoPs varies too, particularly in the European landscape of edge-cloud and network infrastructure [Kamiya2024]. On the other hand, the emerging trends of localised production of green energy can play a major role in driving down energy costs of the edge.

Moving from the edge towards larger clouds, energy unitary costs (e.g. per rack, per sqm, per application) typically decreases. But that comes with the trade-off of longer distance between end-users/-devices (i.e. data generation points) and AI workflow components. This automatically introduce network artefacts, observed through data transmission from generation points to storage/processing counterparts. Consequently, increases the role of the network energy in the trade-off. Although, the opposite does not automatically hold. Meaning, bringing AI workflow to the edge does not automatically decrease exposure to the network. Because CPU and memory storage of edge PoPs are constrained. This increases the likelihood of having to distribute workflow components across multiple PoPs, introducing inter-component data transmission within a workflow.

Our work in [Ejaz2024] provides an elementary benchmark of resource utilisation, when distributing cloud workflows. It showed that distributing CPU-light microservices is likely to degrade the efficiency of a workflow, as the added CPU usage for routing across components outweighs the CPU usage of the components themselves. Therefore, from the perspective of AI workload management, placement/distribution policies need to systematically factor these trade-offs to support AI sustainability. This need, along with potential clustering of AI workflow components in larger data centres, is recognised by the World Economic Forum as strategic techniques to drive energy demand of AI lower [WEF2024].

From a ML perspective, different data ingestion and learning architectures results in different energy profile of the respective AI service. From an architectural perspective, federated and distributed learning can yield significant advantages in distributing and balancing the training load across multiple agents. The breakdown of such a large task into smaller sub-tasks allows for higher flexibility and larger opportunities for optimising the three energy profiles of computation, storage and networking. However, the flexibility is not without a cost that arise in the form of: fractional knowledge loss due to limited training per agent and to model aggregation, higher vulnerability to cyber threats/disruptions and higher management complexity due to the larger number of opened interfaces [Galaz2021]. The trade-off between load balancing, optimisation flexibility and risk exposure are often influenced/constrained by the landscape of infrastructure. In Europe, the higher dispersion of resources and larger number of autonomies fosters larger adoption of federated and distributed ML architectures; compared to international competitors elsewhere.

At the granularity of a single training process, data patterns directly correlate with the number of training cycles, required to balance the 'over-under fitting' trade-off and achieve a certain quality threshold. Higher entropy within datasets may induce a larger number of training cycles, to identify a specific pattern, as shown by [ALNaday2024] for anomaly detection use-cases in cybersecurity. Added to that, hyper parameters used in training are next in influencing the computational demand, as the speed with which a model can learn.

Emerging Large Language Models (LLMs), such as OpenAI and ChatGPT, introduce new variables into their sustainability. Most prominent of which, are: the number of parameters they train upon, the modality of data and the inference decisions they make [Wu2022, Luccioni2024]. These works show the correlation between the number of parameters (i.e. LLM size) and the CO2 emissions. The work of [Luccioni2024] further shows a breakdown of CO2 budget per LLM operation. Inference is found to have a significant contribution into CO2 emissions, in some cases outweighs that of model training and/or fine-tuning. Moreover, image training and inference have much higher CO2 footprint, compared to other data models. These efforts illustrate the trade-off between model size, data modality and computation demand, which present an orthogonal optimisation direction to that of workload management and even ML architecture to enable holistic optimisation of trade-offs that collectively support AI sustainability goals.

References:

[IEA2025] *Energy and AI*, International Energy Agency,Paris. Retrieved from https://www.iea.org/reports/energy-and-ai

[Frymire2024] Luke Frymire and Robi Rahman (2024), "The power required to train frontier AI models is doubling annually". Published online at epoch.ai. Retrieved from: 'https://epoch.ai/data-insights/power-usage-trend'

[Bertoldi2017] Bertoldi, P., Avgerinou, M. and Castellazzi, L., Trends in data centre energy consumption under the European Code of Conduct for Data Centre Energy Efficiency, EUR 28874 EN, Publications Office of the European Union, Luxembourg, 2017, ISBN 978-92-79-76445-5, doi:10.2760/358256, JRC108354.

[Shehenaz2022] Shehenaz Shaik, Sanjeev Baskiyar, "Distributed service placement in hierarchical fog environments," Sustainable Computing: Informatics and Systems, Volume 34, 2022, 100744, ISSN 2210-5379, https://doi.org/10.1016/j.suscom.2022.100744.

[Kamiya2024] Kamiya, G. and Bertoldi, P., Energy Consumption in Data Centres and Broadband Communication Networks in the EU, Publications Office of the European Union, Luxembourg, 2024, doi:10.2760/706491, JRC135926.

