Multi-Partner Project: Artificial Intelligence in Manufacturing Leading to Sustainability and the Consideration of Human Aspects (AIMS5.0)

Anouar Nechi[1] Yasin Ghafourian[2], Belal Abu Naim[2], Thomas Gutt[3], George Dimitrakopoulos[4]
Amira Moualhi[1], Mladen Berekovic[1], Pal Varga[5], and Markus Tauber[2]

[1]*University of Lübeck *, Lübeck, Germany*
[2]*Research Studios Austria *, Vienna, Austria*
[3]*IFAG BEX RDE RDF CFA, Infineon Technologies AG *, Neubiberg, Germany*
[4]*Harokopio University, Athens *, Greece*
[5]*Budapest University of Technology and Economics *, Budapest, Hungary*

***Abstract*—The industrial landscape is undergoing a transformative shift towards Industry 5.0, a paradigm characterized by the convergence of sustainability, digital autonomy, and human-centric design. This article focuses on the adoption, enhancement, and implementation of AI-driven hardware, tools, methodologies, and semiconductor technologies in this progression. We present here a comprehensive strategy from the AIMS5.0 project with the objective of connecting academic developments with practical industrial use, fostering a harmonious relationship between humans and machines to improve efficiency, spur innovation, and enhance adaptability. Hence we show here our global vision, and examples of how the creation of AI-based industrial solutions is supported by novel AI-tool chains, advancements in hardware, and tools supporting human aspects.**

***Index Terms*—Artificial Intelligence (AI), Cyber-Physical System of Systems (CPSoS), Co-pilots, Sustainability, Optimization, Compliance Check**

I. Introduction

Artificial Intelligence (AI) has emerged as a transformative force in the manufacturing industry, with the potential to revolutionize processes, enhance efficiency, and drive sustainability. As we navigate the complexities of the Fourth Industrial Revolution, AI offers a promising avenue to address the challenges of global competition, resource scarcity, and climate change. By leveraging AI-driven technologies, manufacturers can optimize operations, improve product quality, and develop innovative solutions that meet the evolving needs of consumers and society.

Integrating AI effectively into manufacturing requires more than just implementing technology. It needs a strategic approach. This means understanding the industry's specific challenges and opportunities and committing to developing and using AI solutions that are both effective and ethical. AI can contribute significantly to sustainability by optimizing resource utilization and promoting circular economy principles. This can be achieved through AI-driven solutions in areas such as predictive maintenance, energy management, and supply chain optimization, ultimately leading to minimized waste and reduced emissions.

Furthermore, successful AI implementation hinges on recognizing the complementary strengths of humans and machines. AI systems should be designed to augment, rather than replace, human capabilities such as creativity, problem-solving, and critical thinking. This collaborative approach fosters a more productive and engaging work environment. Additionally, bridging the gap between academic research and industrial practice is paramount. Fostering collaboration among researchers, industry professionals, and policymakers can accelerate the development and deployment of practical AI solutions. These solutions should address the industry's specific needs while driving economic growth and societal advancement.

This paper presents a comprehensive framework to achieve this goal, focusing on the following key contributions:

- **Human-Machine Partnership:** Designing AI systems that complement human strengths (creativity, problem-solving) while automating repetitive or hazardous tasks, leading to optimized efficiency and adaptability.
- **AI-Driven Solutions**: Showcasing practical AI applications with proven benefits in manufacturing. Examples include predictive maintenance to reduce downtime, quality control using computer vision, and supply chain optimization for improved logistics.
- **AI Toolchains and Hardware**: Utilizing advanced AI toolchains and hardware to enable the development and deployment of effective AI solutions. These tools offer faster iteration, scalability testing, enhanced security, and cost-effective development.
- **Human-Centered Design:** Prioritizing human factors and ethical considerations in AI development requires careful attention. This includes conducting user research, incorporating human-centered design principles, and ensuring AI augments human capabilities rather than replacing them.

* The authors' organizations are members of the INSIDE Industrial Association.

978-3-9826741-0-0/DATE25/© 2025 EDAA

II. AI-TOOL CHAINS

In this section, we present our toolchain, designed to offer a robust development environment for AI models through an effective combination of Apache Zeppelin [1], Eclipse Arrowhead [2], and IoT components. This integration bridges the gap between development and production-ready testing. We also introduce AI co-pilots to support different engineering stages in the Cyber-Physical System of Systems (CPSoS).

A toolchain is a collection of interconnected software tools that automate and streamline different stages of the development workflow. In software development, for example, these stages include coding, building, testing, and deployment, with each tool in the toolchain fulfilling a specific role. Toolchains are extensively used in DevOps and MLOps, where continuous integration, continuous deployment (CI/CD), and continuous testing are essential for delivering reliable and scalable software and machine learning solutions.

An AI-gym provides a controlled environment for developing, training, and testing AI algorithms on diverse tasks before real-world deployment. Introduced in 2016 [3], OpenAI Gym is an open-source platform with standardized environments for developing and testing reinforcement learning (RL) algorithms. This platform has inspired similar tools like DeepMind Lab [4], Unity's ML-Agents Toolkit [5], and Intel's OpenVINO Toolkit [6], focused on flexible experimentation. In contrast, Apache Zeppelin is an AI-gym platform emphasizing data science and machine learning model development within an interactive environment. It serves as a dynamic AI-gym for developing, training, and refining models through data analysis and visualization tools. Zeppelin supports code execution in multiple languages and provides REST APIs for deploying and testing models against real and simulated data. As a data exploration and prototyping tool, Zeppelin is well-suited for evaluating models across various machine learning frameworks before integration with production environments.

Fig. 1 depicts the system components in our toolchain, including Apache Zeppelin, REST APIs, Eclipse Arrowhead, and IoT data streams. Apache Zeppelin serves as an interactive environment for model development, training, and refinement, supporting various programming languages and frameworks. The REST APIs allow models to be deployed and tested against real-world inputs. Eclipse Arrowhead enhances IoT integration through its secure architecture, facilitating communication between AI models and devices. Finally, IoT data streams provide realistic inputs for model evaluation, ensuring the models are prepared for deployment.

Co-pilots are AI-powered assistants that automate repetitive tasks within development toolchains. These co-pilots assist with functions like coding, debugging, providing recommendations, and streamlining development. Hegedűs and Varga [7] present a framework for developing co-pilots that support different engineering stages in CPSoS, specifically using the Eclipse Arrowhead framework. The authors outline three co-pilot roles with distinct functions across the CPSoS engineering lifecycle:

1) **Arrowhead Expert:** A chat-based assistant embedded within the Arrowhead Framework Wiki, this tool answers

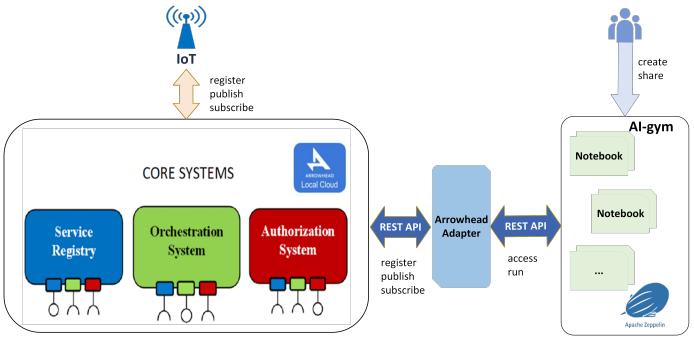

Fig. 1. Toolchain main components

design and integration questions based on documentation. It helps users understand and navigate the Arrowhead architecture using Retrieval-Augmented Generation (RAG) for question-answering.

2) **Arrowhead Management Co-pilot:** This tool interacts with Arrowhead Core Systems in a Local Cloud deployment. It analyzes, monitors, and manages CPSoS environments through API integrations and planning abilities, supporting authenticated users and facilitating operational tasks.

3) **Arrowhead Design Co-pilot:** A conceptual tool integrated with engineering tools (e.g., SysML modeling software) to assist in designing and configuring CPSoS deployments. This co-pilot helps engineers model and plan complex industrial systems.

These co-pilots address industrial automation challenges, including safety, transparency, and process integration. They leverage LLM capabilities like persona prompting, semantic planning, and API integration while considering limitations related to UX, plugin compatibility, and data security. Future improvements include advanced planner engines, improved RAG pipelines, and UX enhancements for industrial applications.

III. AI HARDWARE ACCELERATION

The rise of AI has led to larger models and higher computational costs, increasing the demand for specialized hardware [8]. This has driven advancements in processors and integrated circuits for AI, ranging from single-core to multi-core neural processing systems with unique architectures and improved parallel processing capabilities.