[Ejaz2024] S. Ejaz and M. Al-Naday, "FORK: A Kubernetes-Compatible Federated Orchestrator of Fog-Native Applications Over Multi-Domain Edge-to-Cloud Ecosystems," 2024 27th Conference on Innovation in Clouds, Internet and Networks (ICIN), Paris, France, 2024, pp. 57-64, doi: 10.1109/ICIN60470.2024.10494435.

[WEF2024] https://www.weforum.org/agenda/2024/04/how-to-manage-ais-energy-demand-today-tomorrow-and-in-the-future/

[ALNaday2024] AL-Naday, M., Dobre, V., Reed, M. et al. Federated deep Q-learning networks for service-based anomaly detection and classification in edge-to-cloud ecosystems. Ann. Telecommun. 79, 165–178 (2024). https://doi.org/10.1007/s12243-023-00977-4

[Galaz2021] Victor Galaz, Miguel A. Centeno, Peter W. Callahan, Amar Causevic, Thayer Patterson, Irina Brass, Seth Baum, Darryl Farber, Joern Fischer, David Garcia, Timon McPhearson, Daniel Jimenez, Brian King, Paul Larcey, Karen Levy, "Artificial intelligence, systemic risks, and sustainability," Technology in Society, Volume 67, 2021, 101741, ISSN 0160-791X, https://doi.org/10.1016/j.techsoc.2021.101741.

[Luccioni2024] Luccioni, Sasha, Yacine Jernite and Emma Strubell. "Power Hungry Processing: Watts Driving the Cost of AI Deployment?" Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (2023)

[Wu2022] Wu, Carole-Jean, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga Behram, James Huang, Charles Bai, Michael K. Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin S. Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Kumar Jain, Michael G. Rabbat and Kim M. Hazelwood. "Sustainable AI: Environmental Implications, Challenges and Opportunities." Proceedings of Machine Learning and Systems 4 (MLSys 2022)

4. Role of data architectures for training data and inferences

Flexible data architectures are required to enable AI applications in the manufacturing domain for several key reasons:

- **Integration of Heterogeneous Data Sources** Manufacturing environments generate vast amounts of data from diverse sources, including sensors, machines, enterprise systems (ERP, MES), and digital twins. A flexible data architecture ensures seamless and semantic integration of these heterogeneous data sources.

- **Handling High-Volume and High-Velocity Data** AI applications require access to real-time and historical operational data. Traditional data architectures struggle with large-scale time-series data, making a more adaptable approach necessary.

- **Semantic Enrichment for AI and ML Models** AI-driven insights depend on contextualized data. Flexible architectures, such as RDF-based knowledge graphs, provide semantic descriptions that enhance data interoperability and reuse.

- **Scalability and Extensibility** AI applications evolve over time, requiring additional sensors, new data sources, or expanded analysis. A flexible architecture allows for seamless scaling without requiring major system overhauls.

- **Standardisation and Interoperability** Existing industrial standards (e.g., OPC UA) do not fully support high-frequency, enriched datasets for AI. A flexible data architecture bridges this gap by enabling structured queries (e.g., via SPARQL) or filtered queries through a single API endpoint (e.g., via GraphQL) along with their integration to knowledge graphs.

- **Support for AI-based Control Loops** Real-time AI applications, such as predictive maintenance and process optimisation, require rapid data access and feedback loops. A flexible architecture enables AI models to retrieve, process, and act on data efficiently.

- **Federated Learning and Cross-Company Data Sharing** In collaborative AI projects, such as federated learning for predictive maintenance, a flexible architecture ensures secure and efficient data exchange while preserving sovereignty.

The key components of a **flexible data architecture** for enabling AI applications in manufacturing include:

1. Data Ingestion & Adaption Layer

- Collects data from various sources such as sensors, machines, PLCs, MES, ERP, and cloud services.

- Supports multiple protocols (OPC UA, MQTT, Kafka, Modbus, BACnet, REST APIs).

- Converts raw data coming from heterogenous sources with different data models into standardized formats (e.g., LF-JSON for LinkedFactory).

2. Data Storage & Management Layer

- Time-Series Databases: Store high-frequency sensor and machine data. (e.g., InfluxDB, LevelDB, Parquet)

- Knowledge Graphs (RDF-based): Store semantic metadata, relationships, and contextual information.

- Relational & NoSQL Databases: Handle structured and unstructured data for AI applications.

3. Data Processing & Querying Layer

- SPARQL Query Engine: Merges semantic and time-series data for AI model training and analytics.
- Data Transformation & Preprocessing: Aggregation, normalisation, resampling, and feature extraction.
- Federated Querying: Combines multiple data sources dynamically.

4. AI & Analytics Layer

- Machine Learning Models: Supports predictive maintenance, process optimisation, anomaly detection.
- Streaming Analytics: Real-time AI inference for adaptive control.
- Model Training & Optimisation: Uses historical and live data for continuous learning.

5. APIs & Interoperability Layer

- RESTful & HTTP APIs: Enables easy access to stored data.
- SPARQL Endpoints: Allows querying knowledge graphs.
- GraphQL Endpoint: Provides a flexible and efficient way to query and manipulate data through a single API endpoint, allowing clients to request exactly the data they need.
- Industry Standards Integration: Supports OPC UA, AAS (Asset Administration Shell), IDS (Internation Data Space) and GAIA-X frameworks.

6. Visualisation & Dashboarding

- Grafana or Custom Dashboards: Real-time monitoring of AI insights and production KPIs.
- Semantic Data Exploration Tools: Interactive views for understanding relationships between machines, products, and processes.

7. Edge & Cloud Computing

- Edge Devices: Local AI processing for real-time decision-making.
- Cloud Infrastructure: Scalable storage and compute for AI training and cross-factory data sharing.
- Edge Enabled Dataspace : Facilitates cross-domain pooling of infrastructure and data among coordinated edges (e.g., factories as edge nodes) while maintaining resource sovereignty, traceability, and transparency of pooled resources [EdgeDS2024]. This creates a trusted multi-stakeholder environment that mitigates data silos, enabling mutual resource gains for a win-win business model driven value chain among coordinated organisational edges [Diss2024]. Additionally, resource owners at the edge retain control over local resources for federated and transfer learning-based AI processing methods, supporting cross-domain real-time data or event processing (e.g., in supply chain-enabled multi-stakeholder environments).

  This approach further supports the servitisation model enablement in diverse monetisation models, similar to cloud-driven IaaS (Infrastructure as a Service), PaaS (Platform as a Service), and SaaS (Software as a Service) models, but focusing on organisation-to-

organisation (or org-to-org) collaborative edge-driven Dataspace ecosystem. In this context, International Data Space Standards (IDS) can be leveraged to develop such cross-domain data integration ecosystem to converge modern technologies and integrate data from heterogeneous sources and domains with the required sovereignty and governance in place [Singh2024].

8. Security & Data Governance

- Access Control & Authentication: Ensures data privacy and security.
- Federated Learning & Data Sovereignty: Secure cross-domain AI model training without sharing raw data.
- Compliance with Industry Standards: GDPR, ISO 27001, and other regulatory frameworks.

References:

Flexible data architecture for enabling AI applications in production environments https://nbn-resolving.org/urn:nbn:de:bsz:l189-qucosa2-943149

[Singh2024]Singh, P., Meratnia, N., Beliatis, M. J., & Presser, M. (2024). Navigating the International Data Space To Build Edge-Driven Cross-Domain Dataspace Ecosystem. In International Summit on the Global Internet of Things and Edge Computing (pp. 151-168). Cham: Springer Nature Switzerland.

[EdgeDS2024]Singh, P., Beliatis, M. J., & Presser, M. (2024). Enabling edge-driven dataspace integration through convergence of distributed technologies. Internet of Things, 25, 101087.

[Diss2024] Singh, P. (2025). Enabling Dataspace Value Chain at Edge for Industry 4.0 and Beyond: Technologies Convergence and Cross-Domain Data Integration over Distributed Edge Architectures (Doctoral dissertation, Department of Business Development and Technology, Aarhus BSS, Aarhus University). Retrieved from:https://btech.au.dk/fileadmin/AU_Herning/Pdf/Ph.d.-afhandlinger/PhD_dissertation_Parwinder_Singh.pdf

5. Agentic AI paradigm

Agentic AI represents a paradigm shift in artificial intelligence, focusing on autonomous, goal-oriented behavior. Agentic AI systems are designed to act as independent agents capable of perceiving their environment, making decisions, and executing actions with minimal human intervention: an agent can reason about what it should do next to achieve its ultimate goal, even in the absence of explicit instruction sets [Shinn2023].

In the context of generative AI, agentic AI represents a significant advancement by enabling models to go beyond passive content generation, in fact, one of the most transformative aspects is the ability to use external tools and APIs to enhance its reasoning, decision-making, and content generation capabilities [Schick2023]. Unlike traditional generative AI, which passively generates content based on static inputs, agentic models can actively invoke external tools, fetch real-time data, run computations, and interact with software systems to achieve their goals. In the manufacturing context, AI agents can manage real-world tasks—e.g., coordinating robot arms or mobile manipulators—demonstrated by open-ended embodied agents such as Voyager [Wang2023].

Building upon the Cloud-Edge-IoT continuum discussed in Section 1.1, agentic AI can leverage this architecture to optimize manufacturing processes. High-level planning and computationally intensive tasks can be managed in the cloud, ensuring energy efficiency and compliance. Simultaneously, lightweight sub-agents operate at the edge, handling real-time control loops essential for robotic operations. This distributed approach enhances responsiveness and reliability on the factory floor.