A. Specialized Hardware for AI

AI algorithms demand substantial computing power. While general-purpose CPUs and GPUs can handle these tasks, specialized hardware like FPGAs and ASICs often offer superior performance and energy efficiency [9]. CPUs have been optimized for AI calculations, while GPUs, designed for parallel processing, excel at large-scale computations [10]. FPGAs balance flexibility and efficiency, allowing customized hardware implementations suitable for tasks requiring high parallelism. Tools like OpenCL simplify development on FPGAs [11].

ASICs are specifically designed for specific applications, providing unmatched performance. These specialized chips excel in environments with limited power. FPGAs bridge the gap between general-purpose and specialized hardware, offering greater adaptability than CPUs and GPUs while being more cost-effective than ASICs. However, their lower clock speeds can be a limiting factor. Ultimately, the choice of hardware depends on the specific AI task.

B. Hardware-Oriented AI Optimizations

AI models often require significant computational resources, posing challenges for deployment on resource-constrained devices. Therefore, optimization techniques are essential to reduce complexity without compromising accuracy. These techniques can be broadly classified into model compression and performance optimization.

Model compression techniques aim to reduce the AI's size. Pruning removes unnecessary parameters, which can be done post-training but may reduce accuracy [12]. Integrating sparsity-promoting loss functions into the training process can mitigate this. Quantization reduces memory footprint and computational complexity by lowering the bit width of parameters [13]. Quantization Aware Training (QAT) generally preserves better accuracy [14] [15]. Quantization is particularly important for FPGA-based accelerators. In contrast, performance optimization approaches focus on improving inference speed. Parallelization strategies utilize parallel processing to speed up execution. While parallelization improves performance, it may also increase resource usage [16]. Deploying such workloads on resource-constrained devices necessitates a combination of optimization techniques. Model compression techniques reduce the model's size, while performance enhancement techniques improve performance. The choice of techniques depends on factors such as the target hardware, desired accuracy, and available resources. Effective leveraging of these techniques can improve efficiency without significantly impacting accuracy.

C. AI optimization impact

Model compression can significantly enhance AI performance by reducing latency. Pruning creates smaller, more efficient workloads that support faster inference. Fig. 2-a highlights how pruning impacts latency across various hardware platforms (CPU, GPU, FPGA, and ASIC), with the FPGA demonstrating the lowest latency across all pruning levels. This showcases the potential benefits of this technique in optimizing latency and improving the overall efficiency of AI models, particularly when utilizing dedicated dataflow accelerators.

Quantization further enhances efficiency by reducing the bit representation of model parameters and activations. When combined with pruning, quantization can lead to even greater energy savings due to the combined reduction in model size and computational complexity. Fig. 2-b depicts the positive effects of combining pruning and quantization on energy efficiency for a TX2 GPU, using FP16 and FP32 data types. The results consistently show that FP16 yields superior energy efficiency across all pruning levels, reinforcing the merits of integrating these methodologies.

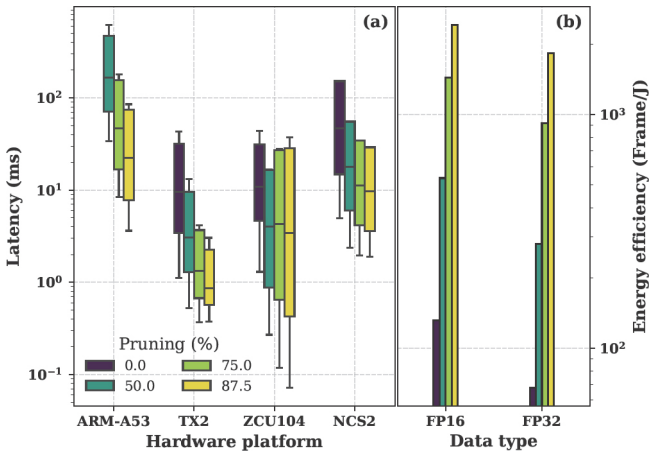

Fig. 2. Impact of pruning on (a) latency and (b) energy efficiency for different hardware platforms and data types. [17]

IV. SELF-ASSEMENT

With the progressive development of standards and regulations, current guidelines and documentation frequently lack alignment with specific user requirements, limiting their practical applicability and diminishing trust in AI within digital workspaces. This misalignment negatively impacts AI adoption in such environments. To address these challenges, the development of diverse tools—such as checklists, self-assessment instruments, guidelines, and recommendations—is essential to enable holistic AI application across varied use cases. In this project, we are taking the necessary steps required to build an AI-based self-assessment tool. Current compliance methods largely rely on manual checklists and subjective human evaluations [18]. Fig. 3 shows a schematic workflow of the envisioned Self-Assessment tool.