In manufacturing scenarios, agentic AI can be designed to treat legacy robotic arms as versatile tools within its operational toolkit. Vision-language-action models, such as DeepMind's Gemini Robotics, enable on-the-fly generation of grasps and trajectories, facilitating brownfield deployments without the need for bespoke programming. This adaptability accelerates integration into existing manufacturing setups [Gemini2025].

A hallmark of agentic AI is its capacity to deconstruct high-level goals into manageable subgoals, iteratively refining strategies based on feedback and new information [Park2023]. This proactive, self-directed nature allows agentic systems to dynamically adapt to evolving circumstances in manufacturing, supply chains, and customer demand. For instance, within a Manufacturing-as-a-Service (MaaS) model [Corti2021], agents can autonomously log machine usage, trigger usage-based invoicing, and dispatch maintenance crews pre-emptively to avert service-level agreement violations.

Agentic AI has the potential to transform business models and redefine operations across various industries, including healthcare, manufacturing, and beyond. Below are some examples of how it is poised to make an impact:

1. **Software Development and Adaptive Software Systems**

   Agentic AI is set to revolutionize the software development lifecycle by autonomously generating substantial portions of code, offering intelligent suggestions, and helping optimize programming practices. Beyond coding, Agentic AI is increasingly integrated into adaptive or "living" software systems that dynamically modify their functionalities based on environmental changes. These systems can rapidly reconfigure themselves through self-learning, improving with each use.

2. **Cybersecurity**

   By examining application code, network activity, user behaviors, and system logs, Agentic AI can detect anomalies and flag potential security threats in real-time.

3. **Manufacturing**

   Agentic AI applications in manufacturing include:

   Production: intelligent production scheduling, predictive maintenance, and real-time quality control, driving efficiency and reducing downtime.

   Inventory: by evaluating historical sales, seasonal trends, marketing activity, economic indicators, and even social media sentiment, Agentic AI can deliver precise demand forecasts and automatically generate optimized purchase orders.

   Supply Chain: AI agents can dynamically optimize transportation logistics, taking into account variables such as traffic patterns, weather conditions, and delivery timelines to ensure efficient routing.

4. **Customer Service**

   In an e-commerce setting, traditional chatbots rely on fixed scripts and keyword matching, often leading to rigid and limited interactions. In contrast, Agentic AI goes further by understanding context, performing real-time sentiment analysis, and dynamically adapting its tone and problem-solving strategies. This allows it to pull in relevant external knowledge and provide more accurate, empathetic, and helpful responses, ultimately improving user satisfaction.

5. **Financial Services**

   In finance, Agentic AI can be used for algorithmic trading, fraud detection, and dynamic portfolio management. It can analyze market trends in real time, assess risks, make autonomous investment decisions, and continuously adjust portfolios based on evolving market conditions and personalized financial goals. This enables more responsive, data-driven financial strategies with minimal human intervention.

6. **Healthcare**

   In medicine, Agentic AI enhances clinical decision-making by analyzing patient records and symptoms to propose potential diagnoses. Beyond diagnostic support, it can continuously monitor patient vital signs, detect patterns indicating possible complications, and autonomously alert healthcare providers. This proactive capability enables timely interventions without human input.

7. **Education**

   Agentic AI in education enables personalized learning by adapting content to each student's needs. It provides real-time, context-aware feedback to enhance understanding and engagement. The technology may support teachers by automating grading and content generation. Agentic AI also supports exploration-based learning by guiding students through complex problems dynamically.

8. **Government**

   Government agencies can leverage Agentic AI to manage inquiries, process applications for services or permits, deliver tailored information to citizens, and assist with navigating complex administrative processes.

AI agents can be categorized into two types:

- **Virtual AI agents** operate within digital environments, functioning as assistants, advisers, or automation tools to autonomously accomplish specific objectives.

- **Embodied AI agents** are integrated into physical systems, such as robots. These agents interpret their surroundings through sensors (e.g., cameras, lidar, radar, microphones) and act using actuators like advanced grippers, enabling complex and adaptive physical tasks.

The influence of AI agents on industrial operations is expected to be profound. In the near future, they are likely to become foundational components of factories worldwide, enabling a shift toward AI-driven, semi- or fully autonomous operations. However, with these advancements come challenges - particularly in areas such as security, legal compliance, social responsibility, and infrastructure readiness.

To navigate this transition effectively, a collaborative approach involving businesses, academic institutions, and policy-makers is essential. Sharing best practices and insights will be key to ensuring the responsible and sustainable adoption of AI technologies in industry.

References:

[Park2023] Park, Joon Sung, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. "Generative agents: Interactive simulacra of human behavior." In Proceedings of the 36th annual acm symposium on user interface software and technology, pp. 1-22. 2023.

[Schick2023] Schick, Timo, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. "Toolformer: Language models can teach themselves to use tools." Advances in Neural Information Processing Systems 36 (2023): 68539-68551.

[Shinn2023] Shinn, Noah, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. "Reflexion: Language agents with verbal reinforcement learning." Advances in Neural Information Processing Systems 36 (2023): 8634-8652.

[Gemini2025] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna et al. "Gemini robotics: Bringing ai into the physical world." arXiv preprint arXiv:2503.20020 (2025).

[Corti2021] Corti, Donatella, Andrea Bettoni, Elias Montini, A. Barni, and E. Arica. "Empirical evidence from the design of a MaaS platform." IFAC-PapersOnLine 54, no. 1 (2021): 73-79.

Google. (2025). Agents Companion, Acharya, D. B., Kuppan, K., & Divya, B. (2025). Agentic AI: Autonomous Intelligence for Complex Goals–A Comprehensive Survey. IEEE Access.

World Economic Forum. (2025). Why should manufacturers embrace AI's next frontier – AI agents – now?

6. Synergies with industrial fields

The pervasive deployment of AI among different manufacturing sectors poses several challenges and opportunities towards best practices in cross-industrial applications. From one hand, the most promising Machine Learning (ML) models which are being exploited in manufacturing are data-driven ones [Elahi2023]. This implies that the datasets used for training are domain-specific and software houses can hardly provide companies with prêt-à-porter solutions but have to consider effort and time-consuming activities to train the models with the domain (or, more frequently company) specific datasets. From the other hand, the software architectures these models are framed into can be easily scaled up into different companies, as they often rely on open-source tools and the configuration effort is usually limited to the Application-to-Program Interfaces (APIs) towards the machine or the existing informative system [Patel2023]. It is also worth mentioning that nowadays "standardized" pipelines and models are growing in popularity. These tools, usually concerning computer vision, are either pre-trained models able to recognise common patterns or figures (e.g., MediaPipe, which recognises keypoints of human body and hands [Lugaresi 2019]) or are such versatile that can be easily retrained with low-size datasets and exploited into various sectors (e.g., the various distributions of YOLO [Redmon2016]).

In any case, the greater opportunity for benefiting from other industries sits probably in the so-called "Industry 5.0", which some scholars see as a resilient measure of European manufacturing against the phenomenon of the population ageing [Katiraee2024]. Under this perspective, the human-centric paradigm which characterises this initiative materialises, at a shopfloor level, with the need of extending the working life of operators, which implies technological measures aimed at relieving these operators from tasks which demand a high level of physical effort. Practical implications of this trend are two-folded (and not mutually exclusive): one approach implies the "empowerment" of the physical capabilities of operators, which can be realised through technological addons such as exoskeletons, collaborative robots, and smart glasses for Augmented Reality [Romero2020]. If the deployment of exoskeletons does not necessarily imply the usage of ML models, the interaction between humans and robots is a fertile ground for the application of several ML tools, in particular the ones belonging to the domains of Large Language and Vision Language Models (LLMs and VLMs [Zhang2024]) which provide operators with the capability of naturally interact with machinery. These tools, combined with the recent advancements in the field of ego-centric perception are supposed to increase the cooperation of human and robot to a seamless level [Wang2024].

The complementary approach is more focused on the study of tasks themselves and to analyse the physical strain produced by the manual work of operators, companies are engaging universities and RTOs to deploy solutions which can automatically quantify this type of strain. These solutions typically rely either on Computer Vision tools (e.g., leveraging on pipelines such as MediaPipe to assess the posture of an operator via a Rapid Upper Limb Assessment, RULA) or exploit devices worn by operators (such as EEGs, ECGs, EMGs) to obtain, in a hyper-dimensional space, from which it can be deduced through ML models, how much a task is physically demanding. These preliminary studies, which draw from the domains of healthcare and neuroscience, can help manufacturing companies in configuring their production lines in order to minimise the physical strain of operators, reducing the risk of injuries and chronical diseases [Syed2024].

In any case, the findings and procedures derived from the healthcare domain should be taken with a grain of salt, instead of being blindly adopted: as a matter of example, recent researches have highlighted how the "pure" healthcare protocol of deriving the patients baselines while the subject is sitting, could induce some biases in a ML solution applied to the manufacturing domains, where operators typically perform their tasks while standing [Syed2025].

Another aspect of crucial importance where ML models can provide benefits to the manufacturing domain is the decision making: even if these tools have been already exploited in factories (e.g., with decision concerning maintenance [Nguyen2022]) the experience of other sectors in embodying these models is driving several innovation actions also in manufacturing: it is, for example, the case of financial sector, where the so-called "sentiment analysis" has been employed for years in the stock investment [Xu2019], while nowadays is starting also being employed in the product design [Islam2024].

At the same time, models that can predict the fluctuations of prices and availability of energy, which in the energy domain are used to optimise the loads on power plants [Kiasari2024], are matching the interests of companies dealing with price fluctuations of raw materials (e.g., metal and wood industry) [Shoushtari2024], even if the difference of time constants of these phenomena require some adaptations too.