Recent studies highlight the potential of Large Language Models (LLMs) for various applications, such as information extraction [19], domain-specific fine-tuning [20], and machine translation [21]. Leveraging an LLM-based tool, our approach ensures compliance through regularly updated standards and provides project managers and developers with a guided checklist. User responses initiate a process where internal policies are aligned with applicable regulations, supporting automated compliance and enhancing organizational co-determination efforts.

The creation of the proposed self-assessment tool and guidelines involves the following steps: (1) Define personas and roles to model user groups and their characteristics through structured interviews and questionnaires. This questionnaire has been launched as part of the project aiming to survey the use of artificial intelligence in the workplace and employees' attitudes toward it. (2) Map the current standards and guidelines landscape by collaborating with project partners to align relevant standards with persona attributes. (3) Develop a tailored Large Language Model (LLM) to meet the tool's specific needs. (4) Implement the LLM as a domain-specific foundation for generating recommendations related to standards and guidelines. These steps are essential not only for developing

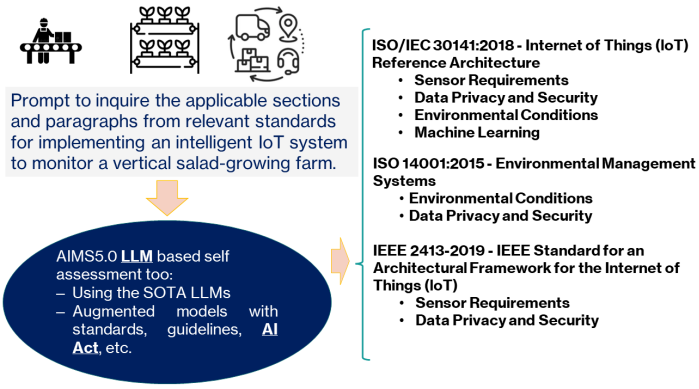

Fig. 3. A Vision of the Self-Assessment tool

a tool that aligns with various standards and regulations but also for enhancing trust and user acceptance of AI systems. They aim to maximize AI's potential in fostering innovation, economic growth, and global competitiveness

V. Conclusions and Future Work

This paper outlines a robust framework that bridges advanced AI capabilities with sustainable and human-centered manufacturing solutions. The work emphasizes the potential of AI-driven technologies to optimize processes, reduce environmental impact, and enhance human-machine collaboration. By integrating toolchains, hardware advancements, and self-assessment mechanisms, the project fosters an adaptable, sustainable manufacturing environment aligned with Industry 5.0 principles.

A key challenge identified by AIMS5.0 partners is the increasing demand for sustainable AI-based solutions. Building on initial results, we envision creating toolchains that support eco-friendly device development, incorporating features like recycled materials and reduced power consumption. The achievements of AIMS5.0 will be applied to semiconductor planning to define an ecological alpha in the operating curve, aiming to reduce the CO2 footprint per transistor function by 20% by the project's end. This establishes a foundation for ongoing improvement, targeting a 10% reduction per year post-project.

This project paves the way for a future where technology and human insight harmonize to build a more sustainable and resilient industrial landscape. Continued collaboration between academia, industry, and policymakers will be vital to translate these innovations into scalable solutions that address economic and ethical imperatives in manufacturing.

Acknowledgment

The AIMS5.0 project is supported by the Chips Joint Undertaking and its members under grant agreement No. 101112089.