The same approach can be applied to food sector and process industry, where both the domains see a chemical seasonality or variability in their row materials, but need to respect organoleptic or physical constraints in their finished products: under this light, the selection of metal scrap to be melted in an Electric Arc Furnace can be framed as the one related to the food mixture for beef cattle [Kannisto2024, Garcìa2023].

In conclusion, while the integration of AI and ML in manufacturing presents significant challenges – in particular for what concerns domain-specific data requirements and the adaptation of models to different industrial contexts – it also paves the way for innovation and cross-sectoral synergy. By drawing from advancements in fields such as healthcare, energy, and finance, and by embodying the human-centric vision of Industry 5.0, manufacturing can evolve towards more resilient, adaptive, and sustainable systems. The key enabler of a virtuous contamination lies in balancing the customisation of AI tools with the reuse of scalable, standardised components, ensuring that technological progress stays aligned with both operational efficiency and human well-being.

References:

[Elahi 2023] Elahi, M., Afolaranmi, S. O., Martinez Lastra, J. L., & Perez Garcia, J. A. (2023). A comprehensive literature review of the applications of AI techniques through the lifecycle of industrial equipment. Discover Artificial Intelligence, 3(1), 43.

[Patel 2023] Patel, D., Lin, S., Shah, D., Jayaraman, S., Ploennigs, J., Bhamidipati, A., & Kalagnanam, J. (2023, September). AI model factory: scaling AI for industry 4.0 applications. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 13, pp. 16467-16469).

[Lugaresi2019] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., ... & Grundmann, M. (2019). Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

[Redmon2016] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

[Katiraee2024] Katiraee, N., Berti, N., Das, A., Zennaro, I., Aldrighetti, R., Dimovski, V., ... & Battini, D. (2024). A new roadmap for an age-inclusive workforce management practice and an international policies comparison. *Open Research Europe*, *4*, 85.

[Romero2020] Romero, D., Stahre, J., & Taisch, M. (2020). The Operator 4.0: Towards socially sustainable factories of the future. *Computers & Industrial Engineering*, *139*, 106128.

[Zhang2024] Zhang, J., Huang, J., Jin, S., & Lu, S. (2024). Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Wang2024] Wang, T., Fan, J., & Zheng, P. (2024). An LLM-based vision and language cobot navigation approach for Human-centric Smart Manufacturing. *Journal of Manufacturing Systems*, *75*, 299-305.

[Syed2024] Syed, D. A., Quadrini, W., Rahmani Choubeh, N., Pinzone, M., & Gusmeroli, S. (2024). Approaching Interoperability and Data-Related Processing Issues in a Human-Centric Industrial Scenario. In *International Summit on the Global Internet of Things and Edge Computing* (pp. 21-34). Cham: Springer Nature Switzerland.

[Syed2025] D. A. Syed, W. Quadrini, M. Pinzone, (2025) "Impact of Sitting vs Standing Baselines on Performance Parameters of Stress Classification Models in Assembly Tasks ",IFAC papers online IN PRINT.

[Nguyen2022] Nguyen, V. T., Do, P., Vosin, A., & Iung, B. (2022). Artificial-intelligence-based maintenance decision-making and optimisation for multi-state component systems. *Reliability Engineering & System Safety*, *228*, 108757.

[Xu2019] Xu, Y., & Keselj, V. (2019, December). Stock prediction using deep learning and sentiment analysis. In *2019 IEEE international conference on big data (big data)* (pp. 5573-5580). IEEE.

[Islam2024] Islam, M. S., Kabir, M. N., Ghani, N. A., Zamli, K. Z., Zulkifli, N. S. A., Rahman, M. M., & Moni, M. A. (2024). Challenges and future in deep learning for sentiment analysis: a comprehensive review and a proposed novel hybrid approach. *Artificial Intelligence Review*, *57*(3), 62.

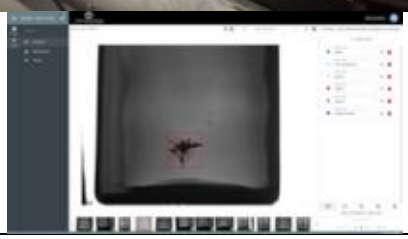
[Kiasari2024] Kiasari, M., Ghaffari, M., & Aly, H. H. (2024). A comprehensive review of the current status of smart grid technologies for renewable energies integration and future trends: The role of machine learning and energy storage systems. *Energies*, *17*(16), 4128.

[Shoushtari2024] Shoushtari, F., Najafi Zadeh, M. S., Ghafourian, H., & Karim Zadeh, E. (2024). Applications of Machine Learning in Financial Accounting for Industrial Engineering: A Case Study on Cost Estimation and Forecasting. *Available at SSRN 4991489*.

[Kannisto2024] Kannisto, P., Kargar, Z., Alvarez, G., Kleimt, B., & Arteaga, A. (2024). Resilient, Adaptive Industrial Self-X AI Pipeline with External AI Services: A Case Study on Electric Steelmaking. *Processes*, *12*(12), 2877.

[Garcìa2023] García, R., Aguilar, J., Toro, M., Pérez, N., Pinto, A., & Rodríguez, P. (2023). Autonomic computing in a beef-production process for Precision Livestock Farming. *Journal of Industrial Information Integration*, *31*, 100425.

7. Use cases/Examples

APICUS - Video Systems	
APICUS demonstrator platform includes physical anthropomorphic robot equipped with a set of machine vision systems dedicated to quality control tasks. The demonstrator uses machine vision technologies powered by deep-learning engine to quality control task of components manipulated by the robot. The pilot demonstrates integration of edge and cloud technologies that uses IIoT to share information for complex Agile manufacturing scenarios simulation. The station is connected to Cloud-based AI learning system based on Docker technology and NVIDIA GPUs accelerations. Different pilots are ready to test on various manufacturing sectors (Hollow Glass, high precision mechanical, foundry, furniture, field of household appliances, automotive.	  
QUESCREM AI at the Edge for Food Sustainability Pilot	
The QUESCREM pilot is part of the AI REDGIO 5.0 project. It showcases how AI-based Edge Computing systems can positively impact and modernise the food manufacturing industry. Located in Galicia, Spain, this initiative is spearheaded by Quescrem, a cream cheese producer, in collaboration with Gradiant technology center. This industrial pilot successfully enhanced product quality and sustainability by integrating AI-driven decision-making directly into the production line. Specifically, leveraging sensor data and real-time streams from Quescrem's manufacturing processes, AI algorithms deployed at the edge can analyse and predict key quality indicators—such as the cream cheese's texture— and adjustments in real time to optimise outcomes and reduce waste. The most innovative element in this pilot is the development of an AI-Edge Continual Learning model enabling the system to dynamically adapt to new data and evolving production conditions. Unlike traditional AI models trained in centralised cloud environments, this edge-based approach supports low-latency, energy-efficient processing and incremental learning. More specifically, the model combines ILVQ-based prototypes with Deep Neural Decision Forests (DNDF), allowing it to learn from new data and improving its knowledge processes. This capability is crucial for maintaining consistent product quality and responding to concept drift in real-time. Importantly, the system follows a Human-In-The-Loop (HITL) paradigm, where AI supports but does not replace human decision-making, aligning with the Industry 5.0 vision of collaborative intelligence.	 

SALF – Consortio Intellimech	
Founded in 1921, SALF specialises in producing injectable pharmaceuticals and medical devices. It uses a complex process that includes multiple production phases, followed by mandatory laboratory tests to assess and certify product quality. This use case aimed to optimise the scheduling of these mandatory analyses to reduce energy consumption, resource consumption (e.g., reagent), and makespan. The scheduling complexities arise because, unlike production, laboratory processes do not have strictly ordered sequences (Open-Shop Scheduling Problem) and require the parallel handling of some analyses, adding complexity to the scheduling and optimisation efforts. This use case highlights the potential of integrating symbolic constraints into a Reinforcement Learning framework for solving the problem considering realistic constraints, such as operator shifts and preparation times. A Proximal Policy Optimization (PPO) agent is trained using curriculum learning, progressively increasing environment complexity. The agent learns to minimise makespan and strategically delay tasks to maximise merge operations across batches with compatible test specifications and reduce material consumption. Moreover, explainable techniques are introduced to support the decision-making process. The best-performing policy achieved a 35% reduction of task preparations and a reduced makespan of 471.5 out of 510 possible hours, demonstrating effective planning and non-trivial scheduling behaviour.	 

8. Conclusions/Recommendations

Artificial Intelligence has transcended its status as an emerging technology to become a transformative force actively reshaping the manufacturing landscape. As demonstrated throughout this analysis, AI's integration encompasses both tactical implementations—including predictive maintenance systems, real-time quality inspection platforms, and automated process optimisation—and fundamental systemic transformations spanning servitisation business models, human-machine collaborative frameworks, and sustainable computational architectures. The convergence of AI with cloud-edge infrastructures, evolving regulatory frameworks, and emerging agentic paradigms represents a defining moment in the evolution of industrial innovation.

The manufacturing sector's AI transformation presents multifaceted challenges that demand strategic attention and coordinated response. Manufacturers, particularly small and medium enterprises, must simultaneously navigate complex regulatory landscapes such as the EU AI Act, invest in explainable and ethically-aligned AI systems, and adapt to rapidly evolving technological environments. Success requires maintaining unwavering focus on human-centric design principles, environmental sustainability objectives, and cross-sectoral collaboration. While these transformation requirements are substantial, they are fundamental to ensuring that AI deployment serves broader societal goals beyond economic efficiency, encompassing social responsibility and long-term industrial resilience.