References

[1] T. A. S. Foundation, "Zeppelin," 23/10/2024. [Online]. Available: https://zeppelin.apache.org/

[2] "Eclipse arrowhead™ – eclipse arrowhead," 04/11/2024. [Online]. Available: https://arrowhead.eu/eclipse-arrowhead-2/

[3] "Openai gym beta," 04/11/2024. [Online]. Available: https://openai.com/index/openai-gym-beta/

[4] C. Beattie, J. Z. Leibo, D. Teplyashin, T. Ward, M. Wainwright, H. Küttler, A. Lefrancq, S. Green, V. Valdés, A. Sadik, J. Schrittwieser, K. Anderson, S. York, M. Cant, A. Cain, A. Bolton, S. Gaffney, H. King, D. Hassabis, S. Legg, and S. Petersen, "Deepmind lab." [Online]. Available: http://arxiv.org/pdf/1612.03801

[5] U. Technologies, "Unity ml-agents toolkit," 05/10/2024. [Online]. Available: https://unity-technologies.github.io/ml-agents/

[6] Intel, "Intel® distribution of openvino™ toolkit," 16/10/2024. [Online]. Available: https://www.intel.com/content/www/us/en/developer/tools/openvino-toolkit/overview.html

[7] C. Hegedűs and P. Varga, "Co-pilots for arrowhead-based cyber-physical system of systems engineering," in *NOMS 2024-2024 IEEE Network Operations and Management Symposium*. IEEE, 2024, pp. 1–6.

[8] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.

[9] A. Shahid and M. Mushtaq, "A survey comparing specialized hardware and evolution in tpus for neural networks," in *2020 IEEE 23rd International Multitopic Conference (INMIC)*. IEEE, 2020, pp. 1–6.

[10] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2704–2713.

[11] A. Nechi, L. Groth, S. Mulhem, F. Merchant, R. Buchty, and M. Berekovic, "Fpga-based deep learning inference accelerators: Where are we standing?" *ACM Transactions on Reconfigurable Technology and Systems*, vol. 16, no. 4, pp. 1–32, 2023.

[12] D. Blalock, J. J. Gonzalez Ortiz, J. Frankle, and J. Guttag, "What is the state of neural network pruning?" *Proceedings of machine learning and systems*, vol. 2, pp. 129–146, 2020.

[13] Z. Cai, X. He, J. Sun, and N. Vasconcelos, "Deep learning with low precision by half-wave gaussian quantization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5918–5926.

[14] Z. Yao, R. Yazdani Aminabadi, M. Zhang, X. Wu, C. Li, and Y. He, "Zeroquant: Efficient and affordable post-training quantization for large-scale transformers," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 168–27 183, 2022.

[15] C. N. Coelho, A. Kuusela, S. Li, H. Zhuang, J. Ngadiuba, T. K. Aarrestad, V. Loncar, M. Pierini, A. A. Pol, and S. Summers, "Automatic heterogeneous quantization of deep neural networks for low-latency inference on the edge for particle detectors," *Nature Machine Intelligence*, vol. 3, no. 8, pp. 675–686, 2021.

[16] T. Ben-Nun and T. Hoefler, "Demystifying parallel and distributed deep learning: An in-depth concurrency analysis," *ACM Computing Surveys (CSUR)*, vol. 52, no. 4, pp. 1–43, 2019.

[17] M. Blott, L. Halder, M. Leeser, and L. Doyle, "Qutibench: Benchmarking neural networks on heterogeneous hardware," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 15, no. 4, pp. 1–38, 2019.

[18] P. Moertl and N. Ebinger, "The development of ethical and trustworthy ai systems requires appropriate human-systems integration: A white paper," InSecTT, White Paper, 2022. [Online]. Available: https://www.insectt.eu/wp-content/uploads/2022/11/Trustworthiness-Whitepaper-InSecTT-Format-v02-1-1.pdf

[19] D. Xu, W. Chen, W. Peng, C. Zhang, T. Xu, X. Zhao, X. Wu, Y. Zheng, and E. Chen, "Large language models for generative information extraction: A survey," *arXiv preprint arXiv:2312.17617 [cs.CL]*, 2023, (or arXiv:2312.17617v1 [cs.CL] for this version). [Online]. Available: https://arxiv.org/abs/2312.17617

[20] C. Jeong, "Fine-tuning and utilization methods of domain-specific llms," *arXiv preprint arXiv:2401.02981*, 2024. [Online]. Available: https://arxiv.org/abs/2401.02981v2

[21] J. Zheng, H. Hong, X. Wang, J. Su, Y. Liang, and S. Wu, "Fine-tuning large language models for domain-specific machine translation," *arXiv preprint arXiv:2402.15061*, 2024.