The technical complexity of implementing AI solutions across distributed manufacturing environments demands new approaches to system integration, data governance, and workforce development. Organisations must develop capabilities spanning AI model lifecycle management, regulatory compliance frameworks, and sustainable computing practices while fostering innovation cultures that embrace human-AI collaboration.

This research represents an initial contribution to what must become a sustained, collaborative dialogue. Our primary objective is to catalyse meaningful discourse across the manufacturing ecosystem—engaging companies, research institutions, technology providers, and policymakers—to collectively define how AI can be responsibly and effectively integrated into European manufacturing's foundational structures. We firmly believe that only through shared exploration, open knowledge exchange, and coordinated collaborative efforts can the manufacturing community develop the comprehensive understanding, standardised practices, and governance frameworks necessary to guide this transformation in ways that benefit all stakeholders.

The complexity and scope of AI's manufacturing transformation exceed any single organisation's capacity to navigate independently. Success requires unprecedented collaboration across traditional industry boundaries, academic disciplines, and regulatory domains. This collective approach must encompass technical standards development, ethical framework establishment, and practical implementation guidance that reflects diverse organisational contexts and capabilities.

We call upon manufacturing stakeholders to actively engage in shaping this transformation through concrete actions:

Industry practitioners should contribute detailed case studies documenting both successful implementations and lessons learned from challenges, providing the community with practical insights that accelerate collective learning and reduce implementation risks.

Research institutions must continue advancing fundamental AI capabilities while addressing manufacturing-specific challenges including explainability, energy efficiency, and human-AI interaction design, ensuring academic research remains aligned with industry needs.

Technology providers should prioritize the development of accessible, ethical AI solutions designed specifically for manufacturing contexts, with particular attention to SME requirements and regulatory compliance capabilities.

Policymakers must foster regulatory environments that balance innovation enablement with responsible deployment, supporting industry transformation while protecting societal interests.

Standards organisations should accelerate the development of comprehensive frameworks addressing AI implementation, interoperability, and governance across manufacturing applications.

The opportunity for proactive engagement is immediate and time-sensitive. The current period represents a critical window for shaping AI's manufacturing integration trajectory. Stakeholders must move beyond passive technology adoption to become active architects of the manufacturing future through responsible, inclusive, and strategically-oriented innovation initiatives.

This transformation requires organisations to simultaneously embrace technological advancement while maintaining focus on human empowerment, environmental stewardship, and societal benefit. The manufacturing sector's response to this challenge will determine whether AI becomes a force for comprehensive positive transformation or merely another efficiency tool with limited strategic impact.

This work initiates what must become an ongoing, dynamic conversation—a collaborative foundation for building shared understanding of AI's evolving role in manufacturing transformation and establishing collective commitment to deploying intelligence in service of human progress, operational excellence, and sustainable development.

We invite the manufacturing community to treat this analysis as a starting point for deeper exploration, critical questioning, and collaborative innovation. The future of manufacturing will be defined not by individual organisational AI adoption but by the sector's collective wisdom in navigating this transformation thoughtfully, inclusively, and with unwavering commitment to serving broader societal objectives.

The conversation begins now. The future of manufacturing depends on our collective response.

Contributors

Editor:

Sergio Gusmeroli	Politecnico di Milano
Alessandro Liani	Video Systems

Reviewer:

Damir Filipovic	AIOTI Secretary General

Contributors:

Alessandro Liani	Video Systems
Claudio de Majo	Gradiant
Davide Cesani	Intellimech
Davide Pasanisi	Intellimech
Francesco Alessandro Cuzzola	Politecnico di Milano
Ken Wenzel	Fraunhofer
Marielena Márquez Barreiro	Gradiant
Mays AL-Naday	University of Essex
Michael Lystbæk	Aarhus University
Michail J. Beliatis	Aarhus University
Mirko Presser	Aarhus University
Parwinder Singh	Aarhus University
Unnikrishnan Radhakrishnan	Aarhus University
Walter Quadrini	Politecnico di Milano

Acknowledgements

About AIOTI

AIOTI is the multi-stakeholder platform for stimulating AI, IoT and Edge Continuum Innovation in Europe, bringing together small and large companies, academia, researchers, policy makers, end-users and representatives of society in an end-to-end approach. We strive to leverage, share and promote best practices in the AI, IoT and Edge Continuum ecosystems, be a one-stop point of information to our members while proactively addressing key issues and roadblocks for economic growth, acceptance and adoption of the AI, IoT and Edge Continuum Innovation in society. AIOTI contributions goes beyond technology and addresses horizontal elements across application domains, such as matchmaking and stimulating cooperation by creating joint research roadmaps, defining policies and driving convergence of standards and interoperability.